Triphone Statistics for Polish Language

Bartosz Ziółko*, Jakub Gałka†, Suresh Manandhar*, Richard C. Wilson*, Mariusz Ziółko†

*Department of Computer Science, University of York
Heslington, YO10 5DD, York, UK
{bziolko,suresh,wilson}@cs.york.ac.uk

†Department of Electronics, AGH University of Science and Technology
al.Mickiewicza 30, 30-059 Kraków, Poland
{jgalka,ziolko}@agh.edu.pl

Abstract

The Polish text corpus was analysed to find information about phoneme statistics. We were especially interested in triphones as they are commonly used in many speech processing applications like HTK speech recogniser. An attempt to create the full list of triphones for Polish language is presented. A vast amount of phonetically transcribed text was analysed to obtain the frequency of triphone occurrences. A distibution of frequency of triphones occuring and other phenomena are presented. The standard phonetic alphabet for Polish and methods of providing phonetic transcriptions are described.

1. Introduction

There is much statistical research at the word and sentence level for different languages (Agirre et al., 2001; Bellegarda, 2000). On the other hand similar research on phonemes is rare (Denes, 1962; Yannakoudakis and Hutton, 1992; Kollmeier and Wesselkamp, 1997). The frequency of phonetic unit presence is an interesting topic itself and can find use in many applications in speech processing, for example speech recognition. It is very difficult to provide proper acoustic data for all possible triphones, but there are methods to synthesise no-recorded ones using data for other triphones and phonological similarities between different phonemes (Young et al., 2005). This means that the list of possible triphones has to be provided for a given language. The triphone statistics can be also used to generate hypotheses used in recognition of out-of-dictionary words.

This paper describes several issues related to phoneme, diphone 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. We sum up the paper with conclusions.

2. Statistics Extraction Scheme

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 for this aim, which is described in the next section. The SAMPA extended phonetic alphabet was applied with 39 symbols and pronunciation rules typical for cities Kraków and Poznań. We altered the PolPhone phonetic alphabet (Table 1) to a 37 symbol version which is used in the largest corpus of spoken Polish (Grocholewski, 1995) and

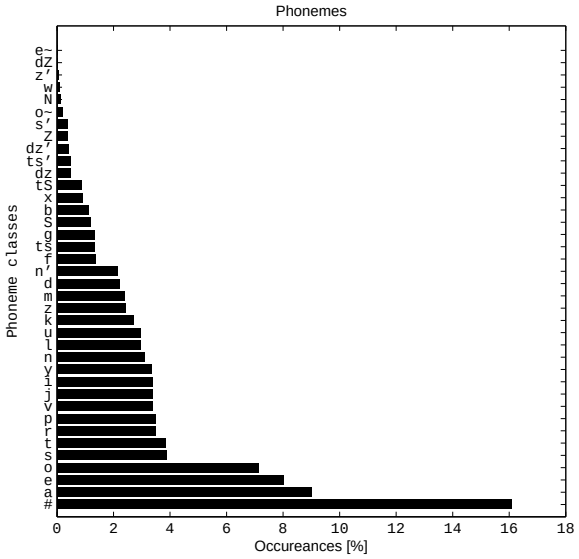

Figure 1: *Phonemes in Polish*

currently recognised as a SAMPA standard. We reduced the number of symbols by changing phoneme c to k and phoneme J to g. We also replaced $w\sim$ to $o\sim$ and $j\sim$ to $e\sim$. These changes were done to work on an offical standard version of SAMPA, which is frequently used, i.e. in the audio corpus with transcription (Grocholewski, 1995), rather than an extended SAMPA used in PolPhone, which is going to be suggested as a new standard. 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. Statistics can be now simply calculated by counting number of occurrences of each phoneme, phoneme pair, and phoneme triple in analysed text, where each phoneme is just one symbol. The analysis of the whole corpus took 3 weeks using PolPhone and scripts written in Matlab.

Table 1: *Phoneme transcription in Polish - SAMPA (Demenko et al., 2003)*

SAMPA	example	transcr.	occurr.	%
#		#	23,810,956	16.086,7
a	pat	pat	13,311,163	8.993
e	test	test	11,871,405	8.020,3
o	pot	pot	10,566,010	7.138,4
s	syk	sIk	5,716,058	3.861,8
t	test	test	5,703,429	3.853,2
r	ryk	rIk	5,171,698	3.494
p	pik	pik	5,150,964	3.48
v	wilk	vilk	5,025,050	3.394,9
j	jak	jak	4,996,475	3.375,6
i	PIT	pit	4,994,743	3.374,4
I	typ	tIp	4,974,567	3.360,8
n	nasz	naS	4,602,314	3.109,3
l	luk	luk	4,399,366	2.972,2
u	puk	puk	4,355,825	2.942,8
k	kit	kitk	4,020,161	2.716
z	zbir	zbir	3,602,857	2.434,1
m	mysz	mIS	3,525,813	2.382
d	dym	dIm	3,267,009	2.207,2
n’	koń	kon’	3,182,940	2.150,4
f	fan	fan	2,030,717	1.372
ts	cyk	tsIk	1,984,311	1.340,6
g	gen	gen	1,949,890	1.317,3
S	szyk	SIk	1,739,146	1.175
b	bit	bit	1,668,103	1.127
x	hymn	xImn	1,339,311	0.904,84
tS	czyn	tSIn	1,285,310	0.868,36
dz	dzwoń	dzvon’	692,334	0.467,74
ts’	ćma	ts’ma	690,294	0.466,36
dz’	dźwig	dz’vik	589,266	0.398,11
Z	żyto	ZIto	536,786	0.362,65
s’	świt	s‘vit	531,402	0.359,02
o∼	wąs	vo∼s	306,665	0.207,18
N	pęk	peNk	184,884	0.124,91
w	łyk	wIk	144,166	0.097,399
z’	źle	z’le	66,518	0.044,94
dZ	dżem	dZem	27,621	0.018,661
e∼	gęś	ge∼s’	1,011	0.000,683
w∼	ciąża	ts’ow∼Za	sampa extension	
j∼	więź	vjej∼s’	sampa extension	
c	kiedy	cjedy	sampa extension	
J	giełda	Jjewda	sampa extension	

3. Text to Phonetic Data Transcriptions

Two main approaches are used for the automatic transcription of texts into phonemic form. The classical approach is based on phonetic grammatical rules specified by human (Steffen-Batóg and Nowakowski, 1993) or automatic machine learning process (Daelemans and van den Bosch, 1997). A second solution utilises graphemic-phonetic dictionaries. In practice both mentioned methods are used in order to cover both 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

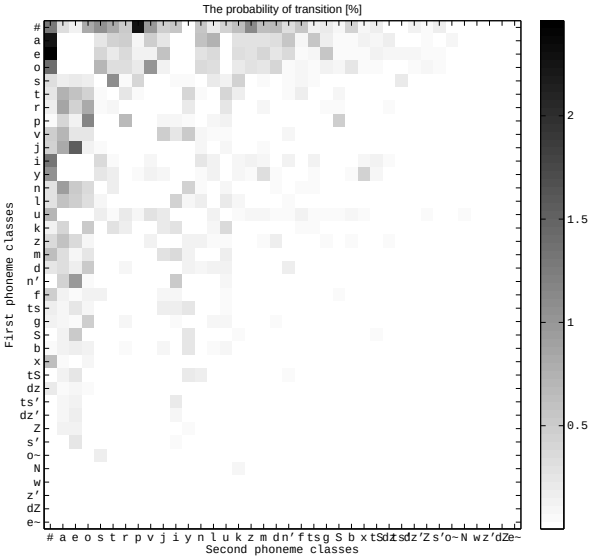

Figure 2: *Diphone probabilities in Polish*

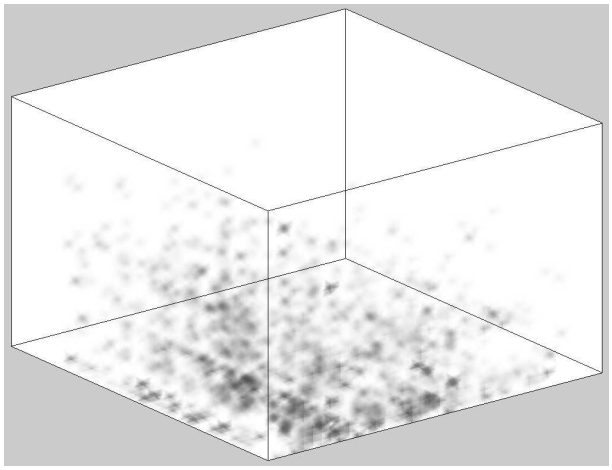

Figure 3: *Triphone probabilities in Polish*

to the use of the high-performance 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 (Table 1) is used, to deal with all nuances of Polish phonetic system. The transcription process is performed by a table-based system, which implements the rules of transcription. A 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 (SAMPA symbols) 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 pro-

ceeding $T[1][1]$ and $T[1][j]$ matches the substring following $T[1][1]$. This basic scheme is extended to cover overlapping phonetic contexts. When 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. Input Data and Results

One of the key uses for this data is speech processing. This is the reason for quite specific choice of analysed texts. Data for statistics were collected mainly from transcriptions of parliament meetings, the Select Committee to investigate corruption in amendment of Act on Radio and Television and Solidarity meetings (more than 90% of spoken language), from literature and an MA thesis.

Table 2: *Most common diphones in the analysed corpus*

diphone	no. of occurrences	percentage
e#	3,640,557	2.460,5
#p	3,379,372	2.284
a#	3,353,504	2.266,5
je	2,321,280	1.568,8
o#	2,094,619	1.415,7
i#	1,987,880	1.343,5
po	1,717,235	1.160,6
#z	1,700,044	1.149
st	1,614,996	1.091,5
y#	1,583,405	1.070,2
#s	1,572,893	1.063
ov	1,535,630	1.0379
#v	1,448,739	0.979,14
n'e	1,443,190	0.975,39
na	1,390,834	0.94
ra	1,306,527	0.883,02
#o	1,236,294	0.835,56
ja	1,236,189	0.835,49
#t	1,208,541	0.816,8
ro	1,195,087	0.807,71
ta	1,128,953	0.763,01
al	1,120,931	0.757,59
os	1,078,738	0.729,07
va	1,043,964	0.705,57
u#	1,033,050	0.698,19
#d	1,019,796	0.689,23
pr	999,628	0.675,6
#m	963,911	0.651,46
m#	959,333	0.648,37

Total number of 148,016,538 phonemes were analysed. They are grouped in 38 categories (including space). Their distribution is presented in Table 1 and in Fig. 1. 1,095 different diphones (Fig. 2 and Table 2) and 14,970 different triphones (Fig. 3) 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 and we are interested in triphone statistics inside words. The list of most common triphones is presented in Table 3. This list seems to be not fully representable because of text choice, specifically vast amount of parliament transcriptions, which caused probably some anomalies. I.e. the most common triphone *#po* and another on the list *pos* are probably related to corpus topic - *poseł* means MP in Polish. The word poseł appeared 141,904 in just its basic form, which is 11% of total appearence of *#po* and 42% of *pos*. Polish is a morphologically rich language so there are other cases of this word, including plural forms, all of them starting with *pos*. Assuming 38 different phonemes (including space) and subtracting mentioned *#* combinations there are 53,503 possible triples. We found 14,970 different triphones which gives a conclusion that almost 28% of possible combinations were actually found as triphones. An average length of words in phonemes can be estimated as 6.22 due to space (noted as #) frequency 16.09.

Fig. 2 shows some symmetry. Of course, the probability of diphone $\alpha\beta$ 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. Data presented in this paper illustrate the well-known fact that probabilities of triphones (presented in Table 3) cannot be calculated from the diphone probabilities (some of them are presented in Table 2). The reason for this is that the conditional probabilities have to be known.

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 expected to receive a very different distribution as very large amount of text was analysed. We hoped to have very few triphones with occurrences smaller than 3 and deduce that they are not real triphones but errors due to foreign names etc. in the corpus. Still even though we added extra text to the corpus several times the distribution did not change much at all. We noted around 1600 triphones which occurred just once, 800 with occurrence 2, 500 with 3, 300 to 400 for 4 to 6 occurrences, 200 for 7 to 9, and up to 100 for 10 or more, every time after we analysed extra text. Such phenomena is nothing unexpected in natural language processing on a level of words or above, where amount of analysed text do not change statistics (considering reasonable large amounts). Still in case of triphones the number of possibilities is much smaller and limited to mentioned 53503. The open question is if we would find distribution we expected if we analysed much bigger corpus or there is no limit in number of triphones lower than number of possible combinations. Every time we analysed extra text we found some new triphones. The new trigrams 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. It is difficult to predict if one can reach a situa-

Table 3: *Most common triphones in the analysed corpus*

triphone	no. of occurrences	percentage
#po	1,273,417	1.026,1
n'e#	925,893	0.746,09
#na	699,608	0.563,75
#pS	660,062	0.531,88
je#	659,674	0.531,57
na#	655,722	0.528,38
#pr	627,962	0.506,02
Ix#	613,589	0.494,43
ej#	602,920	0.485,84
#za	598,060	0.481,92
n'a#	574,708	0.46,31
ova	561,910	0.452,79
ego	558,788	0.450,27
sta	554,876	0.447,12
#do	551,423	0.444,34
go#	551,042	0.444,03
pSe	522,611	0.421,12
pra	492,128	0.396,56
#pa	481,772	0.388,21
#i#	478,500	0.385,58
vje	468,848	0.377,8
#n'e	430,178	0.346,64
#je	421,223	0.339,42
#f#	416,467	0.335,59
#v#	412,967	0.332,77
#vy	407,092	0.328,04
pro	390,429	0.314,61
#sp	357,008	0.287,68
#ko	342,254	0.275,79
#te	341,900	0.275,5
an'e	338,530	0.272,79
pos	337,190	0.271,71
ze#	335,941	0.270,7
ym#	332,437	0.267,88
em#	328,629	0.264,81
rav	318,232	0.256,43
#ze	310,008	0.249,81
ne#	309,151	0.249,12
nyx	307,657	0.247,91
kje	304,426	0.245,31
do#	296,635	0.239,03
ja#	294,220	0.237,08
#st	291,797	0.235,13
s'e#	285,355	0.229,94
#o#	283,500	0.228,45
ki#	282,413	0.227,57
#ro	282,059	0.227,28
to#	272,585	0.219,65
an'a	270,668	0.218,11
mje	266,812	0.215
ktu	265,128	0.213,64
#s'e	257,323	0.207,35
#to	256,113	0.206,38
la#	254,175	0.204,82
#ja	246,452	0.198,59
uv#	244,102	0.196,7
#ma	243,374	0.196,11
pov	242,231	0.195,19

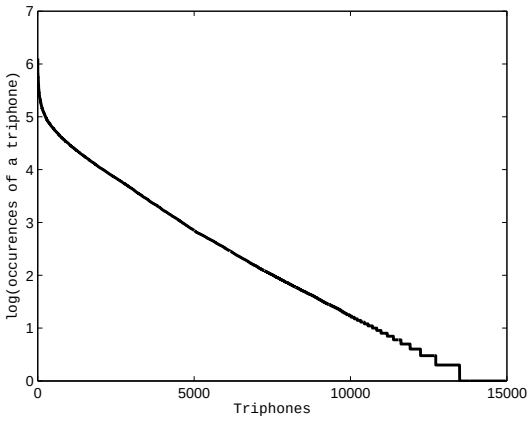

Figure 4: *Distribution of frequency of occurring phonemes in logarithmic scale*

tion new triphones do not appear and distribution of occurrences is changing as a result of more data being analysed. Still it is possible that the large number of triphones with very small occurrence are non-Polish triphones which should be excluded. In our further works we 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; Language Models in Speech Recognition,).

5

5. Conclusions

The statistics of phonemes, diphones and triphones were collected for Polish using a large corpus of mainly spoken formal language. The paper presents summarisation of the data and focus on interesting phenomena in the statistics. Triphone statistics play an important role in speech recognition systems. They are used to improve the proper transcription of the analysed speech segments. 28% of possible triples were detected as triphones, but many of them appeared very rarely. A majority of rare triphones came from foreign or twisted words. The statistics are available on request by an email.

6. Acknowledgements

We would like to thank Institute of Linguistics, Adam Mickiewicz University for providing PolPhone - a software tool to make a phonetic transcription for Polish.

7. References

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

Bellegarda, J. R., 2000. Large vocabulary speech recognition with multispan statistical language models. *IEEE Transactions on Speech and Audio Processing*, 8(1):76–84.

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

Demenko, G., 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).

Denes, P. B., 1962. Statistics of spoken english. *The Journal of the Acoustical Society of America*, 34:1978–1979.

Grocholewski, S., 1995. Założenia akustycznej bazy danych dla języka polskiego na nośniku cd rom (eng. Assumptions of acoustic database for Polish language). *Mat. I KK: Głosowa komunikacja człowiek-komputer, Wrocław*.

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

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

Kollmeier, B. 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.

Language Models in Speech Recognition. www.shlrc.mq.edu.au/masters/students/raltwarg/lmtoc.htm.

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

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

Rabiner, L. R., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

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

Yannakoudakis, E. J. 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.

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