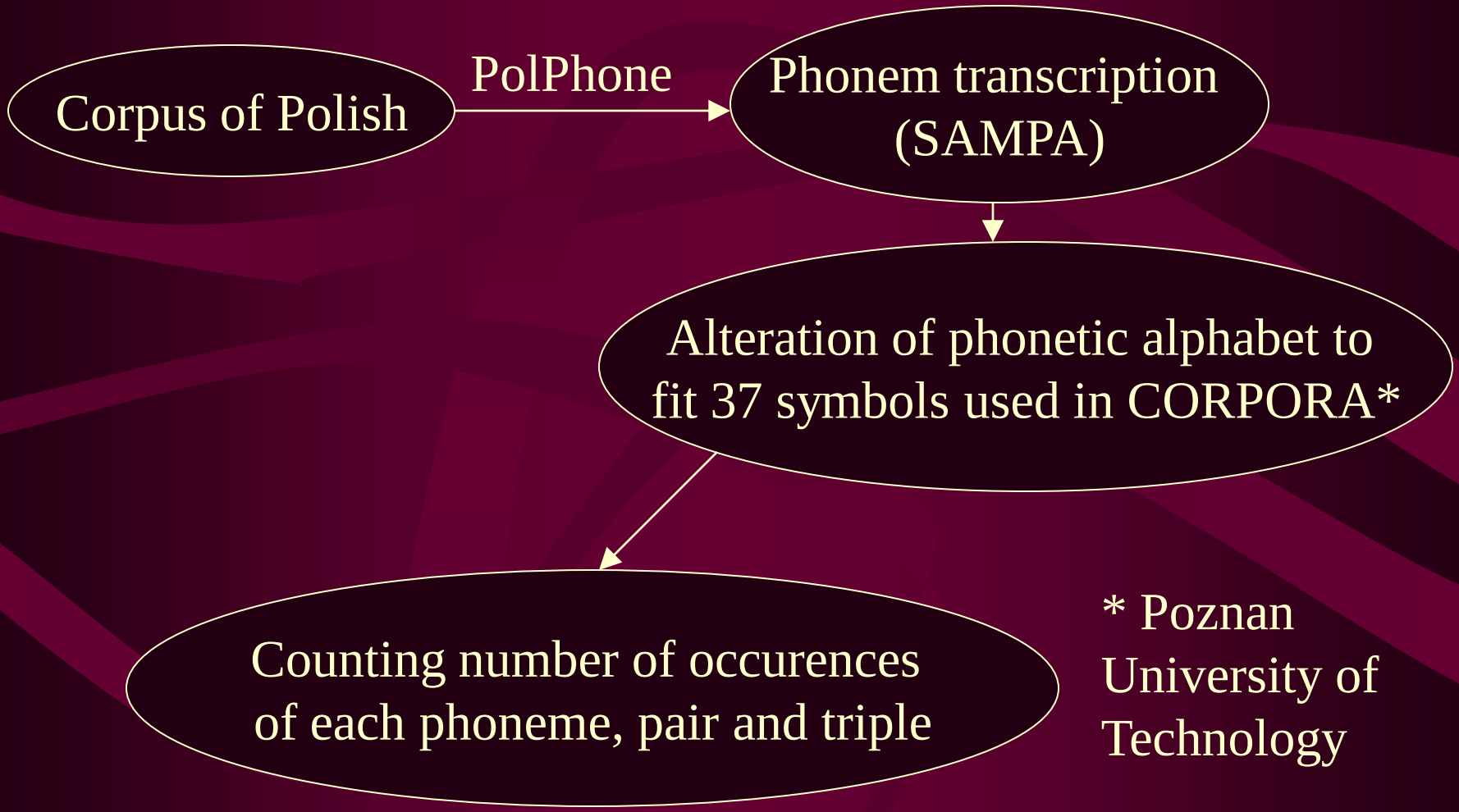
Triphone Statistics for Polish Language

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

THE UNIVERSITY of York

AGH

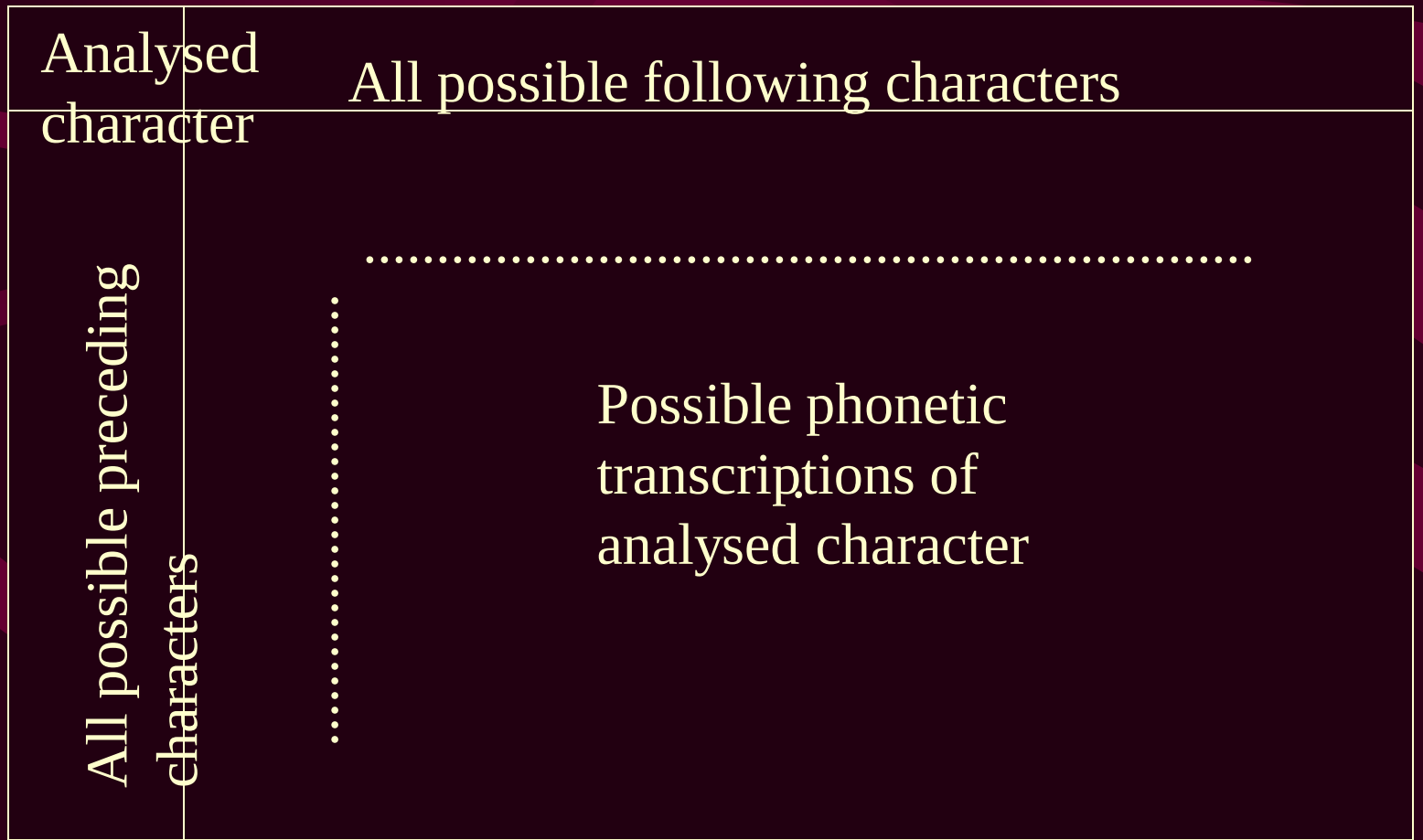
Scheme of the experiment



* Poznan University of Technology

PolPhone (UAM)

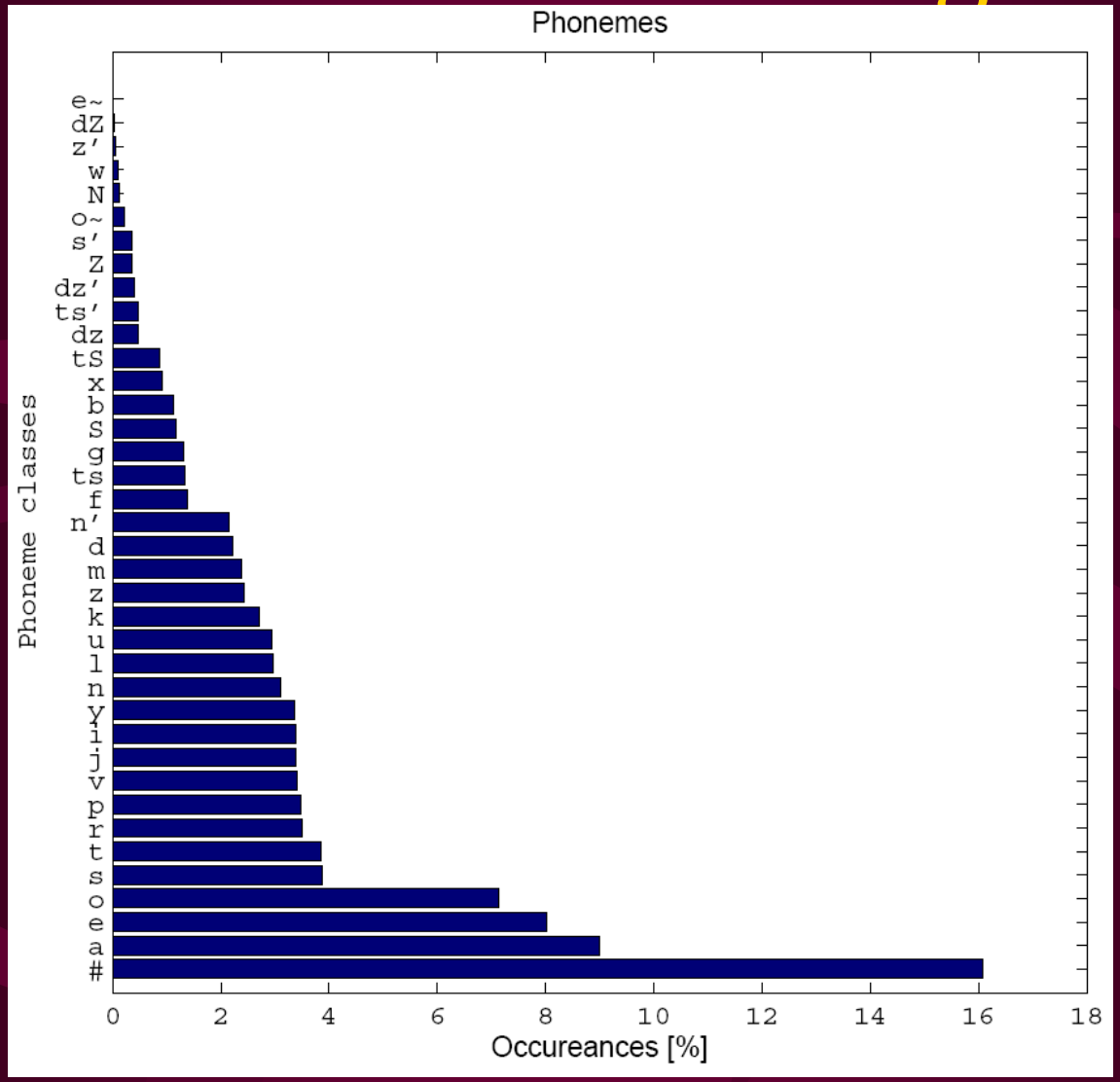
- phonetic grammatical rules specified by human
- machine learning process



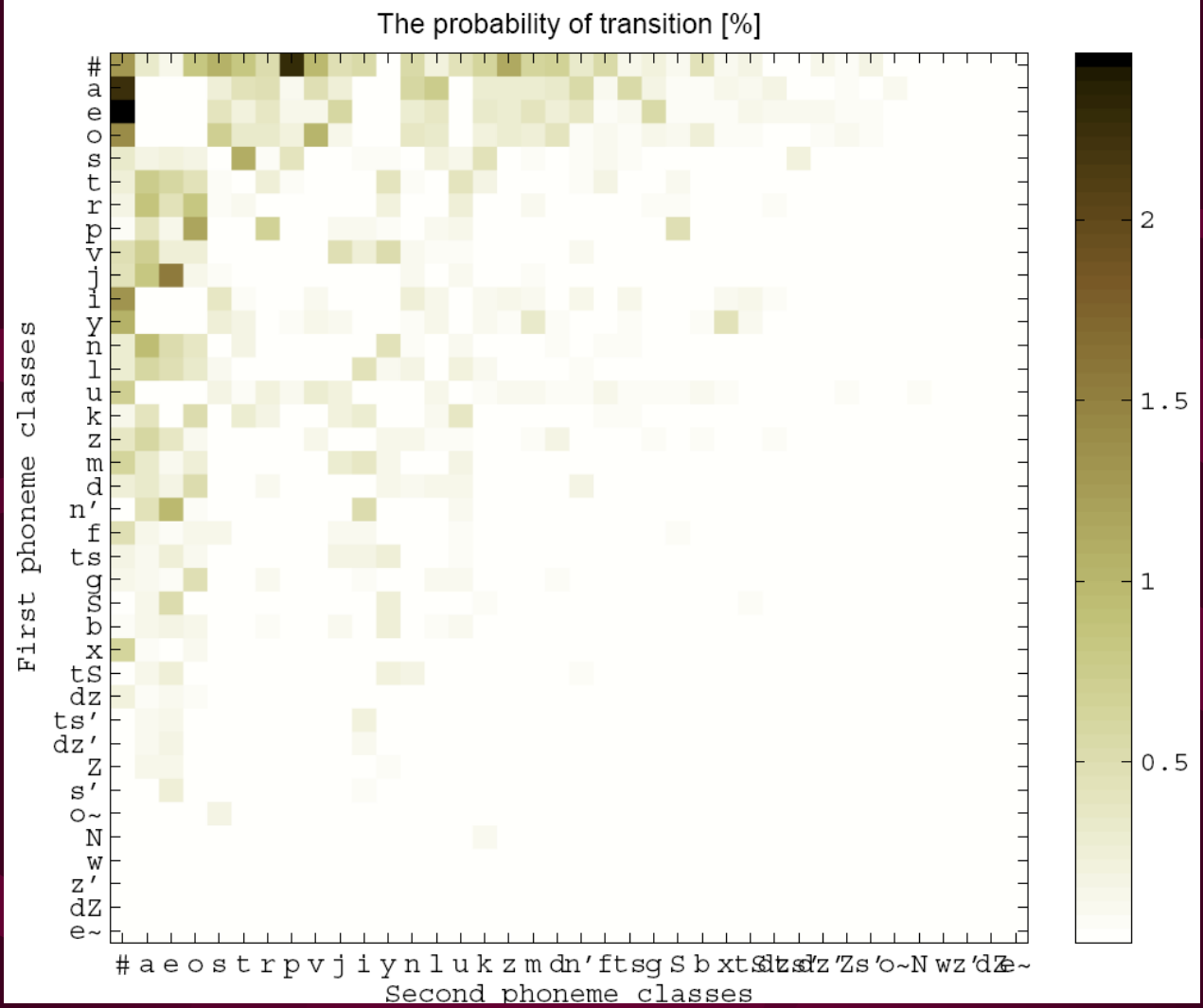
Sampa and phoneme frequencies

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	

Phoneme statistics again



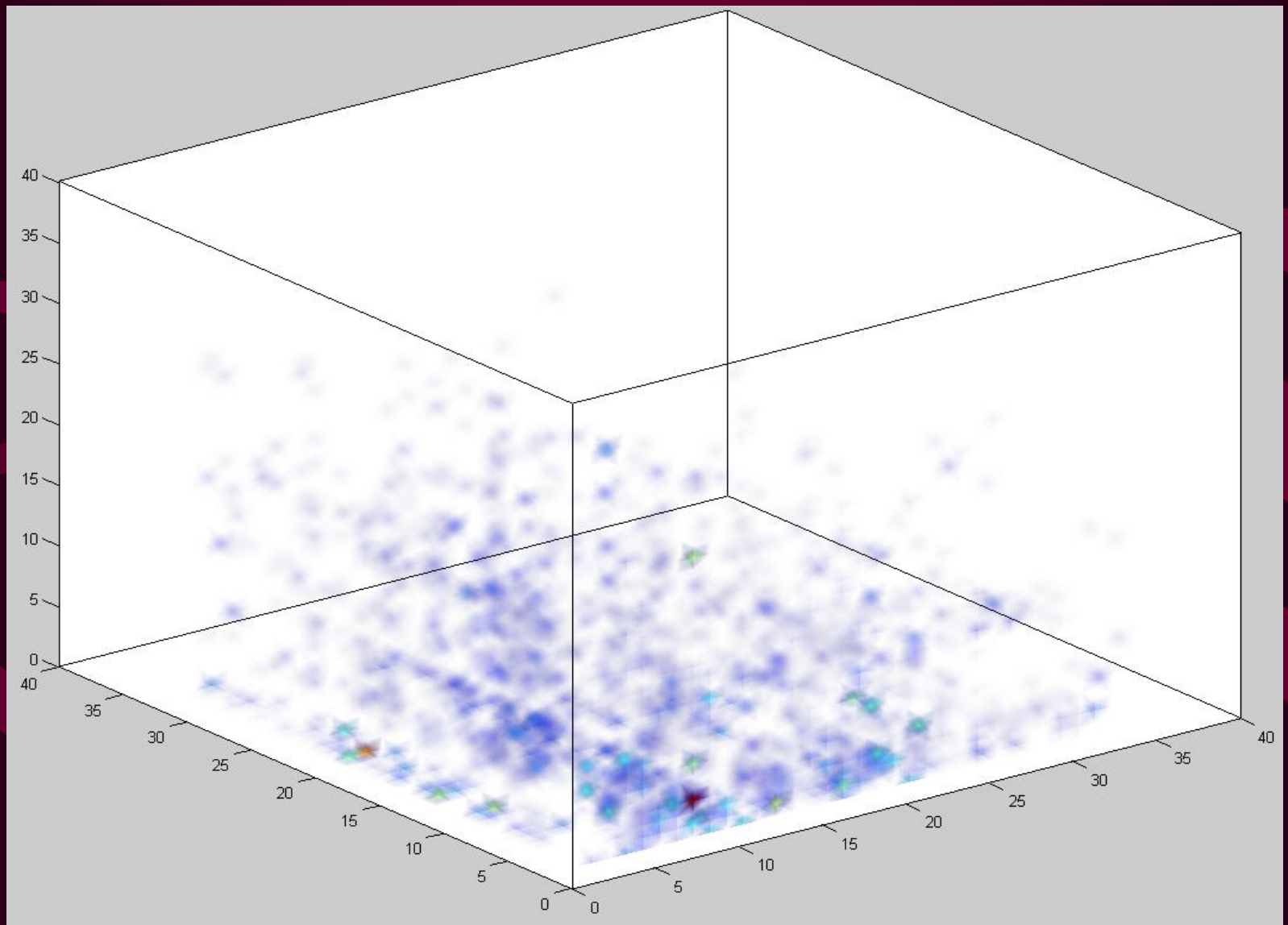
Diphones


The probability of transition [%]
First phoneme classes
Second phoneme classes
a e o s t r p v j i y 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~
2
1.5
1
0.5

Most common diphones

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

Triphones



Most common triphones

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

Input Data

Transcriptions of spoken Polish was the most of the corpus:

-Parliment meetings,

-Select Committee to investigate corruption in amendment of Act on Radio and TV ,

-Solidarność meetings from 80’.

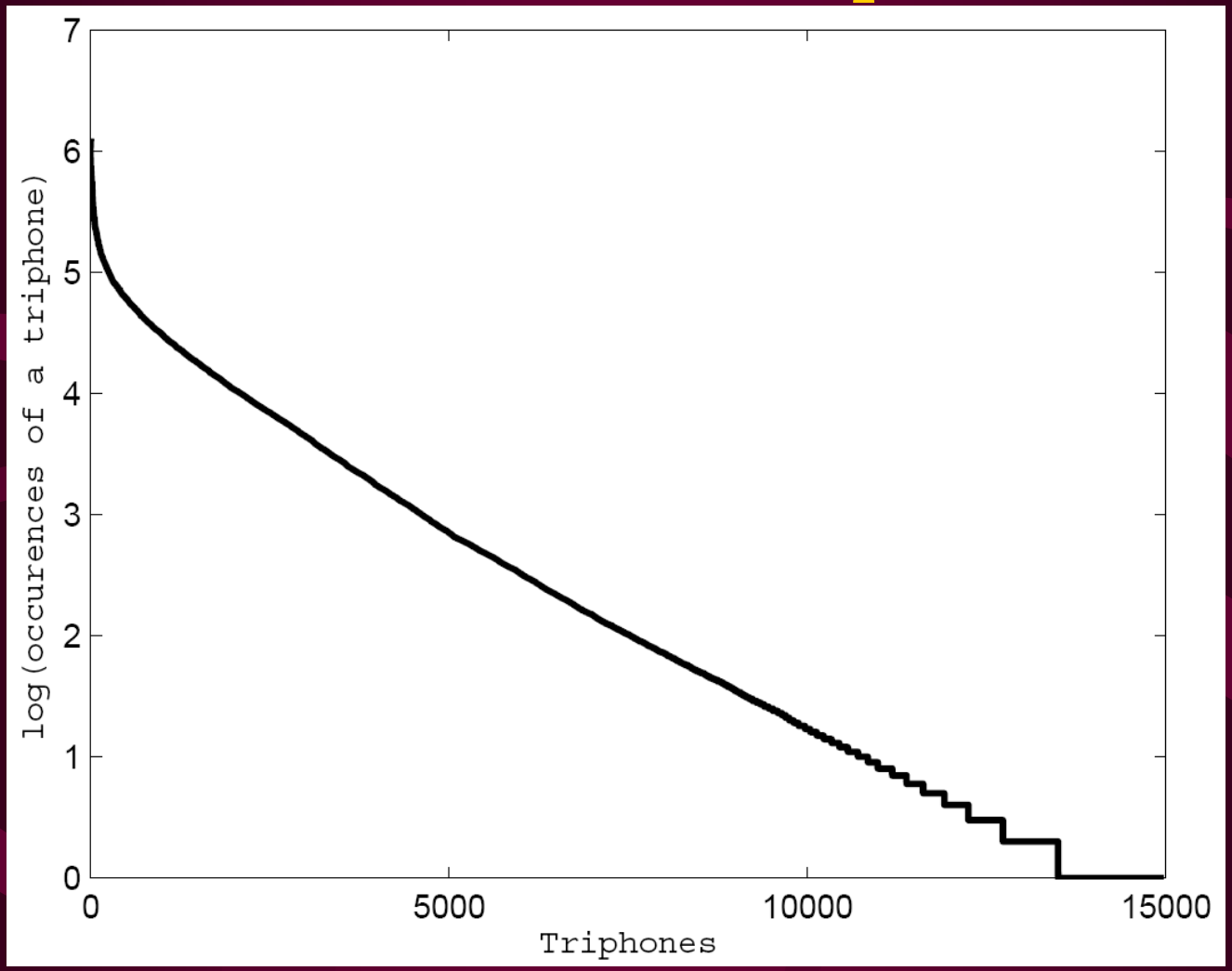
We included also some literature and MA thesis

Total number of 148,016,538 phonemes were analysed, what took 3 weeks using Matlab

Results and some observations

- 1,095 different diphones were detected
- 14,970 different triphones out of 53,503 possible combinations (excluding phoneme space phoneme string) were detected (28%)
- Anomalia in corpus – The word „poseł” appeared 141,904 times in just its morphologically basic form which is 11% of total appearance of *#po* and 42% of *pos*.
- Average length of words in phonemes is 6.2 (space frequency is 16.09 %)

Distribution of frequencies



Conclusions

- The list of triphones is not complete (we observed other triphones in CORPORA and the distribution suggests it) eventhough we analysed 148,016,538 phonemes,
- Triphone statistics play an important role in speech recognition systems,
- 28% of possible triphones were detected,
- Some of them were very rare and came from foreign and twisted words.

Thank You

Statistics are available on request by an email: bziolko@cs.york.ac.uk