AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA W KRAKOWIE

Phoneme Ngrams Based on a Polish Newspaper Corpus

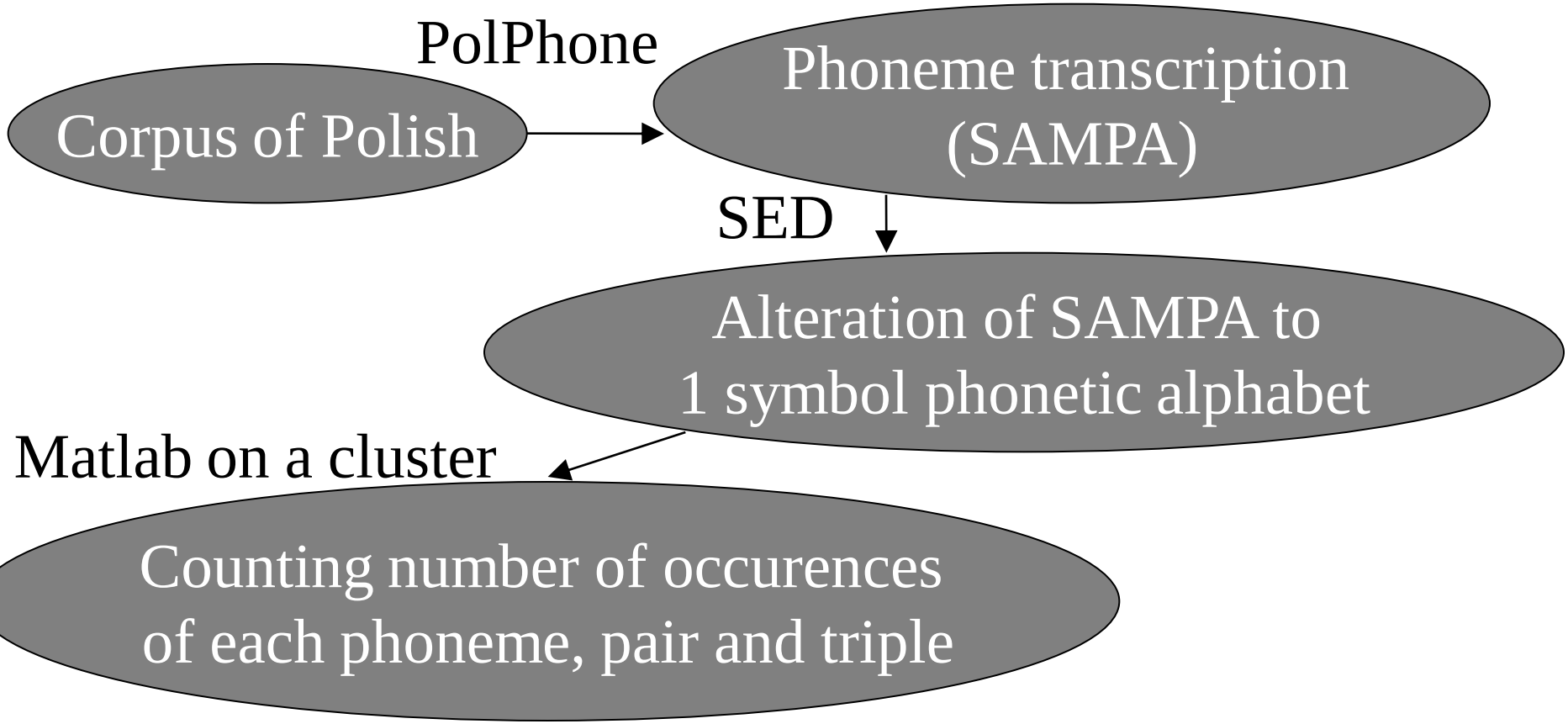
Bartosz Ziółko, Jakub Gałka, Mariusz Ziółko

Wydział Elektrotechniki, Automatyki, Informatyki i Elektroniki
Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie

Signal Processing Group (www.dsp.agh.edu.pl)

WorldComp, Las Vegas, July 2009

General Scheme



Academic Computer Centre cluster instead of PC

Mars - a computer cluster in the Academic Computer Centre CYFRONET AGH, 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.

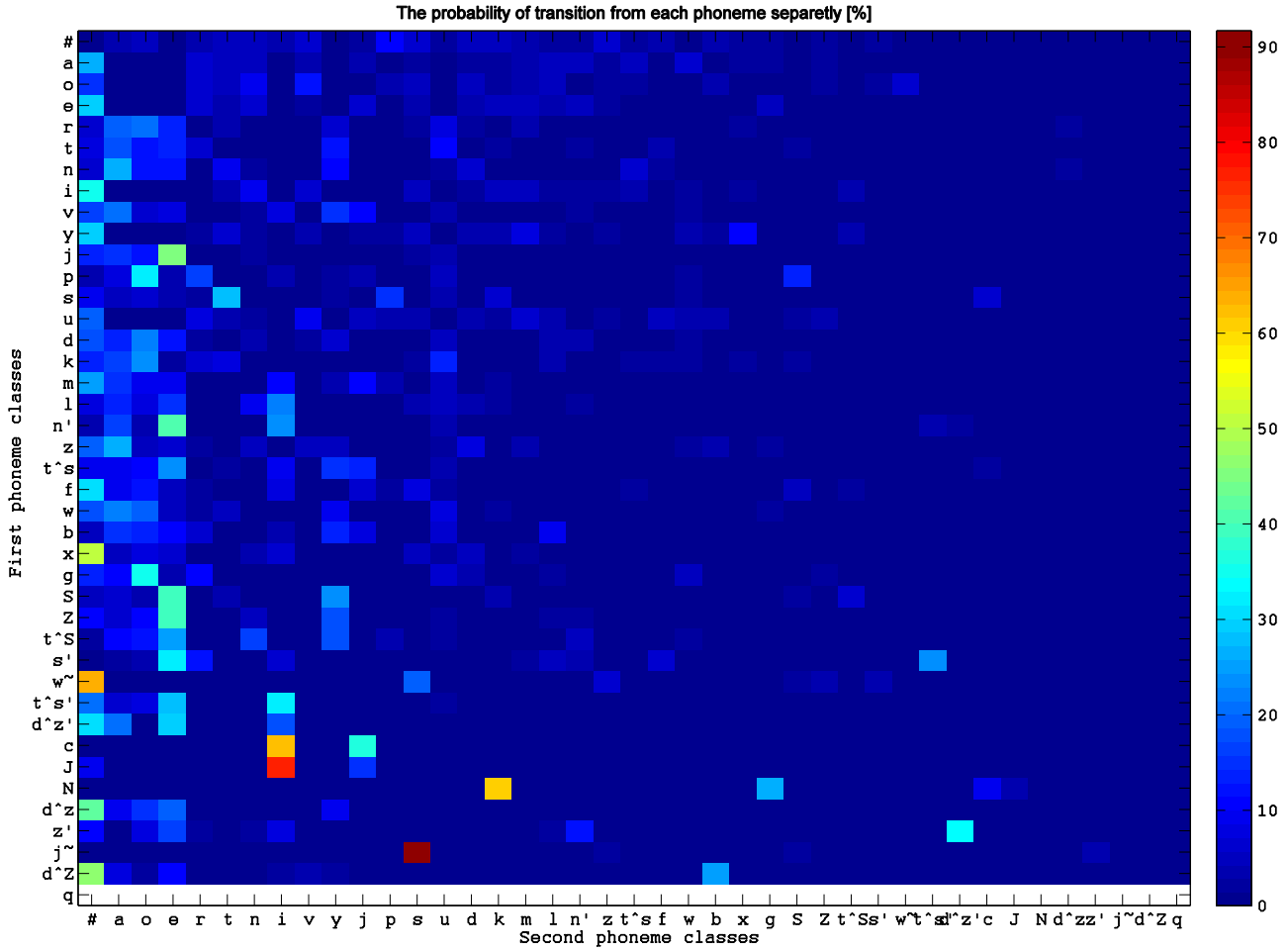
Corpus (Newspaper)

- Articles from Rzeczpospolita from years 1993-2002.
- Mainly political and economic issues.
- High quality language
- Quite many names and places including foreign ones. what may influence the results slightly.
- In total, 879 megabytes of text, which corresponds to around 104 000 000 words.
- Originally in html, problems with removing all tags.

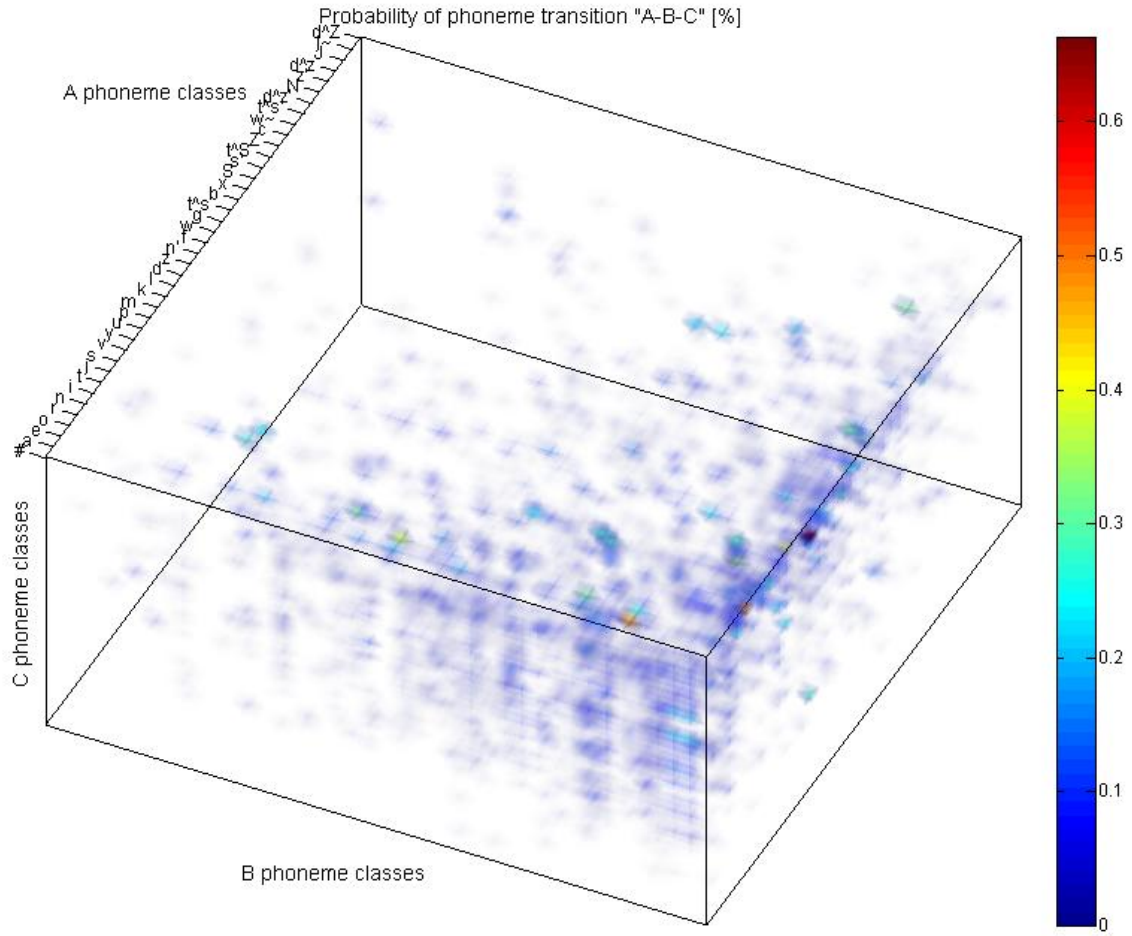
Phone Statistics

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
t^s	cyk	t^sIk	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
t^S	czyn	t^SIn	6 091 250	0.83
s’	świt	s’vit	6 077 420	0.82
w~	ciąża	ts’ow~Za	4 244 488	0.58
t^s’	ćma	t^s’ma	4 206 577	0.57
d^z’	dźwig	d^z’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
d^z	dzwoń	d^zvon’	1 846 929	0.25
z’	źle	z’le	997 176	0.13
j~	więź	vjej~s’	651 376	0.09
d^Z	dżem	d^Zem	218 975	0.03
q	-	-	1	0.00

Biphone statistics



Triphone statistics



Most common biphones

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

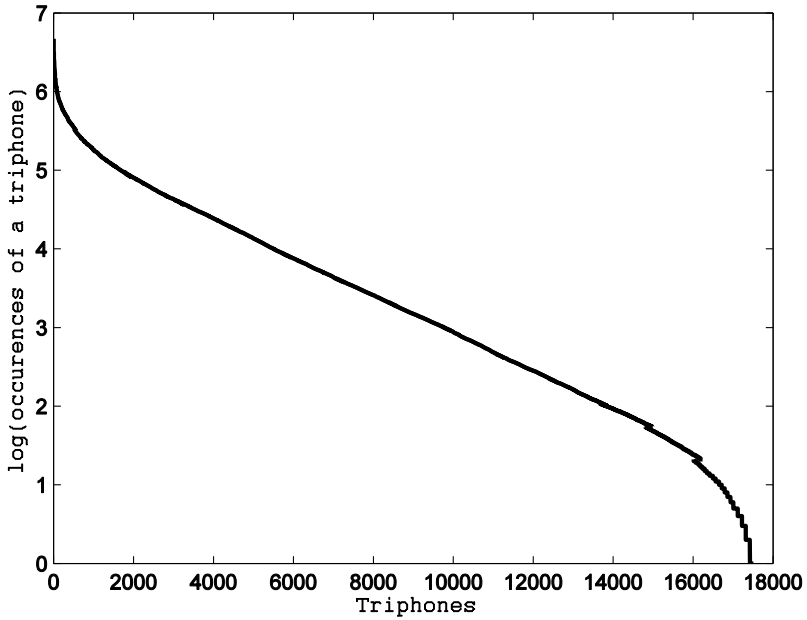
Most common triphones

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

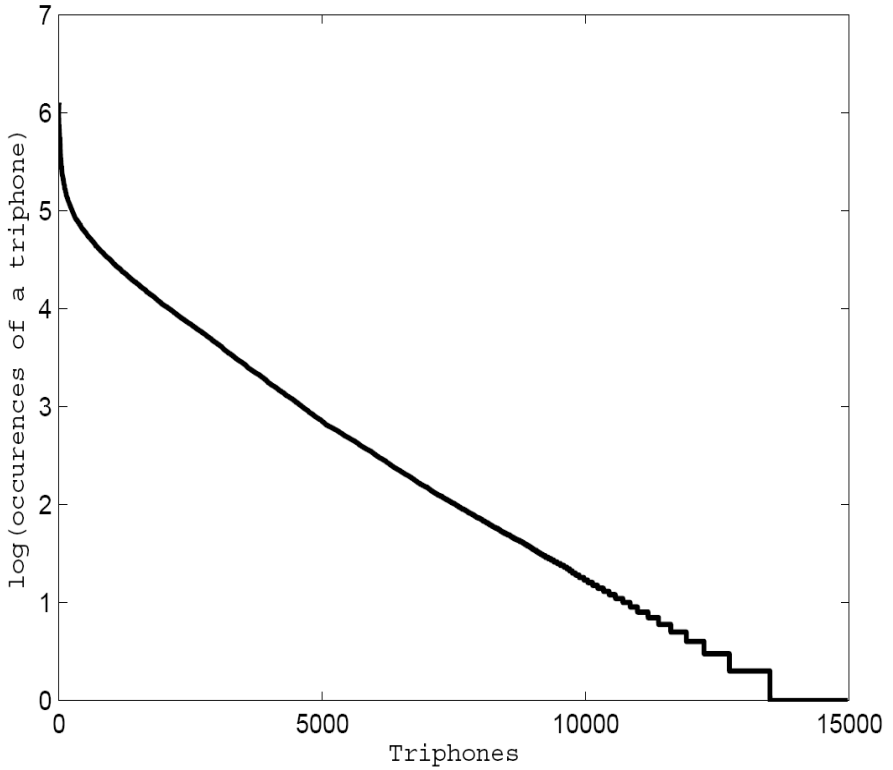
General Results

- 736 715 777 phonemes were analysed.
- 1 149 different biphones for 1 560 possible combinations were found (74 %).
- 17 278 different triphones for 62 479 possible triples (28%).
- Average length of words in phonemes is 6.7.
- Different than for (2007) experiment and than other types of corpora.

Triphones distribution



736 715 777 phones



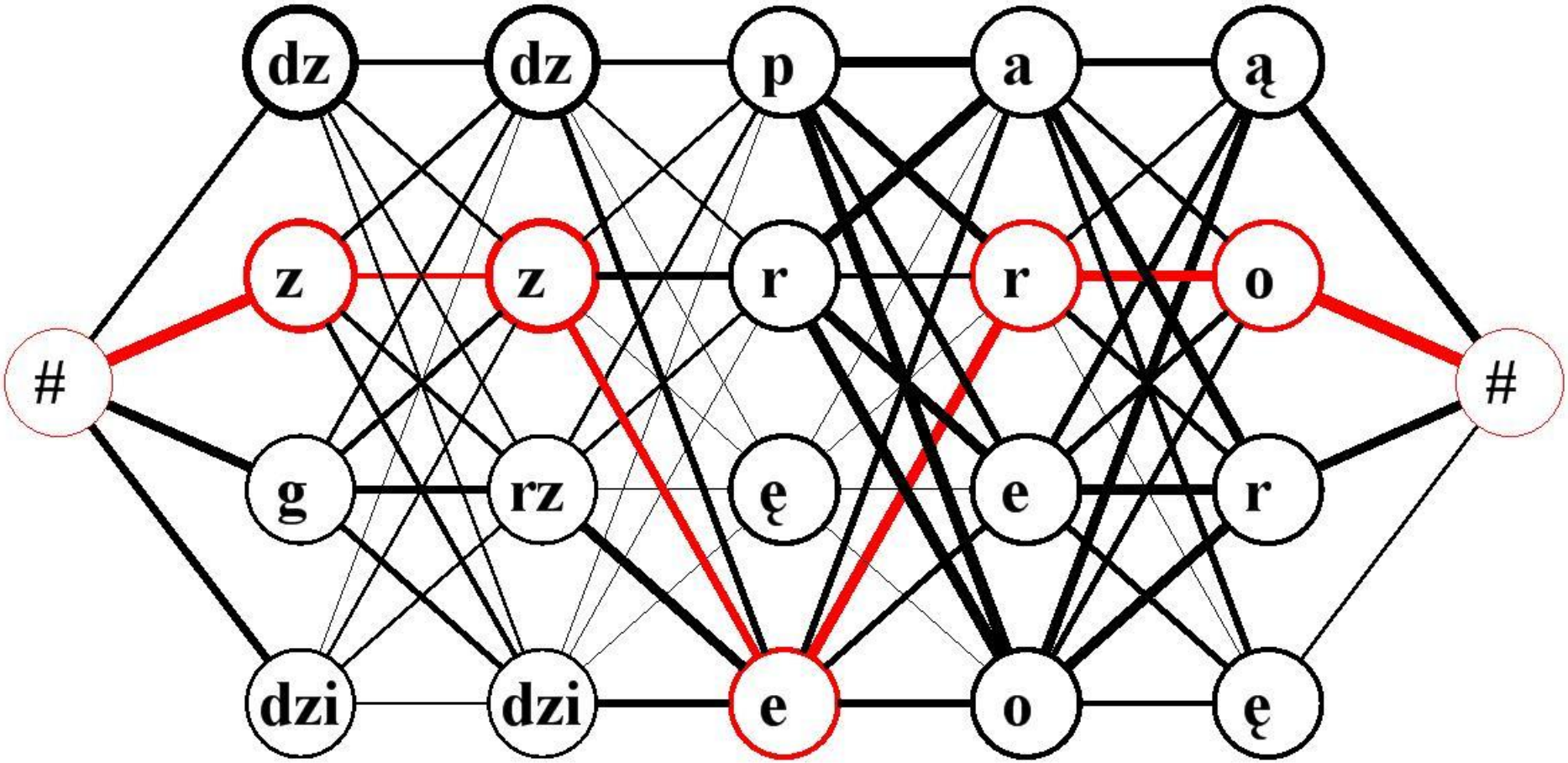
148 016 538 phones

N-gram probability model

- Phonemes vary slightly depending on the context - neighbouring phonemes, due to a natural phenomena of coarticulation.
- Speech recognisers based on triphone models rather than phoneme ones are much more complex but give better results.
- Examples of transcribing word above:
  - phoneme model *ax b ah v*
  - triphone model *-ax+b ax-b+ah b-ah+v ah-v+**

.

Applications



Conclusions

104 000 000 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. 28% of possible triples were detected as triphones, most of them at least several times.

Thank you !!!