ALGORITHMS FOR ACCENTUATION AND PHONEMIC TRANSCRIPTION OF RUSSIAN TEXTS IN SPEECH RECOGNITION SYSTEMS

Yakovenko O.S. (olya.yakovenko@bk.ru), Bondarenko I.Yu. (i.yu.bondarenko@gmail.com), Borovikova M.N. (m.borovikova@g.nsu.ru), Vodolazsky D.I. (daniil.vodolazsky@mail.ru)

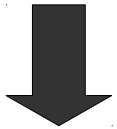
Novosibirsk State University, Moscow Institute of Physics and Technology

Novosibirsk State University

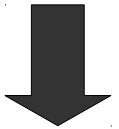
DIALOGUE 2018

MIPT
MOSCOW INSTITUTE OF PHYSICS AND TECHNOLOGY

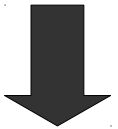
Open-source corpora



SMALL



NOT ALIGNED BY TIME

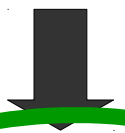


HAVE ONLY WORD MARKUP

What to do?

AUGMENTATION, E-BOOKS, manual markup

FORCE-ALIGNMENT, manual markup



MAKE A PHONEME MURKUP

Goal

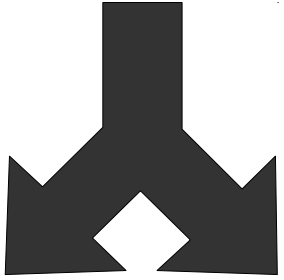
Development of open-source automatic transcription system of Russian texts for speech recognition tasks

Target frameworks

1) CMU Sphinx
2) HTK
3) Kaldi

Algorithm

корова



Accentuation

коро+ва

Transcription

K A R O0 V A

Corpus

Voxforge Russian corpus:

http://www.voxforge.org/ru

Results

SYNTAGM	ETALON	RESULT	PER
отыскал ждановскую набережную	A T Y S K A0 L ZH D A0 N A F S K U J0 U N A0 B0 I R0 I ZH N U J0 U	A T Y S K A0 L ZH D A0 N A F S K U J0 U N A0 B0 I R0 I ZH N U J0 U	0
разбил земную кору	R A Z B0 I0 L Z0 I M N U0 J0 U K A R U0	R A Z B0 I0 L Z0 I M N U0 J0 U K A R U0	0
освещённые окна сарая	A S V0 I SH0 O0 N Y I O0 K N A S A R A0 J0 A	A S V0 I SH0 O0 N Y J0 I A K N A S A R A0 J0 A	10.5

Results

PER 1.7%, Phone Accuracy 98.3%

Results

	russian_g2p	text2dict	epitran
WER	**28,80%**	**27,88%**	**29,80%**
PER	**31,89%**	**32,19%**	**34,36%**

Future work

1) Manual markup of the 20-hour dataset
2) Force-alignment for 20-hour dataset
3) Accents for proper names
4) Machine-learning techniques for out-of-vocabulary words or omographs

Location

https://github.com/nsu-ai/russian_g2p

Acknowledgements

The study was conducted with the support of Botan Investments

Thank you for your attention!