Automatic Detection of Stress
in Russian by Misspellings in Corpus

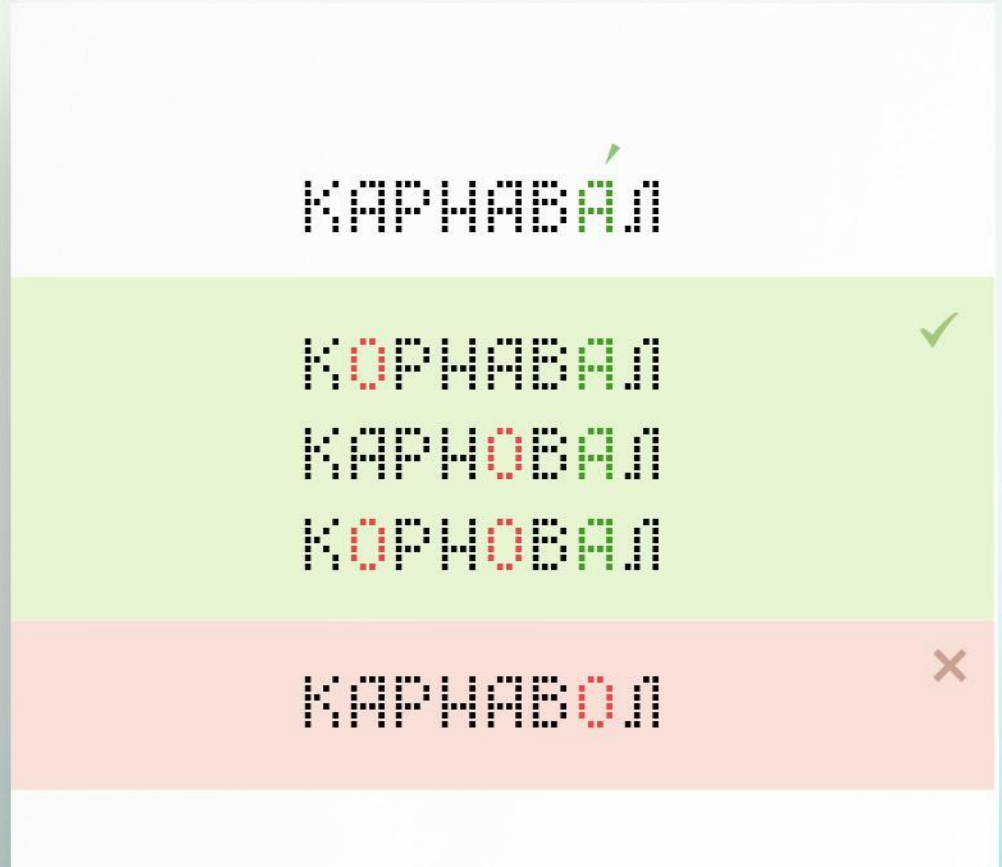
Alexeyevsky D. A., Lipunova A. E.

04.06.2016

The goal

The hypothesis is that mistakes in vowels occur more frequently in unstressed positions than in stressed positions...

..so here we present an implementation of the program to detect stress in corpus with mistakes



Possible use

Stress detection in neologisms for further dictionary adjunction

Analysis of stress distribution within certain communities of people

Stress change detection in usus

Diachronic analysis of stress distribution

Algorithm

Our work relies on detecting mistakes in corpus.
In order for stress detection to function properly a large corpus containing uncorrected texts is required.

Twitter corpus:

17,639,674 tweets, 40 characters minimum, 160’020’610 tokens

Twitter corpus

Preprocessing

Grouping by consonant mask

Group splitting by set of possible substitutions

Stress recognition for words with one unchanged vowel

Consonant mask

Twitter corpus

Preprocessing

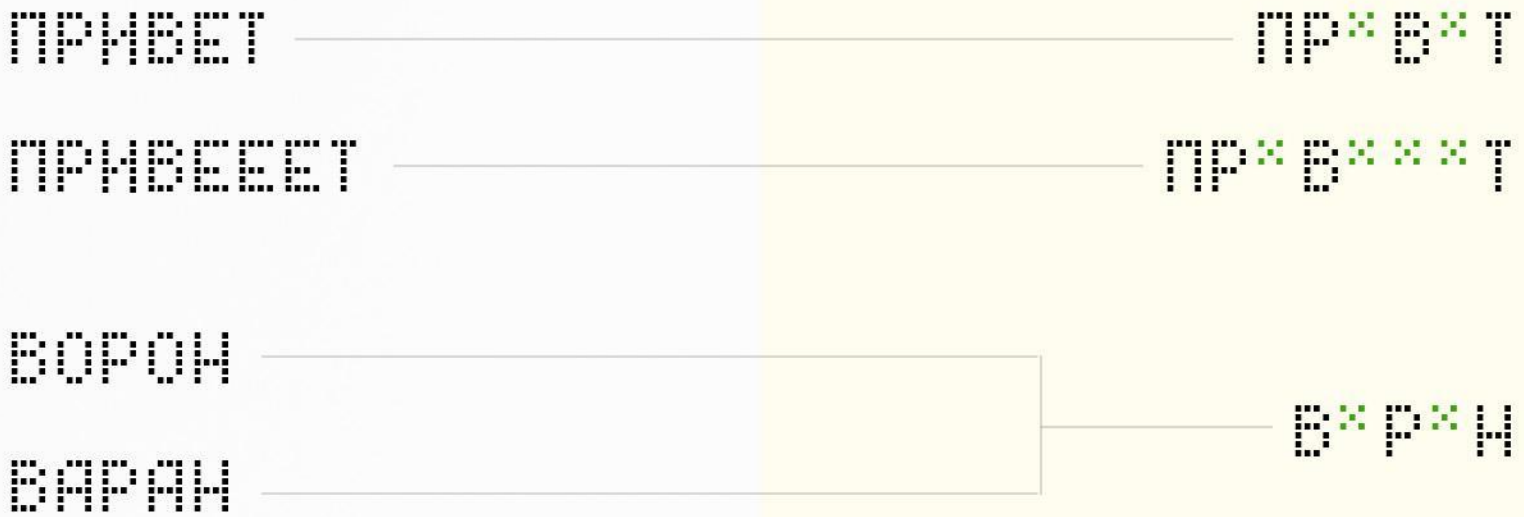
Grouping by consonant mask

Group splitting by set of possible substitutions

Stress recognition

Consonant mask of a token

is the token with each vowel replaced with wildcard character



Set of possible substitutions

Twitter corpus

Preprocessing

Grouping by consonant mask

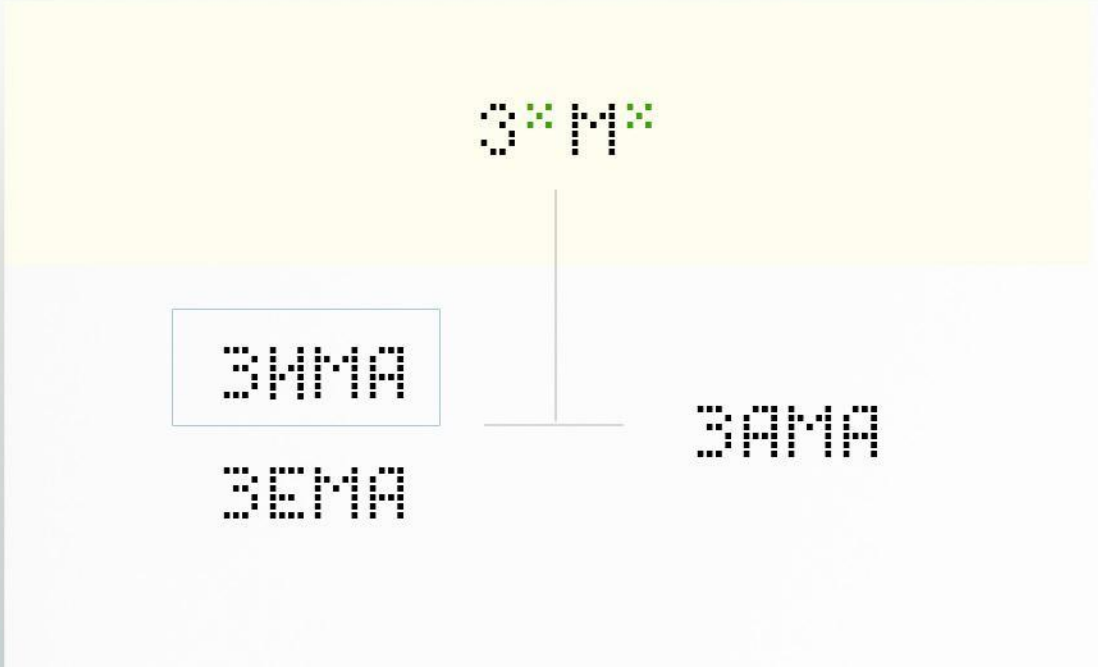
Group splitting by set of possible substitutions

Stress recognition

SET OF RULES

defining the possible substitutions was created

Russian language script is phonematic. Phonematic scripts give rise for orthographical mistakes in cases where one phoneme may be represented in various spellings



Automatic detection of stress in Russian by misspelling in corpus

Results types

Twitter corpus

Preprocessing

Grouping by consonant mask

Group splitting by set of possible substitutions

Stress recognition

All vowels in the nest remain unchanging + monosyllables

Only one word-form variation presented in the nest

ТЕКСТ	9894
АВАНГАРД	1070

Some of the vowels in the nest changing

In this case the program can make an assumption about stress position

АБСОЛЮТНОЕ	289
ОБСОЛЮТНОЕ*	1

Typos or common mistakes of two different word-forms:

ПОБОЛЬШЕ-ПОБАЛЬШЕ*

КВАРТИРАНТАМ-КВАРТИРАНТОМ-КВАРТЕРАНТАМ*

Only one vowel in the nest remain unchanging

Here it is possible to uniquely determine the stress position in the word-form.

ВПОЛНЕ	7312
ВПАЛНЕ*	1
ОТЛИЧНООООО	8
ОТЛИЧНАААААА*	2
АТЛИЧНАААААА*	1

Typos, common mistakes of two different word-forms, two correct word-forms in one nest:

МНОГИК-МНАГИК*

АДМИНАМ-АДМИНОМ-ОДМИНАМ*

СМЕШНО-СМЕШНА

All of the vowels in the nest changing

In this case it is impossible to determine stress position, while it's possible to determine some of them by frequency analysis, which is out of scope for this project

Typos, common mistakes of two different word-forms, homonymy:

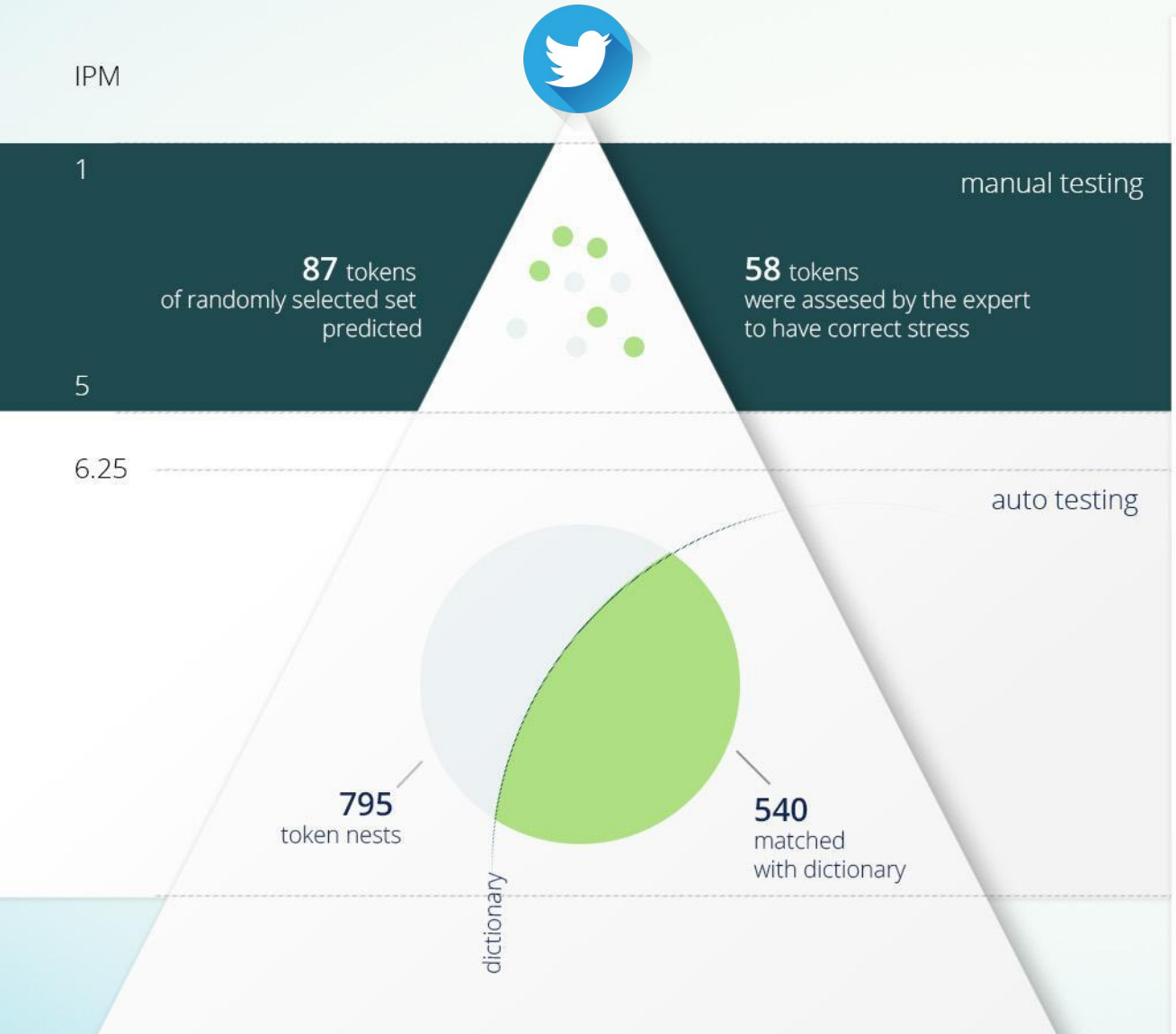
ЗАМОК-ЗАМАК-ЗОМОК*

ВОШЛО-ВОШЛА-ВАШЛО*

ГНЕЗДО-ГНИЗДА*

Automatic detection of stress in Russian by misspelling in corpus

Testing & Results



Two testing sessions:

Autocompare in between stress dictionary and the program results

Manual testing to check the results on random selection of word-forms

Since not all possible vowel misspellings were presented in most word-forms the recall appeared to be rather low (1%).

However, in the first case we get 67% of successful results and in the second case we get the same 67% proven manually, which is a good result for using only one method to detect the stress and may be improved in further investigations while using several methods and additional features combined.

Automatic Detection of Stress
in Russian by Misspellings in Corpus

Alexeyevsky D. A., Lipunova A. E.

Thank you!