STATISTICAL APPROACH FOR RECOGNITION AND DETECTION OF PHONETIC BORROWINGS IN CHINESE (CASE STUDY OF THE RUSSIAN PROPER NAMES)

Kirill Semenov

HSE – Moscow

eReL: May 11th , 2019

OBJECT: IS IT IMPORTANT?..

There are other more popular spheres in modern Chinese lexicography:

- Semantic borrowings: Superman > 超人 (chāorén = “exceed”+ “man”)
- Loanword blending: ballet > 芭蕾舞 (bālěiwǔ = phonetic loan + “dance”)
- “Lettered words”: 卡拉OK (kǎlā ou kei – “karaoke”);

  三G手机 (sān Gē shǒujī – “three” + “Generation [English]” + “mobile phone”)
- Modern Internet neologisms
  - 同志 (tóngzhì) – “comrade” > “homosexual”
  - 1314 (yī sān yī sì) ≈ yī sheng yī shì = “forever”

OBJECT: **IT IS** IMPORTANT!

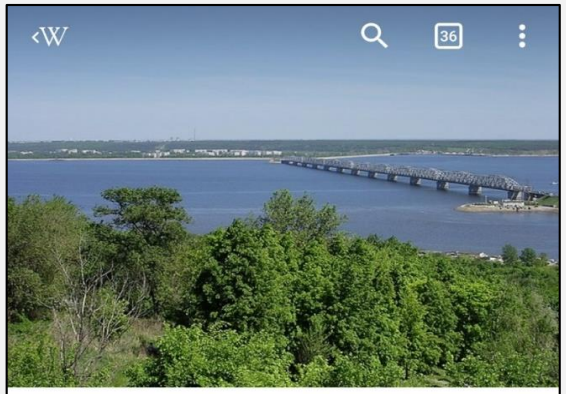
For the fundamental research:

- phonetic adaptation in a language of different phonetic inventory

For the natural language processing (NLP) purposes:

- PoS-tagger
- NER (Named Entity Recognition)
- MT (Machine translation)
- etc

OBJECT



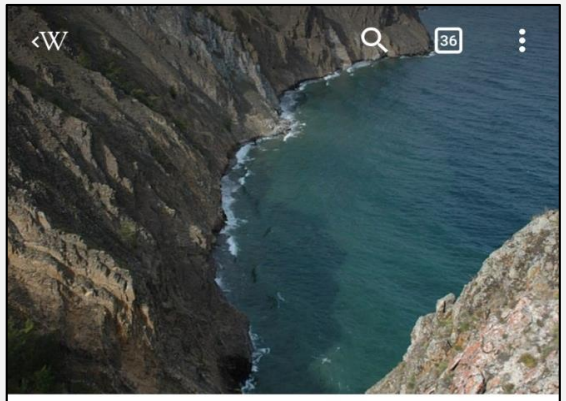
伏尔加河

添加标题描述

伏尔加河（俄语：Волга，又译窝瓦河），位于俄罗斯西南部，全长3,692公里，是欧洲最长的河流[2]，也是世界最长的内流河[2]，流入里海。

快速预览: 国家, 城市 ...

伏尔加河是欧洲流域最广以及流量最大的河流，流经欧洲俄罗斯，是代表型的俄罗斯河



贝加尔湖

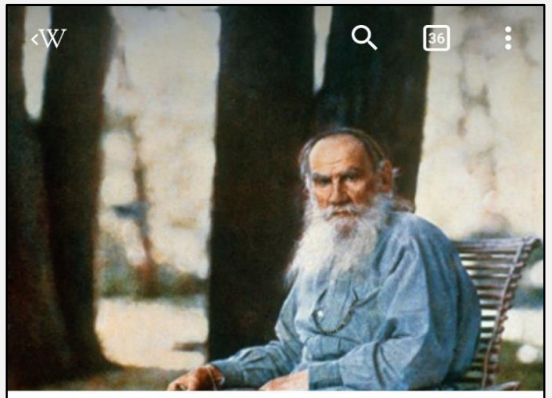
俄罗斯湖泊

贝加尔湖（俄语：óзеро Байка́л，罗马化：Ozero Baykal，IPA：[ˈozʲɪrə bɐjˈkalʲ]；布里亚特语：Байгал нуур, 拉丁转写：Baygal nuur，蒙古语：Байгал нуур；意思是"自然之湖"[3]；一说名称来源于"贝音嘎　嘎拉"（蒙古语意为不灭的火焰）[4]）。汉朝人称之为"**翰海**[5]"，五胡十六国时北朝叫"**于巳尼大水**"，隋唐叫"**小海**"，18世纪初期的《异域录》称之为"**柏海儿湖**"，

莫斯科

俄羅斯首都

莫斯科（俄语：Москва，罗马化：Moskva，IPA：[mɐsˈkva]）是俄罗斯首都与最大都市、以及莫斯科州首府，为俄罗斯全国政治、经济、科学、文化及交通的中心。面积2,510平方公里，与莫斯科州和卡卢加州接壤。城区人口约1200万，是欧洲人口第二多的城市，仅次于伊斯坦布尔，占俄罗斯总人口的1/10。



列夫·托尔斯泰

俄國作家

列夫·尼古拉耶维奇·托尔斯泰（俄语：**Лев Николаевич Толстой**，拉丁化：Lev Nikolayevich Tolstoy；1828年9月9日（儒略历8月28日）–1910年11月20日（儒略历11月7日）），俄国小说家、哲学家、政治思想家，也是非暴力的基督教无政府主义者和教育改革家。他是在托尔斯泰这个贵族家族中最有影响力的一位。

PROBLEMS: PHONETIC ASPECT

A big difference in phonetic inventories of the SAE languages and the Chinese

SAE	Chinese
Voiced VS Voiceless consonants	Aspirated VS Unaspirated consonants
Stress	Tone

A big level of homonymy in Chinese

- shí: 时 (a while), 十 (ten), 石 (stone)…
- shì: 是 (to be), 市 (town), 事 (case), 试 (to try), 世 (generation), 示 (to demonstrate)…

PROBLEMS: GRAPHICAL ASPECT

Absence of spaces between words

- 本宪章不得认为授权联合国干涉在本质上属于任何国家国内管辖之事件，且并不要求会员国将该项事件依本宪章提请解决；但此项原则不妨碍第七章内执行办法之适用。

No set of symbols used specifically for the phonetic loanwords

- 马里 (mǎlǐ) = "Horse mile/Inside horse" > Mali (African country) / Mary (city in Turkmenistan)

Ambiguity of the characters

- 乐 (yuè) = "music" / (lè) "happy"
- 了 (le) = PERF / (liǎo) = "to understand"

Our Hypothesis:

There is a particular pattern of transliteration for the Russian words in Chinese.

Our Aim:

To find it!

OUR TASKS

1. To check whether the transliteration trends of English and German are applicable to the Russian loanwords
   - Case study: to check whether there is influence of a Chinese-Russian pidgin on the current Russian loanwords' adaptation

2. To compare the prescribed transliteration rules (Xinhua) to the real data (Wikidata, dictionaries)

3. To analyze the cases of the partial semantic translation of the Named Entities (Wikidata, dictionaries)

4. To compare the most frequent N-grams in the Russian NE, Chinese NE and the reference corpus (Wikidata)

1. PHONETIC ADAPTATION OF THE RUSSIAN WORDS IN CHINESE: THE OT APPROACH

PHONETIC ADAPTATION OF THE RUSSIAN WORDS IN CHINESE: THE OT APPROACH

Based on the study of the English, German and Italian loanword adaptation in Chinese (Miao 2005)

- Alternation of consonant phonemes
- Transformation of consonant clusters

Made in the Optimality Theory paradigm

PHONETIC ADAPTATION OF THE RUSSIAN WORDS IN CHINESE: RESULTS

The main principles are confirmed on the Russian data:

- The crucial consonant feature is MANNER:
  - s => s (75%) >> ʂ (20%) >> ɕ (5%); *dʒ
  - t => tʰ (70%) >> t (20%) >> *! tsʲ (10%); *n

- The C deletion rate (Coda position):
  - Liquids >> plosives >> fricatives >> /m, n/

PHONETIC ADAPTATION OF THE RUSSIAN WORDS IN CHINESE: RESULTS

Hard VS Soft feature of Russian consonants

In general – does not affect the transformation

Soft velars – c, ɟ, ç:

- In Russian – palatal place and a fricative-like release
- In Chinese – substituted by affricates dʒʲ tsʲ ɕi
- detected in English as well, but more consistent in Russian

Soft dentals - dʲ , tʲ:

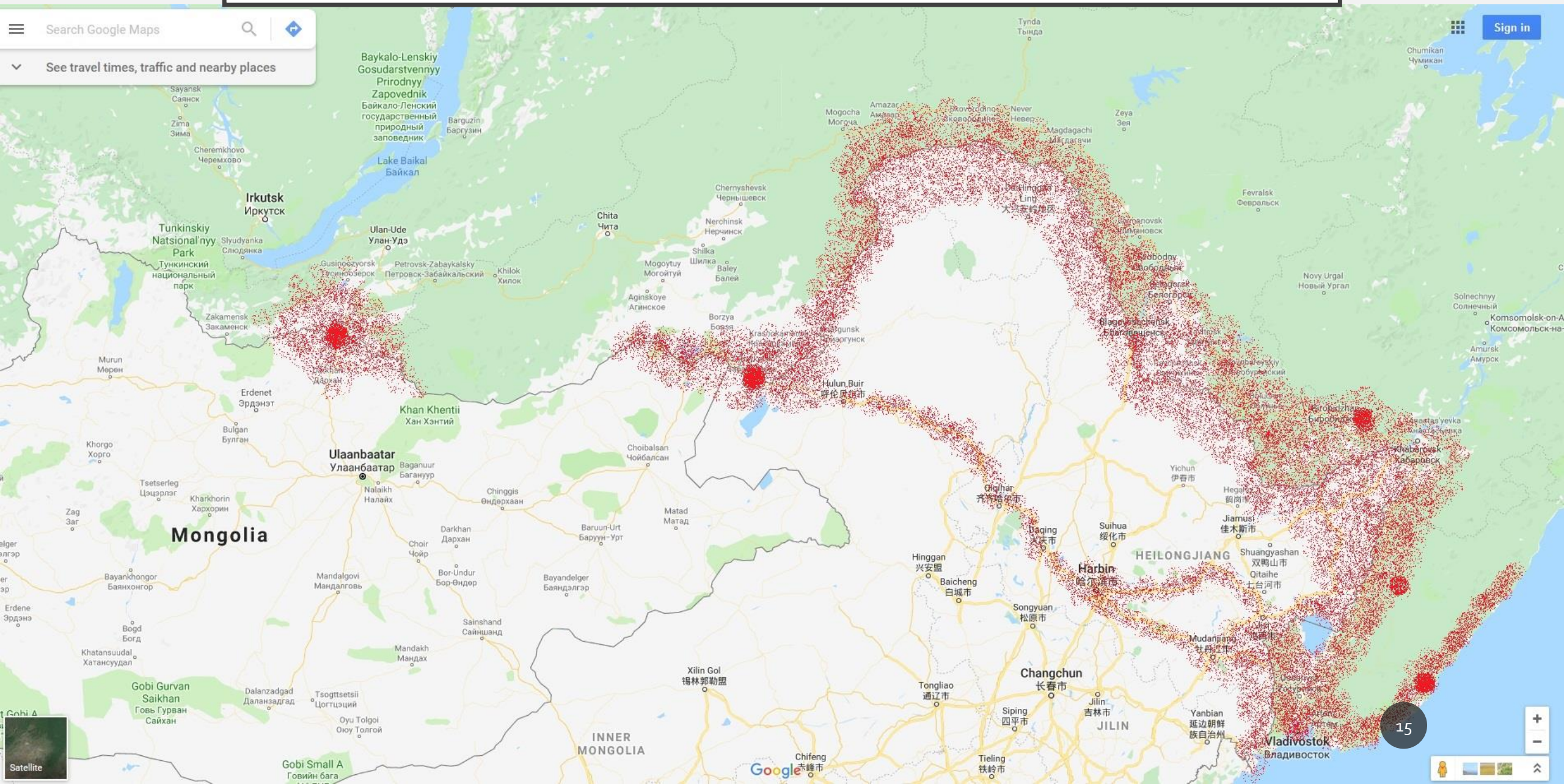
- In Russian - almost affricates dʒʲ tsʲ
- In Chinese - substituted by the corresponding affricates

CASE STUDY: CHINESE-RUSSIAN PIDGIN

CASE STUDY: CHINESE-RUSSIAN PIDGIN

- Russian has a different history of contact with Chinese
- Pidgin has existed since the end of XVIII cent.

CASE STUDY: CHINESE-RUSSIAN PIDGIN



CASE STUDY: HYPOTHESIS

The Chinese-Russian pidgin can influence:

- The phonetic appearance of the exact "old" loanwords
- The phonetic adaptation strategy for the new Russian loanwords

We can state it if:

- The phonetic adaptation differs from the usual Chinese one
- The phonetic adaptation is made according to the dialect phonology

CASE STUDY: RESULTS

The majority of the dialectal features from the standard Chinese is not observed in the loanwords

If it is, it can be usually explained by the OT assumptions:

- 哈拉嗦 hālāsuó instead of halaʂuo – predicted by the variation within one MANNER type

CASE STUDY: RESULTS

Two aspects likely to be transferred from pidgin:

- "ʤa-adaptation" of nouns – a very "pigdinish" feature

Russian	**Pidgin**	**Northern Dialects**	**Modern Chinese**
Купец	kupe-ʤa	谷瘪-子 gǔbiě-zi*	谷瘪-子 gǔbiě-zi*
Халат	Hala-ʤa	哈拉-子 hālā-zi*	哈拉-呢 hālā-nǐ

- The adaptation of a sound [z] is an affricate [ʤ] or [ʤ], which differs from the predicted by OT adaptation (same for pidgin and for contemporary Russian loans)

	Russian	**Pidgin**	**Modern Chinese**
XX cent. Borrowing	колхоз	kaxóʤə	科尔火支 kē'ěrhuǒzhī*
New Borrowing (NE)	Рязань	---	梁赞 liángzàn*

*by official "pinyin" transcription, "z" = [ʤ], "zh" = [ʤ]

2. PRESCRIPTIONS AND USE

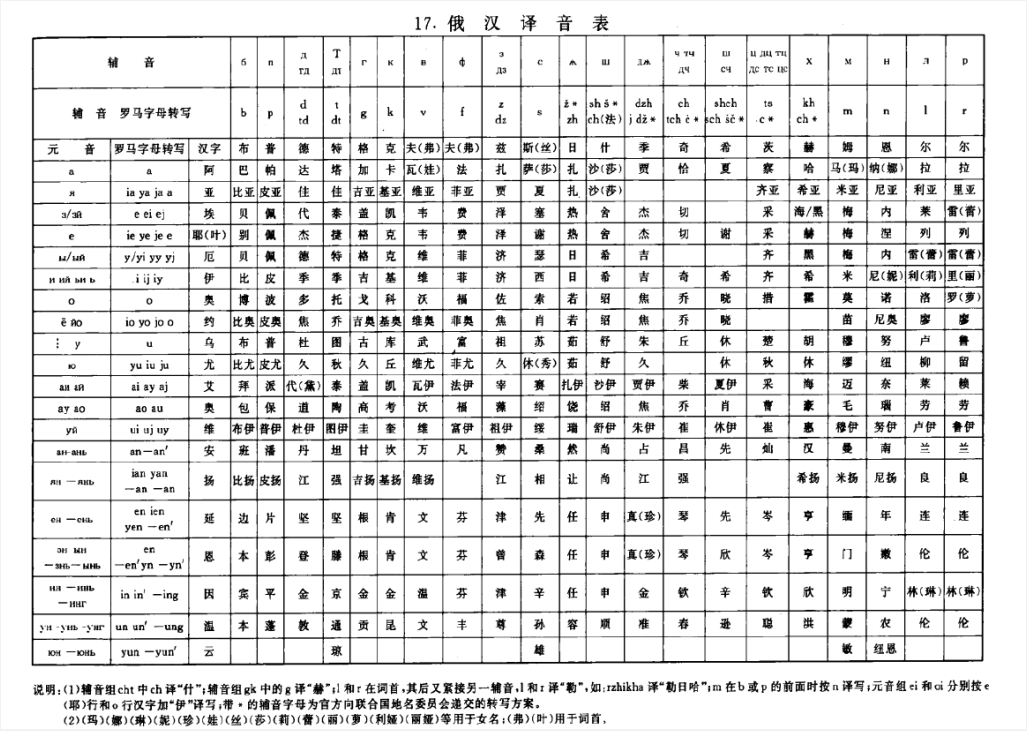
XINHUA PRESCRIPTIONS

XINHUA (新华社) IS THE BIGGEST OFFICIAL NEWS AGENCY IN PRC

RUSSIAN-CHINESE TRANSLATION DICTIONARY (1982)

Крутьковский 克鲁季科夫斯基
Крутяков 克鲁佳科夫
Круус 克鲁斯
Круцко 克鲁茨科
Кручек 克鲁切克
Кручёных 克鲁乔内赫
Крученюк 克鲁切纽克
Кручина 克鲁奇纳
Кручинин 克鲁奇宁
Кручинкин 克鲁钦金
Кручинов 克鲁奇诺夫
Кручинский 克鲁钦斯基
Крушанов 克鲁沙诺夫
Крушеван 克鲁舍万
Крушевский 克鲁舍夫斯基
Крушельницкий 克鲁舍利尼茨基
Крушинский 克鲁申斯基
Крывелев 克雷韦列夫
Крыгин 克雷金

RUSSIAN-CHINESE TRANSLITERATION TABLE (1993)



17. 俄 汉 译 音 表

辅 音			б	п	д тд	Т дт	г	к	в	ф	з дз	с	ж	ш	дж	ч тч дч	щ сч	ц дц тц дс тс цс	х	м	н	л	р
辅 音 罗马字母转写			b	p	d td	t dt	g	k	v	f	z dz	s	ž * zh	sh š * ch(兹)	dzh j dž *	ch tch č *	shch sch šč *	ts c *	kh ch *	m	n	l	r
元 音	罗马字母转写	汉字	布	普	德	特	格	克	夫(弗)	夫(弗)	兹	斯(丝)	日	什	季	奇	希	茨	赫	姆	恩	尔	尔
а	a	阿	巴	帕	达	塔	加	卡	瓦(娃)	法	扎	萨(莎)	扎	沙(莎)	贾	恰	夏	察	哈	马(玛)	纳(娜)	拉	拉
я	ia ya ja a	亚	比亚	皮亚	佳	佳	古亚	基亚	维亚	菲亚	贾	夏	扎	沙(莎)				齐亚	希亚	米亚	尼亚	利亚	里亚
э/эй	e ei ej	埃	贝	佩	代	泰	盖	凯	韦	费	泽	塞	热	舍	杰	切		采	海/黑	梅	内	莱	雷(蕾)
е	ie ye je e	耶(叶)	别	佩	杰	捷	格	克	韦	费	泽	谢	热	舍	杰	切	谢	采	赫	梅	涅	列	列
ы/ий	y/yi yy yj	厄	贝	佩	德	特	格	克	维	菲	济	瑟	日	希	吉			齐	黑	梅	内	雷(蕾)	雷(蕾)
и ий ы ь	i ij iy	伊	比	皮	季	季	吉	基	维	菲	济	西	日	希	吉	奇	希	齐	希	米	尼(妮)	利(莉)	里(丽)
о	o	奥	博	波	多	托	戈	科	沃	福	佐	索	若	绍	焦	乔	晓	措	霍	莫	诺	洛	罗(萝)
ё ЙО	io yo jo o	约	比奥	皮奥	焦	乔	吉奥	基奥	维奥	菲奥	焦	肖	若	绍	焦	乔	晓			苗	尼奥	廖	廖
у	u	乌	布	普	杜	图	古	库	武	富	祖	苏	茹	舒	朱	丘	休	楚	胡	穆	努	卢	鲁
ю	yu iu ju	尤	比尤	皮尤	久	秋	久	丘	维尤	菲尤	久	休(秀)	茹	舒	久		休	秋	休	缪	纽	柳	留
ай ай	ai sy aj	艾	拜	派	代(黛)	泰	盖	凯	瓦伊	法伊	宰	赛	扎伊	沙伊	贾伊	柴	夏伊	采	海	迈	奈	莱	赖
ау ао	ao au	奥	包	保	道	陶	高	考	沃	福	藻	绍	饶	绍	焦	乔	肖	曹	豪	毛	瑙	劳	劳
уй	ui uj uy	维	布伊	普伊	杜伊	图伊	圭	奎	维	富伊	祖伊	绥	瑞	舒伊	朱伊	崔	休伊	崔	惠	穆伊	努伊	卢伊	鲁伊
ан-ань	an—an′	安	班	潘	丹	坦	甘	坎	万	凡	赞	桑	然	尚	占	昌	先	灿	汉	曼	南	兰	兰
ян —янь	ian yan —an —an	扬	比扬	皮扬	江	强	吉扬	基扬	维扬		江	相	让	尚	江	强			希扬	米扬	尼扬	良	良
ен —ень	en ien yen —en′	延	边	片	坚	坚	根	肯	文	芬	津	先	任	申	真(珍)	琴	先	岑	亨	缅	年	连	连
эн ын —энь—ынь	ën —en′yn —yn′	恩	本	彭	登	滕	根	肯	文	芬	曾	森	任	申	真(珍)	琴	欣	岑	亨	门	嫩	伦	伦
ин —инь —инг	in in′ —ing	因	宾	平	金	京	金	金	温	芬	津	辛	任	申	金	钦	辛	钦	欣	明	宁	林(琳)	林(琳)
ун —унь —унг	un un′ —ung	温	本	蓬	敦	通	贡	昆	文	丰	蒋	孙	容	顺	准	春	逊	聪	洪	蒙	农	伦	伦
юн —юнь	yun —yun′	云				琼						雄								敏	纽恩		

说明：(1)辅音组 cht 中 ch 译"什"；辅音组 gk 中的 g 译"赫"；l 和 r 在词首，其后又紧接另一辅音，l 和 r 译"勒"，如：rzhikha 译"勒日哈"；m 在 b 或 p 的前面时按 n 译写；元音组 ei 和 ei 分别按 e (耶)行和 o 行汉字加"伊"译写；带 ′ 的辅音字母为官方向联合国地名委员会递交的转写方案。
(2)(玛)(娜)(琳)(妮)(珍)(娃)(丝)(莎)(莉)(蕾)(丽)(萝)(利娅)(丽娅)等用于女名；(弗)(叶)用于词首。

But are they really valid?..

XINHUA ALGORITHM

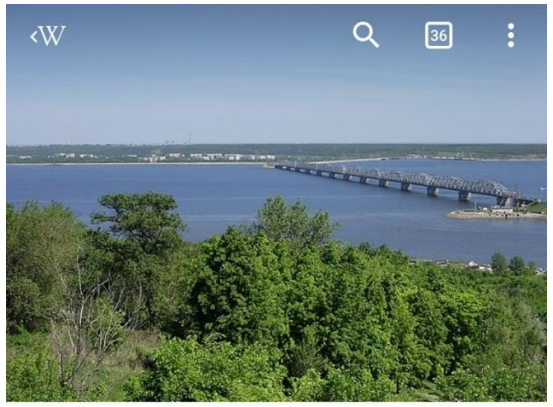
Python code

- Regular expressions
- “if”- “else” constructions
- Pandas library
  - .csv table with prescriptions

ALGORITHM APPLICATION

Wikidata

DATA: WIKIPEDIA OBJECTS



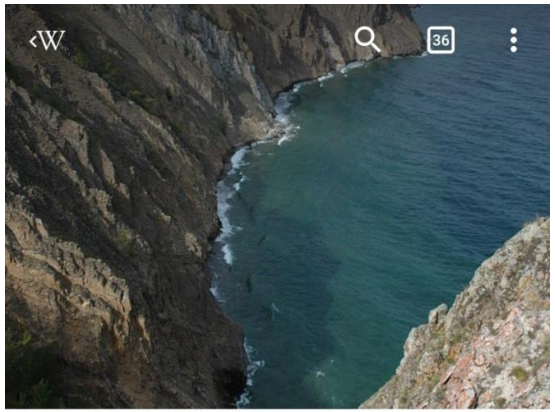
伏尔加河

添加标题描述

伏尔加河（俄语：Волга，又译窝瓦河），位于俄罗斯西南部，全长3,692公里，是欧洲最长的河流[2]，也是世界最长的内流河[2]，流入里海。

快速预览: 国家, 城市 ...

伏尔加河是欧洲流域最广以及流量最大的河流，流经欧洲俄罗斯，是代表型的俄罗斯河



贝加尔湖

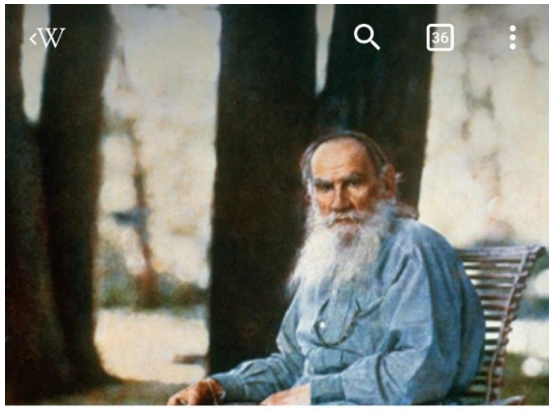
俄罗斯湖泊

贝加尔湖（俄语：óзеро Байка́л，罗马化：Ozero Baykal，IPA：[ˈozʲɪrə bɐjˈkalʲ]；布里亚特语：Байгал нуур, 拉丁转写：Baygal nuur，蒙古语：Байгал нуур；意思是"自然之湖"[3]；一说名称来源于"贝音嘎　嘎拉"（蒙古语意为不灭的火焰）[4]）。汉朝人称之为"**翰海**[5]"，五胡十六国时北朝叫"**于巳尼大水**"，隋唐叫"**小海**"，18世纪初期的《异域录》称之为"**柏海儿湖**"，

莫斯科

俄羅斯首都

莫斯科（俄语：Москва，罗马化：Moskva，IPA：[mɐsˈkva]）是俄罗斯首都与最大都市、以及莫斯科州首府，为俄罗斯全国政治、经济、科学、文化及交通的中心。面积2,510平方公里，与莫斯科州和卡卢加州接壤。城区人口约1200万，是欧洲人口第二多的城市，仅次于伊斯坦布尔，占俄罗斯总人口的1/10。



列夫·托尔斯泰

俄國作家

列夫·尼古拉耶维奇·托尔斯泰（俄语：**Лев Николаевич Толстой**，拉丁化：Lev Nikolayevich Tolstoy；1828年9月9日（儒略历8月28日）–1910年11月20日（儒略历11月7日）），俄国小说家、哲学家、政治思想家，也是非暴力的基督教无政府主义者和教育改革家。他是在托尔斯泰这个贵族家族中最有影响力的一位。

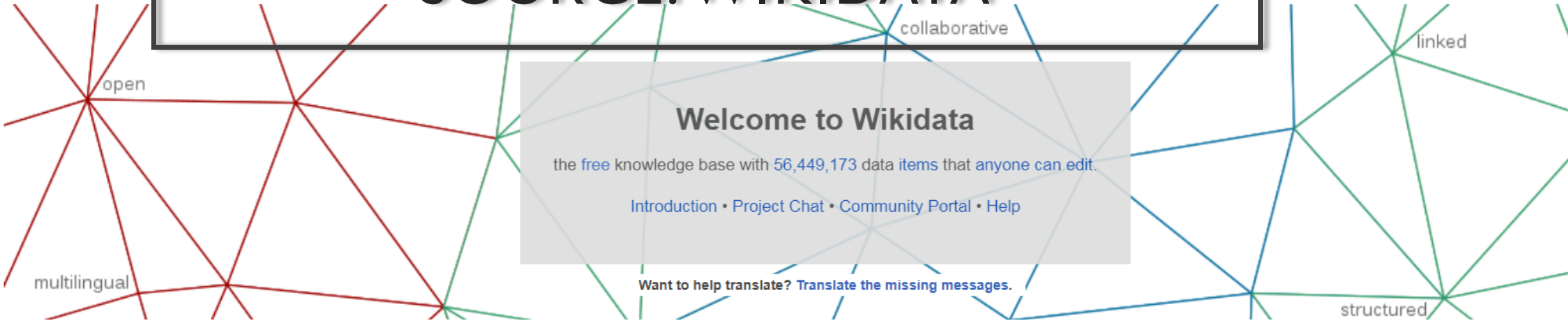
https://www.wikidata.org/wiki/Wikidata:Main_Page

SOURCE: WIKIDATA

WIKIDATA

English | Not logged in | Talk | Contributions | Cre

Main Page | Discussion | Read | View source | View history | Search Wikidata



Welcome to Wikidata

the free knowledge base with 56,449,173 data items that anyone can edit.

Introduction • Project Chat • Community Portal • Help

Want to help translate? Translate the missing messages.

Main page
Community portal
Project chat
Create a new Item
Create a new Lexeme
Recent changes
Random item
Query Service
Nearby
Help
Donate

Print/export
Create a book
Download as PDF
Printable version

In other projects
Wikimedia Commons
MediaWiki
Meta-Wiki
Wikispecies
Wikibooks
Wikimania
Wikinews
Wikipedia
Wikiquote
Wikisource
Wikiversity
Wikivoyage
Wiktionary

Tools
What links here
Related changes
Special pages
Permanent link
Page information
Wikidata item

In Wikipedia

Welcome!

Wikidata is a free and open knowledge base that can be read and edited by both humans and machines.

Wikidata acts as central storage for the **structured data** of its Wikimedia sister projects including Wikipedia, Wikivoyage, Wiktionary, Wikisource, and others.

Wikidata also provides support to many other sites and services beyond just Wikimedia projects! The content of Wikidata is available under a free license, exported using standard formats, and can be interlinked to other open data sets on the linked data web.

Get involved

Learn about Wikidata

- What is Wikidata? Read the **Wikidata introduction**.
- Explore Wikidata by looking at a featured showcase item for author **Douglas Adams** (Q42).
- Get started with Wikidata's **SPARQL query service**.

Contribute to Wikidata

- Learn to edit Wikidata: follow the **tutorials**.
- Work with other volunteers on a subject that interests you: **join a WikiProject**.
- Individuals and organisations can also **donate data**.

Meet the Wikidata community

Learn about data

New to the wonderful world of data? Develop and improve your data literacy through content designed to get you up feeling comfortable with the fundamentals in no time.

Item: *Earth* (Q2) | Property: *highest point* (P610) | custom value: *Mount Everest* (Q513)

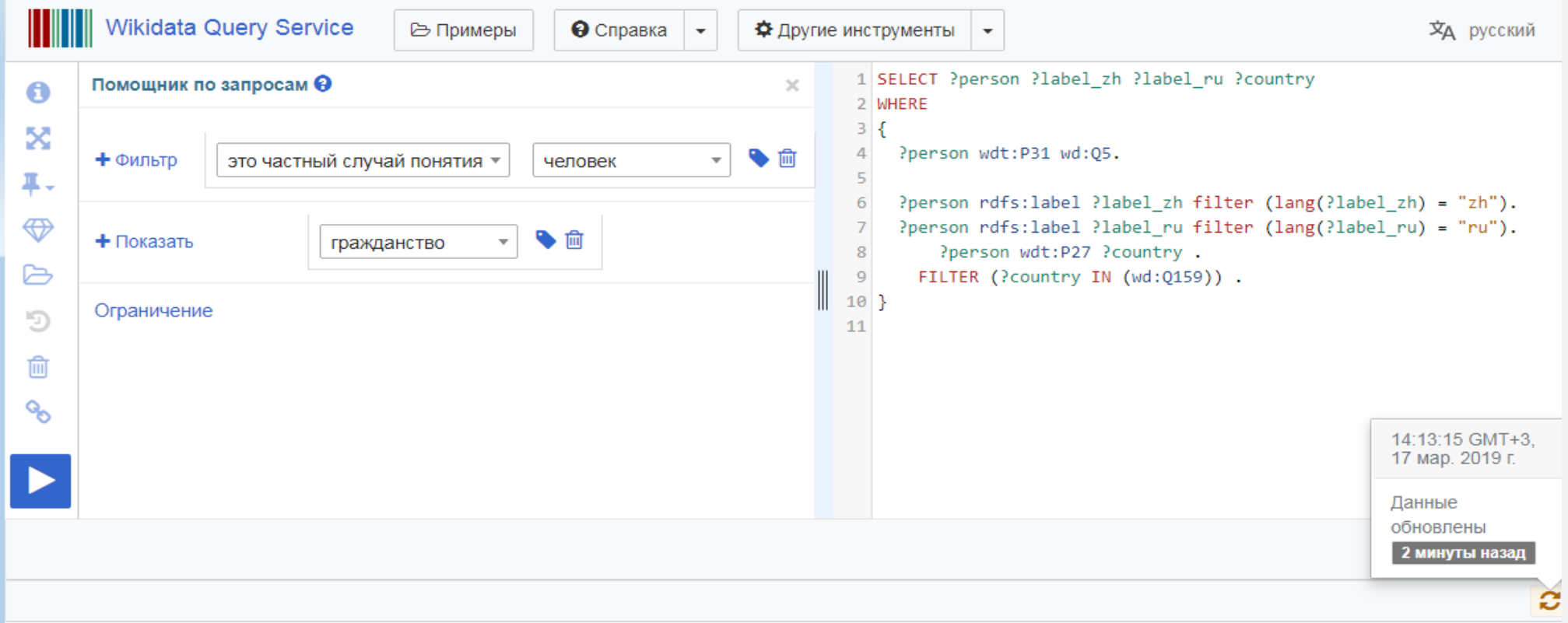
Current highlights

- Avengers: Endgame (Q23781155)
- Red Kelly (Q706565)
- The Visitation (Q63444435)
- Terje Moe Gustavsen (Q4569851)

24

На месте жесткой пос
Шереметьево найдены

snob.ru



2327 результатов за 5114 мс Код Скачать Ссылка

Search

person	label_zh	label_ru	country
wd:Q4864	尤金·卡巴斯基	Касперский, Евгений Валентинович	wd:Q159
wd:Q162646	瓦迪姆·博羅夫謝夫	Вадим Владимирович Бровцев	wd:Q159
wd:Q163256	瓦莲京娜·古尼娜	Гунина, Валентина Евгеньевна	wd:Q159
wd:Q164309	法祖·阿里耶娃	Фазу Гамзатовна Алиева	wd:Q159
wd:Q122425	安德烈·萨金塞夫	Андрей Петрович Звягинцев	wd:Q159
wd:Q166498	維薩里翁	Тороп, Сергей Анатольевич	wd:Q159
wd:Q167113	卡倫·沙赫納扎羅夫	Карен Шахназаров	wd:Q159

DATASET

Parameters:

Categories:

- Country
- Type of an object

Words:

- Russian proper names
- Wikidata transliteration
- Xinhua-based transliteration

Metrics:

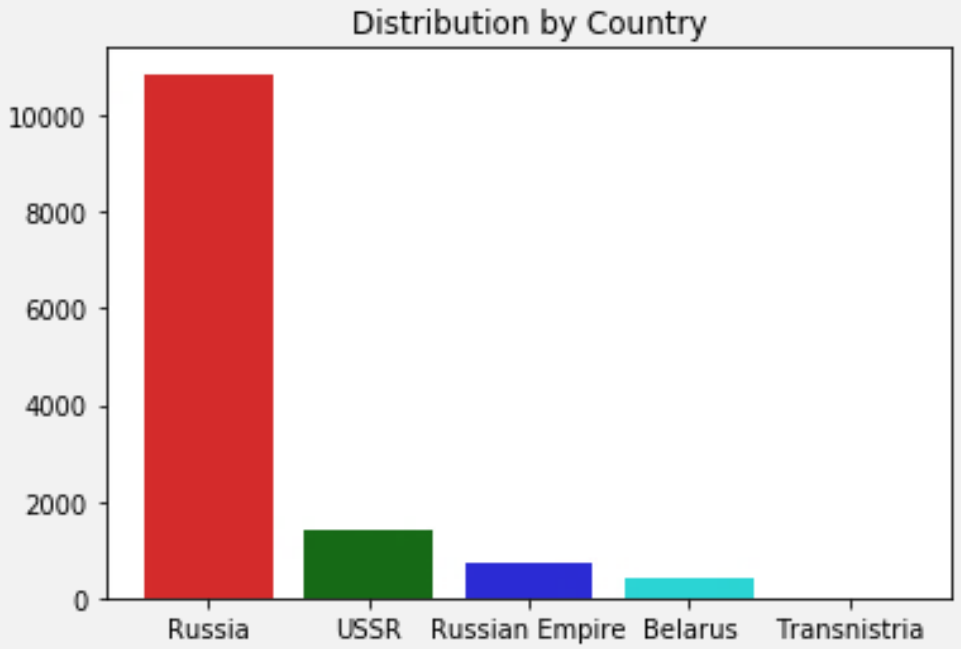
- Absolute Levenstein distance
- Normalized Levenstein distance
- Jaccard index

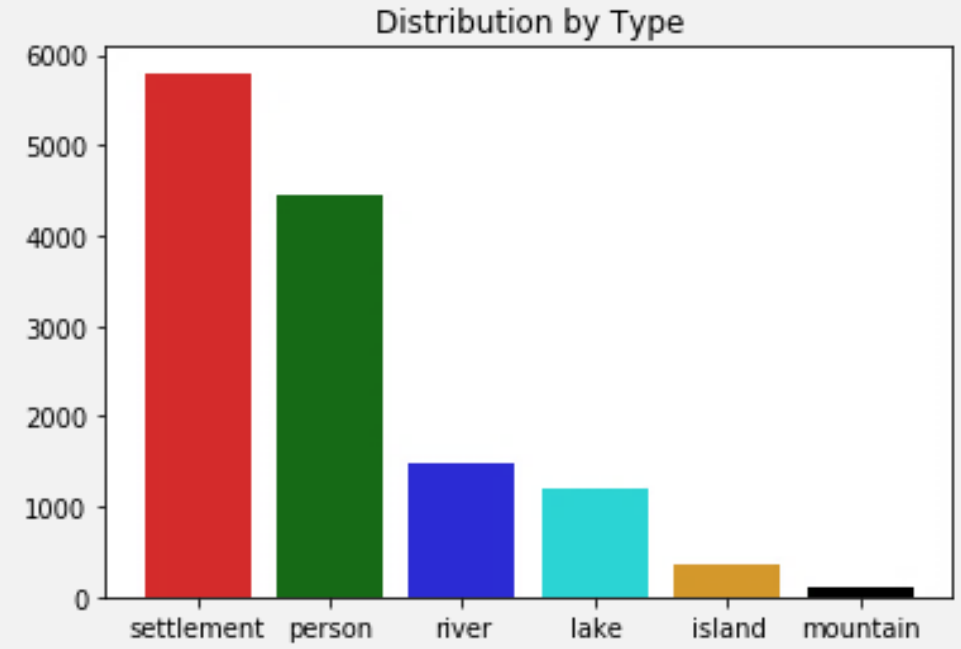
Country sample:

- Russian Federation
- Transnistria (Приднестровье)
- The USSR
- The Russian Empire
- Belarus
- Ukraine – pilot study

Type sample:

- Lakes
- Rivers
- Islands
- Mountains
- Settlements
- Person names





DATASET: OVERVIEW

13410 objects

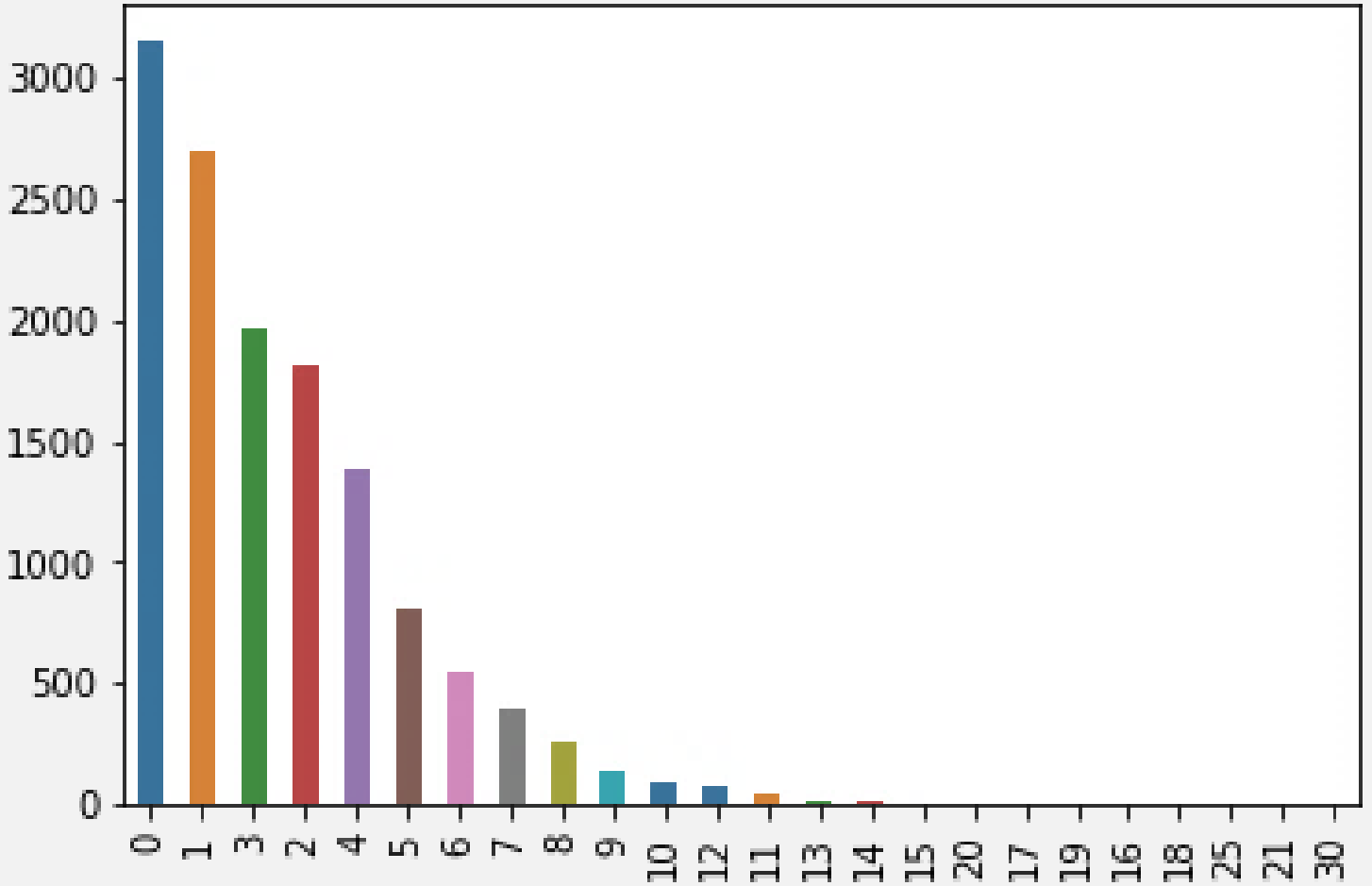
Countries: 81% - Russia, 10% - USSR

Types: 43% - settlements, 33% - names of persons, 11% - rivers

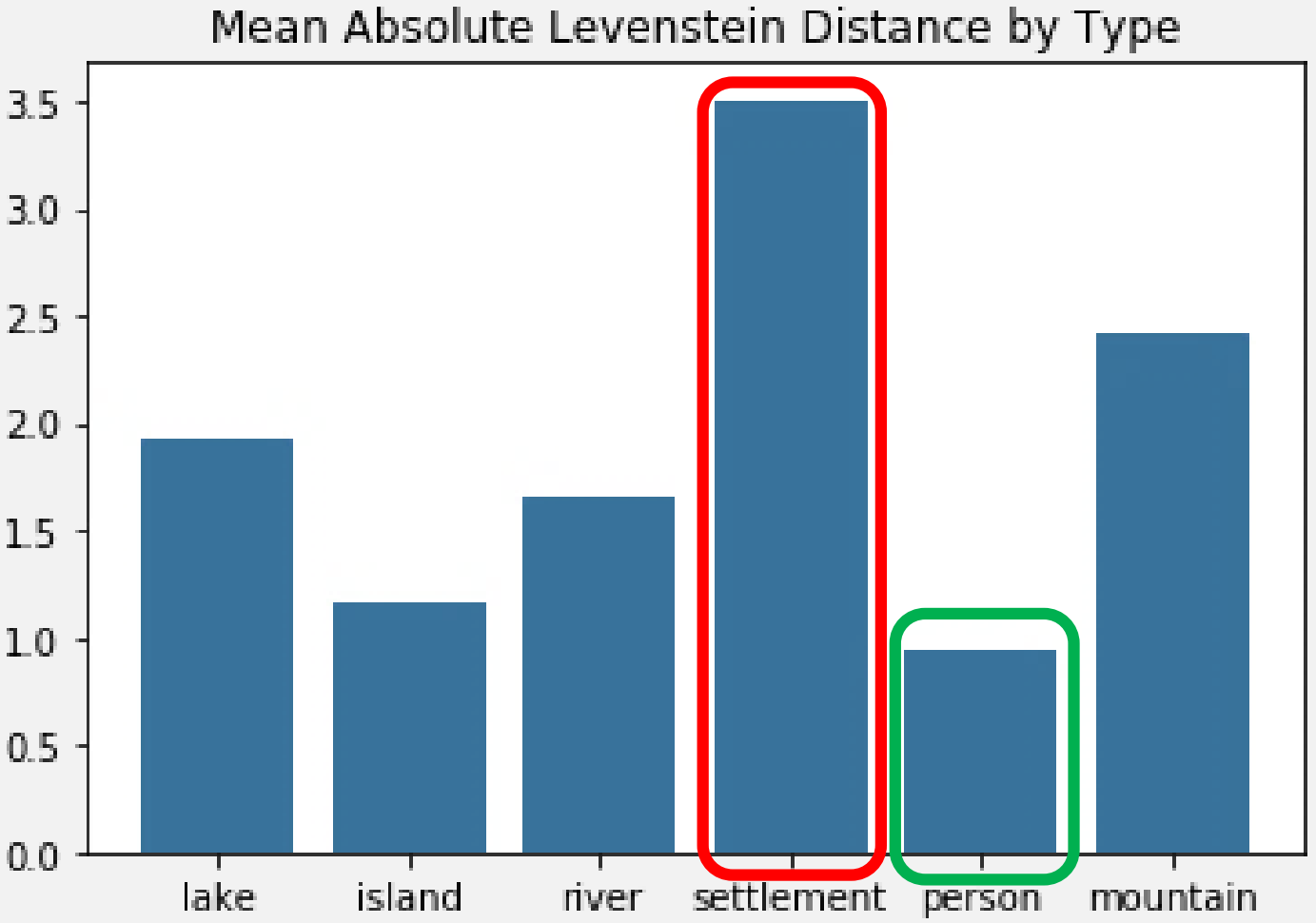
DATASET: STUDY

Levenstein distance:

- Absolute: [0, ∞]; 0 – identical strings,
- 1-∞ - number of symbols to change



DATASET: STUDY



Classifiers?..

DATASET: RESULTS

Xinhua rules are not so bad for personal names…

…but are hardly applicable to other types of objects

There might be some semantic elements which are translated to Chinese…

3. SEMANTIC ELEMENTS IN TRANSLITERATION

CLASSIFIERS: OVERVIEW

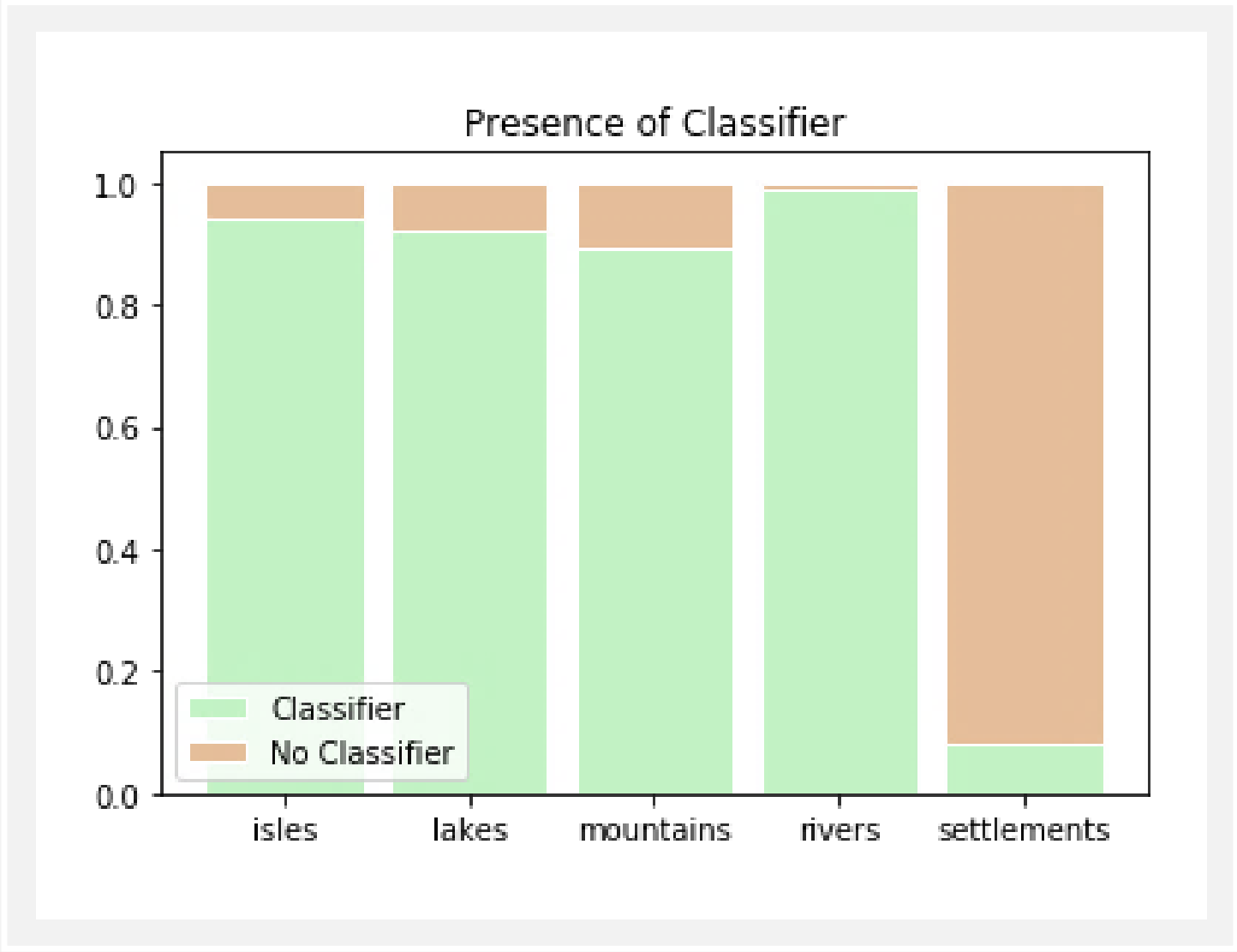
One syllable

In the end of the word

No classifiers for personal names

- 伏尔加河 fú'ěrjiā hé
- 奥涅加湖 àonièjiā hú
- 萨哈林岛 sàhālín dǎo
- 弗拉基米尔市 fúlājīmǐ'ěr shì

CLASSIFIERS: FREQUENCY



IRRESISTIBLE COMPULSION TO SEMANTIC ELEMENTS

Novi Zeland VS New York

Новая Зеландия VS Нью-Йорк

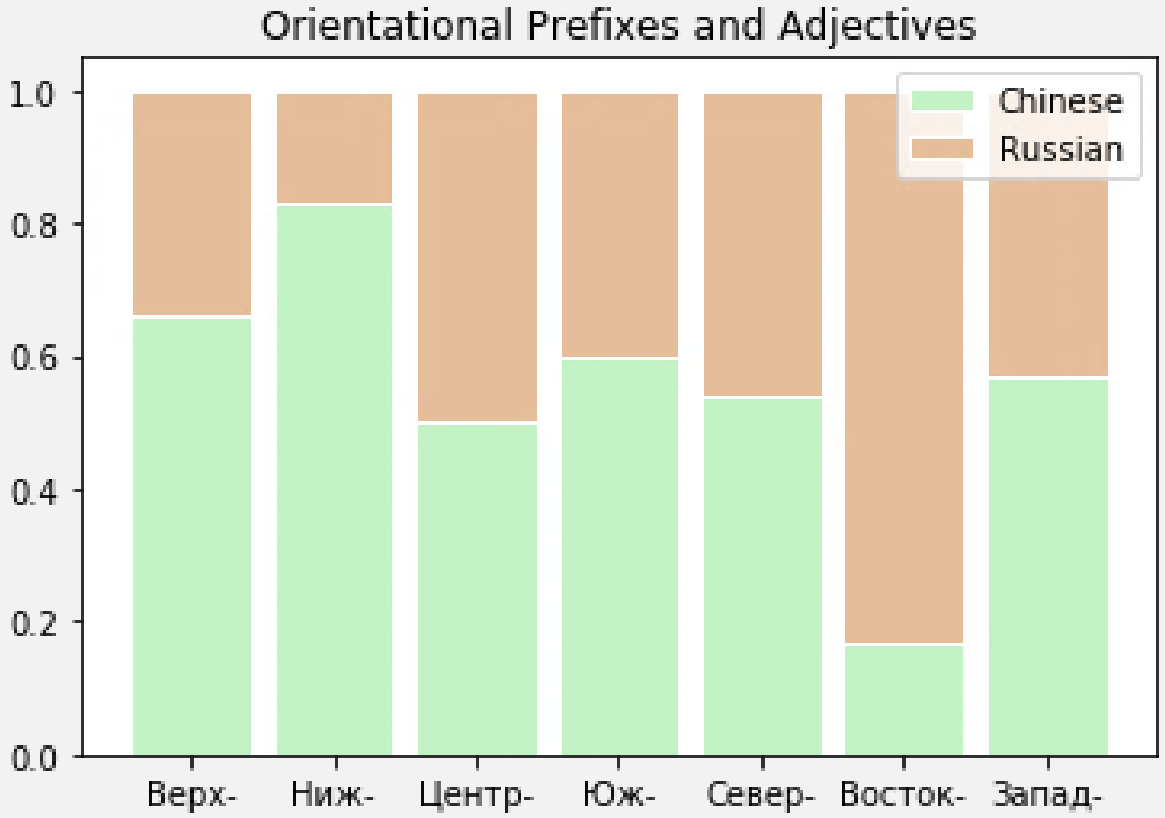
How strong is it in Chinese?

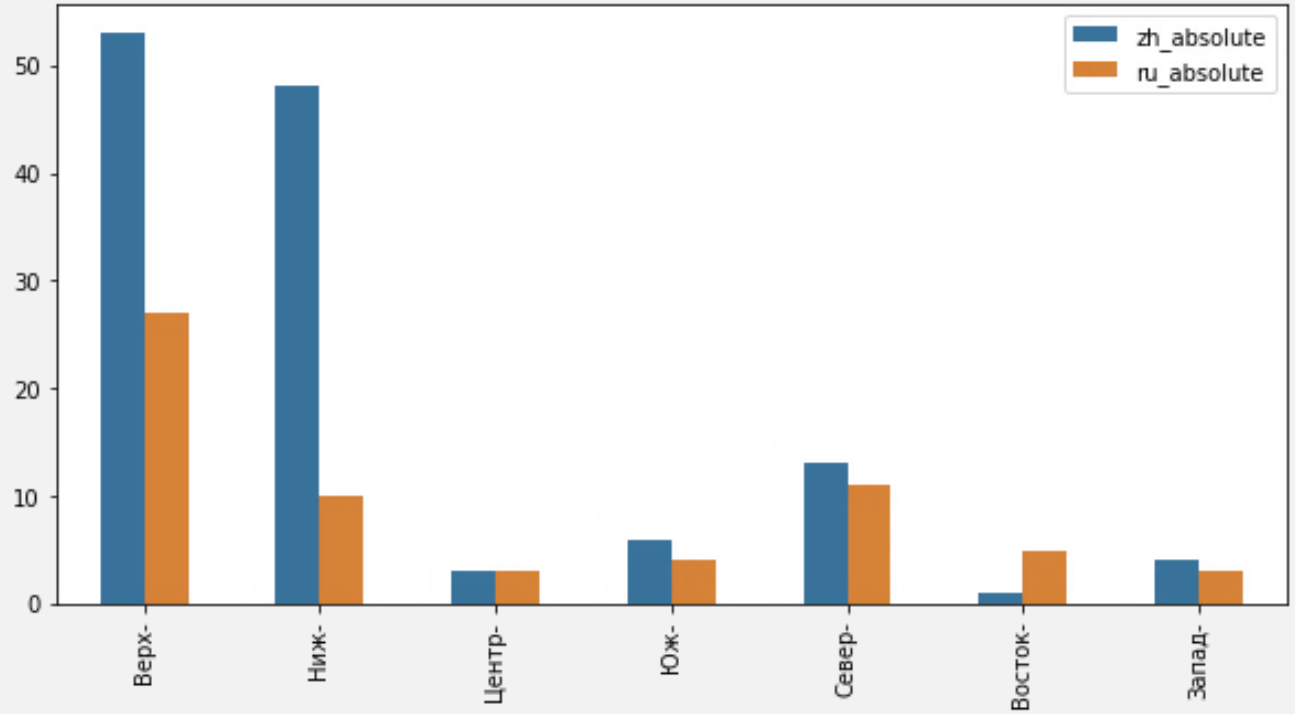
PARTIAL SEMANTIC TRANSLATION

Based on the rough data overview, we split the possible variants into 3 groups:

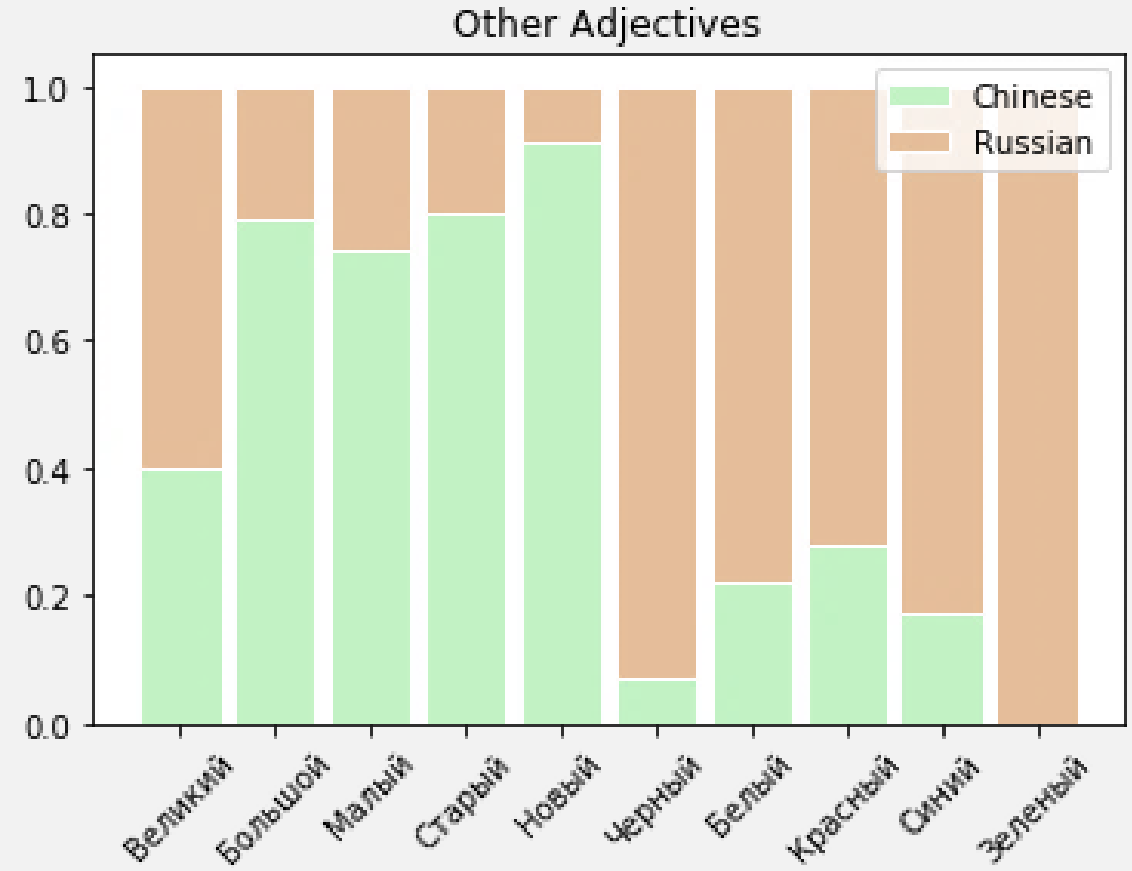
- "Spatial" Adjectives ("Higher", "Southern", "Central" etc.)
- Other Adgectives ("Greater", "New", "Red" etc.)
- Affixes ("Trans-", "-upon-" etc.)

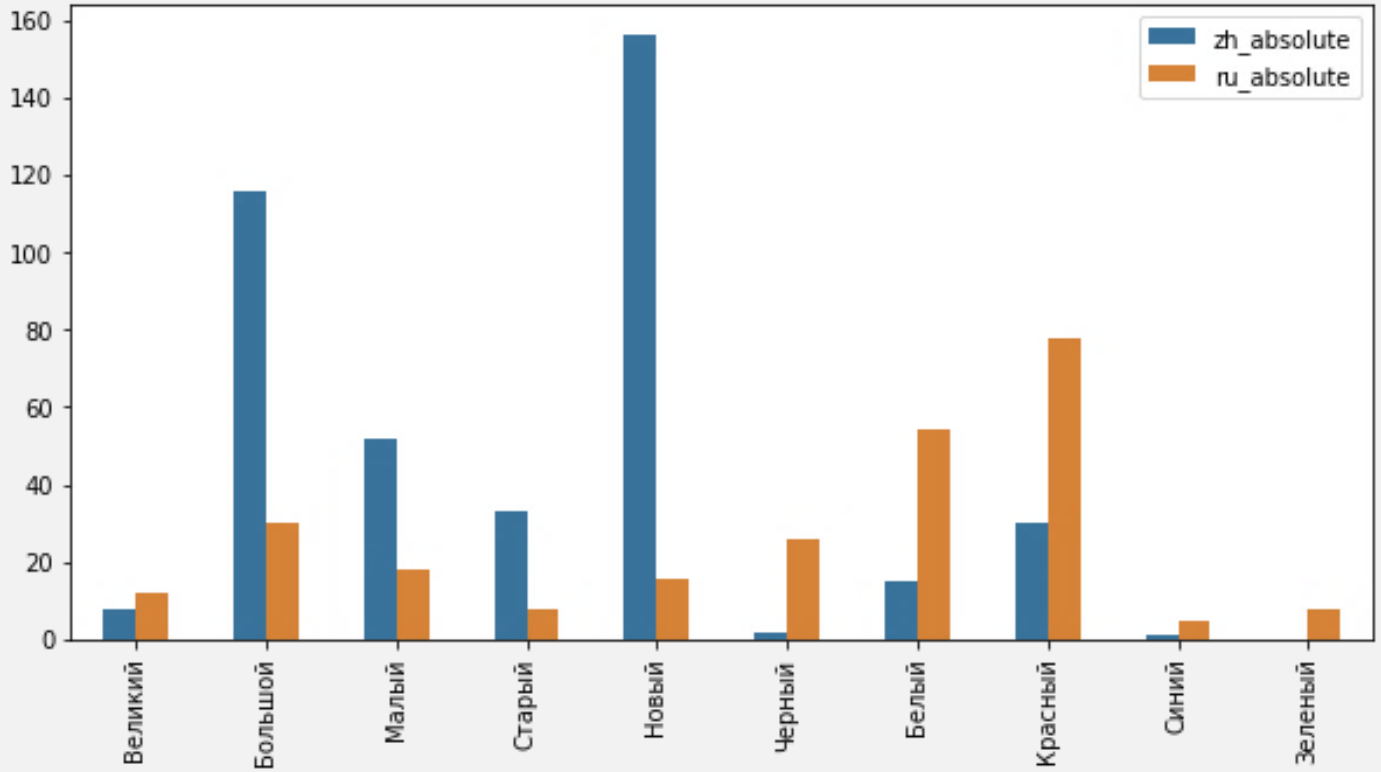
PARTIAL SEMANTIC TRANSLATION: “SPATIAL” ADJECTIVES





PARTIAL SEMANTIC TRANSLATION: OTHER ADJECTIVES





PARTIAL SEMANTIC TRANSLATION: AFFIXES

The least likely to be semantically transformed

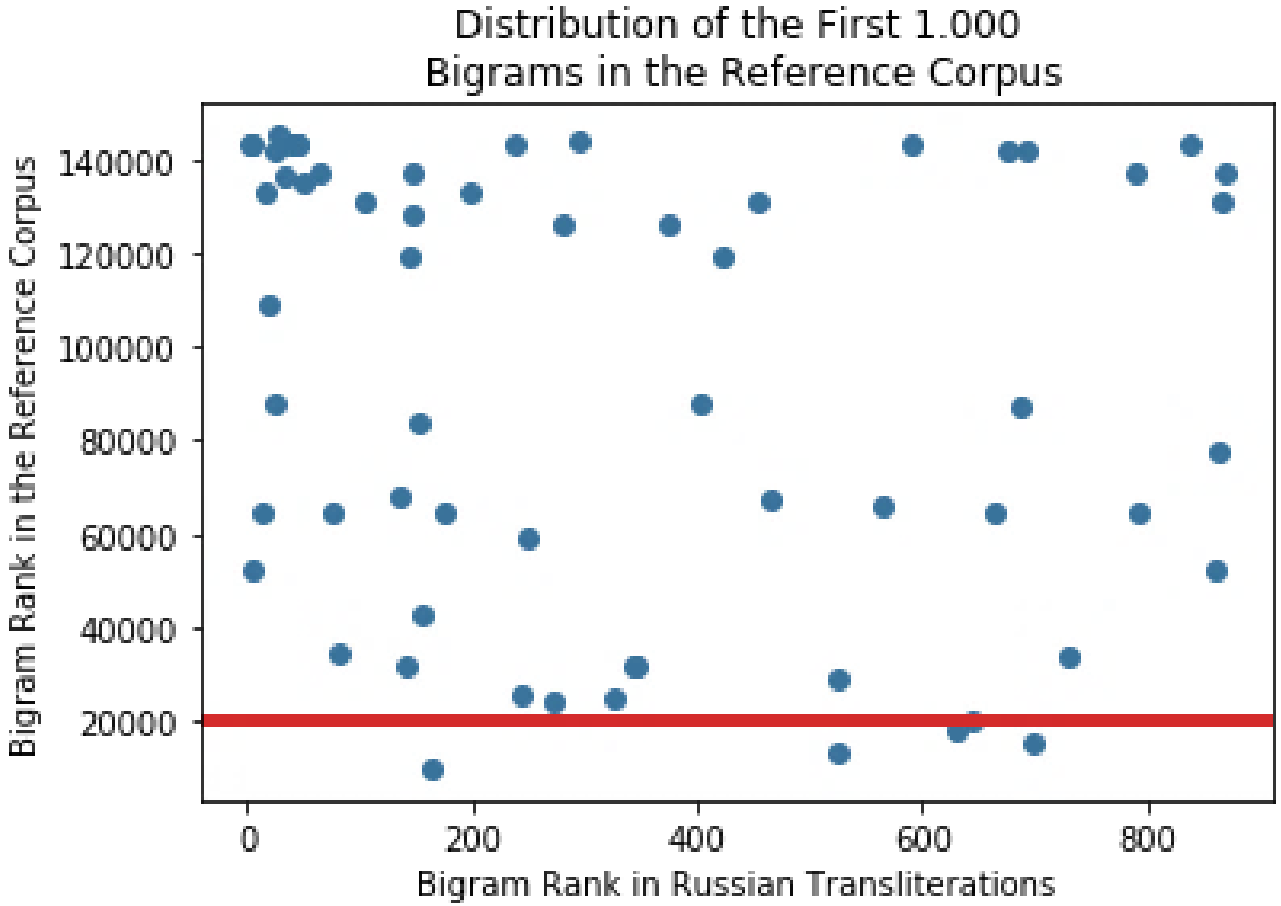
Two occurrences:

- Забайкальск > 后-贝加尔斯克 > hòu (behind) -bèijiā'ěrsīkè
- Ростов-на-Дону > 顿河畔罗斯托夫 > dùnhé pàn luósītuōfū

4. N-GRAM FREQUENCY

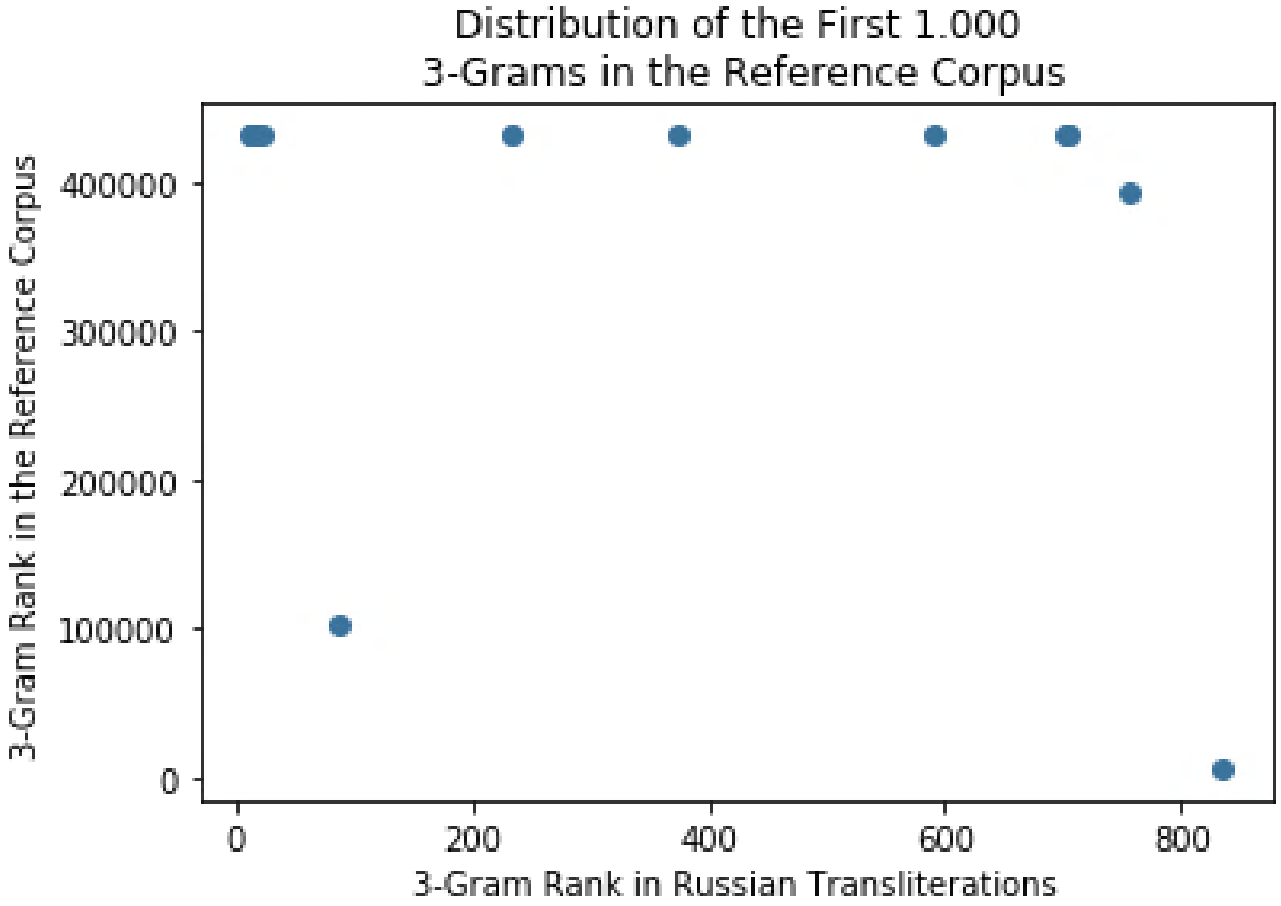
BIGRAMS

936 of Bigrams are not met in the reference corpus at all



3-GRAMS

988 of 3-Grams are not met in the reference corpus at all



RESULTS

- Generally the trends of Russian loanwords' phonetic adaptation are similar to those of other European languages
- But the official (Xinhua) prescriptions do not cover them
- The Chinese meaningful characters can be an efficient hint in limiting the loanwords
  - Adjective-like affixes – from the left
  - Classifiers (except personal names) – from the right
- The majority of N-grams in transliterations are rare in the usual Chinese texts

FUTURE PERSPECTIVES

Implementation of the statistics in Chinese NLP algorithms

- Hidden Markov models
- Seq2Seq Neural networks

Analysis of a bigger dataset

- BaiduPedia (百度百科) – bigger than Wikipedia in 5 European languages altogether
- More oriented on PRC

conduct the same research on other European languages

- Experts in these language are needed – we invite you to take part!

Thank you for your attention!

Hvala na pažnji!

非常感谢！

Kirill Semenov,
HSE – Moscow
kir.semenow@yandex.ru

ADDITIONAL MATERIALS

PROBLEMS: ALL OVER RUSSIA

* 红场 - hóng chǎng – Red Square

红肠 - hóng cháng – Red Guts

PHONETIC ADAPTATION OF THE RUSSIAN WORDS IN CHINESE: THE OT APPROACH

- Sources:
  - 汉语外来词词典 (Chinese Loanword Dictionary), 1984 – 387 words
  - БКРС-Online (Big Chinese-Russian Dictionary Online) – 1494 items

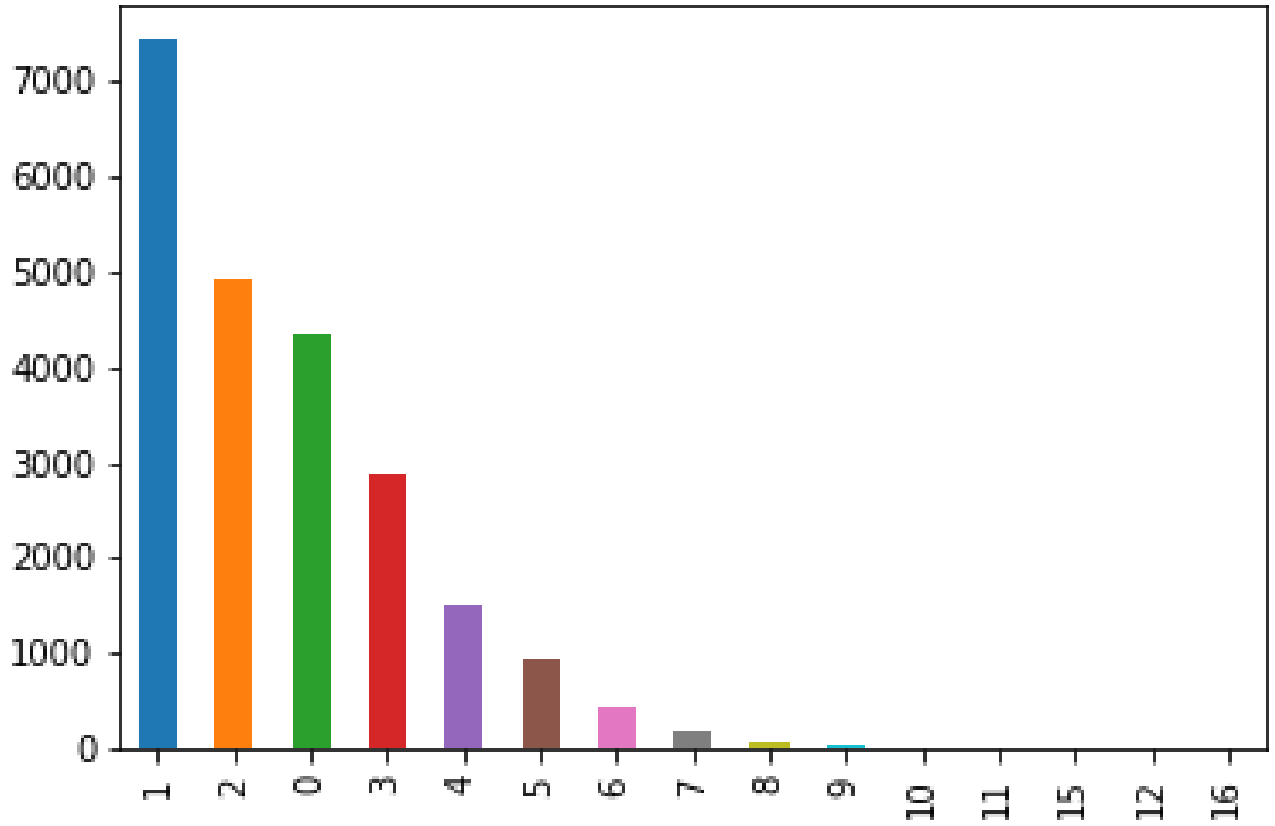
CASE STUDY: MATERIAL

Dictionaries and grammar of Russian-Chinese pidgin (Perekhval'skaya 2008)

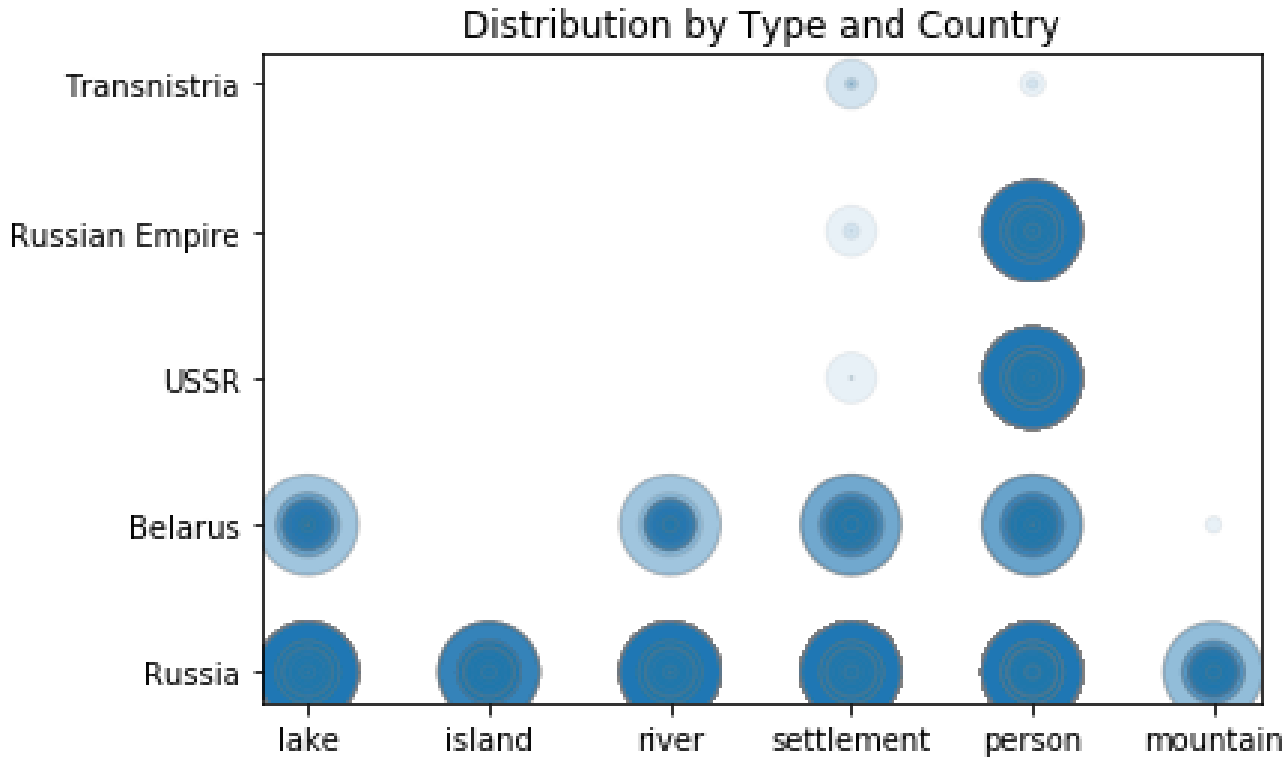
Grammar of the Northern Chinese dialects (Zavyalova 1996)

Dictionaries of the Northern Chinese dialects and of the modern Chinese

LEVENSTEIN DISTANCE FOR UKRAINIAN SETTLEMENTS



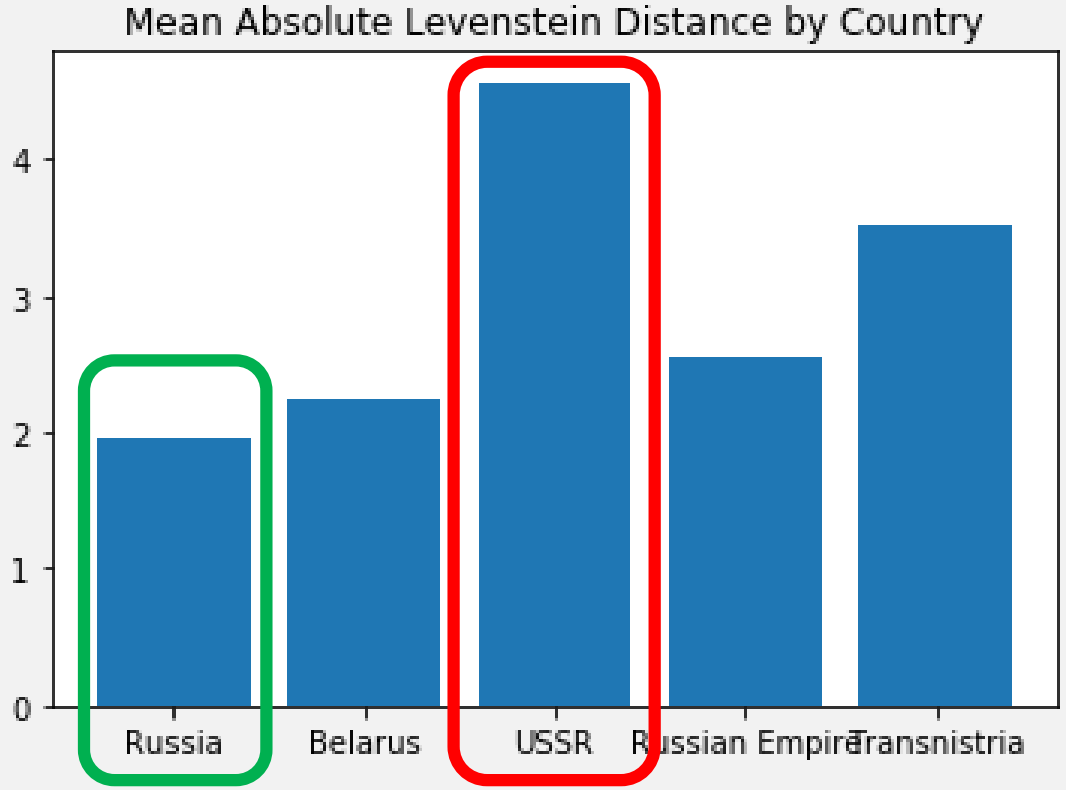
DATASET: OVERVIEW



DATASET: STUDY

Character diversity in Wikidata items	**877**
Character diversity of Xinhua-based transliterations	260

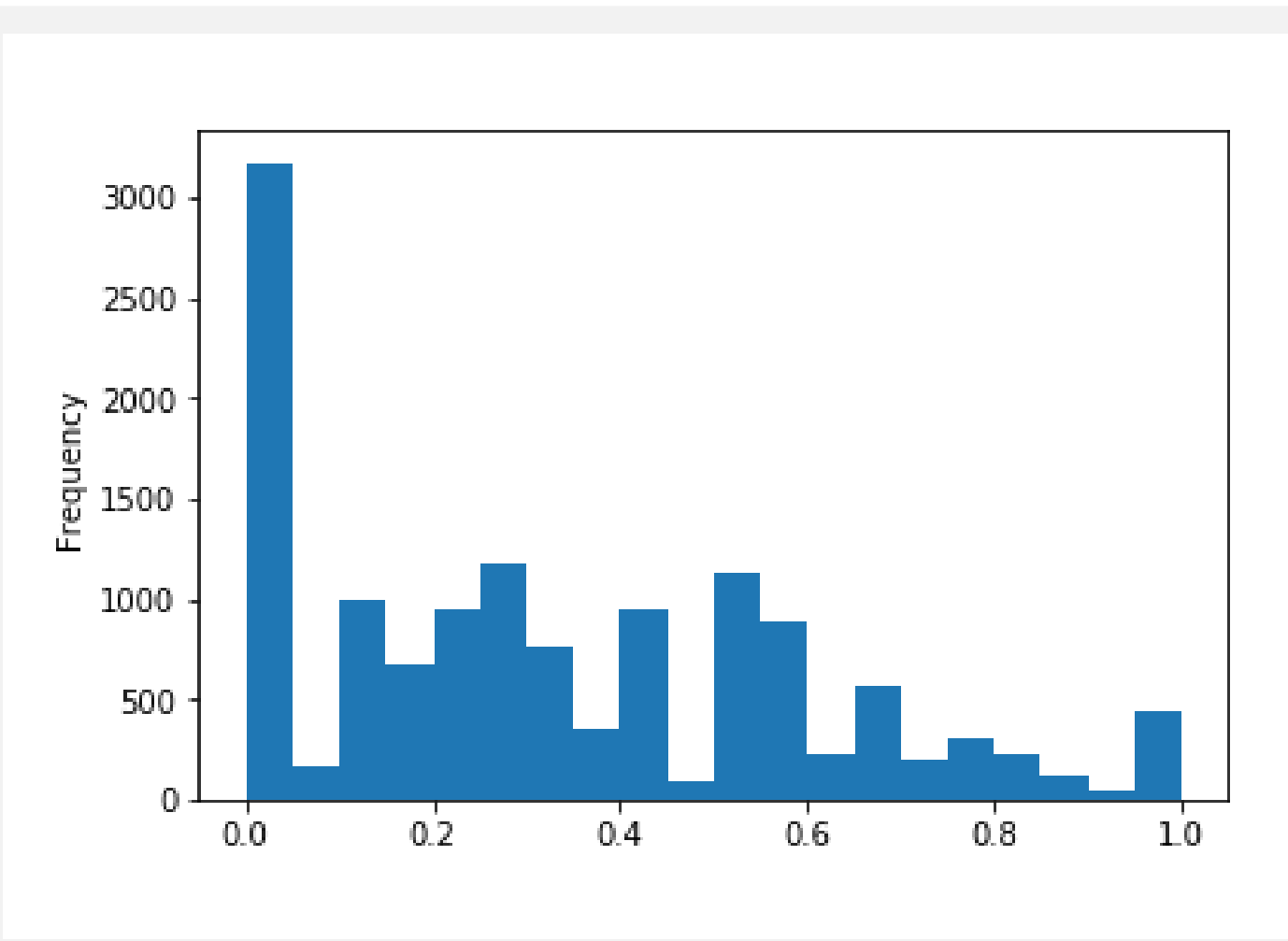
DATASET: DISTRIBUTION BY TYPE



Communist names?..

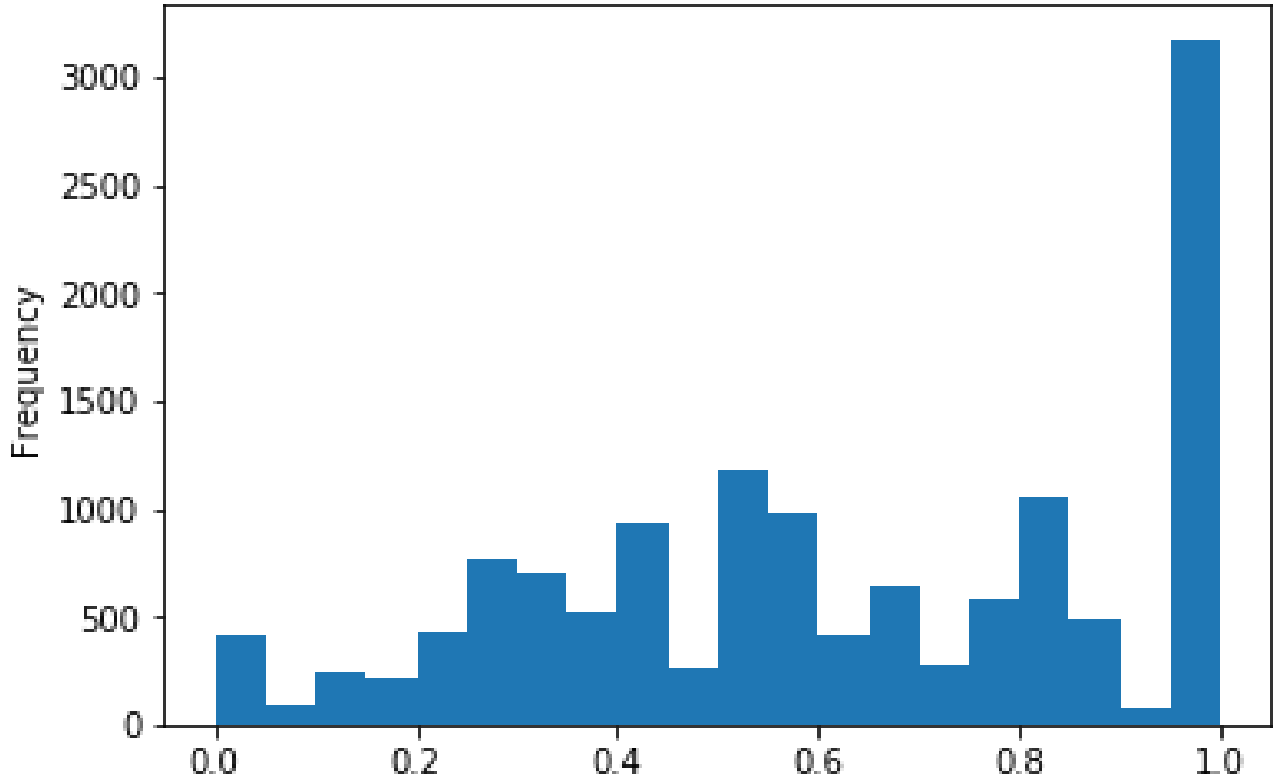
DATASET: STUDY

- Normalized Levenstein distance:
- [0, 1]; 0 – identical strings, 1 – totally different strings

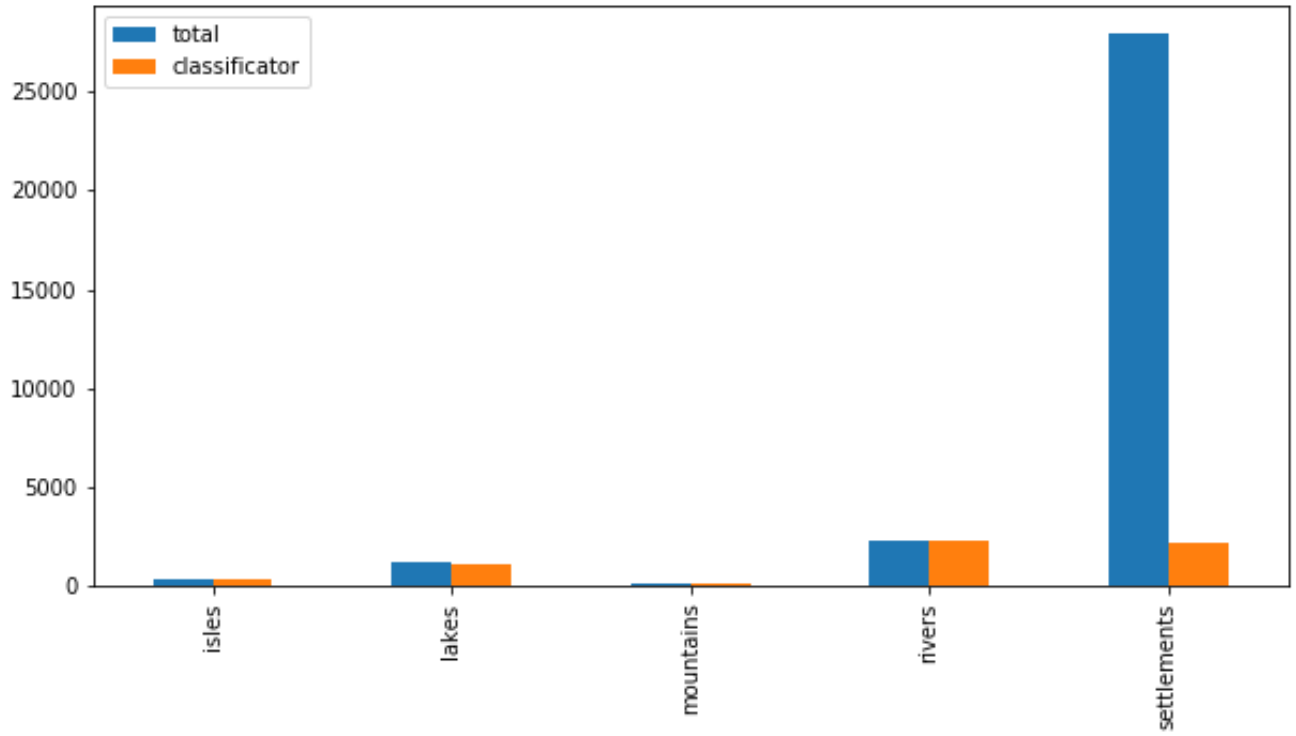


DATASET: STUDY

Jaccard Index: [0, 1]; 0 – the same set of symbols, 1 – no overlap



CLASSIFIERS: ABSOLUTE NUMBERS



TOTAL SEMANTIC TRANSLATIONS

- East Asian Toponyms
  - Especially former Japanese territories – Kuril islads, Sakhalin, etc.
  - On-border settlements:
    - 海参崴 hǎishēnwǎi (Sea cucumber river bend) = Владивосток
- Communist Toponyms
  - Октябрьское, Первомайское, etc.
- Other literal translations:
  - Белая (деревня)
  - Старое (село)
  - Аэропорт (район)
  - Чистые пруды

N-GRAMS: CHINESE WIKIDATA 3-GRAMS

维奇 · 739	历山大 253	诺耶$ 164
夫卡$ 491	诺夫$ 251	$弗拉 160
斯基$ 466	耶维奇 242	洛夫$ 157
夫斯$ 434	山大 · 231	诺维奇 154
科耶$ 410	夫斯基 215	$尼古 148
斯克$ 396	尼古拉 210	拉基米 146
亚历山 355	米哈伊 200	尔盖 · 146
斯科耶 350	耶夫$ 189	弗拉基 145
$亚历 280	谢尔盖 185	罗维奇 142
科夫$ 273	$谢尔 182	米尔 · 137

N-GRAMS: RUSSIAN WIKIDATA 4-GRAMS

кий$ 2550	нско 655	екса 376
ский 2508	ович 583	енск 344
кое$ 2204	андр 536	ий$к 339
ское 2100	лекс 525	евич 330
овск 1385	алек 521	анов 319
вско 937	$але 421	ова$ 318
инск 904	евск 408	ков$ 300
вски 864	ксан 390	новс 295
нски 834	санд 379	ковс 285
вич 753	нико 379	ое$к 279

ALGORITHM APPLICATION: PART I

- Chinese Loanwords' Dictionary (汉语外来词词典), Shanghai, 1984
- 378 words of Russian origin
  - Or borrowed into Chinese via Russian

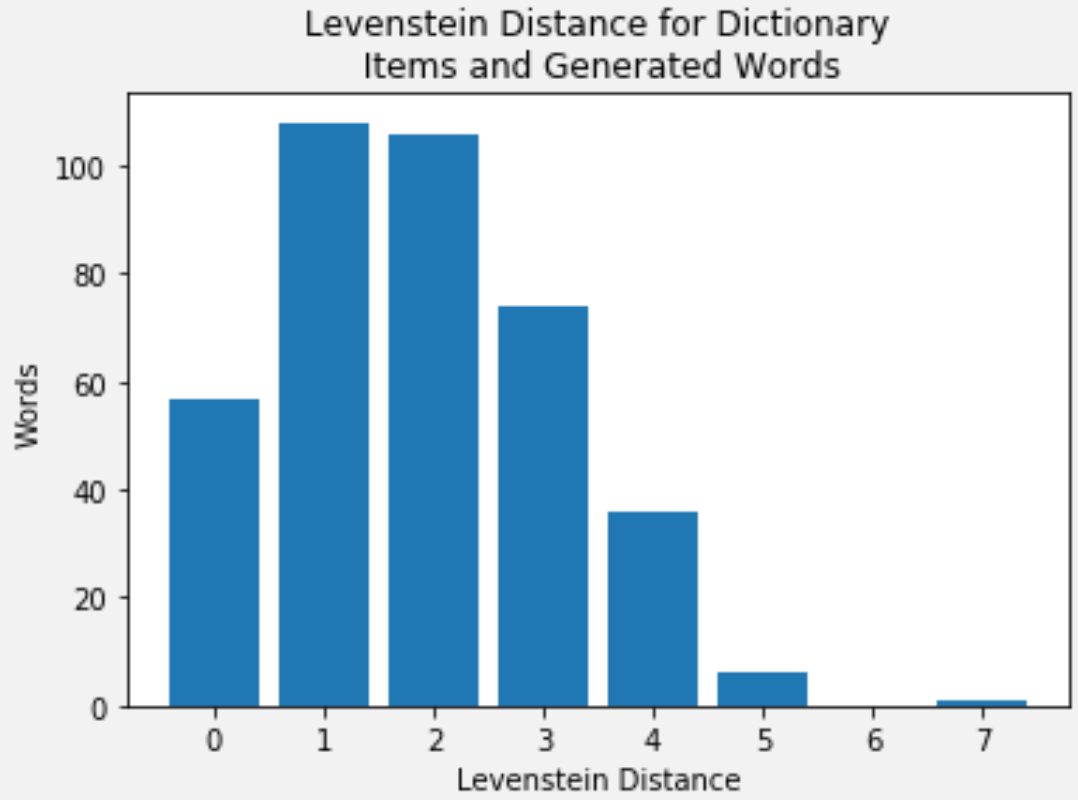
ALGORITHM APPLICATION: PART I

- Dataset:
  - Chinese dictionary occurrence
  - Russian analogue
  - Xinhua-generated transliteration of the Russian word
  - Levenstein distance between the dictionary and Xinhua words

id	word	origin	Xinhua	levenstein_word
0	阿尔申	аршин	阿尔申	0
1	阿尔西非	альсифер	阿利西费尔	3
2	阿法林	афалина	阿法莉娜	2
3	阿留米特	алюмит	阿柳米特	1
4	阿依尔	аил	艾尔	2
5	阿札林	азарин	阿扎林	1
6	艾费勃	офеб	奥费布	2
7	艾米利通	эмиритон	埃米里托恩	4
8	艾木兴	эмшер	埃姆舍尔	4
9	艾特纳	этнэ	埃特内	2
10	艾匹配	эпипэ	埃皮佩	3
11	爱特罗尔	этрол	埃特罗尔	1
12	安诺	анау	阿瑙	2

ALGORITHM APPLICATION: PART II

Character diversity in dictionary items	**319**
Character diversity of Xinhua-based transliterations	168



- Explanation: there is a bigger variance of the Chinese characters in the dictionary, and the problem is in multiple choice of a character based on one phonetic reading:
  - "li": 利 (Xinhua)
    利, 里, 理, 立, 列 (dictionary)