УДК 811.581
DOI: 10.28995/2686-7249-2020-7-30-63

Стратегии преобразования русских фонетических заимствований в китайском языке: фонетические и графические аспекты

Кирилл И. Семенов

Национальный исследовательский университет «Высшая школа экономики», Москва, Россия
kir.semenow@yandex.ru

Аннотация. Статья посвящена рассмотрению фонетических и графических трансформаций, которые претерпевают звуковые заимствования из русского языка. Исследование включает в себя анализ имен нарицательных и имен собственных, кодифицированных в словарях и справочниках и употребляемых в Интернете. Рассмотренные данные позволяют выявить основные тенденции в адаптации согласных, а также локализовать предположительное влияние русско-китайского пиджина на современное употребление русских заимствований в путунхуа. В сфере графики выявлено существенное несоответствие норм транслитерации, предписываемых государственными СМИ КНР, и узуса в Интернете. Кроме того, обнаружена значительная специфичность наиболее частотных иероглифических N-грамм в заимствованиях по сравнению с контрольным корпусом китайских текстов. Мы рассчитываем на то, что работа будет полезна как для фундаментальных лингвистов, занимающихся языковыми контактами и фонетикой, так и для специалистов по автоматической обработке естественного языка.

Ключевые слова: китайский язык, русский язык, языковые контакты, фонетические заимствования, автоматическая обработка естественного языка, N-граммы

Для цитирования: *Семенов К.И.* Стратегии преобразования русских фонетических заимствований в китайском языке: фонетические и графические аспекты // Вестник РГГУ. Серия «Литературоведение. Языкознание. Культурология». 2020. № 7. С. 30–63. DOI: 10.28995/2686-7249-2020-7-30-63

Adaptation strategies of Russian phonetic loanwords in Chinese. Phonetic and graphic aspects

Kirill I. Semenov

National Research University "Higher School of Economics", Moscow, Russia
kir.semenow@yandex.ru

Abstract. This article considers phonetic and graphic transformations of Russian loanwords in Chinese. The study comprises an analysis of both proper and common nouns, as well as both loanwords included in dictionaries and those used in the Internet. The data considered make it possible to detect the main trends in the adaptation of Russian consonants in Chinese, as well as to localize the hypothetical influence of the Russian-Chinese pidgin on current loanword adaptation in Mandarin Chinese. It is noted that there is a dramatic discrepancy between the norms of transliteration prescribed by the PRC media and the usage in the Internet. Furthermore, a significant level of specificity of the hieroglyphic N-grams in the loanwords is revealed, compared to the reference corpus of the Chinese texts. The author expects that the results of the work will be useful for specialists both in phonetic typology and in NLP.

Keywords: Chinese, Russian, language contacts, phonetic loanwords, NLP, N-grams

For citation: Semenov, K.I. (2020), "Adaptation strategies of Russian phonetic loanwords in Chinese. Phonetic and graphic aspects", *RSUH/RGGU Bulletin. "Literary Theory. Linguistics. Cultural Studies " Series,* no. 7, pp. 30-63, DOI: 10.28995/2686-7249-2020-7-30-63

1. Введение

На протяжении всей истории китайского языка (как в нормативном варианте, так и в диалектах) можно увидеть, что он испытывал влияние других языков, типологически и генеалогически очень сильно от него отстоящих. За последние две тысячи лет, согласно устоявшемуся взгляду в западной лингвистике, можно выделить несколько основных волн заимствований в китайский язык из других языковых семей. Первая – массовые заимствования буддийской терминологии при адаптации текстов пали на китайский язык (начиная с I в. н. э.) [Vervaet 2017, p. 23]. Вторая – заимствования из языков Средней Азии в эпоху существования Великого шелкового пути. Третья – христианская лексика, пришедшая в Китай вместе с миссионерами-иезуитами в XVI в. [Vervaet 2017, p. 26]. Четвертая – заимствования в науке и технике из быстро модернизировавшейся Японии, иногда – относящиеся к категории обратных (заимство-

ванных из китайского в японский, затем – наоборот). Это происходило во второй половине XIX в. Последняя волна заимствований, начавшаяся в период правления Дэна Сяопина и продолжающаяся по сей день, включает в себя заимствования из европейских языков, в первую очередь из английского.

Весьма разнообразны стратегии, которые используют носители китайского языка, заимствуя те или иные слова. Традиционным считается разделение заимствований в китайском на фонетические, семантические (кальки) и гибридные (различные комбинации фонетических и семантических элементов). Такую классификацию можно увидеть, например, в [Семенас 2005, с. 210] или в [Vervaet 2017]. Впрочем, западные исследователи нового поколения предлагают более подробное и исчерпывающее деление всех заимствований на 15 типов, включающее все возможные вариации с передачей смысла, звучания и графического облика заимствования [Cook 2018, p. 20][1]. Примечательно, что даже если базовой стратегией адаптации слова стала фонетическая, то семантическая интерпретация все равно будет влиять на финальный фонетический облик этого заимствования, то есть звуковой, тоновый и морфемный состав заимствования будет зависеть от семантически интерпретируемых китайских морфем [Семенас 2005, с. 216], [Nelson 2013, p. 500].

В настоящий момент большинство ученых находят более перспективной сферой анализ графических и семантических заимствований в современном китайском. Это связано с тем, что основные языки, из которых продолжают приходить заимствования в путунхуа, – английский и японский – дают по большей части именно графические заимствования или семантические кальки.

На фоне повышенного интереса к изучению вышеописанных типов заимствований в китайском языке крайне слабо затронутым (как в фундаментальном, так и в прикладном языкознании) остается исследование фонетических заимствований. Те же работы, которые посвящены этой сфере, отличаются либо отсутствием системного взгляда на весь процесс фонетической адаптации слов, либо узкой областью заимствованных слов (ограниченных либо конкретным языком-донором, либо конкретным семантическим полем).

[1] Отдельно стоит отметить феномен «буквенных слов» (lettered words, 字母词 – zìmǔ cí). Пример такого слова – 三G手机 (sāngē shǒujī) – обозначает «мобильный телефон третьего поколения», при этом G здесь – сокращение от английского generation. Большая часть данного лексического класса образовывает пересечение с заимствованиями из европейских языков.

Между тем, согласно [Cook 2018, p. 25], в ближайшие десятилетия фонетических заимствований будет в китайском все больше.

Стоит, тем не менее, упомянуть небольшое количество выдающихся работ, посвященных проблематике фонетических заимствований в китайском языке. Одним из наиболее значимых китайских трудов стал Словарь заимствований в китайском языке, впервые вышедший в 1985 г. – [刘正埮 (liú zhèngtán) 1985]. Основная часть вхождений в нем представляет собой именно фонетические заимствования. Из публикаций на английском языке необходимо отметить две работы. В [Lin 2008] приводится описание основных исследований последних 30 лет в западной научной литературе. Из этого обзора видно, что основной интерес у лингвистов вызывал кантонский диалект китайского, значительно отличающийся от путунхуа. Вторая работа – [Miao 2005] – представляет подробное исследование фонетических заимствований в путунхуа из трех европейских языков. Методика этой работы будет подробнее описана ниже и ляжет в основу одного из разделов настоящей статьи. Что касается работ, посвященных фонетическим заимствованиям из русского языка, то единственным достаточно систематическим описанием этого явления стала статья русских и китайских лингвистов Амурского государственного университета, выпущенная в 2016 г. [Ин, Шипановская 2016]. К сожалению, данная статья отличается малым объемом проанализированных данных (около 50 слов) и отсутствием единообразия в классификации возможных фонетических трансформаций.

Очевидно также, что фонетические заимствования – большая проблема для автоматической обработки китайского языка. Усугубляется она тем, что в китайской системе письменности нет пробелов, а предлагаемое обычными алгоритмами NLP деление на слова достаточно европоцентричное [Huang, Chen 1996, p. 4]. В отличие от семантических калек, часто уже кодифицированных и не представляющих особенных проблем для китайского словоделения, подход к обработке фонетических заимствований машинными алгоритмами развит крайне слабо. В хрестоматийном труде Introduction to Chinese Natural Language Processing можно встретить лишь небольшой пассаж о распознавании транслитерированных заимствований, который является частным случаем распознавания слов, не включенных в словарь (проблема OOV) – [Wong, Xu 2010, pp. 39, 67]. Эвристики, которые применяются в автоматических переводчиках, например Google.Translate или Яндекс. Переводчик, пока достаточно далеки от идеала и в обнаружении фонетического заимствования, и в его правильном обратном переводе. Более досконально автоматический анализ европейских

фонетических заимствований был исследован на материале корейского и японского языков. Несмотря на типологические и генеалогические различия между каждым из трех языков, имеется и ряд принципиальных для нашего исследования сходств между ними: общие черты наблюдаются как в фонетике (в т. ч. жесткие ограничения на структуру слога), так и в графике (иероглифическая и/или слоговая письменность). В числе работ, посвященных японскому и корейскому, можно отметить [Koo 2015] (заимствования из английского в корейский и японский) и [Fujii, Ishikawa 2001] (пара языков: английский и японский).

Настоящая работа предлагает начать систематичное заполнение лакуны в области исследований фонетических заимствований из русского языка в путунхуа. Мы попытались проанализировать основные аспекты в последовательности фонетических и графических трансформаций, которая происходит при транслитерации русских слов в стандартном китайском языке. Методы и результаты настоящего исследования были описаны в курсовой работе, защищенной в Высшей школе экономики в 2019 г. [Семенов 2019].

2. Материал, методика и ограничения исследования

Объектом нашей работы стали фонетические заимствования из русского в путунхуа. В первую очередь это имена собственные, однако в нескольких разделах были использованы заимствования имен нарицательных, зафиксированные в Словаре заимствований в китайском языке [刘正琰 (liú zhèngtán) 1985]. Отметим, что в этом словаре есть вхождения лексем, относящиеся к разным историческим периодам в России – например, к имперскому (слова «мазурка» или «сударь») или к советскому («ГПУ»). Это позволяет нам предположить, что они могли прийти в китайский язык в разные эпохи. Однако, во-первых, в самом словаре отсутствуют пометы, касающиеся времени (или хотя бы источников) заимствования, во-вторых, стоит принимать во внимание, что время использования лексики в русском языке не обязательно соответствует времени его заимствования. Например, если какое-то слово, касающееся имперских реалий, пришло в китайский через пиджин (об этом будет рассказано в разделе 4), можно ожидать, что оно сохранило черты китайского языка или диалектов XIX в.; но если оно пришло в китайский через переводы русской литературы XIX в., это, скорее всего, произошло не раньше середины XX в. В дальнейших исследованиях было бы полезно сгруппировать вхождения словаря по

предположительному времени их появления в китайском; однако на использованных для данного исследования материалах сделать это не представляется возможным. Исключение составляет лишь подвыборка слов, встречающихся одновременно в словарях пиджина и в Словаре заимствований в китайском языке.

Первый этап исследования (раздел 3) посвящен анализу трансформаций, которые происходят с согласными русского языка в процессе адаптации к китайской фонетической системе. Анализ выполнен в парадигме Теории оптимальности и повторяет методику, предложенную в диссертации Жуйцинь Мяо, защищенной в Университете Стоуни-Брук (Нью-Йорк) в 2005 г. [Miao 2005]. В этой работе рассматриваются фонетические заимствования из английского, немецкого и итальянского языков. Анализ для русского языка был проведен на материале имен нарицательных из Словаря заимствований в китайском языке (387 слов).

В разделе 4, посвященном проверке гипотезы о влиянии приграничных пиджинов на современную адаптацию фонетических заимствований из русского в путунхуа, за основу был взят материал словаря сибирского пиджина, собранный Е. Перехвальской в ее книге [Перехвальская 2008]. Необходимо отметить, что в рассмотрение не брался недавно опубликованный Словарь кяхтинского пиджина – [Попова, Таката 2017]. Мы приняли это решение по причине существенных различий в системе нотации в русских работах и в данном словаре.

Следующий этап работы (раздел 5) посвящен графической адаптации фонетических заимствований из русского в путунхуа. Раздел 5.1 посвящен сравнению нормы и узуса в транслитерациях русских имен собственных в китайском сегменте Интернета. В качестве образца нормы была использована таблица для транслитерации названий на русском языке, опубликованная Синьхуа [新华 (xīnhuá) 1993]. В качестве корпуса реальных употреблений имен собственных в китайском Интернете был автоматически собран список из более чем 13 тыс. имен собственных, о которых имеется статья в китайской Википедии. Это было сделано на материале Wikidata – масштабной многоязычной базы знаний, основанной на данных из Википедии (URL: https://www.wikidata.org/wiki/). Данный этап работы был осуществлен при помощи языка Python.

В последнем этапе исследования (раздел 5.2) анализируются частотность иероглифических N-грамм, использованных в транслитерациях заимствований, и их сравнение с N-граммами в контрольном корпусе, составленном из нескольких художественных произведений китайских авторов конца XX в. Данный анализ был также проведен при помощи языка Python.

Необходимо отдельно отметить ограничения последних двух разделов исследования. Во-первых, все статистические исследования были проведены на основе пары «русское слово – иероглифическое заимствование», минуя фонетическую транскрипцию китайского слова. Возможно, при дальнейшем анализе этот слой окажется очень нужен в качестве медиатора между исходным русским словом и иероглифической репрезентацией. Второе ограничение – использование материалов китайской Википедии для анализа узуса носителей путунхуа. Дело в том, что на территории КНР китайская Википедия заблокирована; соответственно основную часть пользователей этой энциклопедии предположительно составляют жители Китайской Республики, Гонконга, Макао, Сингапура и диаспоры. При этом в основной части вышеперечисленных государств, во-первых, не обязательно использование предписаний по транслитерации, выдвинутых Синьхуа, во-вторых, используются другие варианты языковой нормы (гоюй) и традиционная система китайских иероглифов. Следовательно, Wikidata, пользующаяся статьями Википедии, едва ли может претендовать на звание самой большой базы знаний именованных сущностей на китайском, так как в Китае существует своя интернет-энциклопедия Baidupedia (百度百科, URL: https://baike.baidu.com/), объем которой превышает общий объем англоязычной, немецкоязычной, русскоязычной и китайскоязычной Википедии. Это позволяет предположить, что в Baidupedia содержится больше статей о русских именованных сущностях, чем в рассматриваемой нами базе знаний. Избрание не самого большого китайскоязычного набора данных, аудитория которого сильно смещена в сторону Тайваня, Сингапура и Гонконга, могло породить сильный перекос в данных, на основе которых строилось наше исследование.

Почему же в таком случае была выбрана именно Wikidata? В первую очередь из-за интерфейса по автоматической выкачке данных: при помощи сравнительно простого поискового запроса она позволяет сформировать набор данных по всем именам собственным, в оригинале имеющим русское имя, а ныне переведенным на китайский язык. Такого объема вхождений невозможно добиться ни пользуясь бумажными версиями словарей заимствований, ни параллельными русско-китайскими корпусами, ни скачав списки имен собственных, находящиеся в онлайн-словарях (так как обычно их объем исчисляется сотнями). Baidupedia же не предлагает алгоритмов поиска по своей базе знаний или возможности выкачки всей энциклопедии, поэтому на сегодняшний день остается довольствоваться базой знаний Википедии.

3. Анализ фонетических заимствований в китайском языке в парадигме Теории оптимальности

Прежде чем начать сравнительный анализ фонетических заимствований из европейских языков и из русского в китайский, отметим основные параметры различия фонетики т. н. SAE-языков и путунхуа в области консонантизма. В фонетике стандартного китайского языка отсутствуют фонемные противопоставления по звонкости-глухости и по мягкости-твердости. Вместо этого бóльшая часть согласных обладает оппозицией по придыхательности. Кроме этого, структуру слога в китайском можно задать формулой $(C_1)V(C_2)$, где C – согласный, V – гласный или дифтонг. C_1, в соответствии с китаистикой, принято называть инициалью, сочетание $V(C_2)$ – финалью. При этом набор согласных, возможных в позиции C_2, значительно ограничен по сравнению с C_1.

Обратимся теперь к исследованиям фонетических заимствований в стандартном китайском. Как уже было отмечено в обзоре литературы, наиболее общим и последовательным трудом, изучающим адаптацию фонетических заимствований в путунхуа из европейских языков, стала диссертация Жуйцинь Мяо [Miao 2005]. Автор выполнила ее в парадигме Теории оптимальности (далее – ОТ), и основным объектом ее изучения стали трансформации отдельных согласных в китайских заимствованиях, а также стратегии адаптации консонантных кластеров в путунхуа. Мы применили методы ее работы и сравнили результаты, полученные на данных русских заимствований, с результатами, которые были сделаны Мяо для английского, немецкого и (в некоторых вопросах) итальянского языков. Был воспроизведен первый раздел ее работы, мы рассмотрели возможные трансформации согласных (сгруппировав их по способу артикуляции).

Учитывая, что большое количество иероглифов имеет вариативность в прочтении (как в тоне, так и в звуковом составе), сравнивать русские и иероглифические вхождения при отсутствии эксплицитной транскрипции было бы некорректно. Поэтому в рассмотрение был взят ресурс, где транскрипция была дана эксплицитно.

Таков Словарь заимствований в китайском языке [刘正埮 (liú zhèngtán) 1985], где каждому словарному вхождению была приписана принятая в КНР транскрипция пиньинь. Из этого словаря было взято 387 слов, в большинстве своем имен нарицатель-

ных. Необходимо отметить, что не все из них являются исконно русскими, немалая часть из них пришла в русский из языков России и бывшего СССР. Анализ слов из этого ресурса был сделан вручную.

Последнее, о чем следует напомнить перед анализом заимствований из европейских языков, – между Международным фонетическим алфавитом (МФА, IPA) и официальной китайской транскрипцией пиньинь (拼音, pīnyīn) есть существенные различия, при этом обе системы взаимозаменяемы. В данной работе в целях удобства как для специалистов по китайскому языку, так и для лингвистов-типологов всегда будет приводиться и транскрипция пиньинь (в кавычках “ ”), и транскрипция МФА (в косых скобках //).

В работе Мяо были рассмотрены все согласные сегменты, встречающиеся в английских, немецких и итальянских заимствованиях, проанализированные по группам (в зависимости от способа образования). Такое деление было избрано вследствие работы в парадигме P-map (от Perceptual mapping), подразумевающей, среди прочего, что главным качеством согласного, улавливаемым перцептивно, является его способ артикуляции [Miao 2005, p. 90].

Для каждого согласного из языков-доноров была приведена сводная таблица всех соответствующих ему согласных в китайских фонетических адаптациях, где все варианты были упорядочены по частотности (в процентном соотношении). На основе таблиц была составлена иерархия ограничений (в данном случае – faithfullness constraints), выглядящая для большей части согласных (кроме губных и плавных) таким образом:

(1) IDENT (Manner) >> IDENT (Place) >>
>> IDENT (Voice/Asp),

где первое предписывает сохранение того же способа образования, второе – сохранение того же места образования, третье – переход звонкого согласного в непридыхательный, а глухого – в придыхательный [Miao 2005, p. 82].

Мы повторили данное исследование на русском языке и проанализировали частотность всех преобразований для каждого из согласных звуков. Ниже приведена таблица (табл. 1), где представлены обобщенные результаты, показывающие все возможные трансформации каждого из согласных русского слова. Для каждого согласного реализации упорядочены по частоте их встречаемости.

Таблица 1

Иерархии реализаций консонантных сегментов в фонетических заимствованиях

Способ артикуляции	Сегмент	Твердый, onset	Твердый, coda	Мягкий, onset	Мягкий, coda	Пример наиболее частотной реализации
взрывные	/б/, /б'/	“b” /p/ >“p” /p^h/	“b” /p/ >“p” /p^h/	“b” /p/	н./д.	баян > 巴扬 “bāyáng” /pajaŋ/
	/п/, /п'/	“p” /p^h/ >“b” /p/ >“f” /f/	“p” /p^h/ >“b” /p/ > ø	“p” /p^h/ >“b” /p/	н./д.	пуд > 普特 “pǔtè” / $p^hut^hɤ$/
	/к/, /к'/	“k” /k^h/ >“g” /k/ >“j” /tɕ/	“k” /k^h/ >“g” /k/ >“j” /tɕ/	“j” /tɕ / >“g” /k/ >“k” /k^h/	н./д.	катюша > 喀秋莎 “kāqiūshā” /$k^hā:tɕ^həuʂā$/
	/г/, /г'/	“g” /k/ >“j” /tɕ/	“g” /k/ > ø	“g” /k/ >“j” /tɕ/ >“q”, “x” /$tɕ^h$/, /ɕ/	н./д.	ГПУ >格别乌 “gébiéwū” /$kɤ́:piéːū$/
взрывные*	/д/, /д'/	“d” /t/ >“t” /t^h/ >“z” /ʦ/	“d” /t/ >“t” /t^h/ > ø	“d” /t/ >“j” /tɕ/	н./д.	дума >杜马 “dùmǎ” /tù:mǎ:/
	/т/, /т'/	“t” /t^h/ >“d” /t/ >“q” /tɕ/	“t” /t^h/ >“d” /t/ > ø	“q” /$tɕ^h$/ >“t” /t^h/ >“j” /tɕ/ >“d” /t/	“j” /tɕ/, “t” /t^h/	точка >托其卡 “tuōqíkǎ” /$t^hō:$ $tɕ^hí:$ $k^hǎ:$/

Окончание табл. 1

Способ артикуляции	Сегмент	Твердый, onset	Твердый, coda	Мягкий, onset	Мягкий, coda	Пример наиболее частотной реализации
щелевые	/в/, /в’/	“w” /w/ >“f” /f/ >“b” /p/, “u” /-u-/, ø	“f” /f/	“w” /w/ >“ui” /-ui/, “f” /f/	н./д.	ведро >维得罗 “wéidéluó” /wə́itɤlʷó:/
	/ф/, /ф’/	“f” /f/ >“h” /h/	“f” /f/	“f” /f/	“f” /f/	финка >芬卡[帽] “fēnkǎ” /fə̄nkʰǎ:/
	/з/, /з’/	“zh” /ʧ/, “s” /s/ >“z” /ʦ/ >“sh” /ʂ/, “ch” /ʧʰ/	“z” /ʦ/ >“s” /s/	“j̀” /ʨ/ >“q” /ʨʰ/	н./д.	азарин >阿札林 “āzhálín” /ā:ʧǎ:lɨ́n/
	/с/, /с’/	“s” /s/ >“sh” /ʂ/ >“x” /ɕ/	“s” /s/>“z” /ʦ/> “zh” /ʧ/ >“sh” /ʂ/	“x” /ɕ/ >“s” /s/	“x” /ɕ/ >“s” /s/	сарафан >萨腊范 “sàlàfàn” /sà:là:fæn/
	/ж/	“zh” /ʧ/ >“r” /ɻ/, “sh” /ʂ/, “x” /ɕ/, “ch” /ʧʰ/	н./д.	---	---	крыжачок >克雷扎卓克 “kèléizházhuōkè” /kʰɤ̀:lə́iʧā:ʧʰó:kʰɤ̀:/
	/ш/, /ш’/	“sh” /ʂ/ >“s” /s/, “x” /ɕ/	“sh” /ʂ/ >“s” /s/, “x” /ɕ/	н./д.	“sh” /ʂ/ >“x” /ɕ/	шапка >沙普克 “shāpǔkè” /ʂā:pʰǔ:kʰɤ̀:/
	/х/, /х’/	“h” /h/	“h” /h/	“h” /h/, “x” /ɕ/	н./д.	хорошо >ㄏ哈拉嗦 “hālāsuó” /xā:lā:sʷō:/
аффрикаты	/ц/	“ch” /ʨʰ/ >“c” /ʦʰ/ >“sh” /ʂ/	“c” /ʦʰ/	---	---	червонец >切尔逢涅茨 “qiē’ěrféngniècí” /ʨʰē:ɚfəŋnʲè:tsʰz/
	/ч/	---	---	“q” /ʨʰ/ >“ch” /ʧʰ/ >“zh” /ʧ/, “j̀” /ʨ/, ø	“q” /ʨʰ/ >“ch” /ʧʰ/ >“j̀” /ʨ/	ЧК >契卡 “qìkǎ” /ʨʰi:kʰǎ:/

ISSN 2686-7249 RSUH/RGGU Bulletin: Literary Theory. Linguistics. Cultural Studies. 2020. no. 7

носовые	/м/, /м’/	“m” /m/	“m” /m/ >“ng.” /ŋ./>“n.” /n./	“m” /m/	“n.” /n./, “ng.n” /ŋ.n/	МиГ >米格 “mǐgé” /mʲǐ:kɤ́:/
	/н/, /н’/	“.n” /.n/ >“n.n” /n.n/ >“ng.n” /ŋ.n/ >“ng.” /ŋ./	“n.” /n./ >“ng.” /ŋ./	“.n” /.n/ >“n.n” /n.n/ >“ng.n” /ŋ.n/	“n.” /n./ >“ng.” /ŋ./ >“.n” /.n/	НЭП >纳普 “nàpǔ” /nà:pʰǔ:/
плавные	/р/, /р’/	“l” /l/	“r.” /ɚ./ > ø >“i.” /ɻ/ >“ng.” /ŋ./	“l” /l/ >“r.” /ɚ./	“l” /l/ >“r.” /ɚ./	рало >拉罗 “lāluó” /lā:lʷó:/
	/л/, /л’/	“l” /l/ >“n” /n/	“r.” /ɚ./ >ø>“l” /l/	“l” /l/	“l” /l/ >“r.” /ɚ./ > ø	лезгинка >列兹金卡 [舞] “lièzījīnkǎ” /lʲè:tsʑ́:tɕīnkʰǎ:/
глайды	/й/	---	---	“.y” /.j/ > ø >“i.” /ɻ/	“i.” /ɻ/ >“yi” /ji/ > ø	ералаш >叶拉拿西 “yèlānáxī” /jè:lā:ná:ɕī:/

Примечания

Трансформация в ноль звука обозначается символом ø.

Несуществующий в русском языке сегмент обозначается прочерком (---); русские сегменты, примеров которых не нашлось в выборке, обозначены «н./д.».

Символ границы слога – «.». Соответственно обозначение “n.” необходимо трактовать как «звук n в позиции coda», а «.l» – как «звук l в позиции onset».

Сегменты /д/, /д’/, /т/, /т’/ выделены в отдельную группу взрывных, так как их мягкие варианты по способу образования близки к аффрикатам.

Можно увидеть, что основная часть реализаций согласных удовлетворяет иерархии ограничений, построенной для европейских языков. Наличие дополнительной (по сравнению с европейскими языками) оппозиции по мягкости в большинстве случаев также не добавляет исключений.

Так, русское /к/ в позиции onset переходит в /kʰ/ – в 95% случаев, в /k/ – в 3% случаев, а /ф/ в onset реализуется в 93% как /f/, в 7% – как /h/. Это подтверждает гипотезу Мяо о том, что в первую очередь будут выполняться ограничения способа и места, а затем сохраняться переход «звонкий > непридыхательный; глухой > придыхательный».

Тем не менее существует два типа явлений, достойные более пристального внимания. В пункте 3.1 будут рассмотрены случаи адаптации согласных, удовлетворяющие основной иерархии, но обладающие сильной вариативностью. В пункте 3.2 будет рассмотрен ряд согласных, основные реализации которых в заимствованиях противоречат предложенной иерархии.

3.1. Широкая вариативность

3.1.1. Аффрикаты и мягкие взрывные

Первая такая группа – русские аффрикаты /ц/ и /ч/. Например, /ч/ может реализовываться четырьмя аффрикатами – “q” /tɕʰ/, “ch” /ʧʰ/, “zh” /ʧ/ или “j” /tɕ/.

Кроме того, примечательна реализация аффрикат в виде щелевых в наиболее редких случаях:

(2) царь > 沙尔 “shā’ěr” /ʂɑ̄:ɚ/

Этот процесс общий для заимствований из русского и из рассмотренных Мяо европейских языков, и он указывает на то, что аффрикаты – одна из наиболее неустойчивых групп согласных с точки зрения сохранения способа образования. В принципе, это не очень удивительный факт, так как типологически известен эффект спирантизации[2].

Здесь было бы логично упомянуть и о двух группах согласных, в которых часто происходят трансформации, на первый взгляд нарушающие иерархию. Речь идет, во-первых, о мягких сегментах

[2] *Kerstens J., Ruys E., Zwarts J.* Spirantisation [Электронный ресурс] // Lexicon of Linguistics. 1996. URL: https://lexicon.hum.uu.nl/?lemma=Spirantisation (дата обращения 25 марта 2020).

/г’/ и /к’/, а во-вторых, о /д’/, /т’/. Согласные из обеих этих групп часто превращаются в аффрикаты в китайских транслитерациях. Ниже можно увидеть примеры таких трансформаций.

Трансформация мягких заднеязычных в аффрикату:

(3) кисель >吉协力 “jíxiélì” /tɕíːɕéːlìː/

Трансформация мягких зубных в аффрикату:

(4) «Катюша»>喀秋莎 “kāqiūshā” /kʰāːtɕʰəūʂāː/

Представляется достаточно простой интерпретация такого изменения сегментов в сторону аффрикат. Мягкие заднеязычные из-за смещения места образования и долгого шума при взрыве (в силу большой площади задней части языка) перцептивно могут напоминать аффрикативные согласные. Необходимо отметить, что Мяо обращала особое внимание на сочетание «заднеязычный согласный + передний гласный» в своей работе, где она указывала на аналогичное поведение (передачу заднеязычного переднеязычной аффрикатой) в заимствованиях из европейских языков (например, нем. “Kiel” /kiːl/ > кит. 基尔 “jī’ěr” /tɕ͡iɚ/). Было отмечено, что это следствие палатализации [Miao 2005, p. 63]. Это кажется справедливым, а в русском языке с более выраженными процессами палатализации это становится еще более очевидным. Что касается мягких зубных /д’/ и /т’/, то в современном русском их можно вполне считать аффрикатами; отнесение их на письме к их твердым взрывным аналогам можно объяснить скорее традицией и влиянием орфографии, нежели отображением реальных фонетических особенностей этих сегментов. Китайские же заимствования очень явно отображают эту особенность русских согласных.

Таким образом, явления, кажущиеся на первый взгляд исключениями из предложенной Мяо теории в зоне мягких согласных, на самом деле являются лишь ее дополнительным подтверждением.

3.1.2. Плавные

Напомним, что в китайском языке распределение плавных согласных принципиально отличается от обычного для европейских языков. Наименее маркированным является боковой /l/, который может встречаться в позиции onset. Этот звук является «базовым» плавным в китайском языке, аналогично тому, каким в японском является одноударный /ɾ/. Существует также ретрофлексный сегмент, который в позиции onset представляется как аппроксимант /ɻ/, а в позиции coda – как эризованный шва /ɚ/.

Это, видимо, является причиной некоторой асимметрии в преобразовании плавных (как русских, так и английских/немецких/итальянских), заключающейся в том, что в позиции onset оба согласных /р/ и /л/ в абсолютном большинстве случаев преобразуются в /l/, а в позиции coda – в /ɚ/. Примечательно, что в случае с мягкими сегментами /р'/, /л'/ происходит значительное расслоение: в позиции coda примерно половина вхождений /р'/ трансформируется в /ɚ/, в то время как остальные вхождения реализуются как onset /l/ (обычно в слоге /li/). Аналогичный процесс происходит при трансформации /л'/, где как onset /l/ реализуется больше половины вхождений (55%).

Реализация /р'/ в позиции coda в /l/:

(5) сударь >苏达利 “sūdálì” /sū:tá:lʲì/

Реализация /р'/ в позиции coda в /ɚ/:

(6) богатырь >波加的尔 “bōjiādìěr” /pō:tɕā:tʲì: ɚ/

Для описания трансформации плавных Мяо вводит дополнительные ограничения маркированности. Оба из них предлагается ставить в вершине иерархии ограничений. Первое из них – */ɻV/ >> */lV-/ – обозначает, что аппроксимант в onset более маркирован, чем /l/ [Miao 2005, с. 86]. Это оказалось верно. Второе ограничение, предложенное автором, – *-Vl/-Vr” -> /CV/: запрет плавной coda становиться плавным onset в заимствованном слове [Miao 2005, с. 87]. Это ограничение оказывается опровергнуто в случае с русскими заимствованиями на материале мягких согласных.

3.1.3. Носовые

Что касается употребления носовых согласных в onset, то, аналогично заимствованиям из европейских языков, вариация согласных очень мала и в основном губно-губной /м/ и зубной /н/ совпадают со своими аналогами /m/ и /n/. Примечателен в данном случае единственный момент (не затронутый в работе Мяо): немалое количество вхождений в русском языке, где /н/ встречается в позиции между гласными, преобразуется в сочетание коды /n/ или /ŋ/ и следующего за ним в позиции onset /n/, как бы «усиливая» этот согласный (см. пример 7).

(7) Семинар > 习明纳尔 “xímíngnà’ěr” /ɕímíŋnàɚ/

Что же касается позиции coda, то здесь (аналогично тому, как это происходит в европейских языках) примечателен переход /м/ в допустимую в китайском назализованную финаль (/n/ или /ŋ/). Впрочем, этот процесс происходит далеко не во всех случаях (27%), уступая при этом эпентетической вставке гласного после /m/.

3.2. Согласные, трансформация которых противоречит иерархии ограничений

Случаи значимого несоответствия предложенной иерархии ограничений относятся к трансформации сегментов /ж/, /з/ и /з'/. Приведем два примера:

(8) жалейка > 扎列卡[管] “zhálièkǎ” /ʧaˑ:lʲèˑ:kʰǎˑ/

(9) мазурка > 玛组卡[饼] “mǎzǔkǎ” /măˑ:tsŭˑ:kʰǎˑ/

В отличие от господствующей в западных заимствованиях трансформации звука в щелевые (в большинстве случаев в англ. – в /s/, /ʂ/ или /ɕ/), в заимствованиях из русского они преобразуются в аффрикаты /ʧ/ (30%) и /ʦ/ (20%). Ожидаемые же трансформации данных сегментов в щелевые представлены для каждого из звуков в меньшинстве. Мы попробуем проверить связь данного несоответствия с влиянием пиджина на стратегию фонетической адаптации русских слов.

Помимо этого, сегмент /ж/, особенно в позиции onset, имеет тенденцию к реализации в виде ретрофлексного аппроксиманта “r” /ɻ/ (17% на материале Словаря заимствований). Такая тенденция не наблюдается в европейских языках, где сегмент /ʒ/ обычно реализуется как “y” /j/ (из-за близости по месту образования). При этом можно увидеть, что в обоих случаях (и в русском, и в европейском) фрикативные согласные могут меняться на аппроксиманты того же места образования. Это можно объяснить тем, что аппроксиманты и фрикативные различаются только шириной щели [Кодзасов, Кривнова 2001, с. 280]. Кроме того, возможно, на такой способ адаптации повлияла устоявшаяся норма китайско-русской практической транскрипции Палладия: буква «ж» обозначает в транслитерации китайских слов как раз данную ретрофлексную инициаль: например, имя великого китайского писателя 郭沫若 (“Guō Mòruò”, /kʷōˑ:mòɻʷòˑ/) в системе Палладия передается как «Го Можо».

4. Проверка влияния пиджина и северных диалектов на адаптацию русских заимствований в путунхуа

Перед непосредственным обсуждением влияния пиджина на китайские идиомы необходимо локализовать объект нашего исследования. В семействе северных диалектов китайского ученые выделяют группу диалектов гуаньхуа [Завьялова 1996, с. 19]. Нас будут интересовать несколько диалектов из этой группы, носители которых живут в приграничных с Россией и Монголией регионах КНР.

Фонетика диалектов гуаньхуа, во-первых, отличается от фонетики путунхуа, во-вторых, неоднородна между говорами. Особенности консонантизма можно увидеть как в наборе согласных фонем (например, неразличение верхнезубных согласных и переднетвердонёбных), так и в позиционном распределении инициалей и финалей (например, взаимная заменяемость [-ŋ] и [-n] и более продуктивная эризация).

В зоне распространения приграничных северокитайских говоров с конца XVIII в. стал распространяться вариативный русско-китайский пиджин, которым пользовались изначально русские и китайские крестьяне и купцы. Далее этот пиджин распространился на приграничные и даже на внутренние территории России, а также стал языком межнационального общения для малых народов этого региона (маньчжуров, удэгэ и т. д.) с русскими – [Перехвальская 2008, с. 73, 121]. Примечательно, что, несмотря на периоды изоляции между китайским и русским населением во второй половине XX в., а также на почти полное исчезновение русской общины в Маньчжурии в эпохи японской оккупации и Культурной революции в КНР, этот пиджин продолжает существовать (хоть и сильно видоизменившись), на что указывают недавние исследования [Цзе 2007].

Нас интересует в данном случае тот факт, что большое количество как российских, так и китайских ученых отмечают в своих работах, что «транзит» русских слов через пиджин и северные китайские диалекты мог оказаться важным этапом для адаптации русских заимствований в нормативном путунхуа. С вариациями этого предположения можно ознакомиться в [Ин, Шипановская 2016, с. 145]. Тем не менее нам не удалось встретить ни одной научной работы, где было бы доказано прямое влияние пиджина и/или северных диалектов китайского языка на адаптацию русских заимствований в путунхуа. Теперь мы обладаем трактовкой ОТ для трансформаций русских согласных в путунхуа, а также сравнением этих данных с аналогичными трансформациями в европейских

языках. Значит, мы можем попробовать проверить гипотезу о влиянии пиджина или диалектов на трансформацию русских согласных в путунхуа с лучшей верифицируемостью.

Для того чтобы обоснованно предполагать связь между ранними языковыми контактами русских и китайцев и современным процессом русских заимствований в путунхуа, мы предлагаем следующий набор требований, удовлетворение каждого из которых повышает вероятность того, что заимствование пришло в путунхуа при посредничестве пиджина или северо-китайских говоров:

1. Трансформация элементов в этих словах противоречит предсказаниям OT для русских и европейских заимствований.

2. Трансформация элементов в этих словах объясняется фонетикой/грамматикой пиджина или северных диалектов.

3. Существуют словарные вхождения в пиджине или диалектах, которые подтверждают возможность прямого заимствования той или иной трансформации из пиджина/диалекта в путунхуа.

В качестве главного источника по грамматике и фонетике пиджина была использована монография Е.В. Перехвальской, посвященная русским пиджинам и собственно русско-китайскому контактному языку, на котором автор специализировалась [Перехвальская 2008]. В данной монографии также приводится словарь всех найденных ею лексем и словоформ со всеми возможными транскрипциями. В качестве описания фонетики северных диалектов был взят труд О.И. Завьяловой «Диалекты китайского языка», упоминавшийся выше [Завьялова 1996]. Для анализа словарных вхождений северных диалектов были использованы русскоязычные статьи о русско-китайских языковых контактах в этом регионе [Ин, Шипановская 2016], [Ма 2015].

Что касается возможности трансфера диалектных фонетических черт в стратегию адаптации в путунхуа, то оказалось, что ни одна из наиболее знаковых особенностей диалектов не проявляется в словах пиджина, диалектов и путунхуа так, чтобы удовлетворить всем вышеперечисленным требованиям. Например, слово «хорошо» согласно Словарю заимствований в китайском языке выглядит в путунхуа как 哈拉嗦 “hālāsuó” /xāːlāːsʷó/. Можно было бы заподозрить трансформацию /ш/ > /s/ как проявление диалектной особенности (отсутствие оппозиции зубных и передненёбных щелевых), но такая трансформация вполне предсказывается иерархией ограничений OT, рассмотренной в предыдущем разделе. Более того, примеры из английских и немецких заимствований указывают на то, что процент аналогичных чередований в заимствованиях из европейских языков значительно

больше, чем в русских заимствованиях (25% в европейских заимствованиях против 8% в русских). Необходимо также отметить, что большое количество консонантных чередований, описанных в диалектной фонетике, не встречается ни в лексемах пиджина, ни в диалектных заимствованиях, ни в фонетических заимствованиях путунхуа. Это значит, что во всех вышеописанных случаях не следует множить сущности без необходимости и вводить пиджин и диалекты в качестве посредников в процессах фонетической адаптации русских заимствований в путунхуа.

Есть, однако, две особенности, удовлетворяющие всем требованиям к «пиджинному следу», обозначенным нами. Первую из них можно увидеть на закрытой группе лексем, и связана она с процессом первичного лексического заполнения пиджина. Некоторое количество существительных (34 из 193 зарегистрированных во всем корпусе) заканчиваются квази-суффиксом, записываемым обычно как «дза»: «яйдза», «монедза» (монеты, деньги), «купедза» и т. д. [Перехвальская 2008, с. 97]. Согласно наиболее распространенной гипотезе, этот суффикс появился из китайской морфемы 子 “zǐ” /ʦi/, достаточно грамматикализовавшейся в современном китайском. Согласно [Моисеев 2013], в северных диалектах в принципе распространена практика частотного употребления подобных «классификаторов», в то время как для большинства современных заимствований (из русского или из европейских языков) такая стратегия не реализуется. Самое главное здесь то, что мы можем проследить цепочку вхождений слов с таким суффиксом в словаре пиджина, в диалектных словарях и затем в Словаре заимствований в путунхуа. Такова судьба слова «купец»: через пиджинное «купедза» это слово появляется в диалекте и в путунхуа в виде сочетания 谷瘪子 “gǔbiēzǐ” /kŭ:pʲĕ:tsž, где мы можем увидеть реализацию суффикса, исходно прикрепленного к русской псевдооснове. Таких слов (включая позднюю замену родового слова на более подходящее по смыслу) в современном китайском языке, пришедших из пиджина и/или диалектов, сейчас можно встретить пять. Таким образом, мы можем с большой долей уверенности говорить о влиянии исходного морфемного парсинга русских слов носителями китайского на восприятие этих заимствований в нормативном китайском языке. Однако класс этих слов, как показывают остальные данные Словаря заимствований, непродуктивен, поэтому это не является главным объектом для рассмотрения в нашей работе.

Другая особенность трансформации русских заимствований лежит уже в области фонетики. Как уже упоминалось, бóльшая часть современных заимствований, содержащих звуки /ж/, /з/ и /з’/, во-

преки предсказаниям OT трансформирует этот звук в аффрикаты /ʦ/ или /ʧ/. Если же посмотреть на вхождения в словаре пиджина, окажется, что и там абсолютное большинство вхождений русского звука /з/ реализуются так же (напр., “магазин” > [maaʦinə], “паровоз” > [palawoʦə]). Такая же трансформация в пиджине актуальна и для другого фрикативного /ж/. Этот переход, уникальный тем, что противоречит выработанным ограничениям OT, но совпадает с рядом вхождений в корпусе данных пиджина, позволяет предположить, что это и есть одна из немногих продуктивных фонетических трансформаций, которая была заимствована из пиджина в путунхуа для адаптации русских заимствований.

5. Анализ графических преобразований в заимствованиях из русского языка на основе больших данных

Как правило, каждый слог в китайском языке обладает определенным лексическим значением и соответствует определенному иероглифу, составляя таким образом единство слога, морфемы и обозначающего ее графического символа [Хаматова 2003, с. 19]. При этом широко распространены как омофоны, так и омографы. Из-за этого часты случаи написания одного слога разными иероглифами и, наоборот, множественных вариантов прочтения одного и того же иероглифа.

Отсутствие прямого перевода слогов в иероглифы и обратно подразумевает, что для максимально точного анализа следовало бы рассматривать такие данные, где были бы эксплицитно указаны и произношение адаптированного русского заимствования, и его иероглифическая запись. Однако основная часть доступных на китайском языке ресурсов, которая могла бы нам для этого пригодиться, не содержит информации о транскрипции слов. Поэтому данные, которые были проанализированы в практической части работы, содержали только русский оригинал и его иероглифическую транслитерацию.

Как уже было упомянуто в пункте 2, в нижеследующих разделах был использован ресурс Wikidata, связанный со свободной интернет-энциклопедией Википедия. Основным его минусом является его малый объем (по сравнению с Baidupedia), основными плюсами – большое количество данных (по сравнению со словарями и справочниками) и удобный интерфейс для скачивания данных.

Проект Wikidata позволяет осуществить выкачку необходимых данных в разном формате при помощи языка SPARQL на странице

https://query.wikidata.org/. Мы последовательно выполнили несколько запросов на этом ресурсе и сохранили данные в формате .tsv, затем объединив их в один файл .csv. Поисковые запросы собирали следующую информацию об объектах: страну (из набора: Российская Федерация, СССР, Российская империя, Беларусь, Приднестровье), тип объекта (из набора: озера, реки, горы, острова, населенные пункты, персоны), названия объекта на русском и на китайском языках.

В результате поиска было получено 13 410 объектов. Из них 81% относились к РФ, 10% – к СССР. Выборка стран была сделана с учетом того, в каких государствах на русском языке говорит большинство населения и где он является официальным. Что касается распределения по типам объектов, то 43% всех объектов покрывали населенные пункты, далее (33%) шли имена людей, еще по 10–11% были заняты реками и озерами. Эти типы объектов были выбраны, так как ожидалось, что именно они будут содержать наибольшее количество фонетических заимствований (в отличие, например, от названий организаций и экономических компаний, которые, как показал предварительный анализ, в большинстве своем калькировались семантически).

Таблица 2

Описание набора данных для графического анализа транслитераций

Столбец	Легенда и тип переменной	Значение
id_	Индекс объекта; порядковая	Натуральное число (с нуля)
wiki_id	Индекс объекта в системе Wikidata; порядковая	Натуральное число (с нуля)
label_ru	Название объекта на русском языке; номинальная	Любая строка кириллических символов
label_zh	Транслитерация объекта в системе Wikidata; номинальная	Любая строка из иероглифических символов
xinhua	Нормативная транслитерация по Синьхуа; номинальная	Любая строка из иероглифических символов
country	Название страны из выборки; номинальная	Значение из набора {Russia, USSR, Russian Empire, Belarus, Transnistria}
type	Название типа объекта из выборки; номинальная	Значение из набора {settlement, person, river, lake, island, mountain}

Окончание табл. 2

Столбец	Легенда и тип переменной	Значение
levenstein_abs	Абсолютное расстояние Левенштейна между значениями label_zh и xinhua; количественная	Натуральное число (с нуля)
levenstein_norm	Нормированное расстояние Левенштейна между значениями label_zh и xinhua; количественная	Десятичная дробь в интервале [0, 1]
jaccard	Коэффициент Жаккара между значениями label_zh и xinhua; количественная	Десятичная дробь в интервале [0, 1]

Предобработка данных состояла из нескольких шагов и выполнялась над обобщенным файлом .csv при помощи программы на языке Python с использованием библиотек HanziConv, re и pandas.

Названия, содержащиеся в объектах Wikidata, были записаны разными системами иероглифов – традиционной и упрощенной. Для чистоты анализа все иероглифические вхождения были переведены в упрощенную письменность при помощи библиотеки HanziConv.

Далее, при помощи регулярных выражений, были очищены от семантических элементов русские и китайские вхождения. Так, китайские названия были очищены от так называемых родовых слов – семантических морфем, употребляемых после имен собственных (как в примере 10). Русские названия были очищены от поясняющих слов («городское поселение», «река», «вулкан» и т. д.), которые могут стоять по обе стороны от собственно названия. Также были удалены все знаки препинания (включая кавычки) и латинские и числовые элементы в названиях на русском и китайском языках. Наконец, были унифицированы последовательности имени, отчества и фамилии в русских и китайских именах людей.

(10) 伏尔加 -河

“fúěrjiā -hé”

/fú:ə ɚ-tɕā: -xɤ́:/

Волга (фон.) -река (род. слово)

После этого было добавлено четыре столбца, необходимые для одного из разделов статистического исследования: нормативная транслитерация, порожденная созданным нами алгоритмом, а также три метрики близости строк, сравнивающие китайское вхождение из Wikidata и нормативную транслитерацию на китайский язык. Использованные метрики – абсолютное расстояние Левенштейна, нормированное расстояние Левенштейна и индекс Жаккара. Три метрики были высчитаны для каждого объекта при помощи библиотеки textdistances. Подробнее об алгоритме и метриках будет рассказано ниже.

5.1. Сравнение нормы и узуса в китайских транслитерациях

С конца XX в. в Китайской Народной Республике существуют прескрипции к транслитерации иностранных слов. Занимается этими информационное агентство Синьхуа (新华社) – официальное правительственное издание и самое большое новостное агентство в КНР. В 1982 г. оно выпустило Русско-китайский словарь транскрипций (俄汉姓名译名手册). Этот словарь включал в себя все хотя бы единожды употребленные в изданиях имена собственные, взятые из русского языка, и состоял почти из 600 страниц. В 1993 г. Синьхуа выпустило более компактную и удобную в использовании таблицу для транслитерации русских слов, которая умещалась на одном листе и позволяла находить необходимые сочетания русских букв и выбрать подходящие для этих сочетаний иероглифы. Все примечания и исключения были компактно сформулированы в шести примечаниях, идущих под таблицей; бóльшая их часть касалась фонетических поправок, связанных как с русской (например, для транслитерации сочетаний «чн» и «чт» следовало искать сочетания «шн» и «шт»), так и с китайской фонетикой (для транслитераций «мп» и «мб» следовало искать иероглифы, соответствующие сочетаниям «нп» и «нб»).

Мы решили проверить, насколько нормативные предписания, диктуемые агентством Синьхуа, соответствуют узусу, который в нашем исследовании представлен вхождениями из Wikidata. Для этого нами был создан код на языке Python, который бы получал на вход строку с русскими символами и выводил бы иероглифическую транслитерацию этой строки. Несмотря на то что в русскоязычном сегменте Интернета уже существует автоматический транслитератор русских слов (представленный на сайте БКРС: https://bkrs.

info/proper_convert.php), при проверке даже на небольшой выборке он выдает регулярные ошибки, поэтому мы решили сделать собственный транслитератор, создающий максимально «ортодоксальные» транслитерации. В качестве источника он принимает таблицу предписаний Синьхуа 1993 г., переработанную в удобный для считывания компьютером вид и сохраненную в формате .csv. В наборе данных из Википедии наш алгоритм принимал на вход слово из столбца с русским именем собственным и записывал его китайскую транслитерацию в новый столбец, предназначенный для нормативной транслитерации. Отметим, что в процессе транслитерации было найдено 30 объектов, правописание которых не могло быть предсказано созданным нами алгоритмом. Это произошло из-за отсутствия предписаний для конкретных комбинаций русских букв, в основном невозможных с точки зрения русской орфографии (имена собственные на языках России, например «Чыбыда»). Все эти случаи были исключены из рассмотрения.

Для сравнения предсказанной и реальной транслитераций было использовано три метрики близости строк. Первая – абсолютное расстояние Левенштейна. Эта метрика показывает, сколько символов в строке A необходимо вставить, убрать или заменить, чтобы получить строку B. Область значений этой метрики – натуральные числа от 0 (полное совпадение строк) до бесконечности. Вторая – нормированное расстояние Левенштейна – метрика, учитывающая не только количество замененных символов, но и длину исходной строки и показывающая отношение количества исходных символов, которые необходимо было заменить, к длине исходной строки. В данном случае область значений – десятичные дроби от 0 до 1 включительно, где 0 – полное совпадение строк, 1 – полная замена исходной строки. Третья метрика – индекс Жаккара. Эта метрика не учитывает порядка символов, зато рассматривает частотное распределение символов в строке A и в строке B. Чем более соответствуют друг другу символьные наборы и их частотность в двух строках, тем ближе значение индекса к 1, чем меньше – тем ближе значение индекса к 0. При обработке данных были сделаны попытки унифицировать последовательность элементов в русских вхождениях и их китайских переводах (в первую очередь это касается личных имен – они были приведены в порядок: (имя)?-(отчество)?-фамилия). Однако где-то могли остаться строки с несовпадающим порядком морфем или слов, и индекс Жаккара позволяет сравнить такие строки и не помечать как расхождение разный порядок слов в транслитерации Синьхуа и вхождении Wikidata.

Далее мы проанализировали распределение объектов нашего набора данных в зависимости от абсолютного расстояния Левенштейна. Расстояние, равное 0, оказалось у 24% объектов, равное 1 – у 20% объектов, равное 2 – у 14% объектов и равное 3 – у 15% объектов (см. график 1). Мы также рассмотрели среднее расстояние Левенштейна, сгруппировав данные по типам объектов (см. график 2). Оказалось, что наименьшее среднее расстояние свойственно именам собственным (чуть меньше 1), в то время как географические объекты имели среднее расстояние в промежутке между 1 и 2,5, а населенные пункты – более 3. Это позволяет предположить, что личные имена – тип объектов, при транслитерации которого наиболее последовательно соблюдаются прескрипции Синьхуа. Это подтверждает распределение нормированного расстояния Левенштейна на подвыборке личных имен в датасете, где около 80% объектов имеют метрику в интервале (0,0–0,1), а 15% объектов – в интервале (0,1–0,2). Кроме того, мы рассмотрели среднее расстояние Левенштейна, сгруппировав данные по странам (график 3), в результате чего получили наименьшее среднее расстояние по объектам в РФ (около 2), а наибольшее – в СССР (около 4).

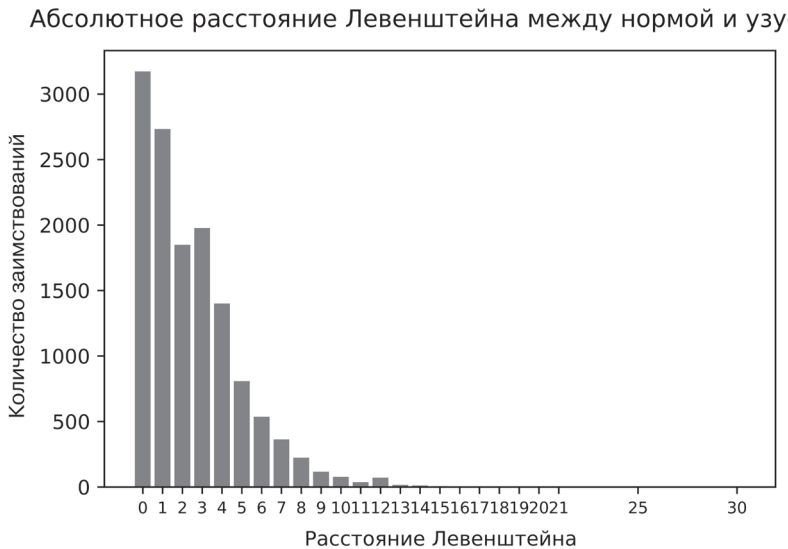

График 1. Частотное распределение абсолютного расстояния Левенштейна

График 2. Усредненное расстояние Левенштейна по типам объекта

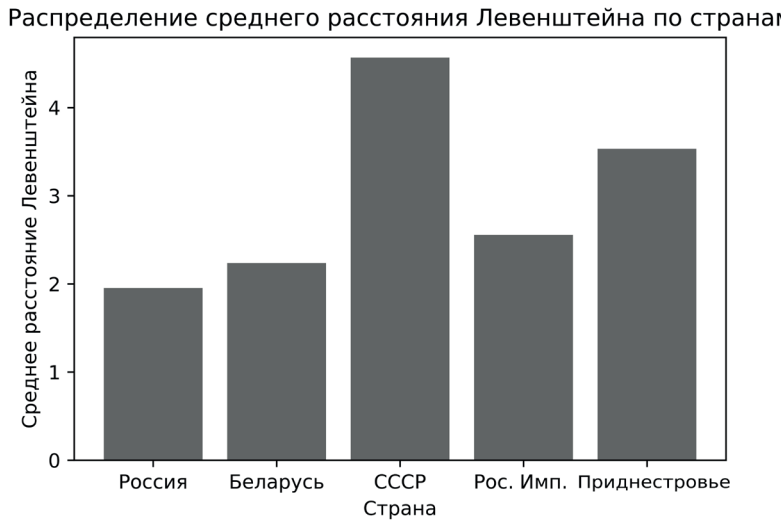

График 3. Усредненное расстояние Левенштейна по странам

Также были проанализированы нормированное расстояние Левенштейна (график 4) и индекс Жаккара (график 5) на всех данных. Гистограммы показывают, что получившиеся распределения «зеркально» отражают друг друга. При этом около 3 тыс. объектов (что составляет около 22%) показывают полное совпадение нормы и узуса по обеим метрикам, а также наблюдается два «всплеска» на уровнях 0,2 и 0,5 (нормированное расстояние Левенштейна) и 0,8 и 0,5 (индекс Жаккара).

Наконец, существует около 500 объектов, полностью не совпадающих по каждой из метрик. Ручной анализ этих вхождений показал, что основная их часть – полные семантические кальки (напр., Первомайское и «五一» “wǔyī” /ǔːjīː/, досл. «первое мая») либо дальневосточные топонимы (напр., Владивосток и «海参崴» “hǎishēnwǎi” /xǎiʂənwǎi/ – «Излучина трепанга»).

Можно сделать вывод о том, что соответствие между предписаниями Синьхуа и реальными данными далеко не полное. Это не позволяет нам, например, автоматически сгенерировать еще несколько десятков тысяч китайских транслитераций на основе всего набора собственных имен, имеющихся в русском разделе Википедии.

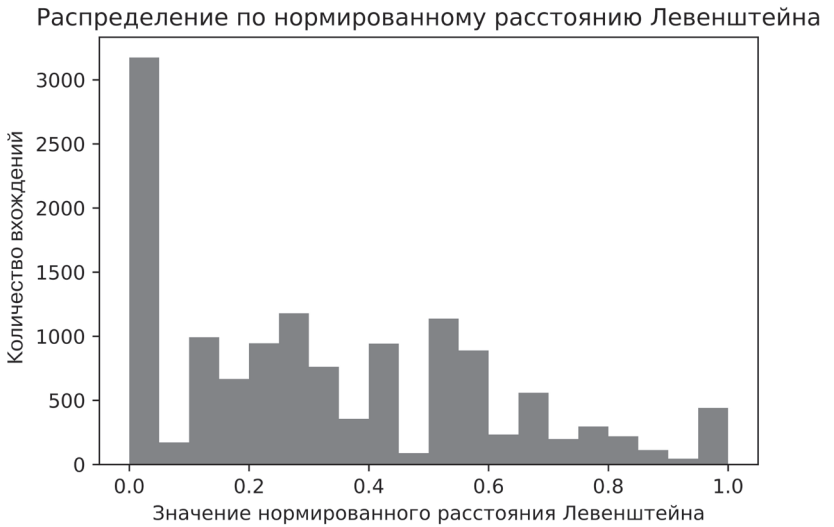

График 4. Частотное распределение
по нормированному расстоянию Левенштейна

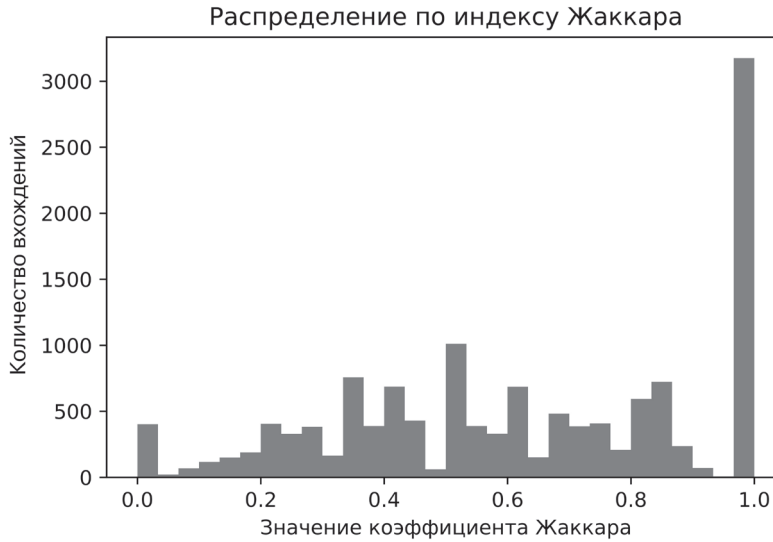

График 5. Частотное распределение по индексу Жаккара

Зависимость между расстоянием Левенштейна и дихотомией «личное имя VS географический объект» позволяет предположить, что в географических объектах, даже при использовании в целом фонетической транслитерации, части названий (прилагательные или морфемы в сложных словах) переводятся на китайский семантически мотивированными морфемами. Эта гипотеза выливается в достаточно обширное исследование, которое также было нами проведено, однако, чтобы не слишком отклоняться от проблематики данной статьи, промежуточные итоги этого исследования здесь приводиться не будут. Покажем лишь один пример такой частично семантической адаптации.

(11) 下- 诺夫哥罗德

“xià-“ “nuòfūgēluódé”

/ɕà-/ /nʷòːfʷūːkɤ̄ːlʷóːtɤ́ː/

низ (сем.) Новгород (фон.)

5.2. Анализ N-грамм в фонетических заимствованиях и в контрольных китайских текстах

В предыдущем разделе мы обнаружили, что нормативные предписания не описывают весь узус, представленный в Интернете. Это подтверждает и следующее наблюдение: было подсчитано символьное разнообразие иероглифов в списке всех переводов из Wikidata и в списке всех транслитераций по Синьхуа. Первых оказалось 877, вторых – 260. Можно ли в таком случае извлечь пользу из такого многообразия иероглифов, не умещающегося в рамки нормы?

Мы попробуем ответить на этот вопрос при помощи анализа частотности иероглифических (символьных) N-грамм. Для этого мы возьмем 1 тыс. самых частотных N-грамм, встреченных в китайских заимствованиях из Wikidata, и проверим по контрольному корпусу современной китайской литературы, насколько часто эти N-граммы встречаются в аутентичных китайских текстах. Мы сделали сравнение при N = 2 (биграммы) и при N = 3 (триграммы). В качестве контрольного корпуса мы взяли художественные произведения трех современных китайских авторов (Мо Янь, Лю Чжэньюнь и Юй Хуа) и создали корпус размером 860 тыс. символов.

Анализ тысячи наиболее частотных биграмм в заимствованиях показал, что 936 из этих биграмм вообще не встречаются в контрольном корпусе, а бóльшая часть оставшихся биграмм имеет ранг выше, чем 10 тыс. Что касается тысячи наиболее частотных триграмм, то уже 988 из них не встречаются в контрольном корпусе вовсе, а все остальные (за исключением одной) имеют ранг выше 400 тыс.

Это позволяет сделать вывод о том, что даже простые модели, построенные на символьных N-граммах, должны будут показать хороший результат при дальнейшей реализации прикладных задач, таких как вычленение заимствованных именованных сущностей.

6. Перспективы исследования

Результаты, полученные в данном исследовании, преждевременно считать окончательными. Например, в разделах 3 и 4 был проигнорирован так называемый графический фактор, о котором не раз упоминала Мяо [Miao 2005, pp. 46, 56, 79], но который с трудом поддается исчислению и воспроизводимости, а следовательно, и верификации. Речь идет о том, что в пиньинь привычные для европейцев символы согласных обозначают звуки с совершенно другой артикуляционной характеристикой. Например, “z”, в боль-

шинстве европейских письменностей обозначающий щелевой, в пиньинь помечает аффрикату /ʦ/. При сегодняшнем культурном доминировании английского языка и латинской письменности нельзя отвести такое предположение: носители китайского, размышляя о возможной транскрипции того или иного русского имени собственного, обращают внимание в первую очередь на английскую транслитерацию русского слова, а затем на соответствие английских букв символам пиньинь. Тогда объяснение аффрикаты в транслитерации, например, названия города Рязань становится более прозаичным, чем через влияние пиджина (см. пример 12). Контраргументом здесь может послужить существование системы прямой китайско-русской транскрипции Палладия, которая более точно передает различия аффрикат и щелевых, обозначая рассмотренный выше звук как «цз»; однако эта система более известна именно в России и нацелена на адаптацию слов в обратном направлении — из китайского в русский.

(12) Рязань > (гипотеза) англ. Ryazan >
кит. 梁赞 “liángzàn” /ljáŋʦæ̀n/

Кроме этого, в первых разделах не было рассмотрено ни адаптации консонантных кластеров, ни адаптации гласных в заимствованиях в путунхуа. Особый интерес в этом разрезе представляет не раз высказанное предположение о том, что ударение в словах языка-донора влияет на тоновое маркирование в полученных словах (одну из вариаций этой гипотезы можно найти в [Ин, Шипановская 2016]). Дальнейшие исследования должны будут завершить описание всех сегментов и супрасегментных единиц русских заимствований в китайском.

Очевидно, найденные фонетические и (в особенности) графические закономерности вдохновляют и на прикладные исследования в сфере автоматической обработки китайского языка. Наиболее важной задачей в этой области видится создание алгоритма, позволяющего выделять в китайских текстах именованные сущности, являющиеся фонетическими заимствованиями из русского языка (возможно, с дальнейшим их обратным переводом на русский). Учитывая неспособность норм Синьхуа предсказать реально используемые фонетические заимствования, следует использовать статистические модели порождения и распознавания заимствований. Для этих целей могут быть использованы как традиционные статистические подходы – например, модель скрытых цепей Маркова, – так и нейросетевые модели. Важным аспектом при составлении работающих моделей для китайского языка станет предва-

рительное ознакомление с уже существующими подходами к выделению именованных сущностей в изолирующих языках, таких как индонезийский [Wibawa, Purwarianti 2016], [Gunawan, Suhartono, Purnomo, Ongk 2018].

В дальнейшем практическом исследовании необходимо будет учитывать и вхождения со вспомогательными (такие, как кавычки) и некитайскими (такие, как «буквенные слова» всех видов и вообще использование латиницы) символами, так как в Интернете именно такие обозначения нередко используются для передачи фонетических заимствований именованных сущностей из других языков. Один из показательных примеров – страница в китайской Википедии, посвященная российскому рэперу LOC-Dog:

(13)	LOC-	狗
	/lok/	“gǒu” /kəu/
	LOC (граф.)-	собака (калька Dog)

Алгоритмы, полученные в результате дальнейшего практического исследования, можно будет оценить на данных русско-китайского параллельного корпуса НКРЯ, в котором содержатся всевозможные виды русских заимствований (названия населенных пунктов, ручьев, уменьшительно-ласкательные русские имена и т. д.).

Благодарности

Автор искренне благодарит Б.В. Орехова (НИУ ВШЭ) за помощь в выработке методологии и О.В. Дерезу (НИУ ВШЭ) за помощь в формулировке работы на начальных этапах. Кроме того, автор выражает признательность Д.А. Скоринкину (НИУ ВШЭ) за консультации в области поисковых запросов на языке SPARQL и Л.С. Холкиной (РГГУ) за консультации по поводу прескриптивных норм Синьхуа.

Acknowledgements

The author is sincerely grateful to B.V. Orekhov (HSE University) for help in developing the methodology and to O.V. Dereza (HSE University) for help in formulating the work at the initial stages. In addition, the author expresses his gratitude to D.A. Skorinkin (HSE University) for consultations in the field of search queries in the SPARQL language and L.S. Kholkina (RSUH) for consultations on the Xinhua prescriptive norms.

Литература

Завьялова 1996 – *Завьялова О.И.* Диалекты китайского языка. М.: Научная книга, 1996. 207 с.

Ин, Шипановская 2016 – *Ин Ц., Шипановская Л.М.* Русские заимствования в китайском языке как результат языковых контактов // Филологические науки. Вопросы теории и практики. 2016. Т. № 7 (61): В 3 ч. Ч. 1. С. 144–152.

Кодзасов, Кривнова 2001 – *Кодзасов С.В., Кривнова О.Ф.* Общая фонетика. М.: РГГУ, 2001. 590 с.

Ма 2015 – *Ма Н.* Харбинский русско-китайский пиджин первой половины XX века и его влияние на русский и китайский языки и культуры // Филологические науки. Вопросы теории и практики. 2015. Т. 3. № 8 (50). С. 122–127.

Перехвальская 2008 – *Перехвальская Е.В.* Русские пиджины. СПб.: Алетейя, 2008. 363 с.

Семенас 2005 – *Семенас А.Л.* Лексика китайского языка. Базовый учебник. М.: АСТ: Восток–Запад, 2005. 310 с.

Семенов 2019 – *Семенов К.И.* Стратегии преобразования иноязычных фонетических сочетаний в китайском языке (на материале заимствований из русского языка): Курсовая работа студента 2 курса бакалавриата. М.: НИУ ВШЭ, 2019. 55 с.

Попова, Таката 2017 – Словари кяхтинского пиджина / Под ред. И.Ф. Попова, Т. Таката. М.: Наука, 2017. 603 с.

Хаматова 2003 – *Хаматова А.А.* Словообразование современного китайского языка. М.: Муравей, 2003. 224 с.

Цзе 2007 – *Цзе Я.* Забайкальско-маньчжурский препиджин: опыт социологического исследования // Вопросы языкознания. 2007. № 2. С. 67–74.

Cook 2018 – *Cook A.* A typology of lexical borrowing in Modern Standard Chinese // Lingua Sinica. 2018. T. 4, no. 1. P. 6.

Fujii, Ishikawa 2001 – *Fujii A., Ishikawa T.* Japanese/English Cross-Language Information Retrieval: Exploration of Query Translation and Transliteration // Computers and the Humanities. 2001. Vol. 35, no. 4. P. 389–420.

Gunawan, Suhartono, Purnomo, Ongk 2018 – *Gunawan W., Suhartono D., Purnomo F., Ongk A.* Named-Entity Recognition for Indonesian Language using Bidirectional LSTM-CNNs // Procedia Computer Science. 2018. Vol. 135. P. 425–432.

Huang, Chen 1996 – *Huang C.-R., Chen K.-J.* Issues and topics in Chinese natural language processing. Taipei: Chinese University Press, 1996. 22 p.

Koo 2015 – *Koo H.* An unsupervised method for identifying loanwords in Korean // Language Resources and Evaluation. 2015. Vol. 49, no. 2. P. 355–373.

Lin 2008 – *Lin Y.-H.* Variable vowel adaptation in Standard Mandarin loanwords // Journal of East Asian Linguistics. 2008. Vol. 17, no. 4. P. 363–380.

Miao 2005 – *Miao R.* Loanword Adaptation in Mandarin Chinese: Perceptual, Phonological and Sociolinguistic Factors. Ph.D. Thesis. Stony Brook University, Stony Brook, NY, 2005. 184 p.

Nelson 2013 – *Nelson L.* English Loan Words in Mandarin Chinese: Phonology vs. Semantics // Proc. of the National Conf. on Undergraduate Research (NCUR) 2013, University of Wisconsin La Crosse, WI April 11–13, 2013. Madison, 2013. P. 497–505.

Vervaet 2017 – *Vervaet R.* English Loanwords in the Chinese Lexicon. MA Thesis. Gent: Universiteit Gent, 2017. 128 c.

Wibawa, Purwarianti 2016 – *Wibawa A., Purwarianti A.* Indonesian Named-entity Recognition for 15 Classes Using Ensemble Supervised Learning // 5th Workshop on Spoken Language Technologies for Under-resourced languages, SLTU, 9–12 May 2016. Yogyakarta, 2016. P. 221–228.

Wong, Xu 2010 – *Wong K.-F., Xu R.* Introduction to Chinese Natural Language Processing. San Rafael, CA: Morgan & Claypool Publ., 2010. 148 p.

刘正琰 (liú zhèngtán) 1985 – 汉语外来词词典 (hànyǔ wàiláicí cídiǎn) / ed. 刘正琰 (liú zhèngtán). 上海 (shànghǎi): 上海辞书出版社 (shànghǎi císhū chūbǎnshè), 1985. 422 p.

新华 (xīnhuá) 1993 – 世界人名翻译大辞典 (shìjiè rénmíng fānyì dà cídiǎn)，新华社译名室编 (xīnhuá shè yìmíng shì biān), 1993. 56 p.

References

Cook, A. (2018), "A typology of lexical borrowing in Modern Standard Chinese", *Lingua Sinica*, vol. 4, no. 1, p. 6.

Fujii, A. and Ishikawa, T. (2001), "Japanese/English Cross-Language Information Retrieval: Exploration of Query Translation and Transliteration", *Computers and the Humanities*, vol. 35, pp. 389–420.

Gunawan, W., Suhartono, D., Purnomo, F. and Ongk, A. (2018), "Named-Entity Recognition for Indonesian Language using Bidirectional LSTM-CNNs", *Procedia Computer Science*, vol. 135, pp. 425–432.

Huang, C.-R. and Chen, K.-J. (1996), *Issues and Topics in Chinese Natural Language Processing*, Chinese University Press, Taipei, Taiwan.

Jie, Y. (2007), "Baikal-Manchurian pre-pidgin: a sociological study", *Voprosy yazykoznaniya*, vol. 2, pp. 67–74.

Khamatova, A.A. (2003), *Slovoobrazovaniye sovremennogo kitayskogo yazyka* [Word formation in Modern Chinese], Muravei, Moscow, Russia.

Kodzasov, S.V. and Krivnova, O.F. (2001), *Obshchaya fonetika* [General phonetics], RGGU, Moscow, Russia.

Koo, H. (2015), "An unsupervised method for identifying loanwords in Korean", *Language Resources and Evaluation*, vol. 49, no. 2, pp. 355–373.

Lin, Y.-H. (2008), "Variable vowel adaptation in Standard Mandarin loanwords", *Journal of East Asian Linguistics*, vol. 17, no. 4, pp. 363–380.

Liu, Zh.-T. (1985), *Hanyu wailaici cidian* [Chinese Loanword Dictionary], Shanghai cishu chubanshe, Shanghai, China.

Ma, N. (2015), "Harbin Russian-Chinese pidgin in the first half of 20th century and its influence on Russian and Chinese languages and culture", *Filologicheskiye nauki. Voprosy teorii i praktiki*, vol. 3, no. 8 (50), pp. 122–127.

Miao, R. (2005), *Loanword Adaptation in Mandarin Chinese: Perceptual, Phonological and Sociolinguistic Factors*, Ph.D. Thesis, Stony Brook University, Stony Brook, NY, USA.

Nelson, L. (2013), "English Loanwords in Mandarin Chinese: Phonology vs. Semantics", *Proc. of the National Conf. on Undergraduate Research (NCUR) 2013*, University of Wisconsin La Crosse, WI April 11–13, 2013, Madison, pp. 497–505.

Perekhval'skaya, E.V. (2008), *Russkiye pidzhiny* [Russian pidgins], Aleteiya, Saint-Petersburg, Russia.

Wibawa, A. and Purwarianti, A. (2016), "Indonesian Named-entity Recognition for 15 Classes Using Ensemble Supervised Learning", *5th Workshop on Spoken Language Technologies for Under-resourced languages, SLTU*, 9–12 May, 2016, Yogyakarta, pp. 221–228.

Semenas, A.L. (2005), *Leksika kitayskogo yazyka. Bazovyy uchebnik* [The Chinese lexicon. Basic Textbook], AST: Vostok-Zapad, Moscow, Russia.

Semenov, K.I. (2019), *Strategii preobrazovaniya inoyazychnykh foneticheskikh sochetaniy v kitayskom yazyke (na materiale zaimstvovaniy iz russkogo yazyka)* [Adaptation Strategies for Foreign Phonetic Borrowings in Chinese (Based on Russian Loanwords)], 2nd year BA term paper, NRU "Higher School of Economics", Moscow, Russia.

Popova, I.F. and Takata, T. (2017), *Slovari kyakhtinskogo pidzhina* [Dictionaries of Kyakhta pidgin], Nauka, Moscow, Russia.

Vervaet, R. (2017), *English Loanwords in the Chinese Lexicon*, MA Thesis, Universiteit Gent, Gent, Belgium.

Wong, K.-F. and Xu, R. (2010), *Introduction to Chinese Natural Language Processing*, Morgan & Claypool Publ., San Rafael, CA, USA.

Xinhua (1993), *Shijie renming fanyi da cidian* [Translation dictionary for world personal names], Xinhua she yiming shi bian, Beijing, China.

Ying, J. and Shipanovskaya, L.M. (2016), "Russian loanwords in Chinese as a result of language contacts", *Filologicheskiye nauki. Voprosy teorii i praktiki*, no. 7 (61), pt. 1, pp. 144–152.

Zav'yalova, O.I. (1996), *Dialekty kitayskogo yazyka* [Dialects of Chinese], Nauchnaya kniga, Moscow, Russia.

Информация об авторе

Кирилл И. Семенов, Национальный исследовательский университет «Высшая школа экономики», Москва, Россия; 101000, Россия, Москва, ул. Мясницкая, 20; kir.semenow@yandex.ru

Information about the author

Kirill I. Semenov, HSE University, Moscow, Russia; bld. 20, Myasnitskaya Str., Moscow, Russia, 101000; kir.semenow@yandex.ru