DOI: 10.24411/2686-7702-2020-10002

Языковая политика и языковые ресурсы в китайском интернете

О.И. Завьялова

***Аннотация*:** Языковая политика, которой китайское государство уделяет значительное внимание на протяжении последних десятилетий, освещается на многочисленных официальных сайтах центрального и регионального уровня. Среди задач, поставленных перед лингвистами, – ежегодный статистический анализ слов и иероглифов, которые употребляются в современном китайском языке, прежде всего, в интернете. Технологии, позволяющие автоматически выделять в иероглифических цепочках слова и определять их принадлежность к той или иной части речи в изолирующем китайском языке, лежат также в основе текстовых корпусов, многие из которых доступны в сети. Исследования, связанные с распознаванием и синтезом устной речи, востребованы при разработках в области машинного устного перевода, в процессе изучения *путунхуа* на специальных сайтах и при создании устных языковых корпусов, в том числе онлайновых диалектных на Тайване и в Гонконге.

***Ключевые слова*:** китайский язык, языковая политика, интернет, *путунхуа*, *вэньянь*, *байхуа*, китайские диалекты, диалекты Миньнань, кантонский диалект, иероглифическая письменность, лексика, статистические исследования, иероглиф года, слово года, языковой корпус, искусственный интеллект, синтез речи, распознавание речи, машинный перевод, языковые ресурсы.

***Автор*:** *Завьялова Ольга Исааковна*, доктор филологических наук, главный научный сотрудник, Институт Дальнего Востока РАН (адрес: 117997, Москва, Нахимовский пр-т, 32). ORCID: 0000-0003-3355-4598; E-mail: olgazavyalova@yahoo.com

Language policy and language resources on the Chinese Internet

O.I. Zavyalova

***Abstract*:** The language policy and planning in the PRC are represented on various official central and local websites. The annual statistical analysis of Chinese words and characters used online are among the tasks, set by the authorities before the linguists. The technologies, allowing to recognize words within the character texts and to mark them as belonging to a particular part of speech in the isolating syllabic Chinese language are applied within the process of creating numerous text corpora, often accessible on the Web. Speech recognition and speech synthesis technologies based on AI are used in machine translation, on the official websites for Standard Mandarin learners, as well as in spoken corpora. Among the latter are dialect corpora created in Taiwan and Hong Kong and available online.

***Keywords*:** Standard Mandarin, Chinese dialects, Minnan dialects, Cantonese, Chinese characters, language policy, language planning, Internet, lexical frequency studies, part-of-speech tagging, language

corpora, artificial intelligence, speech synthesis, speech recognition, machine translation, language resources, character of the year, word of the year.

Author*: *Zavyalova Olga I., Doctor of Sciences (Philology), Leading Researcher, Institute of Far Eastern Studies of the Russian Academy of Sciences (address: 32, Nakhimovsky Av., Moscow, 117997, Russian Federation). ORCID: 0000-0003-3355-4598; E-mail: olgazavyalova@yahoo.com

Лингвистические инициативы периода «реформ и открытости»

Без использования высоких технологий китайский язык, несмотря на более чем миллиардное число носителей и древнейшую письменную, книжную и библиотечную традиции, вряд ли бы занял достойное и конкурентоспособное место среди других языков. Уже в 1970-е гг. в КНР были поставлены инновационные задачи в отношении китайского языка. В 1974 г. к исследованиям, ориентированным на решение теоретических и практических проблем, приступил Институт компьютерных технологий при Пекинском университете (*Бэйцзин дасюэ цзисуаньцзи яньцзюсо*) [Цзисуаньцзи…].

В 1980-е гг. в континентальном Китае и на Тайване были разработаны системы кодирования иероглифов. С помощью обычной клавиатуры, уже изобретённой за пределами китайского мира для алфавитных систем письма, стал возможным ввод на компьютеры тысяч находящихся в употреблении иероглифов – как с принятым в КНР упрощённым написанием некоторых знаков, так и с сохраняющимися на Тайване и в Гонконге традиционными вариантами [Ванло…]. В результате было не только положено начало революции в книгоиздании, но также открыта дорога к созданию в 1994 г. китайского интернета. В то время он насчитывал всего 10 тыс. пользователей, которые получили возможность читать во всемирной сети иероглифические тексты на китайском языке. В конце 1990-х появились два первых сайта на языках малых народов КНР. В 1998 г. был разработан уйгурский сайт с использованием арабско-персидской графики, в 1999 г. – сайт на тибетском языке, сохраняющем письмо, которое восходит к индийскому деванагари. В 2017 г. в китайском сегменте интернета уже активно действовали 1140 сайтов на языках малых народов с двенадцатью вариантами письменных систем, в том числе особых традиционных [Шаошу…][1].

Учреждённый в 1954 г. Комитет по реформе китайской письменности в 1985 г. был переименован в Государственный комитет по работе в области языка и письменности. В 1994 г. он вошёл в состав Министерства образования КНР (в то время – Государственного комитета по образованию). Тогда же были созданы два важнейших подразделения, ведавших проблемами использования (*Юйянь вэньцзы инъюн гуаньлисы*) и информатизации (*Юйянь вэньцзы синьси гуаньлисы*) языка и письменности [Юйянь вэньцзы синьси…]. В 2001 г. вступил в силу первый в истории Китая «Закон об общеупотребительных в государстве языке и письменности». В документе шла речь, прежде всего, о статусе и функциях официального языка *путунхуа*, в основе которого лежит пекинский и отчасти другие северные диалекты, а также о стандартном иероглифическом письме с включением принятых в 1950-е гг. в КНР упрощённых вариантов иероглифов. Предшественник *путунхуа* – «государственный язык» *гоюй* – получил своё название и

[1] Подробнее об исследованиях и информатизации языков малых народов Китая см.: [Завьялова, 2016; Завьялова, 2017].

по сути дела стал официальным ещё в 1909 г., в конце правления династии Цин, в КНР был переименован в *путунхуа*, но на Тайване до сих пор называется «государственным».

В 2002 г., когда число пользователей китайского интернета достигло почти 46 млн человек, появился официальный сайт «Язык и письменность в Китае», который сейчас функционирует внутри сайта Министерства образования [Чжунго юйянь вэньцзы…]. К концу 2005 г. в сети было уже 100 аналогичных официальных региональных сайтов. Больше всего – 18 сайтов – было создано в крупнейшем мегаполисе Шанхае с населением 17 млн человек, по данным на 2000 г. В примыкающей к нему провинции Цзянсу, которая, как и Шанхай, входит в экономически развитый ареал распространения диалектов группы У, было разработано 11 сайтов [Юйянь вэньцзы синьси…; Юйянь вэньцзы гунцзо…]. Параллельно в интернете появлялись многочисленные сайты с разнообразными китайскими языковыми ресурсами: оцифрованными письменными памятниками, доступными пользователям словарями разных периодов и назначения, лингвистическими исследованиями.

Инновационные программы для изолирующего китайского языка

Важнейшая составляющая языковых программ последних десятилетий – исследования, связанные с использованием технологий искусственного интеллекта при распознавании, синтезе и машинном переводе устной речи, а также в процессе обучения *путунхуа*, в том числе на специальных сайтах. В 2016 г. одна из ведущих китайских компаний в области языковых технологий iFlytek сумела создать устройство, которое распознавало и переводило китайскую устную речь на английский, уйгурский, японский и корейский языки. Одновременно перевод демонстрировался в письменном виде на большом экране. Одно из подразделений этой компании, действующее в Урумчи, разрабатывает мобильные устройства, которые переводят китайскую устную речь на языки Синьцзяна и соседних стран Шёлкового пути. В 2017 г. один из таких гаджетов с использованием уйгурского и китайского языков был представлен премьеру Ли Кэцяну во время его встречи в провинции Аньхой с местными депутатами ВСНП [Ян Эрхун, Хоу Минь, 2016; Ли Кэцян лай…].

Особое внимание уделялось также автоматическому делению на слова и выявлению принадлежности выделенных слов к той или иной части речи в иероглифических текстах на *путунхуа* [Ян Эрхун, Хоу Минь, 2016]. Исследования велись с учётом слогового изолирующего характера китайского языка и особенностей иероглифической письменности.

Известно, что морфема в современном китайском языке, как в *путунхуа*, так и в диалектах, как правило, фонетически равна слогу, который, в свою очередь, равен морфеме. На письме каждому слогу соответствует индивидуальный иероглиф, который в большинстве случаев «сопровождал» морфему на протяжении всей истории её существования. Сравнительно малочисленные двусложные или многосложные морфемы-исключения – это, прежде всего, слова-полуповторы и фонетические заимствования из других языков. В случае эризации две морфемы, напротив, представлены одним слогом. В диалектах есть и другие исключения из правила «морфема равна слогу», например, «разделившиеся слова» (分音词 *fēnyīncí*) в диалектах группы Цзинь.

В древнекитайском языке и в основанном на древней лексике и грамматике письменном языке *вэньяне*, который веками использовался в качестве единственного официального, слово, как правило, было образовано одной корневой морфемой и записывалось

соответствующим индивидуальным иероглифом. В *путунхуа* и в современных диалектах – так же, как в предшественнике письменной формы *гоюя* и *путунхуа*, неофициальном «разговорном» языке *байхуа*, который сформировался на основе северных диалектов начиная со средних веков, – слова могут состоять из одной, двух и более морфем-слогов. Чаще всего это корни, реже – аффиксы, словообразовательные или обозначающие число существительного, время и вид глагола. Во многих случаях определить принадлежность того или иного сочетания корневых морфем к слову или словосочетанию очень трудно. В дополнение к этому в цепочке иероглифов, записывающих обычный текст на *путунхуа* или *байхуа*, а также на современных диалектах (например, кантонском), пробелы между словами отсутствуют, все иероглифы – вне зависимости от того, записывают они словосочетание или слово, – находятся на равном расстоянии друг от друга. Автоматически определить границы большей части слов в иероглифическом тексте, таким образом, можно только с помощью особых специально разработанных программ.

Статистические исследования иероглифов и слов

Структурными особенностями китайского языка и соотношением морфема-слог-иероглиф определяются, в частности, методы, которые на протяжении последних лет используются в процессе статистического анализа текстов, ориентированного на разработку норм для *путунхуа* и современной иероглифической письменности. Анализ осуществляется Государственным центром мониторинга и изучения языковых ресурсов (*Гоцзя юйянь цзыюань цзяньцэ юй яньцзю чжунсин*), который был создан ещё в 2004 г. в составе упомянутого выше другого государственного центра, ведающего проблемами информатизации языка и письменности. Основной объект исследований – это публикации в СМИ, в том числе новостные в интернете, а также собственно язык сети, или та особая, по сути, нерегулируемая разновидность китайского языка, которая используется в блогах и чатах. Эта разновидность включает в себя оригинальные, созданные пользователями китайские новые слова и фразеологизмы; иногда диалектизмы, которые в интернете неожиданно становятся общекитайскими; нестандартные заимствования из английского языка. Пользователи интернета могут изобретать также особые иероглифические формы записи собственно китайских слов и оригинальные, нигде больше не употребляющиеся разновидности «буквенных слов» (сокращения иностранных и китайских слов или словосочетаний в латинском написании) [Сбоев].

Итоги статистических исследований сначала публиковали в формате второго тома официального ежегодника «Доклад о языковой ситуации в Китае» (*Чжунго юйянь шэнхо чжуанкуан баогао*) и позже стали прилагать к основному тому на диске (первый ежегодник был выпущен в 2006 г. и посвящён 2005 г.). Ранние обзоры были в значительной степени ориентированы не только на анализ слов или словосочетаний, употреблявшихся в СМИ, интернет-новостях и блогах, но также на статистическую обработку иероглифов. Списки обнаруженных знаков сравнивали с обнародованными в 1988 г. нормативными списками разного уровня, содержащими 2500, 3500 и 7000 иероглифов, в зависимости от частоты их употребления в современном китайском языке. Выявляли 20 реально наиболее употребительных знаков, с одной стороны, и те иероглифы из нормативных списков, которые не были обнаружены в проанализированных текстах, с другой [Баочжи…]. Более

поздние исследования были в значительной степени сосредоточены – возможно, благодаря появлению продвинутых технологий по распознаванию слов в иероглифических цепочках – на лексических единицах. Так, по результатам исследования в 2018 г. были обнародованы четыре списка, посвящённых языку СМИ: 1) всех находившихся в употреблении иероглифов; 2) наиболее частотных слов; 3) неологизмов; 4) обнаруженных в текстах фразеологических единиц *чэнъюев*.

Не меньшее внимание в последние годы уделялось ещё одному статистическому исследованию иероглифов и лексики. В 2006 г. Государственным центром мониторинга и изучения языковых ресурсов, знаменитым издательством «Шанъу иньшугуань», а также интернет-компанией Синьлан (*Sina Corp*) была инициирована «Инвентаризация китайского языка» (*Ханьюй паньдянь*) – аналог западных акций «слово года» (позже в проекте принимали участие другие организации и СМИ). Анализ основан на обработке данных, полученных при опросе многочисленных пользователей интернета, и проходит под лозунгом «Одним иероглифом, одним словом описываем Китай и мир». По результатам опросов назначают один иероглиф и одно слово или словосочетание для Китая и один иероглиф и одно слово или словосочетание для внешнего мира [Ханьюй паньдянь…]. Позже в ходе акции стали дополнительно выбирать десять самых популярных обычных слов/словосочетаний и десять самых популярных неологизмов в СМИ, а также десять наиболее востребованных слов/словосочетаний в языке пользователей интернета.

«Слово года» впервые было названо в Германии в 1971 г., в России – в 2007 г. Акция, посвящённая иероглифу года, в Японии проводится с 1995 г., на Тайване она была инициирована в 2008 г. [Дай Хунлян, Ян Шуцзюнь]. В 2019 г. для КНР был выбран иероглиф (и по сути дела, соответствующее односложное слово) 稳 *wěn* «стабильный» и слово (на самом деле словосочетание) «Я и моя родина» (我和我的祖国 *Wǒ hé wǒ de zǔguó*) – название песни и популярного юбилейного фильма, снятого к 70-летию КНР. В то же время для внешнего мира подобрали менее оптимистичные варианты – иероглиф 难 *nán* «трудный» и словосочетание 贸易摩擦 *màoyì mócā* «внешнеторговые трения» [Ханьюй паньдянь 2019…].

Языковые ресурсы в свете тысячелетних библиотечных традиций

Среди первоочередных задач, решение которых также связано с проблемой автоматического деления текста на слова, – создание языковых корпусов разного объёма и назначения. Эти задачи имеют для Китая особое значение с учётом его давних словарной, книжной и библиотечной традиций. На протяжении столетий в династийных историях были разделы, которые назывались «Сведения о книгах». Первую императорскую библиотеку стали собирать в пределах всей страны уже при династии Хань по указу императора Чэн-ди (32–7 гг. до н.э.). Тогда же был составлен каталог под названием «Семь подразделений» («Ци люэ»), в котором была предложена классификация ещё рукописных в то время книг [Меньшиков, с. 10–113].

В период Троецарствия в государстве Вэй (221–264 гг. н.э.) один из хранителей библиотеки по имени Сюнь Сюй предложил классификацию книг и их распределение в хранилищах по четырём отделам *сы ку* (四库 *sì kù*): конфуцианские классики, исторические сочинения, философские трактаты, изящная литература. При этом буддийские сочинения, которые часто либо содержали элементы разговорного языка, либо вообще были написаны

на *байхуа*, стали включать в официальные, но, как правило, не основные императорские книгохранилища только начиная с V в. В VII в. классификация «по четырём отделам» стала основной в Китае. В своём окончательном виде она была представлена в библиографии книг знаменитой императорской библиотеки «Сы ку цюань шу» – «Полной библиотеки с четырьмя отделами», созданной по распоряжению императора Цяньлуна. Текст «Пояснений к сводному каталогу всех книг по четырём отделам» (*Сы ку цюань шу цзунму тияо)* был завершён в 1781 г. и опубликован в 1790–1794 гг. [Меньшиков, с. 84]. Значительная часть сохранившихся текстов библиотеки «по четырём отделам» уже оцифрована и представлена в интернете.

Первый языковой корпус в Китае был разработан Уханьским университетом в 1979 г. и включал только современные литературные произведения [Ян Эрхун, Хоу Минь, с. 499]. К настоящему времени в КНР созданы многочисленные текстовые корпусы китайского языка (устных проектов сравнительно мало, и они не доступны в интернете). Поисковые системы, как правило, разработаны с учётом не только деления текстов на слова, но также принадлежности выделенных слов к той или иной части речи [Колпачкова]. Китайские лингвисты при этом в процессе анализа текстовых корпусов вместо слова «морфема» часто употребляют слово «иероглиф» даже в статьях на русском языке: «В сложных словах между иероглифами существуют разнообразные связи и отношения» [Лу Исинь].

Корпусы общего характера ориентированы как на современные тексты, так и на письменные памятники прежних эпох на древнекитайском языке, его наследнике *вэньяне* и на *байхуа* разных периодов. Так, инициированный в 1984 г. корпус Китайской академии (лат. Academia Sinica) на Тайване представлен четырьмя хронологическими разделами: древним, средневековым, а также ранними и новейшими текстами на современном языке – с отдельными адресами в интернете [Хуан Цзюйжэнь].

Основные текстовые корпусы в КНР

Наиболее значимыми в континентальном Китае считаются три текстовых корпуса общего характера, все они в той или иной степени доступны в интернете [Xu Jiajin].

1. На первом месте по цитируемости, благодаря, в частности, своему значительному объёму, по данным на 2015 г., стоял языковой корпус Центра китайского языкознания при Пекинском университете [Бэйцзин даюэ CCL…]. Его разработка была инициирована в 2000 г., уже в 2004 г. в сети появился первый вариант [Бэйцзин дасюэ юйляоку…]. В современной части корпуса при этом есть только метаразметка, а синтаксическая и частеречная разметки отсутствуют, поэтому его в значительной степени следует считать базой текстовых данных [Колпачкова]. От других проектов общего характера корпус отличается большим объёмом лингвистических публикаций и обширным китайско-английским разделом, ориентированным на изучающих соответствующие языки китайцев и иностранцев.

Объём раздела на современном китайском языке в корпусе Пекинского университета – 1,2 млрд иероглифов (и, соответственно, морфем), в текстах употребляется 10645 неодинаковых иероглифических знаков. Источники нового периода, начало которому было положено «движением 4 мая» 1919 г. и связанным с ним «движением за *байхуа*», ограничены литературными произведениями. Они составляют лишь небольшую часть современного раздела, хотя именно в результате «движения 4 мая» *вэньянь* постепенно

перестал быть одним из языков художественной литературы, а затем утратил статус официального языка и начал исчезать из других областей жизни, уступая дорогу современному варианту *байхуа*, лежащему в основе письменной формы *гоюя*/*путунхуа*. Большая часть современного раздела, 98,72 %, представлена материалами новейшего периода, появившимися уже после образования КНР в 1949 г., в том числе записями устных текстов, художественными произведениями, образцами языка интернета и официально-деловыми документами.

Объём исторического раздела корпуса – 400 млн иероглифов, число неодинаковых иероглифов в текстах – 18898. Часть материалов объединена хронологически – начиная с периода династии Чжоу и заканчивая периодом Китайской Республики (в последнем случае, очевидно, имеются в виду тексты на *вэньяне*, которые продолжали появляться в этот период параллельно с текстами на *байхуа*). Часть распределена по жанрам и, в частности, содержит написанные на *байхуа* пьесы эпохи Юань.

2. «Сбалансированный корпус современного китайского языка Государственного комитета по работе в области языка и письменности» (*Гоцзя юйвэй сяньдай ханьюй пинхэн юйляоку*) был создан при поддержке Госсовета КНР Центром мониторинга и исследования языковых ресурсов с участием других подразделений Министерства образования. Разработка корпуса была инициирована в конце 1991 г., в 2005 г. он уже содержал примерно 100 млн иероглифов, его базовая часть – около 20 млн [Юйянь вэньцзы синьси…, с. 135].

В своём онлайновом варианте [Юйляоку…] современная составляющая корпуса содержит 9478 текстов и насчитывает 19,5 млн знаков. В это число включены не только иероглифы, но также латинские буквы, знаки препинания и цифры, заимствованные из европейских языков. Как известно, морфемы, обозначающие числа, традиционно и отчасти в современном языке записываются иероглифами. В общей сложности в текстах выделено 162 875 неодинаковых слов, в том числе 151 300 иероглифических китайских.

Хронологически современная составляющая корпуса охватывает период с начала «движения 4 мая» 1919 г. по настоящее время. При этом тексты 1919–1925 гг. составляют всего 5 % от общего объёма по числу знаков, отобраны источники, которые в минимальной степени включают в себя элементы *вэньяня*. На более поздние материалы периода «созревания *байхуа*» вплоть до образования Китайской Народной Республики в 1949 г. приходится 15 %. Большая часть текстов, включённых в корпус, таким образом, появилась уже в период существования КНР. Время становления «нового Китая» (1950–1965 гг.), для которого характерно появление новой «социалистической» лексики, представлено четвертью объёма корпуса. На период «культурной революции» (1966–1976 гг.) с его гонениями на интеллигенцию и особыми, ставшими сейчас историческими, словами и фразеологизмами, приходится всего 5 %. И наконец, половина общего объёма – это тексты периода «обновления китайского языка» с начала курса «реформ и открытости» по настоящее время.

Наряду с современным, на сайте сейчас присутствует обширный исторический раздел объёмом в 70 млн знаков, с источниками от периода династии Чжоу до династии Цин. Он, в частности, включает в себя значительную часть сохранившихся книг упомянутой выше императорской библиотеки «Сы ку цюань шу». В числе текстов – написанные на *байхуа* средневековые романы, чрезвычайно популярные как в прошлом, так и в настоящее время, много раз экранизированные в континентальном Китае и за его пределами.

3. Корпус Пекинского университета языков (англ. Beijing Language and Culture University) характеризуется уникальным для китайского языка объёмом – 15 млрд иероглифов [Xu Jiajin, с. 219]. Тексты периодических изданий насчитывают в составе корпуса 2 млрд иероглифов, по 3 млрд приходится на современные литературные произведения, на публикации блогеров и на научно-технические тексты, 1 млрд – на тексты общего характера (по-видимому, те, которые нельзя отнести ни к одному из перечисленных разделов). Наконец, в корпус включены тексты на древнекитайском языке (очевидно, древнекитайские и более поздние на *вэньяне*) объёмом 2 млрд знаков [Шиюн…].

Новейшие обследования и устные корпусы китайских диалектов

На протяжении последних десятилетий, помимо корпусов с письменными текстами разных периодов, лингвисты из разных научных и учебных учреждений КНР в экспериментальном порядке создавали устные корпусы многочисленных китайских диалектов. Особое значение создание подобных ресурсов приобретает в последние годы, когда диалекты, прежде всего относящиеся к далеко отстоящим от *путунхуа* южным группам в экономически развитых приморских провинциях, все более активно взаимодействуют с официальным языком. Благодаря, в частности, интернету в этих диалектах появляются в разном объёме новые заимствованные из *путунхуа* фонетические, лексические и даже грамматические явления, в том числе через специальные учебные сайты. Примером таких сайтов может служить новейшая учебная платформа по изучению *путунхуа*, созданная с использованием технологий искусственного интеллекта компанией iFlytek при участии Министерства образования и Государственного комитета по работе в области языка и письменности. Платформа появилась в интернете в октябре 2019 г. и ориентирована как на носителей китайских диалектов и языков малых народов в континентальном Китае, так и на говорящих в основном на кантонском диалекте жителей Гонконга и Макао, тайваньцев, а также китайцев и иностранцев, живущих за пределами китайского мира [Цюанцю…].

Очередная общекитайская программа сохранения языков малых народов, китайских диалектов и языковых составляющих локальных культур была инициирована Министерством образования и Государственным комитетом по работе в области языка и письменности в 2015 г. К 2018 г. она была в значительной своей части выполнена, а в некоторых разделах даже перевыполнена. Так, вместо 900 намеченных пунктов с китайскими диалектами было обследовано 1011 пунктов. Полученные материалы будут использоваться, во-первых, при создании диалектных корпусов и, во-вторых, при разработке онлайновой платформы. Её вариант некоторое время был доступен в интернете не только специалистам, но также обычным пользователям, которые могли прослушать и посмотреть записи китайских диалектов и отчасти языков малых народов, сделанные в десятках пунктов по всей стране [Чжунго юйянь цзыюань…, 2018, 2019]. Те диалектные корпусы, которые уже присутствуют в интернете, созданы за пределами континентального Китая.

На Тайване это, прежде всего, корпусы с численно доминирующими здесь диалектами группы Миньнань, которые называют здесь также «тайваньским языком» (*тайюй*). К наиболее значимым проектам с устными записями этих диалектов относят корпус, разрабатываемый с 1999 г. Государственным университетом Чжунчжэн [Тайвань…]. Материалы на втором по численности на Тайване диалекте *хакка* используются

реже. Они, в частности, включены наряду с диалектами Миньнань в устный корпус Государственного университета управления [Чжэнчжи…; Chui Kawai and Lai Huei-ling].

В Гонконге с учётом его особых языковых традиций, языковой ситуации и отличного от континентального законодательства были задействованы проекты по созданию корпусов на кантонском диалекте. Известно, что в период, когда Гонконг был колонией Великобритании, в качестве официального устного и письменного языка здесь использовался только английский. В повседневной жизни гонконгцы большей частью говорили на своём родном кантонском диалекте, который относится к далеко отстоящей от пекинского и *путунхуа* группе диалектов Юэ (Кантон – историческое европейское название города Гуанчжоу).

Действие закона КНР о языке и письменности в соответствии с политикой «одна страна – две системы» на Гонконг (САР Сянган) не распространяется. Здесь официально разрешены два письменных языка: близкий *путунхуа* вариант китайского и английский, а также три устных языковых варианта: кантонский диалект, *путунхуа* и английский. После возвращения Гонконга в состав КНР в 1997 г. только четверть его жителей в той или иной степени знали *путунхуа*. К 2011 г. число владеющих им в разной степени гонконгцев удвоилось, составив 47,8 %, хотя кантонский по-прежнему остаётся языком домашнего общения у 89 % жителей [Гутин]. От других китайских диалектов кантонский отличается своей развитой письменной традицией с не только общекитайскими, но также особыми диалектными иероглифами. Разработка корпусов кантонского диалекта была инициирована в конце 1990-х гг. [Lo, Lee, Ching], существующие варианты сделаны с использованием как устной речи, так и текстов. В качестве примера можно привести доступные в интернете корпусы, перечисленные со ссылками на страницу гонконгской Ассоциации кантонского диалекта [Cantonese…], или корпус, при создании которого были использованы записи речи актёров из гонконгских фильмов середины XX в. [The Corpus…].

Заключение

На протяжении последних десятилетий активная языковая политика в КНР осуществляется с использованием инновационных технологий и освещается на многочисленных сайтах, прежде всего официальных, центрального и регионального уровня. Среди задач, поставленных государством перед китайским научным сообществом, – разработка языковых норм на базе результатов ежегодного мониторинга присутствующих в интернете китайских слов и тысяч употребляющихся в современном китайском языке иероглифов. В основе исследования лежат разработанные китайскими лингвистами технологии, позволяющие в большей части случаев автоматически делить тексты на слова и определять их принадлежность к той или иной части речи с учётом изолирующего слогового характера китайского языка и особенностей иероглифической письменности. Эти же технологии используются при создании, в том числе в интернете, многочисленных текстовых корпусов на *путунхуа* и на его предшественнике – средневековом и более позднем письменном языке *байхуа*. Доступный в интернете корпус современного китайского языка «государственного уровня» был разработан при поддержке Госсовета КНР. Лингвистические программы последних лет включают также исследования, связанные с использованием искусственного интеллекта при распознавании и синтезе устной речи. Соответствующие

технологии востребованы при машинном переводе, в процессе изучения *путунхуа* на соответствующих сайтах, а также при создании устных корпусов на *путунхуа* и диалектах.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

Баочжи, гуанбо дяньши, ванло (синьвэнь) юн цзы, юн цы дяоча : [Анализ иероглифов и слов, употребляющихся в газетах, на радио и телевидении, а также в новостях в интернете] // Чжунго юйянь шэнхо чжуанкуан баогао (2006) : [Доклад о языковой ситуации в Китае в 2006 г.]. Пекин: Шанъу иньшугуань, 2007. Т. 2. С. 1–36.

Бэйцзин дасюэ CCL юйляоку : [Языковой корпус Центра китайского языкознания при Пекинском университете]. URL: http://ccl.pku.edu.cn:8080/ccl_corpus/ (дата обращения: 10.02.2019).

Ванло юйянь чжуанкуан : [Ситуация с употреблением языка в интернете] // Чжунго юйянь шэнхо чжуанкуан баогао (2005) : [Доклад о языковой ситуации в Китае в 2005 г.]. Пекин: Шанъу иньшугуань, 2006. Т. 1. С. 211–226.

Гутин И.Ю. Языковая ситуация в Специальном административном районе Гонконг КНР и политика властей в сфере языка // Международный научно-исследовательский журнал. 2018. № 2 (68). С. 79–83.

Дай Хунлян, Ян Шуцзюнь. Тайвань юйвэнь шэнхо чжуанкуан (2018) : [Языковая ситуация на Тайване в 2018 г.] // Чжунго юйянь шэнхо чжуанкуан баогао (2019) : [Доклад о языковой ситуации в Китае в 2019 г.]. Пекин: Шанъу иньшугуань, 2019. С. 271–279.

Завьялова О.И. Языки Китая: новейшие исследования и открытия // Проблемы Дальнего Востока. 2016. № 5. С. 139–143.

Завьялова О.И. Языки Китая в информационном пространстве // Проблемы Дальнего Востока. 2017. № 2. С. 148–152.

Колпачкова Е.Н. Корпусы китайского языка: современное состояние и основные проблемы // Труды международной конференции «Корпусная лингвистика – 2015». СПБ: Издательство Санкт-Петербургского университета, 2015. С. 278–286.

Ли Кэцян лай Аньхуэйтуань: жэньгун чжинэн сян цзунли «баодао» : [Ли Кэцян встречается с делегатами ВСНП в провинции Аньхуэй: искусственный интеллект «докладывает» премьеру]. URL: http://www.china.com.cn/lianghui/news/2017-03/10/content_40440494.htm (дата обращения: 31.07.2018).

Лу Исинь. Принципы создания корпусов китайского языка // Известия Российского государственного педагогического университета имени А.И. Герцена, СПб. 2016. № 181. С. 22–29.

Меньшиков Л.Н. Из истории китайской книги. СПб: Нестор-История, СПБ ИИ РАН, 2005.

Сбоев А.Н. Анализ лексики китайского интернета с точки зрения словообразования // Гуманитарные исследования в Восточной Сибири и на Дальнем Востоке. 2015. № 3 (33). С. 70–81.

Тайвань Миньнаньюй коуюй юйляоку : [Устный корпус тайваньских диалектов группы Миньнань]. URL: http://lngproc.ccu.edu.tw/SouthernMinCorpus/ (дата обращения: 18.02.2020).

Хань Линьтао, Ян Эрхун. 2016 нянь дэ цзици фаньи : [Машинный перевод в 2016 г.] // Чжунго юйянь шэнхо чжуанкуан баогао 2017 : [Доклад о языковой ситуации в Китае в 2017 г.]. Пекин, 2017. С. 119–123.

Ханьюй паньдянь 2019: юн игэ цзы игэ цы мяошу 2019 дэ Чжунго юй шицзе : [Инвентаризация 2019: одним иероглифом, одним словом описываем Китай и мир в 2019 г.]. URL: http://culture.people.com.cn/GB/22226/430827/index.html (дата обращения: 15.02.2020).

«Ханьюй паньдянь» ходун саомяо (2006–2010) : [Обзор «Инвентаризации китайского языка» в 2006–2010 гг.] // Чжунго юйянь шэнхо чжуанкуан баогао (2011) : [Доклад о языковой ситуации в Китае в 2011 г.]. Пекин: Шанъу иньшугуань, 2011. С. 180–191.

Хуан Цзюйжэнь. Тайвань юйляоку юй юйянь цзыюань цзяньшэ : [Создание языковых корпусов и языковых ресурсов на Тайване] // Чжунго юйянь шэнхо чжуанкуан баогао (2016) : [Доклад о языковой ситуации в Китае в 2016 г.]. Пекин: Шанъу иньшугуань, 2016. С. 259 –168.

Цзисуаньцзи цзыку цзысин чжуанкуан : [Ситуация с компьютерными шрифтами и библиотеками шрифтов] // Чжунго юйянь шэнхо чжуанкуан баогао (2006) : [Доклад о языковой ситуации в Китае в 2006 г.]. Пекин: Шанъу иньшугуань, 2007. Т. 1. С. 121–168.

Цюаньцю чжунвэнь сюэси пинтай : [Всемирная платформа для изучения китайского языка]. URL: www.chinese-learning.cn (дата обращения: 10.01.2020).

Чжань Вэйдун, Го Жуй, Чан Баобао, Чэнь Ижун, Чэнь Лун. Бэйцзин дасюэ юйляоку дэ яньчжи : [Разработка языкового корпуса Пекинского университета] // Юйляоку юйяньсюэ. 2019. № 6 (1). С. 71–86.

Чжунго юйянь вэньцзы ван : [Сайт «Язык и письменность в Китае»]. URL: www.china-language.edu.cn/ (дата обращения: 14.02.2020).

Чжунго юйянь цзыюань баоху гунчэн : [Программа по охране языковых ресурсов Китая] // Чжунго юйянь вэньцзы шие фачжань баогао (2018) : [Доклад о развитии языка и письменности в Китае в 2018 г.]. Пекин: Шанъу иньшугуань, 2018. С. 52–57.

Чжунго юйянь цзыюань баоху гунчэн: [Научная программа по охране языковых ресурсов Китая] // Чжунго юйянь вэньцзы шие фачжань баогао (2019) : [Доклад о развитии языка и письменности в Китае в 2019 г.]. Пекин: Шанъу иньшугуань, 2019. С. 48–52.

Чжунъян яньцзююань сяньдай ханьюй пинхэ юйляоку : [Сбалансированный корпус современного китайского языка Academia Sinica (Китайской академии на Тайване)]. URL: http://lingcorpus.iis.sinica.edu.tw/modern/ (дата обращения: 12.01.2020).

Чжэнчжи дасюэ гоюй коуюй юйляоку : [Устный корпус «государственного языка» Университета управления]. URL: http://spokentaiwanmandarin.nccu.edu.tw/corpus-data.html (дата обращения: 18.02.2020).

Шаошу миньцзу юйянь цзыюань баоху юй цзяньшэ : [Охрана и разработка ресурсов на языках малых народов] // Чжунго юйянь вэньцзы шие фачжань баогао (2018) : [Доклад о развитии языка и письменности в Китае]. Пекин: Шанъу иньшугуань, 2018. С. 58–62.

Шиюн банчжу – BCC сяньдай ханьюй юйляоку : [Инструкция по использованию – Корпус современного китайского языка Пекинского университета языков]. URL: http://bcc.blcu.edu.cn/help#intro (дата обращения: 17.02.2020).

Юйляоку цзай сянь : [Языковой корпус онлайн]. URL: http://corpus.zhonghuayuwen.org/ (дата обращения: 14.02.2020).

Юйянь вэньцзы гунцзо ванчжань минлу : [Список сайтов, посвящённых работе в области языка и письменности] // Чжунго юйянь шэнхо чжуанкуан баогао (2005) : [Доклад о языковой ситуации в Китае в 2005 г.]. Пекин: Шанъу иньшугуань, 2006. Т. 1. С. 435–436.

Юйянь вэньцзы синьси гуаньли гунцзо чжуанкуан : [О работе в области информатизации языка и письменности] // Чжунго юйянь шэнхо чжуанкуан баогао (2005) : [Доклад о языковой ситуации в Китае в 2005 г.]. Пекин: Шанъу иньшугуань, 2006. Т. 1. С. 121–139.

Ян Эрхун, Хоу Минь. Юйянь синьси чули яньцзю : [Исследования в области компьютерной обработки языковых данных] // Дандай чжунго юйянь яньцзю (1949–2015) : [Исследования в области современного китайского языкознания (1949–2015 гг.)]. Пекин: Чжунго шэхуэй кэсюэ чубаньшэ, 2016. С. 475–516.

Cantonese Corpora. URL: https://cantoneselanguageassociation.byu.edu/links/ (дата обращения: 18.02.2020).

Chui Kawai and Lai Huei-ling. The NCCU Corpus of Spoken Chinese: Mandarin, Hakka, and Southern Min // Taiwan Journal of Linguistics. 2008. № 6(2). P. 119–144.

Lo W.R., Lee Tan, Ching P.C. Development of Cantonese spoken language corpora for speech applications. International Symposium on Chinese spoken language processing, Singapore, 1998. URL: https://materi.smkyp17pare.sch.id/archive_open/archive_papers/iscslp1998/DTE5.pdf (дата обращения: 18.02.2020).

The Corpus of Mid-20th Century Hong Kong Cantonese (Phase Two). URL: http://hkcc.eduhk.hk/v2 (дата обращения: 18.02.2020).

Xu Jiajin. Corpus-based Chinese studies. A historical review from the 1920s to the present Chinese // Language and Discourse. 2015. No. 6(2). P. 218–244.

REFERENCES

Baozhi, guangbo dianshi, wangluo (xinwen) yong zi, yong ci diaocha [Survey of words and characters used in newspapers, on radio and TV, as well as in Internet news], *Zhongguo yuyan shenghuo zhuangkuang baogao*: 2006 [Language situation in China: 2006], Beijing: Shangwu yinshuguan, 2007, vol. 2: 1–36. (In Chinese).

Beijing daxue CCL yuliaoku [Corpus of the Peking University Center for Chinese Linguistics]. URL: http://ccl.pku.edu.cn:8080/ccl_corpus/ (accessed: 10 February 2019). (In Chinese).

Cantonese Corpora. URL: https://cantoneselanguageassociation.byu.edu/links/ (accessed: 18 February 2020).

Chui Kawai and Lai Huei-ling. (2008). The NCCU Corpus of Spoken Chinese: Mandarin, Hakka, and Southern Min, *Taiwan Journal of Linguistics*, 2008, 6(2): 119–144.

Dai Hongliang, Yang Shujun. (2019). Taiwan yuwen shenghuo zhuangkuang (2018) [Language situation in Taiwan in 2018], *Zhongguo yuyan shenghuo zhuangkuang baogao*: 2019 [Language situation in China: 2019], Beijing: Shangwu yinshuguan, 2019: 271–279. (In Chinese).

Gutin I.Yu. (2018). Yazykovaya situatsiya v Spetsial'nom administrativnom rayone Gonkong KNR i politika vlastey v sfere yazyka [Language situation in the Hong Kong Special Administrative Region of the PRC and the official language policy], *Mezhdunarodniy nauchno-issledovatel'skiy zhurnal* [*International Research Journal*], 2018, 2(68): 79–83. (In Russian).

Han Lintao, Yang Erhong. (2017). 2016 nian de jiqi fanyi [Machine translation in 2016], *Zhongguo yuyan shenghuo zhuangkuang baogao*: 2017 [Language situation in China: 2017], Beijing: Shangwu yinshuguan, 2017: 119–123. (In Chinese).

Hanyu pandian 2019: yong yige zi yige ci miaoshu 2019 de Zhongguo yu shujie [Chinese language inventory 2019: one character and one word to describe China and the world in 2019]. URL: http://culture.people.com.cn/GB/22226/430827/index.html (accessed: 15 February 2020). (In Chinese).

'Hanyu pandian' huodong saomiao (2006–2010) ["Chinese Language Inventory" activity in 2006–2010], *Zhongguo yuyan shenghuo zhuangkuang baogao*: 2011 [Language situation in China: 2011], Beijing: Shangwu yinshuguan, 2011: 180–191. (In Chinese).

Huang Juren. (2016). Taiwan yuliaoku yu yuyan ziyuan jianshe [Development of language corpora and language resources in Taiwan], *Zhongguo yuyan shenghuo zhuangkuang baogao*: 2016 [Language situation in China: 2016], Beijing: Shangwu yinshuguan, 2016: 259 –168. (In Chinese).

Jisuanji ziku zixing zhuangkuang [Situation with computer fonts and font libraries], *Zhongguo yuyan shenghuo zhuangkuang baogao*: 2006 [Language situation in China: 2006], Beijing: Shangwu yinshuguan, 2007, vol. 1: 121–168. (In Chinese).

Kolpachkova E.N. (2015). Korpusy kitayskogo yazyka: sovremennoe sostoyanie i osnovnye problemy [Chinese language corpora: an overview and major problems], *Proceedings of the international conference "Corpus linguistics-2015"*, St. Petersburg: Saint Petersburg State University, 2015: 278–286. (In Russian).

Li Keqiang lai Anhuituan: rengong zhineng xiang zongli 'baodao' [Premier Li Keqiang meets with the NPC members in Anhui: AI "reports" to the Premier]. URL: http://www.china.com.cn/lianghui/news/2017-03/10/content_40440494.htm (accessed: 31 July 2018). (In Chinese).

Lo W.R., Lee Tan, Ching P.C. (1998). Development of Cantonese spoken language corpora for speech applications, *International Symposium on Chinese spoken language processing*, Singapore, 1998. URL: https://materi.smkyp17pare.sch.id/archive_open/archive_papers/iscslp1998/DTE5.pdf (accessed: 18 February 2020).

Lu Yixin. (2016). Printsipy sozdaniya korpusov kitayskogo yazyka [The guidelines for the Chinese language corpora], *Izvestiya Rossiyskogo gosudarstvennogo pedagogicheskogo universiteta imeni A.I. Gertsena* [*Proceedings of the Herzen State Pedagogical University*], St. Petersburg, 2016, 181: 22–29. (In Russian).

Men'shikov L.N. Iz istorii kitayskoy knigi [On the history of Chinese books], St. Petersburg: Nestor-Istoriya; Saint Petersburg Institute of History, RAS, 2005. (In Russian).

Quanqiu Zhongwen xuexi pingtai [Global platform for studying Chinese]. URL: www.chinese-learning.cn (accessed: 10 January 2020). (In Chinese).

Sboev A.N. (2015). Analiz leksiki kitayskogo Interneta s tochki zreniya slovoobrazovaniya [Derivational patterns of the Chinese Internet lexicon], *Humanities Research in Eastern Siberia and the Russian Far East*, 2015, 3(33): 70–81. (In Russian).

Shaoshu minzu yuyan ziyuan baohu yu jianshe [Protection and development of language resources of ethnic minorities], *Zhongguo yuyan wenzi shiye fazhan baogao* (2018) [Report on the development of language and script in China (2018)], Beijing: Shangwu yinshuguan, 2018: 58–62. (In Chinese).

Shiyong bangzhu – BCC Xiandai Hanyu yuliaoku [User guide – BCC Modern Chinese language corpus]. URL: http://bcc.blcu.edu.cn/help#intro (accessed: 17 February 2020). (In Chinese).

Taiwan Minnanyu kouyu yuliaoku [Corpus of spoken Taiwan Southern Min]. URL: http://lngproc.ccu.edu.tw/SouthernMinCorpus/ (accessed: 18 February 2020). (In Chinese).

The Corpus of Mid-20th Century Hong Kong Cantonese (Phase Two). URL: http://hkcc.eduhk.hk/v2 (accessed: 18 February 2020).

Wangluo yuyan zhuangkuang [Language usage on the Internet], *Zhongguo yuyan shenghuo zhuangkuang baogao*: 2005 [Language situation in China: 2005], Beijing: Shangwu yinshuguan, 2006, vol. 1: 211–226. (In Chinese).

Xu Jiajin. (2015). Corpus-based Chinese studies. A historical review from the 1920s to the present Chinese, *Language and Discourse*, 2015, 6 (2): 218–244.

Yang Erhong and Hou Min. (2016). Yuyan xinxi chuli yanjiu [Studies on the computer processing of language data], *Dangdai Zhonguo yuyan yanjiu (1949–2015)* [Studies on contemporary Chinese linguistics (1949–2015)], Beijing: Zhongguo shehui kexue chubanshe, 2016: 475–516. (In Chinese).

Yuliaoku zai xian [Language corpus online]. URL: http://corpus.zhonghuayuwen.org/ (accessed: 14 February 2020). (In Chinese).

Yuyan wenzi gongzuo wangzhan minglu [List of websites on language and script planning], *Zhongguo yuyan shenghuo zhuangkuang baogao*: 2005 [Language situation in China: 2005], Beijing: Shangwu yinshuguan, 2006, vol. 1: 435–436. (In Chinese).

Yuyan wenzi xinxi guanli gongzuo zhuangkuang [On the informatization of language and script], *Zhongguo yuyan shenghuo zhuangkuang baogao*: 2005 [Language situation in China: 2005], Beijing: Shangwu yinshuguan, 2006, vol. 1: 21–139. (In Chinese).

Zavyalova, O.I. (2016) Yazyki Kitaya: noveyshie issledovaniya i otkrytiya [Languages of China: latest surveys and discoveries], *Problemy Dal'nego Vostoka*, 2016, 5: 139–143. (In Russian).

Zavyalova, O.I. (2017). Yazyki Kitaya v informatsionnom prostranstve [Languages of China in the IT age], *Problemy Dal'nego Vostoka*, 2017, 2: 148–152. (In Russian).

Zhan Weidong, Guo Rui, Chang Baobao, Chen Yirong, Chen Long. (2019). Beijing daxue yuliaoku de yanzhi [Building of the CCL corpus: Its design and implementation], *Corpus Linguistics*, 2019, 6(1): 71–86. (In Chinese).

Zhengzhi daxue guoyu kouyu yuliaoku [NCCU Corpus of Spoken Taiwan Mandarin]. URL: http://spokentaiwanmandarin.nccu.edu.tw/corpus-data.html (accessed: 18 February 2020).

Zhongguo yuyan wenzi wang ["Language and script in China" website]. URL: www.china-language.edu.cn/ (accessed: 14 February 2020). (In Chinese).

Zhongguo yuyan ziyuan baohu gongcheng [Project on the protection of language resources in China], *Zhongguo yuyan wenzi shiye fazhan baogao* (2018) [Report on the development of language and script in China (2018)], Beijing: Shangwu yinshuguan, 2018: 52–57. (In Chinese).

Zhongguo yuyan ziyuan baohu gongcheng [Project on the protection of language resources in China], *Zhongguo yuyan wenzi shiye fazhan baogao* (2019) [Report on the development of language and script in China (2019)], Beijing: Shangwu yinshuguan, 2019: 48–52. (In Chinese).

Zhongyang yanjiuyuan Xiandai Hanyu pinghe yuliaoku [Academia Sinica balanced corpus of Modern Chinese]. URL: http://lingcorpus.iis.sinica.edu.tw/modern/ (accessed: 12 January 2020). (In Chinese).

Поступила в редакцию 24.02.2020 Received 24 February 2020

***Для цитирования*:** Завьялова О.И. Языковая политика и языковые ресурсы в китайском интернете // Восточная Азия: факты и аналитика. 2020. № 1. С. 19–33. DOI: 10.24411/2686-7702-2020-10002
***For citation*:** Zavyalova O.I. (2020). Yazykovaya politika i yazykovyye resursy v kitayskom internete [Language policy and language resources on the Chinese Internet], *Vostochnaya Aziya: fakty i analitika* [*East Asia: Facts and Analytics*], 2020, 1: 19–33. (In Russian). DOI: 10.24411/2686-7702-2020-10002