DOI: 10.31862/2500-2953-2022-4-81-100

И.Л. Корецкая

Институт языкознания Российской академии наук,
125009 г. Москва, Российская Федерация

Корпусы Государственного института японского языка и лингвистики

Корпусные исследования давно завоевали популярность среди лингвистов. Полученные результаты достаточно точны, верифицируемы и широко применимы для разных целей. Однако существуют достойные внимания корпусы, остающиеся неизвестными потенциальному пользователю. Среди них, например, корпусы японского языка, разработанные Государственным институтом японского языка и лингвистики, и цель данной статьи – заполнить этот пробел. Основное внимание статьи уделено сравнению имеющихся на данный момент корпусов, возможностей поиска в них и оценке их репрезентативности; указаны выявленные недостатки с точки зрения сбалансированности материала, а также особенности поиска, исходящие из идей традиции японского языкознания, которые необходимо принимать во внимание в процессе проведения исследования.

Ключевые слова: японская корпусная лингвистика, японские корпусы текстов, Государственный институт японского языка и лингвистики

ДЛЯ ЦИТИРОВАНИЯ: Корецкая И.Л. Корпусы Государственного института японского языка и лингвистики // Рема. Rhema. 2022. № 4. С. 81–100. DOI: 10.31862/2500-2953-2022-4-81-100

DOI: 10.31862/2500-2953-2022-4-81-100

I. Koretskaya

Institute of Linguistics, Russian Academy of Sciences,
Moscow, 125009, Russian Federation

Corpora of the National Institute of Japanese Language and Linguistics

Corpus studies have long gained popularity among linguists. The obtained results are accurate, verifiable, and widely applicable. However, some corpora are noteworthy and yet remain unknown to a potential user. One of the examples is the Japanese language corpora which were developed by the National Institute of Japanese Language and Linguistics, and the purpose of this article is to fill this gap. The article focuses mainly on a comparison of the currently available corpora, their search options, and an assessment of the corpora's representativeness. The article shows identified shortcomings in the data balance and the features of the search based on the ideas of the Japanese linguistics tradition. They must be taken into account when conducting a study based on the data of these corpora.

Key words: Japanese corpus linguistics, Japanese text corpora, the National Institute of Japanese Language and Linguistics

FOR CITATION: Koretskaya I. Corpora of the National Institute of Japanese Language and Linguistics. *Rhema*. 2022. No. 4. Pp. 81–100. (In Rus.). DOI: 10.31862/2500-2953-2022-4-81-100

1. Введение

Государственный институт японского языка и лингвистики (国立国語研究所, официальное название на английском языке – National Institute of Japanese Language and Linguistics, NINJAL) по праву считают организацией, заложившей основы корпусной лингвистики в Японии. С самого момента своего создания он уделял много внимания сбору материала для анализа языка. Здесь появились первые картотеки слов с указанием контекста их употребления, шел сбор как письменных текстов, так и устной речи, включая диалектную. Собранный материал подвергался

статистическому анализу, в результате которого получали данные об актуальном состоянии языка. С развитием технологий старые материалы начали оцифровывать, появился и новый способ получения материала – из электронных версий текстов. Одновременно шла разработка требований к корпусам текстов. На данный момент Институт продолжает вести огромную работу по разработке и поддержке лингвистических корпусов, доступ к которым открыт всем желающим, а также по разработке автоматических систем для обработки японских текстов.

2. Немного истории

Институт был основан в декабре 1948 г. в рамках проводимой государством языковой политики послевоенного времени, когда назрела необходимость понять, как язык действительно используется в обществе, и на основе полученных данных подготовить и провести ряд языковых реформ. Цель создания Института состояла в «проведении научных исследований по родному языку и речевой практике народа и создании твердой базы при рационализации языка», в связи с чем главной задачей стал сбор данных, «могущих послужить базой правильной языковой политики», т.е. способствовать установлению и поддержанию языковых норм, «стандартного языка» [Фельдман, 1956, с. 153].

В основу исследовательской работы Института легли принципы так называемой «теории языкового существования», появившейся в японском языкознании в качестве ответа структурализму и другим направлениям, рассматривающим язык вне его связи с человеком. Аналогов этой теории в европейской лингвистике нет. Если в последней в центре изучения находится язык, а социальные факторы лишь учитываются как значимые при проведении некоторых исследований, то в японской лингвистике в рамках вышеупомянутой теории центральным понятием выступает жизнь человека, а язык – комплементарное понятие, и языковое существование является лишь частью (пусть и значительной частью) жизни человека. Это отмечает, например, один из крупнейших исследователей теории языкового существования С.В. Неверов, определяя «языковое существование» как «бытие человека, проявляющееся в его действиях, связанных с речевым общением, которые и составляют полноценность человеческой жизни и противопоставляют его как существо социальное животному миру» [Неверов, 2005, с. 8]. Сторонники теории утверждали, что необходимо изучить роль языка в жизни человека, для чего требовался сбор количественных данных (различных текстов, опросов людей и т.д.), а это, по сути, является начальным этапом создания корпуса текстов.

Положения теории языкового существования согласовались с вышеупомянутой политикой государства. Поэтому неудивительно, что организатором и первым директором Института стал один из основоположников школы языкового существования Нисио Минору[1]. Сотрудники Института продолжили развитие теории, однако в целом уже не с точки зрения теоретических основ, а скорее практических нужд. Институт много средств, сил и времени вложил в сбор эмпирических данных (как письменных, так и устных текстов), на основе которых делались выводы о языковом существовании японцев. Здесь мы видим подход к анализу данных, схожий с корпусным, хотя о корпусной лингвистике говорить рано. Она оформится гораздо позже, и, как видно из введения, Институт сыграет в этом большую роль.

В 2009 г. Институт вошел в состав Государственных институтов гуманитарных наук. По словам сотрудников Института, в это время произошел «революционный» сдвиг в приоритетах организации. Если изначально его работа была подчинена нуждам государственной языковой политики, то теперь посвящена главным образом проведению лингвистических исследований японского языка и внесению своего вклада в общество на основе полученных в ходе исследований результатов [An introduction, 2019, p. 1]. Целью Института является изучение как современного, так и классического японского языка, а также особенностей различных вариантов языков, на которых говорят в Японии, включая айнский язык, рюкюские языки[2], японские диалекты [NINJAL: Справочник, 2019/2020, p. 2]. Одной из приоритетных задач Института и в настоящее время является разработка и поддержка различных типов корпусов японского языка.

3. Корпусы Института

Работа по созданию компьютезированных корпусов ведется в Институте с конца 1990-х гг. На данный момент на сайте Центра по разработке языковых ресурсов Института[3] открыт доступ к 11 корпусам (в скобках указаны официальные названия корпусов на японском и английском языках и сокращения):

1) Сбалансированный корпус современного японского письменного языка (現代日本語書き言葉均衡コーパス, Balanced Corpus of Contemporary Written Japanese, BCCWJ);

[1] В Японии принято сначала называть фамилию, а затем имя человека. Мы сохраняем этот порядок для всех японских авторов в данной статье.

[2] Согласно другой точке зрения, они представляют собой диалекты.

[3] URL: https://clrd.ninjal.ac.jp/en/ (date of access: 12.03.2022).

2) Интернет-корпус NINJAL (国語研日本語ウェブコーパス, NINJAL Web Japanese Corpus, NWJC);

3) Корпус спонтанной японской речи (日本語話し言葉コーパス, Corpus of Spontaneous Japanese, CSJ);

4) Корпус повседневного общения (日常会話コーパス, Corpus of Everyday Japanese Conversation, CEJC);

5) Корпус устных текстов периода Сёва (昭和話し言葉コーパス, Showa Speech Corpus, SSC);

6) Разговорный корпус Университета Нагоя (名大会話コーパス, Nagoya University Conversation Corpus, NUCC);

7) Корпус бесед на рабочем месте Гэн-Нити-Кэн (現日研・職場談話コーパス, Gen-Nichi-Ken Corpus of Work Place Conversation, CWPC);

8) Диахронический корпус японского языка (日本語歴史コーパス, Corpus of Historical Japanese, CHJ);

9) Корпус японских диалектов (日本語諸方言コーパス, Corpus of Japanese Dialect, COJADS);

10) Международный учебный корпус японского языка как иностранного (多言語母語の日本語学習者横断コーパス, International Corpus of Japanese as a Second Language, I-JAS)[4];

11) Корпус современного японского языка (近代語のコーパス, Corpus of Modern Japanese, CMJ)[5];

12) База данных слов, классифицированных по семантическому принципу (分類語彙表－増補改訂版データベース, Word List by Semantic Principles, WLSP).

Первыми одиннадцатью корпусами можно пользоваться бесплатно онлайн посредством поисковой системы Chuunagon[6], WLSP – только платно. Предполагается использование корпусов в академических и исследовательских целях, но предусмотрен вариант использования большинства из них и в коммерческих целях (обсуждается индивидуально с Институтом).

4. Единицы текста в корпусах Института

В лингвистических корпусах языков, письменность которых предполагает использование пробела, единицей текста, как правило, выступает

[4] За этим названием на самом деле стоят два корпуса: Учебный корпус японского языка как иностранного (中国語・韓国語母語の日本語学習者縦断発話コーパス, Corpus of Japanese as a Second Language, C-JAS) и Международный учебный корпус японского языка как иностранного (多言語母語の日本語学習者横断コーパス, International Corpus of Japanese as a Second Language, I-JAS).

[5] Корпус представлен вместе с другими на сайте Центра по разработке языковых ресурсов Института, но является частью диахронического корпуса CHJ.

[6] URL: https://chunagon.ninjal.ac.jp/ (date of access: 12.03.2022).

словоупотребление, в данном случае являющееся синонимом орфографического слова, т.е. единицы между двумя пробелами или другими знаками препинания. Однако в японском языке, как и ряде других, пробелы на письме не используются, поэтому прежде всего встает вопрос о единицах, на которые будет сегментирован текст для его дальнейшей обработки и организации поиска в корпусе.

Институт пришел к следующему решению. Текст сначала делится на SUW (“short-unit words”, термин NINJAL для единиц, называемых в японской традиции “go”; это минимальные языковые единицы, которые в терминах европейского языкознания могут быть как знаменательными или служебными словами, так и морфемами), здесь важно отметить, что граница между двумя SUW проходит на стыке двух мор[7]. Затем некоторые SUW на основе определенных правил объединяются в LUW (“long-unit words”, термин NINJAL для единиц, соответствующих устойчивым сочетаниям минимальных единиц “go” (подробнее “go”, SUW и LUW см. [Корецкая, 2021, с. 277–278]).

5. Поиск в корпусе

5.1. Возможности поиска

На главной странице сайта (рис. 1) можно, во-первых, осуществить поиск сразу по всем корпусам, доступ к которым был одобрен, а во-вторых, искать по каждому из таких корпусов отдельно. Первый способ позволяет увидеть данные в сравнении (в том числе на гистограммах и круговых диаграммах), однако им стоит пользоваться очень аккуратно ввиду серьезных различий корпусов между собой. Второй позволяет искать в каждом отдельном корпусе, здесь не увидеть сравнения результатов с другими корпусами, зато возможности поиска и выдачи результатов значительно шире.

Остановимся подробнее на втором варианте. На странице корпуса есть возможность осуществлять поиск по нескольким категориям запросов (рис. 2):

1) по SUW (во всех корпусах);

[7] Мора – минимальная тактовая единица японского языка, представляет собой либо сочетание «согласный + гласный», включая сочетание «гортанная смычка + гласный», встречающееся в начале слога, либо гласный после гласного, либо согласный, на который заканчивается слог: например, *hatten* «развитие» состоит из двух слогов (*hat-ten*), но из четырех мор (*ha-t-te-n*). Японская слоговая азбука кана отражает членение слов на моры, а не на слоги (в указанном примере будет 4 знака каны). В японской лингвистической традиции границу между единицами текста проводят на стыке двух мор, тогда как в европейской традиции граница может проходить внутри моры: ср. японское членение *yomi-masu* и европейские варианты *yom-i-mas-u* и *yom-imas-u* (гоноратив настояще-будущего времени глагола «читать»).

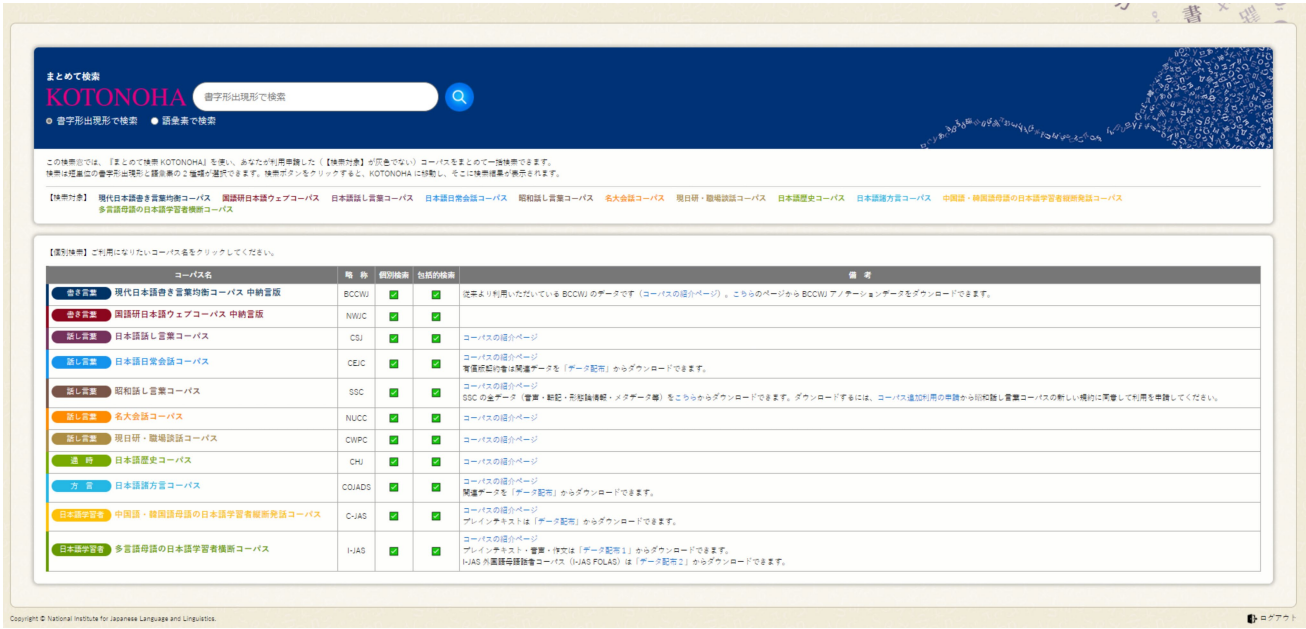

Рис. 1. Главная страница сайта (https://chunagon.ninjal.ac.jp/)

В верхней части страницы находится форма для поиска по всем корпусам одновременно, в нижней – ссылки на каждый из доступных пользователю корпусов

Fig. 1. The main page (https://chunagon.ninjal.ac.jp/)

The form for simultaneous search in all the corpora is at the top of the page and links to the available corpora are at the bottom

ISSN 2500-2953

Рис. 2. Поиск в корпусе (на примере BCCWJ):

1 – SUW; 2 – LUW; 3 – по строке символов; 4 – по месту “go” в корпусе; 5 – поиск в специальной форме; 6 – по отдельным условиям; 7 – по истории поиска

Fig. 2. Search in a corpus (using BCCWJ as an example):

1 – by SUW; 2 – by LUW; 3 – by a string of characters; 4 – by the position of the “go” unit in the corpus; 5 – by using the form; 6 – by specific conditions; 7 – by search history

2) по LUW (только в BCCWJ и ряде подкорпусов CHJ);

3) по строке символов (во всех корпусах): особенно полезен, если статус единицы непонятен, если в корпусе нет возможности искать по LUW, а также если интересует конкретное окружение “go”);

4) по месту “go” в корпусе (во всех, кроме C-JAS и I-JAS).

В каждой из этих категорий запросов есть вариант осуществления поиска в специальной форме, по истории поиска и по отдельным условиям (с использованием РБНФ (Расширенной формы Бэкуса–Наура); недоступно только в C-JAS и I-JAS)).

В данной статье мы расскажем о варианте поиска с помощью формы.

Поиск по SUW и LUW дает возможность искать интересующую нас единицу по различным параметрам. Так, для поиска важным является разграничение “goiso” (語彙素, формы “go”, принятой за начальную, единицы, которую можно сравнить с заголовочным словом в словаре, лексемой[8]) и конкретной формы “go” (語形, “gokei”, близок русскому термину «словоформа»). В первом случае в результатах будут указаны все встретившиеся в текстах формы единицы, во втором – только результаты с указанной в запросе формой.

Далее, можно искать по произношению или написанию “goiso” или формы “go”, что полезно в тех нередких случаях, когда у единицы несколько вариантов произношения или написания, включая широко распространенную в языке омонимию. Также во всех корпусах есть возможность поиска по классу “go” (ваго[9], канго[10], гайрайго[11] и др.), по части речи единицы (включает не только части речи, но и части слова (префикс и суффикс), а также пробел, различные символы и (в C-JAS и I-JAS) пометы «затруднительно дать анализ», «личная информация», «невербальное действие»), а для изменяемых единиц – по типу и форме спряжения. Последние три параметра имеют иерархическую структуру, и в них можно выбрать, насколько подробной должна быть информация о единице.

Наконец, в COJADS, C-JAS и I-JAS есть параметр «тег». В первом корпусе можно искать как по тегам, обозначающим исключительно диалектные особенности (например, региональные ономатопеи, опущенные служебные “go”, а также наречия, союзы, диминутивы, префиксы и гоноративы, аналогов которым нет в «стандартном языке»),

[8] В отличие от заголовочного слова или лексемы, фонетические варианты единицы в случае с удлинением и сокращением звуков (например, *amari* и *ammari* (‘чересчур, очень’) считаются Институтом разными “goiso”.

[9] Исконно японские слова.

[10] Слова, корни которых имеют китайское происхождение.

[11] Слова, заимствованные из западных языков.

так и по некоторым особенностям разговорной речи (словам-филлерам и паузам хезитации). Наборы тегов в C-JAS и I-JAS схожи; учитывая тип корпусов, разработчики добавили теги об ошибках разного рода (фонетических, грамматических и лексических), пометы пауз, удлиненных гласных, междометий-филлеров, неясного произношения единицы или сложности в определении единицы, просьбы объяснить чтение единицы, а также теги отнесения “go” к прииенным или именам собственным.

Отдельно стоит упомянуть о возможности создания своего подкорпуса для поиска по выбранным в форме параметрам. Такая возможность есть во всех корпусах, кроме NUCC, NWJC и CWPC, но ввиду разного наполнения для каждого корпуса список вариантов свой. Например, для диалектного корпуса COJADS важными параметрами, безусловно, являются диалект и префектура, в которой была записана речь, а в C-JAS можно изучить подкорпус речи каждого из шести информантов или нескольких из них, при этом выбрав период записи речи.

Корпусы позволяют пользователю выбрать формат и содержание результатов поиска. Во всех корпусах можно выбрать тип графического разделения единиц в тексте и объем показываемого контекста. Что касается содержания результатов поиска, обязательными для всех корпусов являются параметры, дающие информацию о единице и информацию о корпусе, остальные параметры зависят от типа корпуса (например, информация о тексте и его авторе есть лишь в письменных корпусах BCCWJ и CHJ, а I-JAS дает информацию об изучающем японский язык информанте).

5.2. Примеры поиска

Показанный нами широкий выбор параметров позволяет максимально уточнить запрос, без чего зачастую не обойтись ввиду распространенной в языке омонимии, а также ряда особенностей японской письменности.

5.2.1. Снятие грамматической и фонетической омонимии

Предположим, перед нами стоит задача проанализировать использование существительного 帰り (kaeri, ‘возвращение домой, путь домой, обратно’) в текстах блогов. Оно омонимично одной из форм глагола 帰る (kaeru, ‘возвращаться домой, обратно’). Кроме того, у него много фонетических омонимов, а также есть два варианта написания, в которых различается сочетание иероглифа с окуриганой[12]: 帰る и 帰える, где 帰 – иероглиф, а る (ru) и える (eru) – окуригана.

[12] Окуригана – слоговая азбука, которой записывают часть единицы, идущую вслед за иероглифом.

В корпусе BCCWJ создадим подкорпус текстов блогов (рис. 3), в поисковой форме (рис. 4) выберем поиск по «goiso» и укажем нужную нам единицу 帰り (позволит учесть различные варианты написания единицы, одновременно отсекая ее фонетические омонимы), а в дополнительном условии выберем часть речи «существительное» (избавляемся от грамматической омонимии).

Рис. 3. Создание подкорпуса в корпусе BCCWJ

Галочками помечен подкорпус блогов. Верхняя таблица: левая колонка содержит названия подкорпусов, в средней можно выбрать тексты из ядра корпуса (из текстов, чья разметка была проверена вручную), в правой – из не-ядра (тексты, не проверенные вручную). Ниже можно выбрать жанр и тематику текста, а также год публикации

Fig. 3. Creating a subcorpus in the BCCWJ corpus

The ticks mark the blog subcorpus. The table at the top allows selecting texts from the core corpus with manually checked tagging (the middle column) and from the non-core corpus without one (the right column); the names of subcorpora are given in the left column. The genre and subject of the texts, as well as the year of their publication, can be chosen in the form below

Рис. 4. Заполнение формы поиска в корпусе BCCWJ:

1 – поиск по “goiso”, в поле поиска пишем 帰り; 2 – дополнительное условие (здесь указана часть речи единицы – «существительное»); 3 – выбранные подкорпусы (блоги всех жанров); 4 – графические разделители в контекстах; 5 – длина контекстов; 6 – тип текстов (текст полностью, отрывок текста заданной длины или оба варианта (здесь выбраны оба типа)); 7 – учитывать границы предложений (можно выбрать «не учитывать»)

Fig. 4. The search form in the BCCWJ corpus:

1 – search by “goiso” 帰り; 2 – noun as an additional parameter; 3 – blogs of all genres as selected subcorpora; 4 – graphic marks in the sentences; 5 – length of the contexts; 6 – a type of the texts (a full text, an extract of a given length, or both types as selected here); 7 – whether to take into account the boundaries of the sentences

В результате из подкорпуса общим объемом 12 984 635 “go” (10 194 143 “go” без учета знаков препинания и пробелов) нашлось 1222 контекста, из которых корпус показывает 500 случайно отобранных.

5.2.2. Снятие орфографической омонимии

Допустим, мы хотим изучить все возможные употребления единицы自重 (jichoo, ‘самоуважение’) в интернет-источниках, однако у него есть орфографический омоним (в иероглифическом написании слова) 自重 (jijuu, ‘собственный вес’). Воспользуемся корпусом NWJC (рис. 5).

Рис. 5. Корпус NWJC. Заполнение формы:

1 – главный параметр, указываем “goiso” 自重; 2 – дополнительный параметр, выбираем чтение “goiso” и в поле катаканой записываем jichoo; 3 – примерно те же параметры, что и в BCCWJ, отсутствует только выбор типа текста. В этом корпусе на данный момент нет возможности задать подкорпус

Fig. 5. The NWJC Corpus. The form:

1 – “goiso” 自重 as the main parameter; 2 – the reading of “goiso” (jichoo) written in katakana as an additional parameter; 3 – mostly the same parameters as in the BCCWJ, the only difference is that there is no option of choosing the text type. At the moment, building a subcorpus is not available

В форме поиска выберем, например, “goiso” (единица представляет собой неизменяемое существительное, поэтому можно выбрать другой параметр поиска), в дополнительном параметре укажем чтение “goiso” (обязательно слоговой азбукой катакана) (см. рис. 5).

В результате поиска мы получили 578 контекстов, из которых корпус показывает 500, готовых к анализу. Общим объем корпуса 106 709 834 “go” (86 277 772 “go” без учета знаков препинания и пробелов).

5.2.3. Поиск определенного варианта единицы

Теперь поставим иную задачу – анализ употребленияв повседневной речи единицы やっぱり (yappari, ‘все-таки, всё еще, также, тоже’), представляющей собой разговорный вариант единицы 矢張り (yahari), у которой также несколько вариантов написания. Для примера мы воспользуемся корпусом CEJC (рис. 6).

Для выполнения задачи в качестве главного параметра мы указали “goiso”, дополнительного – его произношение. Корпус нашел 2092 контекста, из которых, как обычно, показал 500. Объем корпуса 2 421 162 “go” (2419171 “go” без учета знаков препинания и пробелов).

5.2.4. Поиск единицы в заданном месте в предложении

Корпусы позволяют искать единицу по номеру ее позиции в предложении. Вернемся к примеру в разделе 5.2.1 и добавим одно условие – единица должна быть третьей с начала предложения. Для этого над главным параметром выбираем расположение единицы от начала предложения и указывает порядковый номер ее места – третий “go” (рис. 7). В результате получили 134 контекста.

6. Репрезентативность и сбалансированность корпусов

Репрезентативность и сбалансированность являются одними из самых важных и одновременно трудновыполнимых требований к корпусу. Репрезентативность обычно подразумевает верное отражение корпусом всех важных для исследования особенностей языка, а сбалансированность – верное соотношение разных типов текстов между собой в корпусе (см., например, [Копотев, 2014]). Соблюдение обоих требований позволяет создать корпус, который представляет собой уменьшенную модель языка (или подъязыка), а значит, при работе с ним можно получить достоверные данные.

Рис. 6. Заполнение формы поиска в корпусе CEJC:

1 – в качестве основного параметра поиска указываем “goiso” 矢張り; 2 – в качестве дополнительного параметра выбираем произношение “goiso” и катаканой вписываем интересующий вариант (yappari); 3 – можно создать подкорпус (выбраны все тексты корпуса); 4 – графические пометы, объем контекстов, учет границы предложений. Нет возможности выбрать тип текстов

Fig. 6. The search form in the CEJC corpus:

1 – “goiso” 矢張り as the main search parameter; 2 – the pronunciation of “goiso” (yappari) written in katakana as an additional parameter; 3 – creating a subcorpus (here, all texts of the corpus are selected); 4 – choosing graphic marks, length of contexts, and whether to take into account the boundaries of the sentences. There is no option to select the text type

Рис. 7. Поиск единицы по номеру ее позиции в предложении в корпусе CEJC:

1 – точка отсчета (начало или конец предложения); 2 – номер позиции ключевой единицы; 3 – выбираем, должна ли единица быть строго на этой позиции (как на рисунке) или в интервале от точки отсчета до указанной позиции

Fig. 7. Searching for a unit by its position number in the sentence in the CEJC corpus:

1 – starting point (from the beginning or end of the sentence); 2 – position number of the key unit; 3 – choose whether the unit should be at the exact position given (as in the figure) or in the range from the starting point to the specified position

Иногда эти понятия совмещаются и используются как синонимы. Например, разработчики Национального корпуса русского языка указывают, что корпус «характеризуется представительностью, или сбалансированным составом текстов», и далее следует пояснение: «создатели стремятся включить в Корпус все типы письменных текстов, представленные в русском языке <…>, и все эти тексты входят в корпус пропорционально их доле в языке соответствующего периода»[13].

[13] URL: https://studiorum.ruscorpora.ru/manual/basic/ (date of access: 12.03.2002).

В данной работе мы количественно оцениваем характеристики материала корпуса, которые могут повлиять на результаты, полученные в ходе работы с ним, а именно соотношение объема текстов разных типов, т.е. говорим в первую очередь о сбалансированности, однако она влияет на репрезентативность.

В идеальном случае соотношение типов текстов в корпусе равно их соотношению в языке, тогда корпус представляет собой «пропорциональное сужение проблемной области» [Баранов, 2017], т.е. языка. Для этого требовалось бы заранее знать искомое соотношение, что до определенной степени осуществимо при условии проведения ряда социологических исследований. Если говорить о корпусах NINJAL, примером такого корпуса является BCCWJ: определение генеральной совокупности и формирование выборки проводились максимально скрупулезно на основе предварительных исследований. Он является самым сбалансированным корпусом NINJAL, впрочем, временные рамки генеральной совокупности в разных подкорпусах разные, что объясняется разной доступностью данных [BCCWJ: Руководство пользователя, 2019, с. 7]. Подробнее о корпусе на английском см. [Maekawa, 2014].

Информации о предварительных исследованиях при разработке других корпусов найти не удалось. В некоторых из них разработчики старались обеспечить репрезентативность за счет увеличения объема текстов того или иного типа, включения текстов других типов и/или соблюдения баланса между информантами. Вот некоторые особенности.

В корпусе CHJ[14] представлены тексты различных эпох, но стоит отметить разный объем подкорпусов по периодам, что, безусловно, придется учитывать при проведении исследования языкового явления в его диахроническом аспекте.

В случае с корпусами устных текстов наиболее репрезентативным (и сбалансированным), по нашему мнению, является CEJC[15]. При сборе материала разработчики контролировали баланс говорящих по полу и возрасту, но в корпусе мало данных по говорящим, не достигшим возраста 20 лет, а также старше 70 лет.

Одним из самых объемных корпусов является CSJ, но сбалансированным его назвать сложно. 95% объема корпуса составляют монологи, и лишь 5% приходятся на диалоги. Разработчики корпуса старались сделать корпус более репрезентативным, включив различные виды речи, однако 41% текстов составляют публичные выступления с научными докладами, 50% – подготовленные рассказы информантов

[14] URL: https://clrd.ninjal.ac.jp/chj/chj-wc.html (date of access: 12.03.2002).

[15] URL: https://www2.ninjal.ac.jp/conversation/report/report06.pdf (date of access: 12.03.2002).

на повседневные темы в относительно непринужденной обстановке, на остальные 9% приходятся интервью с информантами, чтение вслух и др. типы выступлений. На русском языке информацию о корпусе см. [Костыркин, 2009], на английском – [Maekawa, 2003].

В COJADS[16] представлена речь информантов из различных префектур (сбор материала шел с 1977 г. по 1985 г.), но видна большая разница в объеме материала по ним. Продолжительность записей колеблется от 5 минут (0,1% всех записей) в префектуре Тоттори до 572 минут (11,8%) в префектуре Гумма. В разговорах в общем случае чаще участвовали мужчины, чем женщины, хотя в некоторых префектурах (Аомори, Киото, Тояма и др.) статистика обратная. Однако корпус продолжает пополняться данными, и возможны изменения в приведенной оценке в будущем.

В корпусе SSC[17] наблюдается крайне неравномерное соотношение по полу, возрасту, роду деятельности и месту рождения информантов: из 50 информантов монологов только одна женщина, запись ее речи составляет лишь 1,5% от общей длительности монологов; значительно преобладает объем материала от людей в возрасте 45–49 лет (23% от общей продолжительности записей), и вовсе нет речи людей до 30 лет. Также объем диалогов сильно (почти в полтора раза) превышает объем монологов.

В NUCC[18] 85% информантов – женщины, около 50% информантов – в возрасте от 20 до 29 лет; много разговоров между людьми, занимающимися лингвистикой и обучением японскому языку, поэтому часто встречается метаязык.

Таким образом, при проведении исследований на материале корпусов NINJAL стоит помнить о следующих их особенностях: объем разных подкорпусов может отличаться; в корпусах устных текстов часто наблюдается неравномерное соотношение по характеристикам информанта (полу, возрасту, роду деятельности, месту рождения) и по виду текстов (диалог/монолог).

7. Заключение

Государственный институт японского языка и лингвистики, цель основания которого заключалась в изучении языкового существования человека, заложил основы японской корпусной лингвистики и продолжает развивать ее и в настоящий момент, работая над разработкой

[16] URL: https://cojads-data.ninjal.ac.jp/ (date of access: 12.03.2002).

[17] URL: https://www2.ninjal.ac.jp/conversation/showaCorpus/ (date of access: 12.03.2002).

[18] URL: http://telldev.cla.purdue.edu/chakoshi/meidai-chuui.html (date of access: 12.03.2002).

различных программ автоматической обработки текстов и большого числа разнообразных корпусов.

Материал для корпусов собирался с самого основания Института, когда методы сбора были еще только в процессе разработки. В итоге материал (безусловно, ценный) все же не всегда сбалансирован. Материал, полученный позже, как правило, сбалансирован лучше, но есть недостатки. Для получения надежных данных в процессе проведения исследования без учета особенностей корпуса не обойтись.

Также при работе с корпусами Института стоит учитывать, что решения Института основаны на разработанных в рамках японской традиции языкознания представлениях о единицах языка, их структуре и классах, представлениях, несколько отличающихся от принятых в отечественной и западной японистике.

Отчасти соответствующей идеям японского языкознания и отчасти отличающейся от них является принятая Институтом двухступенчатая сегментация текста: SUW соответствует “go”, но LUW не соответствует принятым в японской традиции единицам, однако решает проблему интерпретации составных единиц. Такое решение отлично подходит для корпусов японского языка – оно не только опирается на традицию, но и удобно с практической точки зрения (в первую очередь для автоматизации работы с текстами).

Итак, корпусы японского языка, разработанные Государственным институтом японского языка и лингвистики, дают бесплатный доступ к огромному материалу самой разной тематики. Имея в виду указанные выше особенности корпусов, можно использовать их в исследованиях и получать надежные данные.

Библиографический список / References

Баранов, 2017 – Баранов А.Н. Введение в прикладную лингвистику. Изд. 5-е. М., 2017. [Baranov A.N. Vvedenie v prikladnuyu lingvistiku [Introduction to applied linguistics]. Moscow, 2017.]

Копотев, 2014 – Копотев М. Введение в корпусную лингвистику. Прага, 2014. [Kopotev M. Vvedenie v korpusnuyu lingvistiku [Introduction to corpus linguistics]. Praga, 2014.]

Корецкая, 2021 – Корецкая И.Л. Проблемы японской корпусной лингвистики // Труды международной конференции «Корпусная лингвистика – 2021» (1–3 июля 2021 г., Санкт-Петербург) / В.П. Захаров (отв. ред.). СПб., 2021. С. 272–280. [Koretskaya I.L. Challenges in the Japanese corpus linguistics. *Proceedings of the International Conference* «*Corpus Linguistics–2021*» (*July 1–3, 2021, St. Petersburg*). V.P. Zakharov (ed.). St. Petersburg, 2021. Pp. 272–280. (In Rus.)]

Костыркин, 2009 – Костыркин А.В. Корпус японской разговорной речи // Национальный корпус русского языка: новые результаты и перспективы. СПб., 2009. С. 474–500. [Kostyrkin A.V. Corpus of Spontaneous Japanese. *Natsionalnyi korpus russkogo yazyka: novye rezultaty i perspektivy*. St. Petersburg, 2009. Pp. 474–500. (In Rus.)]

Неверов, 2005 – Неверов С.В. Общественно-языковая практика современной Японии. М., 2005. [Neverov S.V. Obshchestvenno-yazykovaya praktika sovremennoi Yaponii [Sociolinguistic practice in contemporary Japan]. Moscow, 2005.]

Фельдман, 1956 – Фельдман Н.И. О работе Государственного исследовательского института родного языка в Токио // Вопросы языкознания. 1956. № 3. С. 152–156. [Feldman N.I. On the Tokyo National Research Institute for Japanese Language. *Voprosy yazykoznaniya*. 1956. No. 3. Pp. 152–156. (In Rus.)]

An introduction, 2019 – An introduction to The National Institute for Japanese Language and Linguistics: A sketch of its achievements. 6th ed. M. Yamazaki, T. Takada, Yo. Matsumoto (eds.). Tokyo, 2019.

BCCWJ: Руководство пользователя, 2019 – 現代日本語書き言葉均衡コーパス。利用の手引. Tokyo, 2019. [Gendai nihongo kakikotoba kinkoo koopasu. Riyoo no tebiki [The Balanced Corpus of Modern Written Japanese. User Guide]. Tokyo, 2019.]

Maekawa, 2003 – Maekawa K. Corpus of Spontaneous Japanese: Its design and evaluation. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition. 2003.

Maekawa et al., 2014 – Maekawa K., Yamazaki M., Ogiso T. et al. Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation*. 2014. Vol. 48. Pp. 345–371.

NINJAL: Справочник, 2019/2020 – 国立国語研究所: 要覧 2019/2020. Tokyo, 2019. [Kokuritsu kokugo kenkyuusho: Yooran [National Institute for Japanese Language and Linguistics: Guide]. Tokyo, 2019.]

Статья поступила в редакцию 22.05.2022
The article was received on 22.05.2022

Об авторе / About the author

Корецкая Ирина Леонидовна – аспирант Отдела языков Восточной и Юго-Восточной Азии, Институт языкознания РАН, г. Москва

Irina L. Koretskaya – postgraduate student at the Department of Languages of East and Southeast Asia, Institute of Linguistics of the Russian Academy of Sciences, Moscow, Russian Federation

ORCID: https://orcid.org/ 0000-0002-0537-0575
E-mail: koretskayairina@yandex.ru