***И.Л. Корецкая***

**ПРОБЛЕМЫ ЯПОНСКОЙ КОРПУСНОЙ ЛИНГВИСТИКИ**

**Аннотация.** В ходе работы над созданием корпуса текстов его разработчики сталкиваются со многими сложностями, некоторые из которых являются универсальными (обеспечение репрезентативности корпуса, сбор материала и т.д.), а некоторые обусловлены спецификой языка. В данной работе представлен краткий обзор современного состояния японской корпусной лингвистики, а также анализ различных вопросов, решением которых занимаются разработчики японских корпусов.

**Ключевые слова.** Японская корпусная лингвистика, лингвистический корпус, проблема слова в японском языкознании, автоматическая обработка японского языка

1. **Введение**

Современные исследования, проводимые в рамках японской корпусной лингвистики, в основном посвящены изучению различных языковых явлений на материале корпусов, а также проблемам создания новых и усовершенствования имеющихся программ по автоматической обработке языка, особенно с использованием нейронных сетей. Примеры таких работ можно найти, например, в сборниках материалов Семинара по языковым ресурсам[[1]](#footnote-1), на сайте Университета Киото[[2]](#footnote-2) и в др. источниках. Работы по японским корпусам устной речи собраны, к примеру, в [Proceedings of LREC 2018 Special Speech Sessions (2018)]. История развития корпусной лингвистики в Японии подробно описана в статьях сборника «Japanese Linguistics», в том числе Miyajima [2007].

Многие японские исследователи отмечают отставание развития японской корпусной лингвистики от других стран, указывая на то, что долгое время не было объемных репрезентативных национальных корпусов японского языка, создание которых является одной из ее главных задач (см., например, [Gotoo, 2007], [National Institute for Japanese Language and Linguistics: Survey and guide 2020/2021, 2020]). Критика в адрес японской корпусной лингвистики, видимо, вызвана сравнением ее с европейской корпусной лингвистикой. При разработке корпусов японцы ориентировались на идеалы Брауновского корпуса, однако специфика японского языка ставит перед разработчиками корпусов свои задачи.

Исследования по японской корпусной лингвистике вне Японии практически не известны и достаточно редки. И цель данной работы – представить краткий обзор современного состояния японской корпусной лингвистики, а также анализ различных вопросов, решением которых занимаются разработчики японских корпусов.

1. **Развитие японской корпусной лингвистики**

Можно сказать, что зарождение корпусной лингвистики в Японии произошло в 1950-х гг. В то время методология сбора и исследования языкового материала была схожа с методологией современной корпусной лингвистики, хотя сам термин «корпус» еще не использовался. В конце 1960-х – в 1970-х гг. для исследований японского языка начинают использовать ЭВМ, собранные материалы оцифровывают. В 1990-е гг. вместе с ростом популярности персональных компьютеров появляется все больше электронных версий журналов и газет, что облегчает создание электронных баз по их материалам и дает возможность использовать их в отдельных исследованиях по японскому языку, в научный оборот входит термин «корпус». Однако все еще нет сбалансированного корпуса и признанных методов исследования.

По мнению японских исследователей, новый этап развития корпусной лингвистики в Японии начался с создания Центром разработки корпусов Национального института японского языка и лингвистики (国立国語研究所, NINJAL)[[3]](#footnote-3) объемного Сбалансированного корпуса современного японского письменного языка (現代日本語書き言葉均衡コーパス, BCCWJ).

Кроме NINJAL, стоит отметить и Университет Киото[[4]](#footnote-4), работающий над созданием различных корпусов, среди которых текстовый корпус (京都大学テキストコーパス) и веб-корпус (京都大学ウェブ文書リードコーパス), а также над созданием инструментов для автоматической обработки японского языка (морфологического анализатора Juman и его версий, парсера KNP и др.). Международному институту передовых телекоммуникационных исследований ATR[[5]](#footnote-5) принадлежит внушительный список баз данных устной речи, используемых во многих исследованиях: базы спонтанной речи, база речи пожилых людей, база речи детей и др. Национальным институтом информационно-коммуникационных технологий NICT[[6]](#footnote-6) разработаны различные базы данных, включая учебный корпус JLE, содержащий более 1 тыс. записей интервью с носителями японского языка, сдававших устный экзамен на знание английского языка OPI, а также параллельные корпусы (англо-японский, корейско-японский, японско-китайский) и др.

Некоторые организации взяли на себя сбор и распространение корпусов, созданных различными разработчиками. Например, на сайте Национального института информатики (国立情報学研究所)[[7]](#footnote-7) представлено более 40 корпусов устных текстов, на сайте Общества языковых ресурсов GSK (言語資源協会)[[8]](#footnote-8) имеется 11 корпусов письменных и устных текстов. Такие организации помогают получить доступ к необходимому для исследований материалу, ведь не все корпусы широко известны.

1. **Обзор корпусов японских текстов**

В Японии создается большое количество самых разных текстовых корпусов, и их описание заслуживает отдельной статьи. Здесь мы приведем только несколько наиболее важных.

***3.1. Корпусы письменных текстов***

Выше уже упоминался корпус BCCWJ, чей объем составляет около 100 млн единиц[[9]](#footnote-9). Корпус состоит из 3 подкорпусов: подкорпуса публикаций, библиотечного подкорпуса и подкорпуса специализированных текстов.

Объемным является и Текстовый корпус университета Киото (京都大学テキストコーパス), который состоит из 40 тыс. предложений, взятых из газеты «Майнити Синбун» за 1995 г.

Из веб-корпусов стоит отметить разработанный NINJAL Веб-корпус японского языка (国語研日本語ウェブコーパス) и Веб-корпус университета Киото (京都大学ウェブ文書リードコーパス). Первый содержит 10 млрд единиц, для его создания разработчики собирали по 100 млн веб-страниц каждые три месяца в период с октября по декабрь 2014 г. Для создания второго корпуса выбирались по три первых предложения из 5 тыс. веб-документов (общим объемом 15 тыс. предложений) различных жанров и стилей, включая новостные и энциклопедические статьи, блоги и др.

На основе Корпуса истории японского языка (日本語歴史コーパス, CHJ), который пока находится в разработке, планируется создать диахронический корпус, который будет включать в том числе и тексты на старояпонском.

Среди корпусов детской речи выделяется выпущенный Университетом Конан Корпус Kodomo (こどもコーパス), содержащий сочинения детей в возрасте 10–11 лет. В нем можно проследить историю корректуры текстов.

В качестве примера письменных учебных корпусов приведем Корпус JEFLL (The Japanese English as a Foreign Language Learner Corpus), содержащий более 10 тыс. сочинений (свыше 600 тыс. с/у) носителей японского языка, изучающих английский язык. Была создана его версия, где написанные сочинения проверены носителями английского.

***3.2. Корпусы устных текстов***

NINJAL разработал и Корпус спонтанной речи (日本語話し言葉コーパス, CSJ), объем которого составил порядка 7 млн единиц (660 часов записи). Основную часть корпуса составляют «академическая речь» (запись живых академических презентаций) и «симулированная публичная речь» (студийные записи речи на повседневные темы перед небольшой аудиторией и в относительно непринужденной обстановке).

Среди корпусов устных текстов нельзя не отметить упомянутые выше базы данных устной речи ATR разного содержания и разработанный NICT Корпус JLE объемом более 1 тыс. записей.

 Первым японским корпусом, содержащим записи диалектной речи многочисленных регионов Японии, считается разработанный NINJAL Корпус японских диалектов (日本語諸方言コーパス, Corpus of Japanese Dialect). Он содержит 4 тыс. часов записи диалектной речи из более чем 200 мест в 47 префектурах по всей Японии. Сбор материала осуществляло Агентство по культурным вопросам с 1977 по 1985 год.

1. **Сложности при создании корпуса японских текстов**

Одной из главных сложностей является обеспечение репрезентативности корпуса, что в полном объеме зачастую невыполнимо. Для получения максимально репрезентативного корпуса можно, например, уравнивать количество текстов по важным для изучаемого объекта параметрам, следя за тем, чтобы корпус отражал все значимые характеристики объекта. Подобный принцип был использован при создании Базы данных устного повседневного общения (談話資料 日常生活のことば) и Корпуса повседневного общения (日常会話コーパス, CEJC), когда при выборе информантов учитывали их пол и возраст, контролируя таким образом выборку. Разработчики Корпуса спонтанной речи (日本語話し言葉コーパス, CSJ) включили в корпус тексты разных типов, стараясь таким образом сбалансировать материал. При создании корпуса BCCWJ для обеспечения репрезентативности применялась случайная выборка текстов.

Большинство сложностей связаны с автоматической обработкой языка. Несмотря на значительные успехи, программы допускают ошибки при анализе текстов. В связи с этим исследователям приходится идти на определенные жертвы. Можно получить либо объемный корпус, но не очень тщательно размеченный, либо хорошо размеченный, но небольшой по объему. NINJAL использует следующий принцип. В нескольких корпусах разработчики выделили «Ядро», которое составляет небольшую часть корпуса, где разметка проведена наиболее детально и отредактирована вручную. Остальная часть корпуса размечена автоматически.

Сложности с автоматической обработкой языка обусловлены в первую очередь спецификой самого языка. В случае с японским языком многие из них связаны с особенностями японской письменности, среди которых:

1. больший набор символов (иероглифы, две слоговых азбуки, латиница, арабские цифры, знаки пунктуации и др.) и несовместимость имеющихся наборов символов [Костыркин 2004: 21-22];
2. различные варианты написания одного и того же слова: оно может быть записано каной, иероглифами или сочетанием иероглифа с каной, причем каждый из этих способов может иметь несколько вариантов;
3. отсутствие пробелов в тексте, что приводит к сложностям с сегментацией текста, а она, в свою очередь, выливается в проблему морфологического и синтаксического анализа единиц текста.

Сегментация японского текста зависит от выбора его базовой единицы. Японская лингвистическая традиция выделяет единицы, которые получили название «*go*». В европейской японистике этот термин обычно переводят как «слово». В статьях японских авторов на английском языке можно встретить термины «словоподобные единицы» (word-like units) и «морфемоподобные единицы» (morpheme-like units) с замечанием о том, что понятие слова для японского языка плохо разработано и что в значении термина «слово» часто используется термин «морфема», хотя и понимается этот термин иначе, чем в европейской лингвистике. На наш взгляд, критика не совсем справедлива. Дело не в недостаточной проработанности понятия, а лишь в иной его трактовке. Как отмечает Алпатов В.М., самостоятельным go соответствуют единицы, которые могут состоять из корней и словообразовательных показателей, несамостоятельным *go* – как служебные слова, так и аффиксы [Алпатов 2018: 59]. Таким образом, налицо несовпадение японского «*go*» с европейскими «словом» и «морфемой».

Другая единица, выделяемая в японском языке, – *bunsetsu*. Она соответствует самостоятельному *go* со всеми примыкающему к нему формантами, если таковые имеются. Деление текста на *bunsetsu* кажется носителям языка наиболее естественным.

Традиции японского языкознания очевидным образом прослеживаются в идеях NINJAL. Институт предложил интересное с практической точки зрения решение для обработки текстов, состоящее в сегментации текста на «короткие слова» (SUW, short unit words), некоторые из которых затем при помощи статистических методов объединяются в «длинные слова» (LUW, long unit words)[[10]](#footnote-10). LUW составляются из SUW в основном в случае составных существительных, глаголов и частиц, т.е. в тех случаях, когда вопрос о проведении границ внутри слова представляется исследователям японского языка наиболее спорным. Например, составное существительное 公害紛争処理法 («акт об урегулировании конфликтов по вопросам загрязнения окружающей среды») состоит из существительных 公害 («загрязнение окружающей среды»), 紛争 («диспут, конфликт»), 処理 («урегулирование, разрешение») и 法 («закон, акт»); частица における (локатив) – из частицы に (датив), глагола おけ («быть, находиться») и вспомогательного глагола る (настоящее время)[[11]](#footnote-11) [Maekawa et al. 2014: 355]. Как видно из примера, SUW – это и не слова, и не морфемы в европейском понимании этих терминов, SUW наиболее близки традиционно выделяемым базовым единицам. LUW –- единицы, которые могут быть больше SUW, но меньше *bunsetsu*. *Bunsetsu* в основном используются разработчиками разных корпусов для парсинга (но, например, в проекте Universal Dependencies[[12]](#footnote-12) для японского языка используют синтаксическое слово, которое не равно *bunsetsu*).

Для морфологической разметки корпусов NINJAL использует морфологический анализатор MeCab, работающий на словаре UniDic. Этот словарь был разработан относительно недавно как универсальный словарь, решающий большинство проблем его предшественников, одной из которых был непоследовательный подход к членению текста.

В упомянутой выше работе авторы дают сравнение четырех словарей (Juman, IPA, Yahoo! и UniDic) и показывают, что одно и то же составное слово разбивается этими словарями по-разному. Словарь UniDic отличается от остальных словарей тем, что проводит наиболее дробное и единообразное членение текста [Maekawa et al. 2014: 354], что служит не только более эффективному автоматическому анализу текста, но и получению более надежных данных при поиске в корпусе текстов.

В случае с автоматической сегментацией текста программа имеет дело с письменным текстом, а значит, в этом случае речь идет именно об орфографических словах, т.е. единицах между двумя пробелами или пробелом и знаком препинания. Однако перед компьютерной лингвистикой стоит и другая, не менее сложная задача, а именно распознавание и дешифровка речи. В речевом потоке слова сливаются в некоторую последовательность звуков, происходит искажение фонетического облика слов, а значит, для распознавания речи нужны более длинные единицы, как минимум, фонетические слова. Чем длиннее единица, тем более объемным должен быть словарь, но в случае с использованием более длинных единиц снижается количество ошибок в работе программы.

На наш взгляд, для распознавания японской устной речи наилучшим решением является *bunsetsu*. В японском языке высока частота омонимии и эллиптических конструкций, соответственно, для корректного распознавания нужен контекст. Более того, *bunsetsu* часто является и фонетическим словом, имеющим единый тоновый контур. Тоновый контур *go* в японском языке может отличаться от диалекта к диалекту и меняться даже в рамках одного диалекта в зависимости от окружения единицы в предложении, от индивидуальных характеристик говорящего и даже от ситуации общения, что осложняет процесс автоматизации расшифровки устных текстов. Поэтому так важно рассматривать фонетическое слово целиком.

1. **Заключение**

Японская корпусная лингвистика, зародившись в 1950-х гг., стремительно развивается. Появляется все больше представительных корпусов, проводится большое число исследований как на материале уже созданных корпусов, так и для усовершенствования имеющихся и разработки новых корпусов.

В ходе работы над созданием корпуса текстов разработчики сталкиваются со многими сложностями, некоторые из которых являются универсальными (обеспечение репрезентативности корпуса, сбор материала и т.д.), а некоторые обусловлены спецификой языка. В случае с японским языком они связаны с особенностями японской письменности, супрасегментными характеристиками японской речи, высокой степенью омонимии, разногласием по различным теоретическим вопросам (сегментного членения текста, морфологического анализа языковых единиц и пр.). Эти проблемы значительно осложняют аннотирование (а в случае с устной речью – и распознавание) японского текста. При создании корпуса текстов не обойтись без автоматической обработки текстов, но несмотря на развитие программ по обработке японского языка, машине все еще нужна помощь человека для проведения полностью корректного анализа текстов.

**Литература:**

1. *Алпатов В.М. (2018)*, Слово и части речи. М.

2. *Кострыкин А.В. (2004)*, Японские корпусные проекты. Научно-техническая информация, серия 2, № 9, с. 20-30.

3. *後藤 斉 [Gotoo Hitosi] (2007),* コーパス言語学と日本語研究 [Corpus linguistics and Japanese language studies]. 日本語科学 [Japanese Linguistics], 22, pp. 47－58, available at doi/10.15084/00002182 (дата обращения: 22 мая 2021)

4. *Proceedings of LREC 2018 Special Speech Sessions (2018),* available at http://doi.org/10.15084/00001908 (дата обращения: 22 мая 2021)

5. *Maekawa K., Yamazaki M., Ogiso T., Maruyama T., Ogura H., Kashino Wa, Koiso H., Yamaguchi M., Tanaka M., Den Ya. (2014)*, Balanced corpus of contemporary written Japanese. Language Resources and Evaluation, Vol 48, pp. 345–371, available at: link.springer.com/article/10.1007/s10579-013-9261-0 (дата обращения: 12 марта 2021)

6. *宮島 達夫 [Miyajima Tatsuo] (2007)*, 語彙調査からコーパスへ　[From vocabulary statistics to corpus-based studies].　日本語科学 [Japanese Linguistics], 22, pp. 29－46, available at doi.org/10.15084/0000218 (дата обращения: 12 марта 2021)

7. *国立国語研究所要覧 2020/2021 [National Institute for Japanese Language and Linguistics: Survey and guide 2020/2021] (2020)*, available at http://id.nii.ac.jp/1328/00002825/ (дата обращения: 22 мая 2021)

**References:**

1. *Alpatov V.M. (2018)*, Slovo i chasti rechi [Word and Parts of Speech]. Moscow.

2. *Kostyrkin A.V. (2004)*, Yaponskie korpusnyie proekty [Japanese Corpus Projects]. Nauchno-tekhnicheskaya informatsiya [Scientific and Technical Information], 2, № 9, pp. 20-30.

3. *後藤 斉 [Gotoo Hitosi] (2007),* コーパス言語学と日本語研究 [Corpus linguistics and Japanese language studies]. 日本語科学 [Japanese Linguistics], 22, pp. 47－58, available at doi/10.15084/00002182 (accessed on May 22, 2021)

4. Proceedings of LREC 2018 Special Speech Sessions (2018), available at http://doi.org/10.15084/00001908 (accessed on May 22, 2021)

5. *Maekawa K., Yamazaki M., Ogiso T., Maruyama T., Ogura H., Kashino Wa, Koiso H., Yamaguchi M., Tanaka M., Den Ya. (2014)*, Balanced corpus of contemporary written Japanese. Language Resources and Evaluation, Vol 48, pp. 345–371, available at: link.springer.com/article/10.1007/s10579-013-9261-0 (accessed on March 12, 2021)

6. *宮島 達夫 [Miyajima Tatsuo] (2007)*, 語彙調査からコーパスへ　[From vocabulary statistics to corpus-based studies].　日本語科学 [Japanese Linguistics], 22, pp. 29－46, available at doi.org/10.15084/0000218 (accessed on March 12, 2021)

7. *国立国語研究所要覧 2020/2021 [National Institute for Japa-nese Language and Linguistics: Survey and guide 2020/2021] (2020)*, available at http://id.nii.ac.jp/1328/00002825/ (accessed on May 22, 2021)

**CHALLENGES IN THE JAPANESE CORPUS LINGUISTICS**

I.L. Koretskaya

**Abstract.** When building a text corpus, the developers face various problems, some of which are universal (e.g. ensuring the corpus representativeness, collecting material, etc.) and some are caused by some specific features of the language. In case with the Japanese language, most problems are related to the Japanese writing system. The research on the Japanese corpus linguistics is not very well known outside Japan, as well as quite rare. In this paper, we give a brief overview of the current Japanese corpus linguistics, its development and the main Japanese corpora (both written texts and speech corpora) and analyze different issues the Japanese corpora developers deal with.

**Keywords.** Japanese corpus linguistics, linguistic corpus, word in the Japanese linguistics, Japanese language processing

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Корецкая Ирина Леонидовна**

Институт языкознания РАН (Россия).

**Koretskaya Irina**

Institute of Linguistics, Russian Academy of Sciences (Russia).

E-mail: koretskaya\_irina@iling-ran.ru

1. Сборники за 2016 – 2020 гг. можно найти в архиве NINJAL https://repository.ninjal.ac.jp/ [↑](#footnote-ref-1)
2. https://nlp.ist.i.kyoto-u.ac.jp/EN/?Publications [↑](#footnote-ref-2)
3. http://pj.ninjal.ac.jp/corpus\_center/ [↑](#footnote-ref-3)
4. https://nlp.ist.i.kyoto-u.ac.jp/EN/ [↑](#footnote-ref-4)
5. http://www.atr-p.com/products/sdb.html　 [↑](#footnote-ref-5)
6. https://www.nict.go.jp/ [↑](#footnote-ref-6)
7. http://research.nii.ac.jp/src/en/list.html [↑](#footnote-ref-7)
8. https://www.gsk.or.jp/en/catalog/ [↑](#footnote-ref-8)
9. В качестве единиц NINJAL использует SUW, о которых будет сказано ниже. [↑](#footnote-ref-9)
10. Для устной речи выделяют также «средние слова», представляющие собой «короткие слова» с добавлением супрасегментных характеристик. [↑](#footnote-ref-10)
11. Cохранена терминология авторов статьи. [↑](#footnote-ref-11)
12. https://universaldependencies.org/introduction [↑](#footnote-ref-12)