

Научная статья
УДК 811.523
<https://doi.org/10.24866/2542-1611/2023-4/91-99>

Развитие компьютерной лингвистики и технологий обработки естественного языка (NLP) в РК и КНДР: возможности для российского корееведения

Марина Петровна КУКЛА

Дальневосточный федеральный университет, Владивосток, Россия, kukla.mp@dvfu.ru

Юлия Сергеевна КУЗЮРА

Дальневосточный федеральный университет, Владивосток, Россия, kuziura.ius@students.dvfu.ru

Аннотация. В статье исследуются достижения в сфере технологий обработки естественного языка в Республике Корея и КНДР. Дана характеристика общедоступных корпусов корейского языка, созданных в различных институтах Республики Корея. Поднимается проблематика машинного перевода северокорейских текстов и даётся обзор международных исследований северного варианта корейского языка по методологии компьютерной лингвистики. Исследуются возможности использования корпусов корейского языка для развития преподавания корейского языка в российских вузах.

Ключевые слова: КНДР, Республика Корея, обработка естественного языка, корпусная лингвистика, корпусы корейского языка

Для цитирования: Кукла М.П., Кузюра Ю.С. Развитие компьютерной лингвистики и технологий обработки естественного языка (NLP) в РК и КНДР: возможности для российского корееведения // Известия Восточного института. 2023. № 4. С. 91–99. <https://doi.org/10.24866/2542-1611/2023-4/91-99>

Original article
<https://doi.org/10.24866/2542-1611/2023-4/91-99>

The Development of Computational Linguistics and Natural Language Processing Technologies (NLP) in the Republic of Korea and the DPRK: Opportunities for Korean Studies in Russia

Marina P. KUKLA

Far Eastern Federal University, Vladivostok, Russia, kukla.mp@dvfu.ru

Yuliya S. KUZUYURA

Far Eastern Federal University, Vladivostok, Russia, kuziura.ius@students.dvfu.ru

Abstract. The article examines achievements in natural language processing technologies in Republic of Korea and Democratic People's Republic of Korea. An overview of institutes conducting natural language processing research is provided along with an assessment of the work done by ROK and DPRK scientists. The issue of machine translation of DPRK texts based on existing parallel corpora and machine translation tools is discussed. The potential for applying corpus linguistics methodology (the corpus method) to improve the teaching of Korean language in Russian universities is investigated. The paper aims to comprehensively map the current state of natural language processing research and applications in the two Koreas and explore possibilities for cooperation and knowledge transfer that could accelerate progress.

Keywords: DPRK, Republic of Korea, Natural language processing (NLP), corpus linguistics, Korean language corpus

For citation: Kukla M. P., Kuzuyura Yu.S. The Development of Computational Linguistics and Natural Language Processing Technologies (NLP) in the Republic of Korea and the DPRK: Opportunities for Korean Studies in Russia // Oriental Institute Journal. 2023. № 4. P. 91–99. <https://doi.org/10.24866/2542-1611/2023-4/91-99>

Получившие в последнее время большое развитие компьютерная лингвистика и обработка естественного языка (Natural Language Processing, NLP), не являясь изначально родственными научными направлениями, во многом стали соприкасаться и решать смежные задачи. Корпусная (компьютерная) лингвистика предполагает информатизацию языковых исследований в понятной для человека форме, основанной на подобранных, обработанных по определённым правилам совокупностях текстов естественного языка. С другой стороны, обработка естественного языка включает три основных направления: распознавание речи, пони-

мание и генерация естественного языка – и тесно связана с машинным обучением. Сферу обработки естественного языка и компьютерную лингвистику объединяет использование больших данных естественного языка (устного, письменного) [2, с. 105].

Достижения корпусной лингвистики и NLP в Республике Корея

С распространением методов машинного обучения в обработке естественного языка создание языковых наборов данных и их публикация стали краеугольным камнем продвижения исследований. Технологичными NLP в Южной Корее занимаются государственные учреждения, крупные вузы, учёные-энтузиасты, целью которых является создание открытых данных корейского естественного языка.

Национальный институт корейского языка, входящий в структуру Министерства культуры, спорта и туризма Республики Корея, является ведущим учреждением в стране, определяющим нормы корейского языка, а также осуществляющим работу по созданию корпусов и наборов данных корейского языка, стимулируя тем самым лингвистические исследования прежде всего в самой Республике Корея. Результатом работы института является Национальный корпус корейского языка (ModuCorpus), содержащий 39 корпусов и наборов данных общим объёмом 7 млн 583 тысячи документов: газетных текстов, устной речи, протоколов Национальной Ассамблеи и др. Есть также корпус оценки языковых возможностей искусственного интеллекта [18].

В 2021 г. Институт опубликовал доклад о завершении работы над параллельными (двухязычными) корпусами, содержащими тексты на корейском языке и их переводы на иностранные языки: вьетнамский, индонезийский, тайский, хинди, кхмерский, тагальский, русский и узбекский. Эти восемь языков представляются корейским лингвистам значимыми с точки зрения продвижения корейской массовой культуры. В докладе отмечено: "Поскольку "Корейская волна" вышла за рамки дорам и к-попа и охватила различные области, такие как корейский язык, фильмы, еда и индустрия красоты, крайне важно иметь языковые ресурсы для дальнейшего развития отношений с Россией и странами Центральной Азии с точки зрения культуры, экономики и социальной сферы". Новые параллельные корпусы содержат как текстовые материалы, так и устную речь: более 1 млн фраз для каждого языка. Поиск в Общем корпусе Национального института корейского языка, в том числе по параллельным корпусам, возможен лишь для зарегистрированных пользователей, что ограничивает его использование за рубежом [11].

Есть в Национальном институте корейского языка РК и открытый корпус, представляющий большие возможности для исследователей и преподавателей, в том числе русскоязычных. Речь идёт о корпусе изучающих корейский язык, который является полностью открытым [23]. Этот корпус состоит из текстовых и разговорных материалов иностранцев, изучающих корейский язык – представителей 142 стран, говорящих на 93 языках. Работа над составлением корпуса велась в течение 5 лет начиная с 2015 г. В текстах корпуса намеренно допускаются грамматические и орфографические ошибки, что имеет значение для анализа процесса обучения и создания учебных материалов. На настоящий момент в базе данных корпуса присутствует массив текстов 3 784 091 словоупотреблений. Общее количество документов корпуса – 26 152, из них 93,1% приходится на письменные тексты, 6,9% – на устные [1].

Большими достижениями в сфере корпусной лингвистики обладает университет Ёнсе, где работы по составлению корпусов ведутся с 1986 г. На специализированной страничке университета представлен список из 26 корпусов объёмом более 1 млрд 148 млн словоупотреблений. Репрезентативность обеспечивается разнообразием корпусов, представляющих различные жанры литературы, газетные тексты, материалы социальных сетей, учебные материалы школьных курсов, устную речь и другое. Особенностью системы корпусов Университета Ёнсе является возможность свободного поиска: на специализированной странице представлены 4 корпусных менеджера, каждый из которых обеспечивает поиск по нескольким связанным друг с другом корпусам [20].

Репрезентативным и открытым является также словарь и корпус Седжона, созданный в Сеульском национальном университете [14]. Корпус Седжона представляет собой крупномасштабный размеченный NLP-инструмент для таких задач, как синтаксический или грамматический анализ. Кроме того, недавно были выпущены размеченные корпуса объёмом около 300 миллионов словоупотреблений, предназначенные для решения задач машинного обучения.

Наряду с компьютерными корпусами, созданными под запросы человека, в Южной Корее ведётся большая работа по созданию наборов данных для машинного обучения, в который вовлечены не столько гуманитарные, сколько технические вузы и научные организации.

Корейский институт передовых технологий (KAIST) разработал и разместил в свободном доступе морфосинтаксический аннотированный корпус объёмом около 70 миллионов словоупотреблений. Он включает корпуса художественной литературы, публицистики, специально профессиональной литературы, параллельные корпуса (для китайского и английского языков) и т.д. [8]. Корейский корпус KAIST основан на независимо собранных 30000 предложений, аннотированных в соответствии со схемой древовидного тегирования для корейского языка. Корпуса KAIST доступны для скачивания только после регистрации, онлайн-менеджер в них отсутствует.

Наряду с языковыми корпусами ведётся работа по созданию наборов данных для машинного обучения. Они размещаются на различных платформах, в частности созданной Национальным агентством по вопросам информационного общества (NIA) платформой AI HUB [5]. Наборы данных создаются в рамках различных государственных задач, чтобы способствовать развитию искусственного интеллекта. Ведутся разработки моделирования разговорной речи. Также в свободном доступе находится открытый словарь NIAdic, предоставленный K-ICT Big Data Center [9].

Наборы данных готовят как исследователи-энтузиасты, так и крупные институты. Обучение искусственного интеллекта переводу и ответам на вопросы требует большого объёма информации, представленных в виде параллельных корпусов и наборов данных, содержащих вопросы и ответы. В целях развития машинного обучения собирают и речевые корпуса, включающие разговоры по телефону, аудиокниги и даже пхансори. Эти наборы данных также размещаются на различных платформах сообществ IT-специалистов –Github, Kaddle [10].

Таким образом, в настоящее время как в свободном доступе, так и в ограниченном находятся большие объёмы данных в виде многочисленных корпусов южного варианта корейского языка, которые могут быть использованы как для исследований и преподавания, так и для машинного обучения.

Обработка естественного языка в КНДР

О достижениях информационных технологий в КНДР по понятным причинам известно немного. Тем не менее нет сомнений, что такие исследования ведутся, и важной задачей остаётся их интеграция в международные исследования. По причине существенных языковых различий использовать для обработки северокорейского варианта языка указанные выше корпуса южного варианта корейского языка не вполне корректно (хотя и такие исследования ведутся, о чем будет сказано ниже).

Разработки в сфере искусственного интеллекта в КНДР стартовали более 25 лет назад с программирования компьютерных игр. В 1997 г. была разработана программа для игры в го "Серебряная звезда", благодаря которой северокорейская команда одержала победу на международных соревнованиях в 1998 г. В это время "Серебряная звезда" считалась лучшей программой для игры в го в мире, в середине 2000-х гг. северокорейские разработчики продали её Южной Корее [12].

Активизация работ в сфере обработки естественного языка в КНДР относится к концу 2006 г., когда Ким Чен Ир поручил команде учёных из университета имени Ким Ир Сена создать программу машинного перевода. Первые результаты были представлены в 2008 г. в виде программы "Рённамсан" (получившей название по

имени района, где расположен Университет). Есть большие основания полагать, что работа эта получила развитие, и в настоящее время инструменты машинного перевода активно используются в северокорейском интранете или национальной операционной системе "Красная звезда". 28 августа 2014 г. в "Нодон синмун" вышла статья о том, что исследователь Ким Кванхёк из университета имени Ким Ир Сена получил почётную молодежную награду имени Ким Чен Ира за разработку программы перевода, основанной на технологии ИИ [12].

Крупнейший вуз КНДР, Университет имени Ким Ир Сена, является лидером цифровых разработок. В нём действует несколько кафедр и центров, занимающихся вопросами информационных технологий, а в научном журнале Университета выходят статьи на тему компьютерной лингвистики.

В одной из работ проведен анализ морфологических и синтаксических характеристик устной речи северного варианта корейского языка, таких как опущение и изменение порядка слов, а также предложен метод автоматического создания корпуса устной речи. Языковые модели, используемые в эксперименте, были построены из набора данных "Нодон синмун" (письменный стиль) и сборников кинолитературы и драматических произведений (устный стиль). Упоминаемые в статье наборы данных письменных и устных текстов содержат 98 000 и 18 000 словоупотреблений соответственно [17, с. 11]. Считать их полноценными корпусами ввиду небольшого объёма некорректно, однако автор использует слово "корпус" для их обозначения.

В другой статье того же автора рассказывается о методах создания корпуса спонтанных высказываний через корпус письменного и устного языков. Отмечается, что создание большого корпуса устного языка, как правило, нецелесообразно из-за дороговизны сбора информации и записи. Поэтому корпусы, представляющие характеристики конкретной задачи, например, газеты, могут использоваться в сочетании с корпусами, представляющими характеристики спонтанной речи [16, с. 3].

Другая группа исследователей КНДР предлагает метод фильтрации для параллельного корпуса английского и корейского языков, в результате эксперимента получая более высокий уровень аккуратности перевода [13]. В другой работе исследователи университета Ким Ир Сена предлагают методы обработки естественного языка, а именно построение глагольного каркаса в предложении, таким образом занимаясь вопросом генерации текста [22].

Важную роль в развитии информационных технологий играет также Пхеньянский университет компьютерных технологий. В июне 2019 г. появилось сообщение о том, что кафедра программной инженерии этого вуза была переименована в кафедру интеллектуальных информационных технологий. С 2016 г. команда ученых этого университета ведёт разработку применения технологий ИИ в голосовом поиске.

В КНДР регулярно проводятся соревнования и выставки, где представлены достижения в сфере ИТ. В 2013 г. прошли 24-е соревнования по программированию и выставка, на которых был представлен Исследовательский институт ИИ Бюро Информационных технологий. Результатом работы этого исследовательского института является развлекательная площадка "Рюгён" для северокорейского интранета, где можно поиграть в го, шахматы и другие игры, а также программа проверки норм орфографии пхеньянского литературного языка [12].

Крупным событием в сфере информационных технологий в КНДР, связанным с ИИ в КНДР, является национальная выставка достижений информатизации (*чонгукчонбохвасонгвачолламхве*), которая проходит в Пхеньяне. Известно, что такие выставки проходили осенью 2019 и 2022 года.

Темой выставки 2019 г. была "цифровая экономика". На выставке проходили соревнования в программировании ИИ, на которых северокорейские предприятия и организации состязались в различных областях, которые кроме всего прочего включали программы распознавания лиц, программы голосового распознавания, программы текстового распознавания, программы машинного перевода,

программы распознавания отпечатков пальцев, программы распознавания номеров машин [12].

В октябре 2022 г. выставка проходила в виртуальной форме на базе Бюро информационных технологий Министерства информационной индустрии. На ней было представлено 1400 образцов, в том числе программы для операционной системы "Красная звезда" (*пультыгиппёль*), облачные вычисления, технологии больших данных, интернет вещей. Научно-исследовательский институт информационных технологий Института развития передовых технологий Университета Ким Ир Сена представил комплексную систему управления производством программного обеспечения и систему распознавания биометрической информации [21].

Таким образом, исследования и практические разработки в сфере лингвистики и обработки естественного языка, равно как и международное сотрудничество в этой сфере может стать для российских университетов одним из направлений сотрудничества с северокорейскими вузами.

Зарубежные исследования северного варианта корейского языка

С момента разделения Кореи прошло почти 80 лет, и за эти годы языки государств Корейского полуострова стали существенно отличаться друг от друга. Актуальность приобретает вопрос перевода с одного языка на другой. Появляются онлайн словари, осуществляющие такой перевод. Известным и доступным переводчиком-словарём является "Кыльтонму" (*кор.* "товарищ по языку") [15]. Аналогичные приложения-переводчики разработаны и для смартфонов, например, "Нарамаль понёкки" (*кор.* "переводчик языка нашей страны"), выпущенное в 2018 г. и обновлённое в 2021 г. Оба эти приложения механически меняют слова, принятые в одном варианте языка, на другие либо так же заменяют грамматические конструкции. Такие приложения эффективны для простых бытовых фраз, однако для сложных текстов не применимы.

Для обработки естественного языка требуется большое количество языковых наборов данных. Для языка КНДР по многим причинам такие ресурсы дефицитны, что очень затрудняет разработку моделей перевода или генерации текста.

Большинство языковых исследований пытаются обойти проблему дефицита ресурсов либо полагаясь на ручное чтение экспертами предметной области статей о КНДР, либо сосредотачивая внимание на англоязычных статьях, предоставленных Центральным телеграфным агентством Кореи (ЦТАК). Использование методов NLP во всех этих исследованиях, как правило, построены на поиске ключевых слов (наиболее часто используемых слов в тексте).

Проблему нехватки ресурсов северокорейского варианта корейского языка поднимают в своих работах зарубежные специалисты. Например, учёные из Японии занимаются тем, что также создают доступный корпус северокорейского языка, а также проводят эксперименты по созданию машинного выравнивания предложений. Они, в частности, отметили проблему, что северокорейские слова с начальным "ㄹ" (например, *립장, 룡군*) не могут быть адекватно переведены современными средствами машинного перевода [7]. Они же посвятили свою работу по составлению параллельного корпуса северокорейского варианта языка на базе сайта Комитета по мирному объединению родины (известному как "*Уриминждоккири*"), которая публикует статьи на нескольких языках (английском, китайском, русском и японском).

Специалисты из Токийского Университета, обращаясь к проблеме нехватки лингвистических ресурсов северного варианта корейского языка, предлагают большой набор данных, используют его для обучения языковой модели BERT. Актуальными задачами, с которыми выступают исследователи, остаются поиск инструментов для веб-скрапинга ресурсов КНДР, создание языковых моделей, параллельных корпусов [6].

Любопытные результаты получили исследователи из Токийского университета, которые методом веб-скрапинга создали базу текстов газеты "Нодон синмун" с 2018 г. и включающую около 50 тысяч статей [6, с. 5596]. Они получили набор данных из 27401 документов общим объемом 471 417 предложений. Они поль-

зовались не сайтом газеты, а сайтом ЦТАК, который ежедневно печатает тексты "Нодон синмун" как на корейском, так и на английском языках. Второй набор данных, которые изучали исследователи, состоял из новогодних обращений руководителей КНДР 1946 – 2019 гг. в текстовом формате (за исключением 1957 г.). Этот набор данных составил 73 документа объёмом 5709 предложений.

В результате исследования авторы пришли к выводу о том, что использование корейского языка в новогодних речах руководителей КНДР гораздо больше напоминает южный вариант корейского языка по сравнению с новостными статьями "Нодонсинмун". С другой стороны, в новостных статьях "Нодонсинмун" корейский язык используется гораздо более индивидуально и разнообразно [6].

К вопросу о корпусном преподавании в корееведении

В российском корееведении, на наш взгляд, пока недостаточно представлены научные и педагогические работы, основанные на методологии корпусной лингвистики. При этом доступность корпусов корейского языка даёт немало возможностей для применения корпусного подхода не только в преподавании, но и в проектной и научной работе студентов. Такой опыт в ДВФУ уже имеется. Пионером корпусной лингвистики в корееведении ДВГУ был преподаватель кафедры корейской филологии В.Н. Кожемяко, который в конце 1990-х – начале 2000-х гг. вёл активную работу по внедрению в учебный процесс новых информационных технологий, в том числе корпусной лингвистики. Несмотря на то что цели по созданию системы автоматизированного перевода не удалось достичь, корпусный подход показал себя эффективным и действенным способом вовлечь студентов в языковые исследования, повысить мотивацию к учёбе.

Корпусные данные дают возможность получить доступ к обширному массиву аутентичных материалов, увидеть язык в действии, верифицировать или уточнить идеи о языковых закономерностях и многое другое. При этом корпусный подход в преподавании особенно актуален для наглядного представления таких аспектов языка, как историческая, географическая и социальная вариативность, а также для анализа тенденций развития языковой системы.

Корпусная лингвистика может быть применена в преподавании для разработки учебных материалов, заданий для обучающихся, а также в научно-исследовательской и проектной работе студентов. Эксперименты показывают, что использование корпусов как учебного инструмента показало себя очень эффективным [4].

Определённый задел для учебной и исследовательской работы создан на базе Национального корпуса русского языка. На сайте НКРЯ размещены двуязычные пары параллельных корпусов, в том числе для таких восточных языков, как китайский и корейский. Объём корейского корпуса небольшой – 185 документов, 12 300 предложений, 73 752 слова. Составлен он из публицистических текстов южнокорейского варианта корейского языка и их переводов на русский язык. Параллельный корпус представляет собой важный инструмент для научных исследований, например, в области типологии, контрастивной лингвистики, внутриязыкового варьирования, а также для исследований по теории и практике перевода [3].

Заключение

В Южной Корее исследования в сфере обработки естественного языка идут в большом масштабе. Различные направления работы: создание корпусов естественного языка, в том числе параллельных, для преподавания и исследований, создание наборов данных для машинного обучения – ведут государственные исследовательские институты, университеты, инициативные группы исследователей. В настоящее время существует несколько обширных корпусов корейского языка, доступ к которым полностью открыт. Их можно использовать как в обучении корейскому языку на аутентичных материалах, так и в изучении различных аспектов корейского языка.

Что касается КНДР, то здесь ведутся исследования распознавания речи, создаются наборы данных для машинного обучения, однако в очень скромном масшта-

бе. Зарубежные учёные, прежде всего в Южной Корее и Японии, также работают с проблемой машинного перевода северокорейских текстов на базе корпусов южнокорейского варианта языка. Актуальными задачами остаются поиск инструментов для веб-скрапинга ресурсов КНДР, создание языковых моделей и параллельных корпусов

Литература

1. Коврова С.Н. О Корпусе изучающих корейский язык // Гуманитарные науки. Студенческий научный форум. Электронный сборник статей по материалам XXXVI студенческой международной научно-практической конференции. Москва: МЦНО. 2021. № 1 (36). С. 50–59. URL: [https://nauchforum.ru/archive/SNF_humanities/1\(36\).pdf](https://nauchforum.ru/archive/SNF_humanities/1(36).pdf)(дата обращения: 18.11.2020).
2. Кузьмина А.А. и др. Методы компьютерной лингвистики и обработки естественного языка: возможности и ограничения для задач психологии личности/ А.А. Кузьмина, М.А. Лифшиц, В.Ю. Костенко // Современная зарубежная психология. 2022. Т. 11. № 1. С. 104–115. <https://doi.org/https://doi.org/10.17759/jmfp.2022110110>
3. Параллельный корпус (корейский) // Национальный корпус русского языка. URL: <https://ruscorpora.ru/new/search-para.html?lang=kor> (дата обращения: 26.04.2023).
4. Хорошилова С.П. Корпусное преподавание как средство совершенствования исследовательской компетенции студентов// Актуальные проблемы филологии и методики преподавания иностранных языков. 2022. Т. 16. № 2. С. 112–117.
5. AIHub. URL: <https://www.aihub.or.kr/aihubdata/data/list.do?currMenu=115&topMenu=100&srchDataRealmCode=REALM002> (дата обращения: 18.11.2022).
6. Akdemir Arda et al. Developing Language Resources and NLP Tools for the North Korean Language/ AkdemirArda, Jeon Yeojoo, Shibuya Tetsuo // Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), Marseille. 20–25.06.2022. P. 5595–5600. URL: <http://www.lrec-conf.org/proceedings/lrec2022/index.html> (дата обращения: 26.04.2023).
7. Kim Hwicheon et al. Learning How to Translate North Korean through South Korean / Kim Hwicheon, Moon Sangwhan, Okazaki Naoaki, Komachi Mamoru // Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), Marseille. 20–25.06.2022. P. 6711–6718. URL: <http://www.lrec-conf.org/proceedings/lrec2022/index.html> (дата обращения: 26.04.2023).
8. KAIST Corpus // SemanticWebResearchCenter. URL: http://semanticweb.kaist.ac.kr/home/index.php/KAIST_Corpus (дата обращения: 18.11.2022).
9. NIA Data Intelligence Cloud // K-ICTBigDataCenter. URL: https://kbig.kr/portal/kbig/knowledge/files/bigdata_report.page?bltnNo=10000000016451 (дата обращения: 18.11.2022).
10. Cho Won Ik et al. Open Korean Corpora: A Practical Report /Cho Won Ik, Sangwhan Moon, Youngsook Song // ACL Anthology. URL: <https://aclanthology.org/2020.nlposs-1.12.pdf> (дата обращения: 18.11.2022). <https://doi.org/10.48550/arXiv.2012.15621>
11. 2021년. 한국어-외국어병렬말뭉치구축사업 // 국립국어원. 2022 = 2021 год. Проект по созданию параллельного корпуса корейского и иностранного языков // Национальный институт корейского языка. 2022. URL: https://korean.go.kr/front/reportData/reportDataView.do?mn_id=207&searchOrder=years&report_seq=1090&pageIndex=1 (дата обращения: 26.04.2023).
12. 강진규. 조선인공지능인민공화국?... 북한은 이미 AI 열풍 // NK경제, 2020.01.20 = Кан Чингю. Корейская информационная технологическая республика?... Бум ИИ в Северной Корее // NKEconomics. URL: <http://www.nkeconomy.com/news/articleView.html?idxno=2415> (дата обращения: 26.04.2023).
13. 김준규, 신혁철. 병렬코퍼스자동구축체계에서영-조병렬코퍼스러과의한가지방법/김일서종합대학학보. 주체110(2021)년. 제67권. 제1호. 55–58쪽= Ким Чунгю, Син Хёкчхоль. Один из способов автоматического построения англо-корейского параллельного корпуса в системе автоматического построения параллельных корпусов // Вестник Университета имени Ким Ир Сена, 2021. Т. 67. № 1. С. 55–58. URL: <http://www.gyongnamsan.edu.kp/univ/plugins/pdfviewer/web/viewer.html?file=618faa1728eb2ef6e3733645273ab145koj> (дата обращения: 26.04.2023).
14. 꼬꼬마세종말뭉치활용시스템 // 꼬꼬마= Система использования корпуса Маленький Седжон// KKMA. URL: <http://kkma.snu.ac.kr/#/> (дата обращения: 18.11.2022).
15. 남북한어번역기 // 글동무= Северо- и южнокорейский переводчик // Кьльтонму. URL: <https://www.geuldongmu.org/> (дата обращения: 18.11.2022).
16. 리혁철. 조선어입말체본문코퍼스생성을위한글말체와입말체토사이의대응결정방법/김일서종합대학학보. 주체 111(2022)년제68권. 제2호. 3–7 쪽= Ри Хёкчхоль. Способ определения соответствия между формальным и неформальным стилем для создания корейского корпуса с неформальным разговорным стилем // Вестник Университета имени Ким Ир Сена, 2022. Т. 68. № 2. С. 3–7. URL: <http://www.gyongnamsan.edu.kp/univ/plugins/pdfviewer/web/viewer.html?file=bb4abc56ac2093f48c7c26980ec4a4c0koj> (дата обращения: 26.04.2023).
17. 리혁철. 조선어자연발화음성인식에서토의생략과어순바꿈을모방한입말체본문코퍼스의생성방법. 김일

