МРНТИ 16.21

Мадиева Г.Б.[1], Мансурова М.Е.[2],
Аубакиров С.С.[3], Ермеков Ж.С.[4],

[1]д. ф. н. профессор, [2]к. физ.-мат. н., и.о. профессора, [3]докторант PhD, [4]студент 4 курса
Казахского национального университета имени аль-Фараби, г. Алматы, Казахстан,
e-mail: gbmadiyeva.kz@gmail.com

К ВОПРОСУ О ПОДГОТОВКЕ МЕДИАКОРПУСА КАЗАХСКОГО ЯЗЫКА

Особое место в современной корпусной лингвистике занимают медиа-корпусы. В базу медиа-текстов включены на основе приема сплошной выборки новостные тексты, опубликованные в средствах массовой информации. Безусловно, медиа-корпус является весьма ценным источником по сбору, анализу какой-либо новостной информации для широкого круга потребителей, которые могут задавать поиск по различным основаниям. Он может быть и обучающим инструментом для будущих специалистов-журналистов, обозревателей, политиков, специалистов любой медиа-сферы.

Целью данной работы является разработка медиа-корпуса казахского языка на платформе Казахского национального университета имени аль-Фараби. На настоящий момент фактические данные для медиа-корпуса собираются с 44 казахоязычных сайтов, из них 10 порталов по чрезвычайным ситуациям, 11 новостных порталов, 13 образовательных порталов, 10 развлекательных ресурсов. Разрабатываемый авторами медиа-корпус казахского языка будет представлять собой публичный веб-портал, который станет новым инструментом для исследования, анализа, изучения, преподавания казахского языка, предназначенный для широкого круга потребителей на отечественной и мировой арене.

Ключевые слова: корпус, корпусная лингвистика, национальный корпус, национальный корпус казахского языка, медиа-корпус, медиа-тексты, медиа-сфера, новостной портал.

Madieva G.B., Mansurova M.E., Aubakirov S.S., Ermekov Zh.S.

To the question of preparation of media-corpus of the kazakh language

A special place in modern corpus linguistics is occupied by media corpora. The media texts are based on the reception of a continuous sample of news texts published in the media. Undoubtedly, the media corpus is a very valuable source for collecting, analyzing any news information for a wide range of consumers who can search on different grounds. It can also be a training tool for future journalists, observers, politicians, experts of any media sphere. The purpose of this work is to develop a media corpus of the Kazakh language on the platform of the Kazakh National University named after al-Farabi. At the moment, actual data for media corpora are collected from 44 Kazakh-language sites, including 10 emergency portals, 11 news portals, 13 educational portals, and 10 entertainment resources. The media corpus of the Kazakh language, developed by the authors, will be a public web portal, which will become a new tool for researching, analyzing, studying, teaching the Kazakh language, intended for a wide range of consumers in the domestic and world arena.

Key word: corpus linguistics, national corpus, national corpus of the Kazakh language, media corpora, media texts, media sphere, news portal.

Мадиева Г.Б., Мансурова М.Е., Аубакиров С.С., Ермеков Ж.С.

Қазақ тілі медиа-корпусты даярлау мәселесі

Қазіргі корпустық лингвистикада медиа-корпус ерекше орынды алады. Медиа-мәтіндер деректеріне бұқаралық ақпарат құралдарында жарияланған жаңалық мәтіндері жаппай іріктеу тәсілдері негізінде қамтылды. Сөзсіз, медиа-корпус әртүрлі негізге бағытталған сұрауға іздеу арқылы кең ауқымдағы тұтынушылар үшін жаңалық ақпараттарын талдауға, жинақтауға өте құнды дереккөз болып табылады. Ол болашақ медиа-саласының кез-келген мамандарына, саясаткерлерге, шолушыларға, журналист-мамандарға оқу құралы бола алады. Аталмыш жұмыстың мақсаты әл-Фараби атындағы Қазақ ұлттық университеті платформасында қазақ тілінің медиа-корпусын жасау. Қазіргі уақытта медиа-корпусқа нақты деректер 44 қазақ тілді сайттарынан, оның ішінде 10 портал төтенше жағдайлар бойынша, 11 жаңалық порталы, 13 білім беру порталы, 10 ойын-сауық ресурстарынан жинақталуда. Қазақ тілінің медиа-корпусын құрастырушы авторлары отандық және әлемдік аренадағы кең ауқымды тұтынушыларына қазақ тілін оқытуда, меңгертуде, талдауда, зерттеуге арналған жаңа құрал болып табылатын жалпыхалықтық веб-порталды ұсынады.

Түйін сөздер: корпус, корпустік лингвистика, ұлттық корпус, қазақ тілі ұлттық корпусы, медиа-корпус, медиа-мәтіндер, медиа-аясы, ақпараттық портал.

Введение

Национальный корпус языка – ценный инструмент, который позволяет за минуты найти необходимую справочную информацию и тем самым существенно сократить затраты на техническую работу по изучению различных языковых явлений. В свою очередь, корпус казахского языка – это современный инновационный инструмент, справочно-информационная база по казахскому языку, позволяющая получать ответы на многие вопросы, возникающие как перед отечественным, так и зарубежным исследователем, студентом, потребителем, изучающим и исследующим казахский язык.

Особое место в современной корпусной лингвистике занимают медиа-корпусы. В базу медиа-текстов включены на основе приема сплошной выборки новостные тексты, опубликованные в средствах массовой информации. Безусловно, медиа-корпус является весьма ценным источником по сбору, анализу какой-либо новостной информации для широкого круга потребителей, которые могут задавать поиск по различным основаниям (ключевым словам, интересующим рубрикам, темам и т.п.). Он может быть и обучающим инструментом для будущих специалистов-журналистов, обозревателей, политиков, специалистов любой медиа-сферы.

Целью данной работы является разработка медиа-корпуса казахского языка на платформе Казахского национального университета имени аль-Фараби. На настоящий момент фактические данные для медиа-корпуса собираются с 44 казахоязычных сайтов, из них 10 порталов по чрезвычайным ситуациям, 11 новостных порталов, 13 образовательных порталов, 10 развлекательных ресурсов. Разрабатываемый авторами медиа-корпус казахского языка будет представлять собой публичный веб-портал, который станет новым инструментом для исследования, анализа, изучения, преподавания казахского языка, предназначенный для широкого круга потребителей на отечественной и мировой арене.

Развитие тюркской корпусной лингвистики

В рамках казахского языкознания и прикладной лингвистики в Казахстане исследование и разработка Национального корпуса казахского языка представляет особый интерес, что определяется недостаточной разработанностью проблематики в данной области. Несмотря на достижения в этой области (попытка составления корпуса с необходимой разметкой, наличие научных исследований в виде монографий, диссертаций, учебников всех стилей казахского языка, работы сопоставительного характера, анализирующие отличия разговорного и литературного языков, исследования отдельных его аспектов), границы исследований не выходят за рамки традиционного языкознания, что ограничивает исследовательские усилия по разработке корпуса или сводит их к механистическому выявлению лексических, фонетических и других отличий казахского языка.

Казахский язык,являясь агглютинативным языком, характеризуется, как известно, последовательным присоединением различных формообразующих суффиксов или окончаний, несущих грамматическое значение, к неизменяемому корню или основе, являющихся носите-

лями лексического значения. Априори, что порядок добавления аффиксов в агглютинативных языкахстрого определен. Например, в казахском языке для имен существительных к основе слова вначале добавляется суффикс и далее окончание множественного числа, затем притяжательное окончание, далее следует падежное окончание и последним окончание формы спряжения (добавляется только к одушевленным существительным) [1; 2].

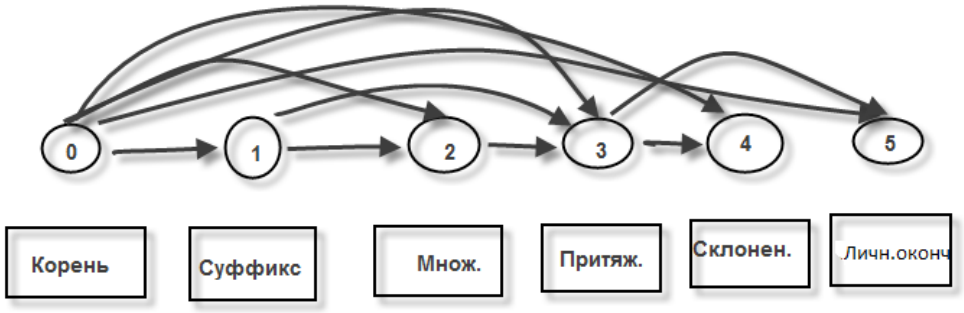

Рисунок 1 – Правило присоединения аффиксов для имен существительных

Тюркская корпусная лингвистика начала интенсивно развиваться лишь с 1990 годов, поэтому проекты создания общедоступных корпусов тюркских языков особенно актуальны. На сегодняшний день имеется небольшое количество репрезентативных корпусов текстов на тюркских языках, к которым относятся:

1) Турецкий национальный корпус объемом 50 миллионов словоупотреблений, который является сбалансированным и репрезентативным корпусом современного турецкого языка. Он состоит из образцов текстовых данных в широком разнообразии жанров, охватывающих период в 20 лет (1990-2009) [3].

2) Башкирский поэтический корпус объемом более 1,8 миллионов словоупотреблений. Он является вторым в мире поэтическим корпусом. Его особенность заключается в том, что корпус состоит из произведений башкирских поэтов XX и начала XXI века [4].

3) Письменный корпус татарского языка объемом более 116 миллионов словоупотреблений при числе различных словоформ – около 1,5 миллиона [5].

Опыт разработки корпусов тюркских языков положительно повлиял на разработку корпуса казахского языка. Однако проблема создания Национального корпуса казахского языка до сих пор остается актуальной. Из реально существующих и функционирующих разработок корпуса казахского языка следует назвать размещенный на портале государственного языка Комитета по языкам Министерства культуры и информации Республики Казахстан Корпус казахского языка [6], Корпус казахского языка, созданный силами сотрудников Национальной лаборатории Астана (NLA) Евразийского университета им. Л. Гумилева [7]. Следует отметить составленный на основе юридических текстов так называемый англо-казахский параллельный корпус, выполняемый Т.Е. Калдыбековым [8], а также Казахский национальный корпус [9], который является еще одной попыткой создания полноценного корпуса казахского языка и может считаться одним из первых корпусов, однако это небольшой по объему не имеющий аннотирования корпус, который также не относится к доступным открытым корпусам. Имеются попытки создать новый ресурс казахского языка с лингвистической аннотацией, это исследование ученых из Назарбаев Университета [10], скорее всего, это инструмент грамматических разметок, который также находится в стадии разработки; совместный проект Г. Алтынбек и W.Xiao-long, которые разработали корпус казахского языка в Xinjiang университете (2010). Однако информация о последнем корпусе, как отмечает Т.Е. Калдыбеков, отсутствует [8]. Близким к полифункциональному корпусу является Алматинский корпус казахского языка [11].

Архитектура медиа-корпуса казахского языка

Компонентная архитектура программного обеспечения – это парадигма программиро-

вания, существенно опирающаяся на понятие «компонент» [12]. Компонентная архитектура подразумевает создание системы, состоящей из компонентов в качестве многократно используемых узлов. Это позволяет отделять разработку отдельного компонента от разработки системы в целом, что дает возможность разрабатывать и поддерживать компоненты разным командам независимо друг от друга.

Информационная система, построенная на основе компонентной архитектуры, удовлетворяет следующим требованиям [13; 14]:

- Работает автономно, не требует оператора;
- Компоненты системы не зависят друг от друга;
- Все компоненты системы горизонтально масштабируются;
- Система устойчива к сбоям, перезагрузкам и отключениям.

При этом каждый компонент обладает следующими четырьмя свойствами: возможностью многократного использования; взаимозаменяемостью; расширяемостью; компонуемостью [11]. В силу приведенных доводов представленная в работе информационная система была построена на основе компонентной архитектуры.

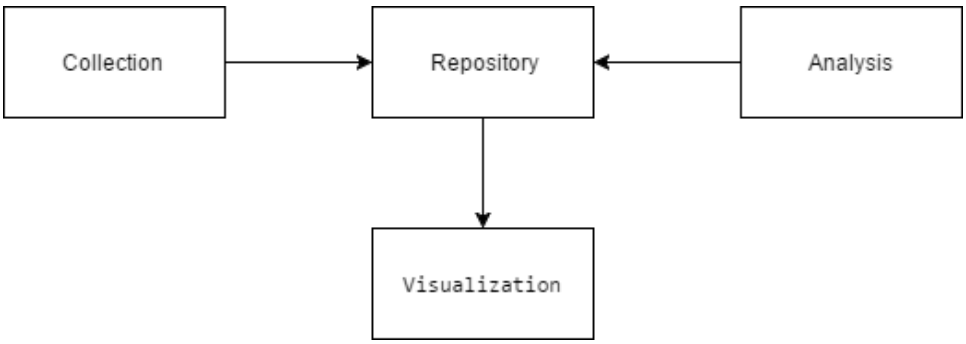

Рисунок 2 – Компонентная архитектура информационной системы

Для автоматизации сбора, хранения и анализа медиа-текстов на казахском языке была спроектирована и реализована информационная система. Данная система состоит из четырех компонентов:

1) компонент сбора информации,
2) компонент хранения данных,
3) компонент анализа данных,
4) компонент визуализации данных.

Компонентная архитектура информационной системы показана на рисунке 2. Использование очередей позволяет системе быть легко масштабируемой и устойчивой к сбоям.

Формат хранения данных

В медиа-корпусе для размеченных текстов применяется язык eXtensibleMarkupLanguage (XML). В компоненте сбора информации происходит извлечение релевантного текста из кода HTML страниц с использованием библиотеки Jsoap. Далее данные проходят обработку в компоненте анализа данных. Также на этом этапе проводится морфологический разбор текстов на казахском языке. Морфологический анализатор получает на вход простой текст, а на выходе отдаёт текст в формате XML, с которым в дальнейшем удобно работать, к примеру, легко преобразовать в JSON формат. Формат XML определен при помощи XML SchemaDefinition (XSD). XSD позволяет эффективно конвертировать данные в любой другой формат, что упрощает обмен данными между системами.

Морфологическая разметка и пост обработка данных

Корпус содержит особую разметку, представляющую собой дополнительную информацию о свойствах входящих в него текстов. Разметка – главная характеристика корпуса; она отличает корпус от простых коллекций (или «библиотек») текстов. Чем богаче и разнообразнее разметка, тем выше научная ценность корпуса [15].

Для постобработки слов в случае неполной морфологической разметки и наличия омонимии разработан специальный интерфейс, с помощью которого эксперт-лингвист может выбрать пра-

вильный вариант разбора, или выполнить полную ручную разметку для конкретного слова.

Заключение

Разработанный медиа-корпус казахского языка позволит:

- предоставлять открытый доступ всем желающим;
- осуществлять поиск по морфологическим параметрам;
- Использовать корпус для решения задач NaturalLanguageProcessing;
- проводить частотный анализ текстов;
- осуществлять обучение языку, используя переводы слов.

Данная работа выполнена при частичной поддержке МОН РК (проект ГФ4/5033 «Разработка интеллектуальной высокопроизводительной информационно-аналитической поисковой системы обработки слабоструктурированных данных», 2015-2017).

Литература

1 Бекманова Г.Т. Соме приближается к проблемам автоматических изменений слова и морфологического анализа на казахском языке// Бюллетень Восточно-Казахстанского государственного технического университета им. Д. Серикбаева, №1, 2009. – С. 192-197.

2 Койбагаров К.Ч., Мусабаев Р.Р., Калимолдаев М.Н. Разработка лингвистического процессора текстов на казахском языке // Проблемы информатики. 2014. № 3. – С. 64-72.

3 Национальный корпус турецкого языка, http://www.tnc.org.tr/index.php/en/, дата обращения 2017/03/03.

4 Башкирский поэтический корпус, http://web-corpora.net/bashcorpus/search/?interface_language=ru, дата обращения 2017/03/03. 2017/03/03.

5 Письменный корпус татарского языка, http://corpus.tatar/, дата обращения 2017/03/03.

6 Portal of the state language of the Committee on languages of the Ministry of culture and information of the Republic of Kazakhstan, http://til.gov.kz/wps/portal/!ut/p/, датаобращения 2017/03/03.

7 Корпус казахского языка, созданный силами сотрудников Национальной лаборатории Астана Евразийского университета им. Л.Гумилева //http: //kazcorpus.kz/klcweb/en/, lastaccessed 2017/03/03.

8 Калдыбеков Т.Е. Англо-Казахский параллельный корпус для статистического машинного перевода // Молодой ученый. – 2014. – №6. – С. 92-95.

9 Қазақстан Республикасы Мемлекеттік Тіл порталы, http://dawhois.com/www/til.gov.kz.html, дата обращения 2017/03/03.

10 Макажанов О.А., Махамбетов О.Е. и др. Разработка морфологического, синтаксического и лексического наборов меток для грамматической разметки текстов на казахском языке // Филология и культура. Philologyandculture. – Казань: Казанский университет, 2014. – № 2 (36). – С. 37-39.

11 Алматинский корпус казахского языка, http://web-corpora.net/KazakhCorpus/search/?interface_language=ru, дата обращения 2017/03/03.

12 C. Szyperski: Component Software: Beyond Object Oriented Programming, Addison-Wesley Professional, 1997.

13 Aubakirov S.S., Akhmed-ZakiD.Zh., Trigo P.S., News Classification using Apache Lucene, KazNU Bulletin Mathematics, Mechanics, Computer Science Series, #3(91), pp. 59-65 Almaty (2016)

14 Len Bass, Paul Clements, Rick Kazman. Software Architecture in Practice (SEI Series in Software Engineering) Addison Wesley; 3rd edition (2012)

15 Национальный корпус русского языка, http://ruscorpora.ru/corpora-intro.html, дата обращения 2017/03/03.

References

1 Bekmanova G.T. Some priblizhaetsja k problemam avtomaticheskih izmenenij slova i morfologicheskogo analiza na kazahskom jazyke// Bjulleten’ Vostochno-Kazahstanskogo gosudarstvennogo tehnicheskogo universiteta im. D. Serikbaeva, ¹1, 2009. – S. 192-197.

2 Kojbagarov K.Ch., Musabaev R.R., Kalimoldaev M.N. Razrabotka lingvisticheskogo processora tekstov na kazahskom jazyke // Problemy informatiki. 2014. ¹ 3. – S. 64-72.

3 Nacional’nyj korpus tureckogo jazyka, http://www.tnc.org.tr/index.php/en/, data obrashhenija 2017/03/03.

4 Bashkirskij pojeticheskij korpus, http://web-corpora.net/bashcorpus/search/?interface_language=ru, data obrashhenija 2017/03/03. 2017/03/03.

5 Pis’mennyj korpus tatarskogo jazyka, http://corpus.tatar/, data obrashhenija 2017/03/03.

6 Portal of the state language of the Committee on languages of the Ministry of culture and information of the Republic of Kazakhstan, http://til.gov.kz/wps/portal/!ut/p/, dataobrashhenija 2017/03/03.

7 Korpus kazahskogo jazyka, sozdannyj silami sotrudnikov Nacional’noj laboratorii Astana Evrazijskogo universiteta im. L.Gumileva //http: //kazcorpus.kz/klcweb/en/, lastaccessed 2017/03/03.

8 Kaldybekov T.E. Anglo-Kazahskij parallel'nyj korpus dlja statisticheskogo mashinnogo perevoda // Molodoj uchenyj. – 2014. – ¹6. – S. 92-95.

9 Қазақстан Respublikasy Memlekett³k T³l portaly, http://dawhois.com/www/til.gov.kz.html, data obrashhenija 2017/03/03.

10 Makazhanov O.A., Mahambetov O.E. i dr. Razrabotka morfologicheskogo, sintaksicheskogo i leksicheskogo naborov metok dlja grammaticheskoj razmetki tekstov na kazahskom jazyke // Filologija i kul'tura. Philologyandculture. – Kazan': Kazanskij universitet, 2014. – ¹ 2 (36). – S. 37-39.

11 Almatinskij korpus kazahskogo jazyka, http://web-corpora.net/KazakhCorpus/search/?interface_language=ru, data obrashhenija 2017/03/03.

12 S. Szyperski: Component Software: Beyond Object Oriented Programming, Addison-Wesley Professional, 1997.

13 Aubakirov S.S., Akhmed-ZakiD.Zh., Trigo P.S., News Classification using Apache Lucene, KazNU Bulletin Mathematics, Mechanics, Computer Science Series, #3(91), pp. 59-65 Almaty (2016)

14 Len Bass, Paul Clements, Rick Kazman. Software Architecture in Practice (SEI Series in Software Engineering) Addison Wesley; 3rd edition (2012)

15 Nacional'nyj korpus russkogo jazyka, http://ruscorpora.ru/corpora-intro.html, data obrashhenija 2017/03/03.