ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

МРНТИ 81.93.29

Н. Хаирова¹, А. Колесник², О. Мамырбаев³, К. Мухсина³

¹Национальный технический университет «Харьковский политехнический институт»,
Харьков, Украина

²Институт Информационных и вычислительных Технологий, г. Алматы, Казахстан

³Казахский национальный университет им. аль-Фараби, г. Алматы, Казахстан
khairova@kpi.kharkov.ua, kolesniknastya20@gmail.com, {morkenj, kuka_ai}@mail.ru

ВЫРОВНЕННЫЙ КАЗАХСКО-РУССКИЙ ПАРАЛЛЕЛЬНЫЙ КОРПУС, ОРИЕНТИРОВАННЫЙ НА КРИМИНАЛЬНУЮ ТЕМАТИКУ

Аннотация. В настоящее время разработка высококачественных параллельных текстовых корпусов является одним из наиболее актуальных и передовых направлений современной лингвистики. Особое внимание уделяется созданию параллельных многоязычных корпусов для языков с низким уровнем ресурсов, таких как казахский язык. В ходе исследования мы исследовали тексты с четырех казахских двуязычных новостных сайтов и создали параллельный казахско-русский корпус текстов, в основе которых лежит криминальная тематика. Для выравнивания корпуса мы использовали набор лексических соответствий и значения POS-тегов обоих языков. 60% наших корпусных предложений автоматически выровнены правильно. Наконец, мы проанализировали факторы, влияющие на процент ошибок.

Ключевые слова: криминальная тематика, POS-тегирование, казахско-русский параллельный корпус, лексические соответствия.

Введение

На сегодняшний день лингвистические ресурсы являются не только частью любого лингвистического исследования, но и важной основой для разработки любых приложений NLP. Такие ресурсы обычно содержат словари, тезаурусы, лингвистические онтологии, одноязычные и многоязычные корпуса. Для создания таких лингвистических ресурсов проводятся лексикографические исследования, анализ лексической структуры языков, изучение характеристик текста и аналогичные исследования.

Проектирование и создание, разработка и использование высококачественных текстовых корпусов являются одним из наиболее актуальных и передовых направлений современной лингвистики [1]. Такая обработка и систематизация с помощью Concordancer корпуса позволяет хранить большой объем текстовой информации, необходимой для статистического анализа языковых явлений и диахронических изменений в устной и письменной речи.

Существует множество типов корпусов. Есть специализированные корпуса (категория, время, место), общие корпуса, многоязычные корпуса, корпуса учащихся, исторические или диахронные корпуса, корпуса мониторов и многоязычные корпуса. Многоязычные корпуса, в свою очередь, делятся на сравнительные (comparable corpus) и параллельные или корпус переводов (translation corpus). На наш взгляд, параллельные текстовые корпуса особенно важны при изучении языка и особенностей перевода, различного синтаксического анализа, задач распознавания речи и др. Например, в задачах обучения иностранному языку такие корпуса позволяют находить возможные эквиваленты анализируемой лексики, отслеживать ее значения и функции в некоторых

контекстах. Кроме того, понятие параллельного корпуса является неотъемлемой частью более широкого и более сложного понятия, такого как – машинный перевод. Известно, что машинный перевод до сих пор остается нерешенной задачей компьютерной лингвистики, несмотря на бурный рост различных программных и эмпирических ресурсов. В некоторых случаях качество машинного перевода также зависит от количества параллельных предложений, используемых в обучении.

За последнее десятилетие в мире было создано множество двуязычных и многоязычных корпусов, среди которых, на наш взгляд, наиболее интересными являются:

- EUROPARL – 20.000.000 словоупотреблений, открытый корпус Европейского парламента на 11 языках (автор - Филипп Козн)
- CHEMNITZкорпус– 1.000.000 словоупотреблений (тексты – академическая проза, политика, туризм)
- KACENKA корпус – 3.000.000 словоупотреблений
- OPUS корпус – (5 языков).
- English-French Canadian Hansard [2];

В то же время, несмотря на существование большого количества параллельных многоязычных корпусов, для языков с ограниченными ресурсами, таких как казахский язык, задача создания параллельных корпусов является жизненно важной. Задача усложняется, когда мы говорим о разработке параллельных корпусов для не похожих языков, языков из разных семейств. Например, один язык принадлежит к тюркской языковой семье, а другой относится к индоевропейской языковой семье, как казахский и русский.

В нашем исследовании мы изучили тексты на двух языках (на русском и казахском) с казахских двуязычных новостных сайтов и создали параллельный казахско-русский корпус на основе текстов этих сайтов. Более того, тексты нашего параллельного корпуса не относятся к художественной литературе или другой широкой теме; они сосредоточены на уголовном предмете, который делает их ограниченными. Таким образом, мы смогли применить метод словаря для выравнивания корпуса. Кроме того, для повышения качества выравнивания предложений мы сделали POS-тегирование текстов на обоих языках, а затем использовали маркировку. Наконец, мы вычисляем процент правильных выровненных предложений и анализируем факторы, влияющие на процент ошибок.

1 Смежные направления исследований

Параллельные корпуса содержат текст оригинала и его перевод на какой-либо другой язык. Кроме того, эти два текста не просто противопоставлены друг другу, они должны быть выровнены: отдельные фрагменты исходного текста должны совпадать с соответствующими фрагментами перевода. Можно сказать, что параллельный корпус полезен только тогда, когда он выровнен. В большинстве исследований два уровня выравнивания явно или неявно различаются: выравнивание предложений и лексическое выравнивание. Как правило, задача автоматического сравнения предложений или слов в одном тексте с их эквивалентами в переводе является очень трудоемкой, так как эта согласованность между словами или предложениями иногда не является "один к одному". Например, несколько абзацев на исходном языке могут соответствовать одному абзацу на выходном языке, при переводе некоторые слова могут быть удалены или заменены очень отдаленными синонимами или фиксированными фразами, которые могут быть абсолютно разными на разных языках и т. д.

Мы можем классифицировать методы выравнивания предложений на три категории. Методы первой категории основаны на использовании длин предложений и абзацев [2]. Этот подход использует гипотезу в том, что длина предложения в оригинале и в переводе приблизительно совпадает. Вторая группа методов использует лексическую информацию,

которую можно получить из корпуса [3], [4]. К сожалению, эти методы применяются крайне редко, что связано с недоступностью двуязычных словарей и трудностями автоматического морфологического анализа для взаимной идентификации слов в словарях. На сегодняшний день в большинстве приложений, основанных на этой группе методов, используются только тексты специализированных предметов, например, тексты парламентов и юридические тексты. Использование методов словаря для художественных текстов встречается редко, поскольку даже в подобном жанре наблюдается высокий процент неоднозначности лексики в сравниваемых предложениях.

Третья группа алгоритмов выравнивания текстов в параллельных корпусах основана на POS-тегах, которые содержатся в аннотированном корпусе или используют орфографическое сходство [5].

Однако использование любого метода этих групп имеет ряд неточностей и недостатков. Соответственно, в настоящее время постоянно растет интерес к разработке систем, в которых применяется сочетание всех подходов. Например, Варга и др. в работе [6] описали гибридный метод параллельного выравнивания текста. Они использовали метод выравнивания, который сочетает в себе метод на основе длины с некоторым подобием на основе перевода. Основу исследования составили венгерский, румынский и словенский языки. Vondricka [7] сделал обзор специального приложения InterText для выравнивания параллельных корпусов, основанного на некоторых гибридных методах выравнивания. Этот источник существует в двух формах: InterText Server (сервер, основанный на системе управления текстом с веб-интерфейсом редактора) и InterText editor (приложение для персонального компьютера). Оба являются программами с открытым исходным кодом. Это же приложение использовалось для создания казахско-английских текстовых корпусов в исследовании Жуманова и соавт. [8]. Кроме того, авторы использовали инструменты Bitextor и halign для сканирования сайтов, содержащих одинаковые тексты на нескольких языках, и их выравнивания. Статья Рахимовой и соавт. [9] посвящена решению аналогичных задач. Они рассмотрели принципы использования такого приложения, как Bitextor, которое генерирует память переводов, используя многоязычные веб-сайты в качестве источника корпусов. Он загружает весь веб-сайт и применяет набор эвристик (основанных главным образом на структуре тегов HTML и длине текстового блока) для поиска параллельных текстов (bitexts).

Авторы статьи [10] выравнивали свой набор данных на уровне предложения, это означает, что они использовали строгую пунктуацию для сегментации текста. Но такой подход требует проверки правильности текста и близости языков. Поэтому требуется ручное управление. В результате все медицинские и все литературные тексты в польско-украинской паре были выровнены и проверены, в то время как только часть текстов на французском и английском языках все еще используется. В то же время финская и русская версии корпуса Arapea и подкорпус газетного корпуса русского национального корпуса и корпуса финской национальной библиотеки выровнены достаточно хорошо [11].

Создание выровненных параллельных корпусов сталкивается с дополнительной проблемой поиска текстовых ресурсов для параллельного корпуса. В настоящее время проводится много повторных поисков, связанных с получением параллельных предложений из непараллельных или сопоставимых данных. Например, такие лингвистические данные могут быть получены из Википедии. Это чрезвычайно ценный ресурс для извлечения параллельных предложений, поскольку выравнивание документов уже обеспечено, и статьи в Википедии на одну и ту же тему могут быть на разных языках. Кроме того, они связаны через ссылки “interwiki”, которые аннотируются пользователями [12]. Однако Википедия еще полностью не изучена [13].

Следовательно, можно сделать вывод, что проблема выравнивания параллельных корпусов до сих пор не решена до конца и не найден универсальный метод. Кроме того, на сегодняшний день в большинстве исследований считается, что выбор метода выравнивания зависит от исследуемой языковой пары, тематической направленности текстов и типов документов, представленных в корпусе [14].

2 Подготовка набора данных

Задача создания параллельного корпуса состоит из нескольких частей. Первой и одной из важнейших задач является сбор текстового материала для таких корпусов. Несмотря на то, что Интернет содержит большое количество сайтов, которые являются двуязычными и многоязычными, выбор необходимых двуязычных ресурсов является важной частью разработки параллельного корпуса. Эта задача усложняется тем, что мы обрабатываем такие разные языки, как казахский и русский. Кроме того, необходимо использовать специализированные инструменты и различные методики не только для обработки языка, но и для сбора необходимого материала для корпусов.

Мы считаем, что разбор веб-сайтов – это лучший способ автоматизировать процесс сбора и сохранения информации. Для нашего исследования мы разработали собственное специальное программное обеспечение для автоматического анализа веб-сайтов, которое позволяет анализировать веб-сайты, которые могут быть похожи по дизайну, содержанию и структуре.

Для разработанного парсера были выбраны четыре двуязычных сайта: zakon.kz, caravan.kz, lenta.kz, nur.kz. Выбранные сайты представляют собой известные и надежные новостные сайты Казахстана, которые могут переключать текстовую информацию сайта между этими двумя языками. Кроме того, веб-сайты содержат большое количество статей по криминальной информации, что является одной из целей нашего исследования.

В результате выполнения программы мы получили общий набор из 3000 текстов на двух языках: русском и казахском. Из них мы вручную отобрали тестовый набор для создания выровненного параллельного корпуса русско-казахских текстов по криминальной информации. Объем корпуса составляет более 50410 слов, из них около 24800 – на русском языке и около 25600 – на казахском.

На следующем этапе мы определили структуру организации корпуса. В настоящее время такая структура может быть очень разнообразной, в зависимости от прагматических целей ее создателей или пользователей:

- в виде традиционного текста со ссылкой на переводы,
- в табличной «зеркальной» форме, более удобной для восприятия и сравнения, в виде базы данных.

Для нашего исследования мы выбрали структуру базы данных, так как это наиболее удобный способ хранения большого объема данных с возможностью его дальнейшего увеличения.

Все новостные статьи хранятся в таблице базы данных, которая включает их идентификатор, название, адрес сайта и текст статьи. На рисунке 1 показан фрагмент общей таблицы текстов на двух языках.

На следующем этапе мы провели POS-тегирование корпуса. Для маркировки русского корпуса мы выбрали пакет Python `py morphology2`, специально разработанный для морфологического анализа русских и украинских текстов. Библиотеки пакета используют словарь `OpenCorpora` и делают гипотетические выводы для нераспознанных слов.

id	head	url	text
1334	2040 жылға қара	https://www.inform.kz/	АСТАНА. ҚазАқпарат - Орташа өмір сү
1335	Керчьтегі колле	https://www.inform.kz/	КЕРЧЬ. ҚазАқпарат - Керчь қаласындағ
1336	Еуроодақ халқын	https://www.inform.kz/	БРЮССЕЛЬ. ҚазАқпарат - Еуропа одағы
1337	Армения үкіметі	https://www.inform.kz/	ЕРЕВАН. ҚазАқпарат - Армения премье
1338	Мароккода пойы	https://www.inform.kz/	АСТАНА. ҚазАқпарат - Марокко астана
1339	Техник байқама	https://www.inform.kz/	АСТАНА. ҚазАқпарат - Бельгия әскери-
1340	Кельнде адамда	https://www.inform.kz/	АСТАНА. ҚазАқпарат - Германияның К
1341	Иракта ауыз суд	https://www.inform.kz/	АСТАНА. ҚазАқпарат - Ирактың Басра
1342	АҚШ-та жол ап	https://www.inform.kz/	АСТАНА. ҚазАқпарат - 14 қазан күні Ло
1343	Түркияда жүк кө	https://www.inform.kz/	АНКАРА. ҚазАқпарат - Түркияның баты
1344	Индонезияда су	https://www.inform.kz/	АСТАНА. ҚазАқпарат - Индонезия құтқа
1345	Мужчина погиб	https://patrul.kz/novos	ДТП со смертельным исходом произо
1346	В Алматы в резу	https://patrul.kz/novos	ДТП с пострадавшими произошло в ю
1347	Женщина погиб	https://patrul.kz/novos	В аварии, произошедшей на трассе «Ал
1348	Молодой водите	https://patrul.kz/novos	Виновником ДТП в южной столице ока
1349	Полицейские СК	https://patrul.kz/novos	В ходе профилактических рейдов сотру

Рисунок1 Фрагмент базы данных с текстами на двух языках

В свою очередь, сложность, структурная и типологическая характеристика казахской маркировки связана с тем, что она относится к агглютинирующим языкам. Структура этого языка довольно сложна и необычна, если ваш родной язык является флективным. Агглютинативное образование противоположно флективному, где каждый формант имеет сразу несколько неразделимых значений (например, падеж, род, число и т.). По этой причине мы делаем POS-тегирование казахских текстов с помощью теггера регулярных выражений, основанного на классе `RegexTagger` пакета `nlk Python`. Например, мы можем определить некоторые типы существительных в казахских текстах с помощью следующего списка регулярных выражений:

```
patterns=[(r'.*бен$', 'NN'), (r'.* пенен$', 'NN'), (r'.* басшылық$', 'NN'), (r'.* іпқону$', 'NN'), (r'.* тармен$', 'NN'), (r'.* герлермен$', 'NN'), (r'.* здар$', 'NN')].
```

Кроме того, для повышения отзыва и точности наших POS-тегирования казахских текстов мы объединяем регулярные выражения с системой правил. Например, "Если слово следует за словами из специального списка-слово обозначается как глагол".

Автоматическое выравнивание корпуса

На первом этапе автоматического выравнивания нашего корпуса мы руководствовались знаками препинания, заглавными буквами и абзацами. На следующем этапе можно выбрать два основных подхода к выравниванию предложений. Первый подход, который обеспечивает значительно более высокую производительность, основан на длине предложения. Во втором, более ресурсоемком подходе, лексические соответствия задаются методом выравнивания слов. При детальном изучении данной области было выявлено, что для языков, принадлежащих к разным языковым группам, а в дальнейшем и для специализированных, тематических текстов лучше всего применять словарный метод выравнивания. На основе этого вывода мы используем набор лексических соответствий и значения POS-тегирования, полученные на предыдущих этапах подготовки. Основной причиной, по которой мы не смогли использовать первый более простой подход к выравниванию предложений, является огромная разница между синтаксисом и семантикой казахского и русского языков.

В качестве набора лексических соответствий мы используем собственный казахско-русский словарь, основанный на англо-казахско-русском словаре, который содержит около 50 000 записей. На рисунке 2 показан фрагмент англо-казахско-русского словаря, который мы используем в качестве фонового. Словарь содержит около 50 000 записей. На

рисунке 3 показан фрагмент казахско-русского словаря, который мы применяем для выравнивания параллельного казахско-русского корпуса.

```
native###_туасынан, ·жаратылысынан, ·тумысынан###_прирождённый·
(прирожденный)¶
native###_табиғи, ·жаратынды###_природный¶
native###_тұнық, ·ұқыпты, ·кірсіз, ·пәк, ·мәлдір, ·кіршіксіз, ·бейкүнә, ·
әйнектей, ·ақ, ·мұнтаздай, ·таза, ·шайдай·ашық, ·мәлдір·бұлақ, ·дақсыз, ·саф, ·
ашық, ·адал, ·шын, ·нағыз, ·айна·дай, ·бәкізе, ·ғилман, ·жазықсыз###_чистый¶
native###_жергілікті·адам, ·туып-өскен, ·туған###_уроженец¶
native###_туған, ·бір·туған, ·тіған-туыскандар, ·туыстық, ·туып-өскен, ·
қарағым, ·шырағым###_родной¶
native·of·a·country·(·aborigine·)###_абориген###_абориген¶
nativity·(the·Nativity)###_рождество###_рождество¶
nativity###_дүниеге·келу, ·жаралу###_рождение¶
north·atlantic·treaty·organization·(NATO)###_°###_североатлантический·
союз·(НАТО)¶
North·Atlantic·Treaty·Organization·(NATO)###_°###_Североатлантический·
союз·(НАТО)¶
nato·(NATO;·North·Atlantic·Treaty·Organization)###_°###_нато·(НАТО;·
```

Рисунок2 Фрагмент используемого базового англо-казахско-русского словаря

kz	ru
қылмыстық	криминал
мәйіті	трупы
мәліметтерге	данные
министрлігімен	с министерством
Оңтүстік	юг
орындарын	мест
өзара	взаимное
өзінің	его собственный
пиғылды	недобросовестный
Рудныйда	в Рудном
сату	продажа
сәттері	моменты

Рисунок 3 Фрагмент базы данных хранилища текстов на двух языках

Для улучшения выравнивания предложений, помимо словарного метода, мы используем знания о POS-тегировании слов в предложениях. Такой подход позволит улучшить результаты словарного метода, так как казахские слова имеют несколько вариантов перевода. Благодаря правильной разметке слов в текстах можно улучшить лучший переводной эквивалент.

Эксперименты и результаты

Наш параллельный казахско-русский корпус содержит тексты с некоторых казахстанских новостных сайтов за период 2018-2019 гг. В состав корпуса входит более 50410 слов, из них около 24800 на русском и около 25600 слов на казахском языке.

Для оценки правильности выровненных предложений мы используем мнение экспертов, которые являются носителями как казахского, так и русского языков.

Хорошо разработанное специальное приложение позволяет экспертам выбирать текст на любом (русском или казахском) языке и автоматически загружать параллельный текстовый файл. При работе с корпусом эксперт может отмечать тексты, сохранять их с маркировкой и выравнивать их вручную. На рисунке 4 показан пользовательский интерфейс нашего специального универсального приложения для работы с выровненными параллельными корпусами.

В результате оценки по меньшей мере трех экспертов было установлено, что около 60% предложений в нашем параллельном казахско-русском корпусе автоматически правильно выровнены. Остальные предложения должны быть выровнены вручную.

Рисунок 4 Пользовательский интерфейс нашего специального универсального приложения для работы с выровненными параллельными корпусами. Жирный шрифт применяется к тем предложениям, которые не имеют параллельных эквивалентов после автоматического выравнивания.

На наш взгляд, такой процент связан со следующими факторами.

1. Не полное совпадение по количеству предложений в корпусах. В связи со сложностью синтаксических структур казахского языка некоторые предложения не соответствуют по структуре русским эквивалентам и делятся на несколько предложений.

2. Сложность создания словарной базы. В основе словарного метода лежит качественно разработанный словарь. Поскольку русский и казахский языки находятся в удаленных языковых группах, при создании такого словаря возникают трудности с точным переводом.

3. Сложность и ограниченность использования сравнительной грамматики для казахского и русского языков в нашем исследовании. Анализ и дальнейшее развитие этого подхода позволит улучшить результат выравнивания.

4. Словарный метод не учитывает имена собственные. Тексты на криминальную тематику содержат большое количество таких слов, особенно в заголовках.

Все эти несоответствия могут существенно повлиять на результаты выравнивания.

Выводы и дальнейшая работа

Параллельные корпуса используются для решения различных задач, таких как разработка и настройка систем машинного перевода, сравнительное изучение языков, языковое обучение. Разработка таких корпусов особенно важна для языков с низким уровнем ресурсов и для пар языков, относящихся к разным группам, например, для русского и казахского языков.

Параллельный казахстанско-русский корпус разработан на основе четырех многоязычных новостных сайтов Казахстана, из которых была отобрана специальная криминальная информация. Корпус содержит 50410 слов, из которых 25600 относится к казахскому языку, а 24800 - к русскому.

Корпус выровнен с использованием специально сконфигурированного словаря и знанием POS-тегов обоих текстов. Экспертная оценка автоматического выравнивания составляет 60% правильно выровненных текстов. Кроме того, корпус снабжен специальным программным приложением, позволяющим добавлять специальную информацию в корпус. Разработанный выровненный казахско-русский параллельный корпус может быть использован в качестве обучающих данных для машинного перевода,

идентификации и извлечения текстов, связанных с преступностью, а также для различных задач NLP.

Следующим этапом нашего исследования является более активное вовлечение в этап информационного выравнивания с использованием POS-тегов, что в настоящее время ограничено сложностью такой полной маркировки для казахских текстов.

Благодарность

Это исследование поддержано Комитетом по науке Министерства образования и науки Республики Казахстан (проект № AP05131073 - Методы, модели поиска и анализа криминальной информации, содержащейся в полуструктурированных и неструктурированных текстовых массивах).

СПИСОК ЛИТЕРАТУРЫ

- [1] Rizun Nina, Waloszek Wojciech. Methodology for Text Classification using Manually Created Corpora-based Sentiment Dictionary. Proceedings of the 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2018) – Volume 1: KDIR, pages 212-220.
- [2] Gale W.A., Church K.W., 1993. A program for aligning sentences in bilingual corpora. Computational Linguistics, 19(1):75–102.
- [3] Kay M., Roscheisen M., 1993. Text translation alignment. Computational Linguistics, 19(1):121–142.
- [4] Fung P., McKeown K., 1994. Aligning noisy parallel corpora across language groups: word pair feature matching by dynamic time warping. Proceedings of the First Conference of the Association for Machine Translation in the Americas (AMTA-94), pages 81–88, Columbia, Maryland, USA.
- [5] Simard M., Foster G.F., Isabelle P., 1992. Using cognates to align sentences in bilingual corpora. Proceedings of the Fourth International conference on theoretical and methodological issues in Machine translation, pages 67–81, Montreal, Canada.
- [6] Daniel Varga, Peter Halacsy, Andras Kornai, Viktor Nagy, Laszlo Nemeth, Viktor Tron, 2007. Parallel corpora for medium density languages. In MREC, H-1111 Budapest MetaCarta Inc., U of Edinburgh.
- [7] Pavel Vondricka, 2014. Aligning parallel texts with InterText. In Institute of the Czech National Corpus, Charles University in Prague Faculty of Arts.
- [8] Zhandos Zhumanov, Aigerim Madiyeva, Diana Rakhimova, 2017. New Kazakh parallel text corpora with online access. Al Farabi Kazakh National University, Laboratory of Intelligent Information Systems, Almaty, Kazakhstan.
- [9] Diana Rakhimova, Zhandos Zhumanov, 2017. Complex Technology of Machine Translation Resources Extension for the Kazakh Language. Al Farabi Kazakh National University, Laboratory of Intelligent Information Systems, Almaty, Kazakhstan.
- [10] Natalia Grabar, Olga Kanishcheva, Thierry Hamon, 2018. Multilingual aligned corpus with Ukrainian as the target language. CNRS, Univ. Lille, National Technical University Kharkiv Polytechnical Institute, Kharkiv, Ukraine, Universite Paris-SaclayOrsay, France.
- [11] Juho Harme, 2018. Last year but not yesterday? Explaining differences in the locations of Finnish and Russian time adverbials using comparable corpora. University of Tampere, Finland.
- [12] Jason R. Smith, Chris Quirk and Kristina Toutanova, 2010. Extracting Parallel Sentences from Comparable Corpora using Document Level Alignment. In Center for Lang. and Speech Processing Johns Hopkins University and Microsoft Research.

[13] Lewoniewski, W., Węcel, K., Abramowicz, W. Quality and importance of Wikipedia articles in different languages. In International Conference on Information and Software Technologies. 2016, p. 613 – 624.

[14] Alexandr Rosen, 2005. In search of the best method for sentence alignment in parallel texts. In Computer treatment of Slavic and East European languages. Third international seminar. Bratislava, Slovakia, p. 174-185.

КРИМИНАЛДЫ ТАҚЫРЫПҚА БАҒЫТТАЛҒАН ҚАЗАҚ-ОРЫС ПАРАЛЛЕЛЬДІ КОРПУСЫ

Н. Хаирова¹, А. Колесник², О. Мамырбаев³, К. Мухсина³

¹Харьков политехникалық институты ұлттық техникалық университеті, Харьков, Украина

²Ақпараттық-есептеу технологиялары институты, Алматы қ., Қазақстан

³Атындағы Қазақ ұлттық университеті әл-Фараби, Алматы қ., Қазақстан,

khairova@kpi.kharkov.ua, kolesniknastya20@gmail.com, {morkenj, kuka_aiaisure@mail.ru

Аңдатпа. Қазіргі уақытта жоғары сапалы параллель мәтіндік кейстерді дамыту заманауи лингвистиканың өзекті және алдыңғы қатарлы бағыттарының бірі болып табылады. Қазақ тілі сияқты ресурстардың деңгейі төмен тілдер үшін параллельді көптілді денелерді құруға ерекше назар аударылады. Зерттеу барысында біз төрт қазақ тілді жаңалықтар сайтынан мәтіндерді зерттедік, қылмыстық тақырыпқа негізделген параллельді қазақ-орыс мәтіндерін құрдық. Корпусты туралау үшін лексикалық сәйкестіктер жиынтығын және екі тілдегі POS белгілерінің мағынасын пайдаланылды. Біздің корпустық ұсыныстарымыздың 60% автоматты түрде дұрыс тураланған. Соңында қателіктердің пайыздық деңгейіне әсер ететін факторларға талдау жасалды.

Түйін сөздер: қылмыстық тақырыптар, POS-белгілеу, қазақ-орыс параллель корпусы, лексикалық сәйкестік.

ALIGNED KAZAKH-RUSSIAN PARALLEL CORPUS, FOCUSED ON THE CRIME THEME

N. Khairova¹, A.Kolesnik², O. Mamyrbaev³, K.Mukhsina³

¹National Technical University "Kharkov Polytechnic Institute", Kharkov, Ukraine

²Institute of Information and Computing Technologies, Almaty, Kazakhstan

³Kazakh National University named after al-Farabi, Almaty, Kazakhstan

khairova@kpi.kharkov.ua, kolesniknastya20@gmail.com, {morkenj, kuka_aiaisure@mail.ru

Abstract. Currently, the development of high-quality parallel textual cases is one of the most relevant and advanced areas of modern linguistics. Particular attention is paid to the creation of parallel multilingual bodies for languages with a low level of resources, such as the Kazakh language. In the course of the study, we examined texts from four Kazakh bilingual news sites and created a parallel Kazakh-Russian corpus of texts based on criminal topics. To align the body, we used a set of lexical correspondences and the meaning of POS tags in both languages. 60% of our package offers are automatically aligned correctly. Finally, we analyzed the factors that influence the percentage of errors.

Keywords: criminal topics, POS-tagging, Kazakh-Russian parallel corpus, lexical correspondences.