Санкт-Петербургский государственный университет

ПЕТРУШЕНКО Лада Сергеевна

Выпускная квалификационная работа

Разработка алгоритма для выравнивания предложных конструкций в параллельном русско-чешском корпусе

Уровень подготовки: бакалавриат

Направление: 45.03.02 «Лингвистика»

Образовательная программа: СВ.5106 «Прикладная, компьютерная и математическая лингвистика (английский язык)»
Профиль: «Прикладная, компьютерная и математическая лингвистика»

Научный руководитель:
доцент, Кафедра
математической лингвистики,
Захаров Виктор Павлович

Рецензент:
заместитель заведующего кафедрой,
Карлов университет,
Rosen Alexandr

Санкт-Петербург

2021

Аннотация

В данной работе исследуется проблема выравнивания предложных конструкций в параллельном русско-чешском корпусе. Целью исследования является разработка алгоритма для извлечения предложных конструкций из русскоязычной части корпуса и поиска их переводных эквивалентов в чешском сегменте при помощи семантических, грамматических и синтаксических маркеров словоформ.

В работе рассмотрены теоретические вопросы, связанные с областью выравнивания текстов, особенности существующих подходов и сферы их применения. Помимо этого, изучены определения предлога и предложной конструкции, а также особенности русского и чешского языков, которые позволяют выявлять и выравнивать подобные конструкции.

В практической части исследования даны описания разработанного алгоритма и эксперимента по проведению процедуры выравнивания предложных конструкций. Методика проанализирована и оценена, высказаны предположения по дальнейшему развитию алгоритма.

Ключевые слова: автоматическое выравнивание, выравнивание предложных конструкций, параллельные корпусы, русско-чешский параллельный корпус.

This graduation qualification paper addresses the issue of aligning prepositional constructions in parallel Russian-Czech corpora. In the paper, we aim at developing an algorithm for extracting prepositional constructions in Russian and finding the correct equivalents for them in Czech on the basis of the given text corpus.

The paper describes the theoretical issues in the field of alignment along with the peculiarities of different existing approaches. Their scopes of application are touched upon. In the theoretical part of the paper, the definitions of the preposition and the prepositional construction are investigated. We overview various linguistic

aspects of Russian and Czech that allow us to extract and align prepositional constructions.

In the practical part of the paper, we give the description of our algorithm. We present the results of the experiment aimed at aligning prepositional constructions in the parallel Russian-Czech corpus. These results are analyzed and evaluated. Further possibilities for development are discussed.

Keywords: automatic alignment of corpora, automatic alignment of prepositional constructions, parallel text corpora, the Russian-Czech parallel corpus.

Оглавление

Введение

Процедура выравнивания, заключающаяся в установлении соответствия между фрагментами оригинального текста и его перевода, является важным предварительным этапом во многих задачах, связанных с корпусными исследованиями или использованием параллельных корпусов текстов. Выравнивание можно рассматривать как этап разметки, который необходим при составлении терминологических словарей, извлечении двуязычной лексики, построении поисковых систем и непосредственно при аннотировании параллельных текстов [Потемкин, Кедрова 2008]. Помимо этого, выравнивание необходимо в статистическом машинном переводе, где параллельные тексты используются для обучения моделей и проверки качества перевода, а также при нейронном переводе как потенциальный способ улучшения качества работы нейронных моделей.

От качества проведения данной процедуры зачастую зависит общий результат работы и успех в достижении поставленной цели. В связи с этим, наблюдается потребность во внедрении качественных инструментов для выравнивания в различные программы по обработке естественного языка. Однако создание подобных инструментов и сама задача выравнивания вызывают затруднения у исследователей и разработчиков: поскольку выравнивание является дополнительным этапом разметки, аннотирование параллельного корпуса сложнее, чем разметка одноязычного корпуса, где выравнивание не предусматривается и достаточно лингвистической информации и метаданных.

Существуют три основных подхода к выравниванию: эвристический, или основанный на правилах, статистический и работающий на нейронных сетях. Каждый из этих подходов можно рассматривать применительно и к родственным, и к далеким друг от друга языкам. В рамках каждого подхода разработаны инструменты, однако эффективность их применения

варьируется в зависимости от рассматриваемой языковой пары, жанра текста и от единицы, на уровне которой проводится процедура. Так, например, разработанные лингвистические правила могут быть применены для выравнивания одной языковой пары, но не могут быть применены для другой в силу особенностей языков и степени их родства. В то же время, для одной и той же языковой пары могут хорошо показать себя разные методики, и их эффективность будет зависеть от жанра произведений в изучаемом корпусе. Статистические модели и модели, работающие на нейронных сетях, обычно справляются только с частотными конструкциями, требуют большого количества текстовых данных для обучения и не всегда учитывают структурную разницу между языками. В тех случаях, когда у нас нет возможности использовать большой корпус текстов, нужен инструмент, который смог бы справиться и с собраниями текстов относительно малого объема.

Таким образом, **актуальность** рассматриваемой проблемы обусловлена тем, что параллельные корпусы необходимы во многих задачах обработки естественного языка — для машинного перевода и проверки качества перевода, составления терминологических баз данных, в информационном поиске, компаративных лингвистических исследованиях и т. д., — а качественные инструменты для выравнивания могут помочь при разметке подобных корпусов и улучшить качество выполнения работы. Современные инструменты, разработанные в рамках каждого из подходов — эвристического, статистического и нейронного — обладают своими недостатками, что дает возможность для их доработки. Для улучшения качества выравнивания важно уделять дополнительное внимание межъязыковым соответствиям и несоответствиям, которые проявляют себя в отдельных конструкциях и поэтому могут быть изучены на их примере. Именно по этой причине в нашей работе мы занимаемся проблемой выравнивания предложных конструкций.

Таким образом, **целью** данного исследования является разработка алгоритма для выравнивания двуязычного (русско-чешского) корпуса на уровне предложных конструкций. Для достижения поставленной цели нами были определены следующие **задачи**:

1. Рассмотреть понятие выравнивания, основные подходы к данной процедуре, их особенности и сферы применения;
2. Описать понятие предлога и обозначить границы термина «предложная конструкция»;
3. Выявить особенности русского и чешского языков и закономерности при выявлении предложных конструкций в них;
4. Разработать алгоритм для выравнивания предложных конструкций;
5. Оценить качество работы алгоритма на примере собранного нами чешско-русского корпуса и проанализировать результаты.

Объектом нашего исследования выступают предложные конструкции в параллельных русско-чешских корпусах текстов. **Предметом** исследования являются методы выявления и выравнивания предложных конструкций. **Материалом для исследования** и проверки алгоритма служит параллельный русско-чешский корпус, насчитывающий 200 предложений и 342 русскоязычных предложных конструкций, для которых были найдены соответствующие переводные эквиваленты. Корпус состоит из предложений из произведений художественного жанра, экономических и новостных сводок и субтитров.

Теоретическая значимость работы заключается в разработке лингвистически гибридного подхода для выравнивания предложных конструкций, опирающегося с одной стороны на морфологические и синтаксические маркеры словоформ, а с другой стороны — на степень семантической схожести между словами, выявленной при помощи предобученных выровненных векторных представлений слов (word

embeddings). **Практическая значимость** работы заключается в возможности применения разработанной методики на параллельных корпусах текстов для пополнения терминологических словарей и решения других лингвистических задач и в возможности доработки алгоритма с целью его использования при работе с другими парами родственных языков.

Данная работа состоит из введения, трех глав, списка литературы и трех приложений. В первой главе нами рассмотрены теоретические вопросы, связанные с сопоставлением сегментов параллельных корпусов, а также приведены современные методики для выполнения данной процедуры и изучены их особенности. Во второй главе мы уделяем особое внимание теоретическим обоснованиям определений «предлог» и «предложная конструкция», а также рассматриваем особенности русского и чешского языков, которые позволяют нам выявлять подобные конструкции в текстах. В третьей главе нами описана разработанная методика и приведены результаты эксперимента по применению созданного алгоритма на собранном параллельном русско-чешском корпусе.

Глава 1. Выравнивание в задачах обработки естественного языка

1.1. Понятие выравнивания и обобщенная формула

При выравнивании мы работаем с несколькими текстами (двумя или более), которые являются переводами друг друга. Обычно есть один первоисточник, оригинальный текст, который мы в данной работе обозначим как исходный текст или текст на языке оригинала, и один или несколько текстов, которые являются его переводами, что в данной работе мы будем называть текстами на целевом языке или текстами на языке перевода. В самом общем смысле выравнивание двуязычного корпуса подразумевает под собой сопоставление некоторого текста с его переводом с целью показать, как переводятся его отдельные сегменты в документе на целевом языке [Kay и Ruoscheisen 1993]. Данное определение может быть распространено и на многоязычные корпусы текстов.

Чаще всего выравнивание проводится на уровне отдельных абзацев, предложений, словосочетаний и словоформ. В результате процедуры исследователь получает упорядоченный набор пар языковых единиц в составе корпуса, где первый элемент пары — некоторая единица в исходном тексте, а второй — его переводной эквивалент, выявленный при анализе текста на языке перевода. Если лишь некоторые элементы исходной языковой цепочки могут быть обнаружены в подкорпусе на целевом языке, то такое выравнивание называется частичным («*partial*»). Если же языковые цепочки полностью соответствуют друг другу, такое выравнивание называется полным («*complete*») [Kay и Ruoscheisen 1993].

При выравнивании исследователь обычно имеет дело с корпусами большого объема. Разметка крупного собрания текстов не может быть проведена вручную вследствие своей ресурсозатратности. Параллельные

тексты выравниваются исключительно при помощи существующих или специально разработанных автоматических инструментов. В большинстве случаев, при сопоставлении единиц длиной меньше, чем целое предложение, изначально проводится выравнивание на уровне предложений, а уже затем исследователь приступает к более глубокому анализу, например, сопоставляет корпусы текстов на уровне отдельных конструкций или лексем.

Несмотря на разнообразие единиц, на уровне которых проводится изучаемая процедура, можно представить обобщенную формулу для выравнивания. Авторы F.J. Och и H. Ney (2003) рассматривают произвольную пару параллельных фрагментов текста из корпуса. Обозначим совокупность лексем, входящих в состав цепочки на исходном языке «*L1*» как $s_{L1}^{J} = w_{L1}^{1}, \ldots, w_{L1}^{j}$, где J — конечное число словоформ в исходной языковой цепочке, а множество лексем в составе цепочки на языке перевода «*L2*» как $s_{L2}^{I} = w_{L2}^{1}, \ldots, w_{L2}^{i}$, где I — конечное число словоформ в цепочке на целевом языке. Теперь нам необходимо определить выравнивание A для данных двух цепочек.

С учетом того, что процедура сопоставления параллельных корпусов обычно осложнена различиями в организации текстов (например, пропусками, вставками или перестановками лексем при переводе, а также возникновением соответствия «один ко многим» (one-to-many correspondance) между фрагментами двуязычного или многоязычного корпуса), выравнивание A можно обобщенно определить как подмножество Декартова произведения позиций слов и выразить в следующей формуле:

$$A \subseteq \{(j, i): j = 1 \ldots, J;\ i = 1, \ldots, I\} \quad (1)$$

Столь обобщенную модель сложно воссоздать на практике, в связи с чем в используемых алгоритмах для выравнивания обычно накладываются

некоторые ограничения, такие как установление взаимно однозначного соответствия между отдельными словами [Och, Ney 2003].

1.2. Проблемы при сопоставлении параллельных корпусов

Как мы уже упоминали ранее, задача выравнивания параллельных корпусов решается только при помощи автоматических инструментов. Хотя такой подход является эффективным с точки зрения скорости обработки данных, он также увеличивает вероятность возникновения ошибок.

Можно выделить три основные группы факторов, которые вызывают затруднения при выравнивании:

1. естественная структурная разница между языками;
2. реорганизация элементов текста при переводе;
3. выбор способов оформления языковых единиц в тексте.

Эти аспекты стоит учитывать при разработке и применении различных методик выравнивания.

Структурная разница между языками. С точки зрения лингвистики, сложность могут представлять следующие явления:

- *Различия в синтаксисе*. Языки могут сильно отличаться по этому аспекту, причем не только в том случае, если не являются родственными. Традиционным примером могут послужить отделяемые приставки у глаголов в немецком языке. Так, если мы используем такой глагол, как «*vorstellen*», приставка «*vor*» по правилам займет последнее место в простом предложении, в то время как «*stellen*» пойдет на второе место в нем. Если мы работаем с немецко-английским параллельным корпусом и найдем переводной эквивалент рассматриваемого предложения на английском языке, мы обнаружим, что глагольная форма во втором случае не будет разбита на отдельные составные

элементы, а это значит, что выравнивание не будет взаимно однозначным, а порядок слов не будет линейным.

- *Разница в словарном составе и способах лексического выражения понятия*. Примером этого явления могут послужить такие сложные случаи, как перевод идиоматических выражений и служебных слов [Och, Ney 2003], которые чаще всего выражают синтаксические связи между знаменательными словами и выражают скорее морфологическое значение. Помимо этого, если переводчик сталкивается с явлением лексической лакуны, он может заменить исходное краткое словесное выражение более крупной конструкцией на целевом языке (ср. нем. «*Geschwister*», русск. «*брат и сестра*» при переводе с немецкого на русский), и наоборот. Ничто в немецком не мешает нам лексически организовать такую единицу, как «*Bruder und Schwester*», однако «*Geschwister*» сильнее закреплена на уровне языка и появляется в корпусной статистике намного чаще. Такие случаи указывают нам на частое отсутствие взаимно однозначного соответствия между сегментами корпусов. На материале собранного нами корпуса, процесс составления которого мы описываем в разделе 3.2., мы также смогли зафиксировать подобные случаи. Так, например, нами было отмечено, что конструкция «*на улице*» чаще всего переводится на чешский как «*venku*» (дословно: «*снаружи*») при указании на место протекания того или иного процесса («*На улице темнело, на небе показались первые звезды*»: «*Venku se setmělo a na nebi se ukázaly první hvězdy*»). И даже хотя лексически есть возможность сказать «*na ulici*» или «*v ulici*», наиболее частотным вариантом все равно становится «*venku*». Подобные случаи мы стараемся учесть в нашем алгоритме.

- *Различия в морфологии*. Существует ряд языков с богатой морфологией, например, венгерский и финский. Если в качестве целевого языка перевода выступает такой язык, где аффиксы исходного выражаются отдельными словами, например, предлогами, возможность взаимно однозначного выравнивания по словам маловероятна. Однако даже в рассматриваемых нами родственных языках появляются проблемы при работе с морфологическими структурами: так, например, форма среднего залога у глаголов в русском языке образуется при помощи постфикса -*ся*, в то время как в чешском языке указанием на так называемую "возвратность" служит синтаксически отделенная частица «*se*». Этот случай важно учитывать, чтобы добиться правильности и полноты при поиске переводного эквивалента. В качестве еще одного примера можно привести разницу в реализации причинной связи в конструкциях: в тех случаях, когда в чешском языке она выражается формой творительного падежа без предлога («*vykřikla úlekem*»), в русском для уточнения связи между главным и зависимым словом будет введен предлог «*от*» в сочетании с формой родительного падежа («*вскрикнула от испуга*»).

Реорганизация текста переводчиком. Зачастую при переводе не сохраняется авторская организация текста. В редких случаях это может касаться разбиения текста по главам, и намного чаще — членения по предложениям и более мелким структурным единицам в тексте. Так, например, если переводчик переставляет предложения местами, не включает какой-то отрывок или слово в перевод, объединяет несколько коротких предложений в одно длинное или наоборот разбивает длинное предложение, порядок следования элементов в тексте будет нарушен, и такая нелинейность может представлять сложность для моделей, основанных на поиске соответствий между позициями словоформ.

Выбор средств оформления единиц внутри текста. Выбор средств оформления текста помогает определить границы фрагментов, которые должны быть сопоставлены. Данный аспект особенно актуален для алгоритмов, основанных на правилах, которые учитывают информацию о типах границ тех или иных языковых цепочек. Так, исследователь Д.В. Сичинава отмечает, что, «в разных языках (а иногда и в разных изданиях) приняты различные способы графического оформления, что иногда затрудняет определение границ предложения в автоматическом режиме» [Сичинава 2015: p. 210]. К последнему пункту исследователь относит, например, способы оформления переходов от прямой речи к словам автора.

Однако даже использование одних и тех же знаков пунктуации не всегда гарантирует высокую точность выравнивания. Так, точка — наиболее частотный знак для обозначения границ предложения при письме. Тем не менее, согласно статистике, приведенной W. Gale и K. Church, около 10% точек в Брауновском корпусе используются как разделители в числах и датах или в сокращениях (например, «*Mr.*» и т. д.); для сравнения, в подкорпусе деловой газеты «The Wall Street Journal» количество таких случаев достигает 47% от общего числа [Gale, Church 1993]. Таким образом, при разработке алгоритма, который учитывает и анализирует знаки препинания, есть возможность отдельно прописать подобные сложные случаи.

1.3. Обзор подходов, применяемых для выравнивания

Как мы кратко упомянули ранее, на данный момент существует три основных группы подходов, применяемых для решения задачи выравнивания: эвристический, или основанный на правилах, статистический и нейронный. В соответствии с этой классификацией, для сопоставления параллельных текстов разрабатываются и используются эвристические, статистические или нейронные модели. В реальной лингвистической практике обычно лучше применять гибридные алгоритмы на стыке

нескольких подходов: например, достаточно часто в статистические модели внедряется информация о синтаксисе.

Наиболее простыми являются эвристические модели, и с исторической точки зрения они первыми получили развитие. С конца 80-х гг. большую популярность приобрели статистические модели, разрабатываемые в основном для области статистического машинного перевода. Подобные модели позволили более эффективно выравнивать единицы длиной меньше, чем предложения. На смену статистическому машинному переводу пришел нейронный, по отношению к которому выравнивание сначала не рассматривалось, а потом стало потенциальным способом улучшения качества работы подобных моделей.

При выборе алгоритма для выравнивания параллельных текстов следует помнить о том, что успешность его применения зависит в первую очередь от степени родства между языками. Помимо этого, разные литературные жанры могут требовать разных средств и подходов. Так, например, наборы грамматических признаков, синтаксические деревья и двуязычные словари редко применяются для выравнивания художественных текстов вследствие разнообразия лексики и стилистических средств, которые при переводе могут отличаться от средств, используемых в первоисточнике. В то же время этот подход может быть достаточным для выравнивания специализированных и научных текстов. Статистические и нейронные модели в свою очередь хорошо обучаются лишь на больших корпусах текстов и зачастую удовлетворительно выравнивают только частотные единицы. В дополнение к этому, они могут плохо работать на тех параллельных корпусах, где рассматриваемые языки не являются родственными.

1.3.1. Эвристические модели, или модели, основанные на правилах

При построении **эвристической модели** исследователь напрямую обращается к количественным показателям, которые можно извлечь при

анализе корпуса текстов, и/или к лингвистическим тегам, которыми размечен текст. Таким образом, при эвристическом подходе мы можем работать как с аннотированным, так и с необработанным корпусом. Выбор зависит от аспектов, которые исследователь учитывает при выравнивании.

Существуют эвристические алгоритмы, в теории опирающиеся исключительно на количественные показатели. Они чаще всего применяются для выравнивания на уровне предложений. Так, например, W. Gale и K. Church (1993) представили алгоритм, выравнивающий предложения на основании сравнения их относительных длин, при этом длина исчисляется символами. По утверждению самих исследователей, они опираются на тот факт, что длинные предложения переводятся столь же длинными предложениями в целевом тексте, и наоборот — короткие предложения переводятся столько же короткими. Похожий алгоритм для выравнивания по предложениям был представлен P. Brown и его коллегами двумя годами ранее (1991), однако здесь длина предложения исчисляется количеством слов.

Доработанная версия алгоритма Гейла-Черча лежит в основе автоматической программы HunAlign [Varga и др. 2005]. Данная система для выравнивания была разработана для венгерско-английского параллельного корпуса (the Hunglish Corpus: http://mokk.bme.hu/resources/hunglishcorpus/) и выравнивает предложения на основании сопоставления их относительных длин и с использованием двуязычного словаря, либо поданного на вход самим пользователям вместе с корпусом, либо составляемого автоматически после первой итерации алгоритма. Спустя некоторое время программа стала применяться в Национальном корпусе русского языка (НКРЯ: https://ruscorpora.ru/new/) как основа для их собственной системы для выравнивания «Евклид» [Сичинава 2015].

Стоит отметить, что в алгоритмах такого типа обычно не учитываются случаи, когда предложения из оригинального текста переставляются местами в тексте на целевом языке или даже исключаются переводчиком. Если мы

хотим применить такой алгоритм, предложения в тексте на языке перевода должны идти в том же линейном порядке, что и в первоисточнике. Такой подход может быть недостаточным при анализе текстов художественного или близкого к нему жанра.

Что касается лингвистических правил, основным объектом их применения являются синтаксические деревья и наборы грамматических признаков словоформ. Так как при выравнивании исследователь в любом случае ставит перед собой цель выявления и сопоставления некоторых синтаксически оформленных единиц (простых и сложных предложений, словосочетаний и т. д.), использование лингвистических правил подразумевает под собой более глубокий анализ элементов, образующих синтаксически связанную последовательность. При подобном подходе мы часто переходим от чисто синтаксического уровня к лексическому, а затем к морфологическому и в редких случаях семантическому.

Основная сфера применения алгоритмов, основанных на лингвистических правилах, — выравнивание единиц меньше, чем предложение. Так, например, E. Pianta и L. Bentivogli (2004) представили программу KNOWA для выравнивания на уровне слов в параллельных англо-итальянских корпусах. Алгоритм в основе программы опирается на двуязычный словарь и морфологическую разметку. Набор частеречных тегов помогает определить набор потенциальных переводных эквивалентов для каждого слова и в английском, и в итальянском. Затем все слова ранжируются от самых вероятных до менее вероятных кандидатов; для выявления наиболее вероятных кандидатов также используется алгоритм, основанный на правилах. Наконец, для каждого слова окончательно определяется его переводной эквивалент, и результат выдается пользователю. Двуязычные словари помогают определить те случаи, когда слово переводится несколькими единицами в тексте на целевом языке. Авторы сравнивают результат работы своего алгоритма с результатом выдачи статистической

программы *GIZA++* и показывают, что модель, специально подстроенная под правила конкретной языковой пары, способна выдавать более качественный результат по сравнению с обычной статистической моделью.

Лингвистические правила, а именно их применение на синтаксических деревьях, лежит в основе алгоритма для выравнивания на уровне слов, предложенного Y. Ma и его коллегами (2008). Модель, предложенная исследователями, также использует базовые статистические показатели совместной встречаемости, вычисленные при помощи таких эвристических метрик, как метрика отношения правдоподобия (log-likelihood ratio) и коэффициент Дайса (the Dice coefficient). По заявлениям авторов, внедрение их подхода в существующие системы могло бы способствовать улучшению качества автоматического китайско-английского перевода.

1.3.2. Статистические модели

При **статистическом подходе** к выравниванию и переводу мы подаем большой массив текстов на вход алгоритму, который должен статистически выявить закономерности в этих данных и выдать наиболее вероятный перевод. Предложения в тексте чаще всего разбиваются на более мелкие единицы: слова, биграммы, триграммы, намного реже — тетраграммы и т. д. Вероятности их возникновения подсчитываются, наиболее частотные соответствия языковых сегментов выдаются как окончательный перевод. Подобные модели могут работать как с использованием лингвистической информации, так и без нее. Без лингвистического аннотирования улучшение качества перевода может быть достигнуто путем увеличения корпуса. Однако существуют исследования, которые доказывают, что предварительная разметка текстов улучшает качество статистических моделей без пополнения корпуса, о чем мы поговорим чуть позже в этом разделе.

Данный подход получил развитие в конце 80-х гг. XX века наравне с областью статистического машинного перевода. Статистические алгоритмы

позволили более эффективно выравнивать единицы длиной меньше, чем предложения, в параллельных корпусах. Наиболее успешными из первых статистических моделей стали пять моделей, представленные исследовательским центром IBM (IBM Model 1-5) [Brown и др. 1993]. Применение в области выравнивания нашли также скрытые модели Маркова, став одним из наиболее действенных алгоритмов до появления нейронных сетей [Vogel и др. 1996].

В работе F. Och и H. Ney (2003) подробно изучен статистический подход к выравниванию и переводу. Целью статистического машинного перевода, по утверждению исследователей, является моделирование вероятности перевода $P(s_{L1}^{J} \mid s_{L2}^{I})$, где, согласно нотации, представленной ранее, s_{L1}^{J}— некоторая цепочка на исходном языке «*L1*», а s_{L2}^{I}— некоторая целевая цепочка на языке «*L2*», которая является потенциальным переводным эквивалентом для s_{L1}^{J}. Выравнивание в статистических моделях выражается вероятностью $P(s_{L1}^{J}\, a_1^{J} \mid s_{L2}^{I})$, где переменная выравнивания a_1^{J} является по сути отображением из исходной позиции j в целевую позицию a_j. Таким образом, отношение между моделью перевода $P(s_{L1}^{J} \mid s_{L2}^{I})$ и моделью выравнивания $P(s_{L1}^{J}\, a_1^{J} \mid s_{L2}^{I})$ может быть выражено в следующей формуле:

$$P(s_{L1}^{J} \mid s_{L2}^{I}) = \sum_{a_1^{J}} P(s_{L1}^{J}\, a_1^{J} \mid s_{L2}^{I}) \qquad (2)$$

Как отмечается исследователями, в тех случаях, когда некоторому слову невозможно подобрать какой-либо переводной эквивалент, a_1^{J} может содержать элемент выравнивания $a_j = 0$, т.е. так называемое «нуль-выравнивание». Статистические модели для выравнивания и перевода работают с набором параметров Θ, извлеченных из тренировочных входных

данных. Параметры определяются путем максимизации вероятности на тренировочном корпусе [Och, Ney 2003].

Статистические алгоритмы выравнивания получили программную реализацию в таких широко используемых системах, как *GIZA++* [Och, Ney 2003], *fastalign* [Dyer и др. 2013] и *efmaral*, где скрытые марковские цепи объединены с байесовской моделью [Östling, Tiedemann 2016].

Стандартные статистические модели могут обладать некоторыми достоинствами по сравнению с алгоритмами, основанными на правилах, но обычно они касаются производительности и возможности применения на большом наборе языковых пар без обязательной разработки отдельных лингвистических правил. Тем не менее, такие подходы не лишены недостатков: качество выравнивания и перевода напрямую зависит от размера корпуса, в связи с чем статистические модели обычно хорошо справляются лишь с частотными языковыми единицами в составе корпуса. Помимо этого, для неродственных языковых пар чисто статистические модели показывают относительно низкие результаты. На практике успешнее показывают себя гибридные модели, которые содержат информацию о лингвистических правилах и могут помочь подобрать более правильные переводные эквиваленты и выявить некорректные случаи выравнивания.

Так, например, S.J. Ker и J.S. Chang (1997) предлагают учитывать семантические классы и морфо-синтаксические маркеры при работе с парой английский-китайский на уровне слов. Для выявления лексической категории каждого слова авторами подхода используются особые словари. Среди преимуществ предложенного подхода можно выделить достаточно высокую точность, которая достигает 90%, а также увеличение вероятности выравнивания редких слов, которые встречаются один или несколько раз в составе корпуса. Помимо этого, исследователи отмечают, что разработанная система может быть адаптирована к другим языковым парам.

Еще одним примером может послужить алгоритм K. Yamada и K. Knight (2010), который опирается на синтаксические деревья. В работе исследователей рассматривается языковая пара английский-японский и подчеркивается важность учета различий в синтаксисе при работе с языками с разными порядками: например, *SVO* (subject-verb-object) — в английском языке и *SOV* (subject-object-verb) — в японском языке.

1.3.3. Модели, работающие на нейронных сетях

Нейронные модели стали активно применяться для решения лингвистических задач в начале 10-х гг. XXI в. Зачастую при машинном перевода речь идет о рекуррентных нейронных сетях (RNN). Архитектуры подобного типа характеризуются двумя сетями: «кодером» и «декодером» («encoder-decoder»). На вход подобной модели подается большое собрание текстов, из которых она выявляет характеристики, представленные в виде векторов. «Кодер» преобразует тексты или единицы в нем в последовательности чисел, векторов, которые неявным образом отображают синтаксические, семантические и морфологические характеристики рассматриваемых единиц. В связи с этим нет потребности в предварительно размеченных данных. «Декодер» с свою очередь преобразует векторное представление назад в текст на целевом естественном языке.

Проблемой при нейронном переводе является то, что при наличии очень длинных предложений в корпусе сложно "уместить" всю релевантную информацию в вектор установленной длины. Для решения этой проблемы в качестве аналога выравнивания стали предлагать механизм внимания [Bahdanau и др. 2014]. С его помощью, алгоритм разбивает длинное предложение на несколько векторов и при переводе выбирает то подмножество векторов, в котором сосредоточена наиболее релевантная информация. Таким образом, при каждой генерации слова на целевом языке модель начинает искать те позиции в предложении на языке оригинала, которые вероятнее всего содержат такую информацию. Данный подход стал

широко распространенным и получил дальнейшее развитие у других исследователей [Mi и др. 2016; Liu и др. 2016]. Однако, в связи с тем, что механизм внимания оказался непригодным в некоторых моделях, в частности, типа *Transformer*, представленной в работе [Vaswani и др. 2017], в недавних исследованиях подчеркивается важность изучения выравнивания и в эпоху нейронных моделей [Li и др. 2019]. По утверждениям X. Li и его коллег, алгоритмы выравнивания могут помочь при понимании принципа работы нейронных сетей и в исправлении ошибок при переводе.

На данный момент использование нейронных сетей в области выравнивания считается экспериментальным подходом, в основном за счет того, что для данной методики нужны большие корпусы текстовых данных. Как и статистические модели, нейронные обычно хорошо справляются с частотными конструкциями. В последнее время в качестве одного из альтернативных решений появилась идея использовать предобученные векторные модели для выравнивания. По такому принципу работает программа *simalign*, которая может использовать для выравнивания такие многоязычные векторные модели, как *mBERT* [Sabet и др. 2020]. По заявлению самих исследователей, использование предобученных моделей оправдано при корпусах малого объема.

1.4. Выводы

Выравнивание — процедура по соотнесению сегментов исходного текста с его одним или несколькими переводами для обнаружения переводных эквивалентов. Выравнивание может быть проведено на разных языковых уровнях, и актуально для целого ряда задач, таких как машинный перевод, извлечение двуязычной лексики и терминологии, информационный поиск и т. д.

Выравнивание параллельных корпусов может быть осложнено несколькими факторами: разницей между языками, что включает в себя

несоответствия синтаксических, лексических и морфологических структур, реорганизацией текста переводчиком и различиями в оформлении.

Можно выделить три основных подхода к выравниванию: эвристический, или работающий на правилах, статистический и основанный на нейронных сетях. На практике часто лучше показывают себя гибридные алгоритмы, которые находятся на стыке нескольких возможных подходов. Разработка лингвистических правил обычно ресурсозатратна, так как правила отличаются для разных языковых пар. Статистические и нейронные модели традиционно применяются на больших корпусах текстов и выравниваниют лишь частотные конструкции. Статистические модели к тому же часто нуждаются в доработке при применении к языкам с разным синтаксисом.

Глава 2. Особенности предложно-падежных конструкций в русском и чешском языках

2.1. Понятие и описание предлога

Классическим определением предлога считается определение, предложенное в «Русской грамматике» под редакцией Н.Ю. Шведовой: «Предлог — это служебная часть речи, оформляющая подчинение одного знаменательного слова другому в словосочетании или в предложении и тем самым выражающая отношение друг к другу тех предметов и действий, состояний, признаков, которые этими словами называются» [Русская грамматика 1980: с. 704]. Данное определение признается некоторыми исследователями недостаточным [Азарова и др. 2018]. Так, по наблюдениям И.В. Азаровой, В.П. Захарова и А.Д. Москвиной, здесь упускается синтаксическое поведение предлога как части речи, а именно то, что предлоги «оформляют предложно-падежную форму, т.е. могут «оформлять подчинение» только падежных знаменательных слов: в первую очередь, имен существительных, во вторую очередь, местоимений-существительных, числительных и, возможно, прилагательных при их окказиональной субстантивации» [Азарова и др. 2018: с. 10].

Чтобы лучше понять, чем характеризуются предлоги, стоит рассмотреть еще несколько определений. Так, в «Энциклопедическом словаре Брокгауза и Ефрона» предлог представлен как неизменяемая частица, которая уточняет значение глагола или падежа [ЭСБЕ 2012]. Малый академический словарь (МАС) в свою очередь предлагает следующее определение: «Предлог (грамм.) — служебное слово (например: *в*, *к*, *на*, *при*), которое, сочетаясь с существительными, местоимениями и числительными, указывает на синтаксические отношения их к другим словам» [МАС 1999: с. 365]. Объединяя оба определения, можно охарактеризовать предлоги, как слова,

указывающие на синтаксические отношения одних слов, представленных знаменательными частями речи (в первую очередь существительных, затем местоимений-существительных, субстантивированных прилагательных и числительных), с другими знаменательным словам. Зависимые слова, выраженные знаменательными частями речи, при этом стоят в определенной падежной форме, а предлог как неизменяемая частица уточняет значение отношения между зависимой падежной формой и управляющим словом.

Из всех определений, представленных выше, в явном виде следует грамматическая характеристика предлога. Действительно, в связи с особой функцией данной категории в языке, заключающейся в оформлении отношения между знаменательными словами, предлог нельзя считать самостоятельной единицей, так как лексическое значение не присуще предлогам в том виде, в котором мы видим его у знаменательных слов. Грамматическое значение предлога традиционно признается всеми лингвистами, в то время как наличие лексического компонента исторически считается спорным вопросом. Так, например, исследователи-лингвисты А.М. Пешковский и А.А. Шахматов выделяли только грамматический компонент, в то время как В.В. Виноградов и его последователи поддерживали идею о том, что у предлогов можно выделить некоторый семантический компонент [Виноградов 1972]. Согласно «Академической грамматике», все предлоги обладают некоторым лексическим значением, однако оно в разной степени проявляется у каждого отдельного предлога: «…в любом случае предлог имеет лексическое значение, различна лишь степень его абстрактности <...> семантически „пустых“ предлогов не существует» [Русская грамматика 1980].

В.П. Захаров в подтверждение гипотезы о том, что у предлогов можно выделить семантический компонент, приводит ряд противопоставлений, образованных заменой предлога в предложных словосочетаниях: например, «*на столе*» и «*под столом*», «*переводить с английского*» и «*переводить на*

английский» [Захаров 2018]. Даже эти два примера показывают нам, что предлоги способны передавать не только синтаксические, но и семантические отношения между двумя знаменательными словами. В.П. Захаров показывает, что выражаемые отношения при этом могут быть самыми разными: пространственными, временными, отношениями сопоставления и причины и т. д. Таким образом, исследователь приходит к тому, что семантическое значение у предлогов действительно отличается от значения знаменательных слов, и причиной этого явления можно считать отсутствия прямого денотата у экземпляров первой группы; тем не менее, говорить о полном отсутствии семантического компонента у предлогов нельзя: оно реализуется в конкретных предложных конструкциях.

В соответствии с этими положениями, в данной работе мы принимаем точку зрения, согласно которой у предлогов можно выделить лексическое значение, однако сила его выраженности отличается для каждого конкретного предлога. Семантика предлогов и минимальных единиц синтаксиса, в состав которых они входят, наиболее подробно изучена и описана в «Синтаксическом словаре» Г.А. Золотовой [Золотова 2011]. При автоматической обработке естественного языка, "усредненное" семантическое значение предлогов можно выявить путем просмотра их векторных представлений. Подобные векторы обычно не зависят от конкретного контекста предложения или конструкции, которые мы рассматриваем в определенный момент. Алгоритмы в основе таких моделей "усредняют" все контексты, что позволяет в дальнейшем выявить самые частотные синонимы или переводные эквиваленты для них.

2.2. Классификация предлогов в русском языке

Как было выяснено нами ранее, с точки зрения синтаксиса предлог оформляет отношения между словами, выраженными знаменательными частями речи и служащими для обозначения некоторого действия или

объекта. Традиционно в русском языке предлоги разбиваются на следующие большие группы: первообразные и производные, однословные и составные:

- ***Производные*** предлоги мотивированы знаменательными частями речи, обычно в сочетании с ***первообразными*** предлогами: «*в качестве*», «*под видом*» и т. д.
- ***Однословные*** предлоги состоят из одного слова, могут быть представлены как немотивированными первообразными, так и производными предлогами: «*в*», «*без*», «*от*», «*посредством*», «*включая*», «*исключая*» и т. д.. ***Составные*** предлоги состоят и двух и более слов, причем данную группу представляют исключительно производные предлоги: «*в течение*», «*вблизи от*» и т. д.

Предлоги как часть речи относятся к открытому классу в русском языке. Их количество строго не фиксировано, и группа постоянно пополняется, в частности разряд производных предлогов. Это происходит не только за счет функционирования единиц знаменательной части речи как предлогов, что в конечном счете приводит к их грамматикализации, но и за счет приобретения устойчивости у сочетаний из нескольких слов (см. «*наравне с*», «*в соответствии с*»). В данной работе мы стараемся охватить максимальное количество *однословных* предлогов.

Как только мы определили предлог с точки зрения мотивированности и количества элементов, входящих в его состав, можно рассмотреть, на какие группы делятся предлоги с точки зрения семантики. Согласно классификации, представленной в работе В.П. Захарова (2018), здесь исследователь имеет дело с двумя большими группами: немногозначными и многозначными предлогами:

- ***Немногозначный*** предлог почти не зависит от управляемого слова, его значение конкретно само по себе. Основную часть этой

группы представляют составные производные предлоги, значение которых уже определено за счет знаменательных слов, входящих в состав конструкции (например, «*по причине*», «*с целью*»).

- Значение ***многозначного*** предлога зависит от конкретного контекста. Чаще всего представителями этой группы становятся первообразные однословные предлоги.

2.3. Определение предложной конструкции

Объектом нашего исследования выступает такая синтаксически связная единица языка, как предложная конструкция. Формально, предложную конструкцию составляет непосредственно сам предлог в сочетании с его «минимальным контекстом»: главным и зависимым словом, парой «хозяин»-«слуга», в состав которой входят знаменательные слова [Захаров, Михайлова 2018: с. 60]. В предложных конструкциях анализу подлежат отношения между главным и зависимым словами, уточненные при помощи предлога, при этом управляющее слово — «хозяин», а зависимое — «слуга» [Захаров 2018].

В отношении определения предложных конструкций есть несколько спорных вопросов: так, например, не совсем ясно, составляет ли предлог часть падежной формы. Традиционно считается, что предлог управляем зависимым знаменательным словом, обычно существительным, но могут существовать и другие мнения на этот счет. По Г.А. Золотовой, предлог в сочетании с падежной формой составляет минимальную нечленимую синтаксическую единицу, называемую *предложно-падежной синтаксемой* [Золотова 2011]. В предложении или словосочетании подобные синтаксемы выступают в качестве конструктивно-смысловых компонентов.

С опорой на исследования Г.А. Золотовой, М.В. Всеволодова (2014) изучает понятия синтаксемы и синтаксической формы слова. Как отмечается автором, в морфологических исследованиях обычно рассматривается

непосредственно падежная форма слова, без учета предлогов и частиц, совместно с которыми она выступает в тексте. В качестве примера М.В. Всеволодова приводит словосочетание «*войти в избу*», где можно выделить три морфологических единицы: инфинитив «*войти*», предлог «*в*» и форму существительного в винительном падеже «*изба*». С морфологической точки зрения, данные единицы не взаимосвязаны. Тем не менее, если мы рассматриваем функционирование слова на уровне синтаксиса, такие служебные категории, как предлоги и частицы, составляют синтаксическую форму слова и являются частью падежа.

Синтаксическая форма слова, по утверждению исследователя, может быть рассмотрена и описана с разных точек зрения. Так, мы можем рассматривать непосредственно *словесную форму*, которая определяется наиболее абстрактно: «Словесная форма слова — это морфологическая форма части речи в конъюнкции со служебными словами (предлогами, частицами), если они есть, без учёта формирующей ее лексики» [Всеволодова 2014: с. 21]. Синтаксическая форма слова может включать в себя *словоформу* как конкретную реализацию слова в тексте. Для синтаксических исследований особую актуальность имеет третий уровень описания синтаксической формы: *синтаксема*, т. е. «словоформа с учетом ее категориального (общего и частного) значения и ее синтаксических потенций» [Всеволодова 2018: с. 22]. Синтаксема наименее абстрактна и включает в свой состав и морфологические формы, и конкретные лексические единицы.

2.4. Сопоставление русского и чешского языков

Целью нашей работы является разработка алгоритма для выравнивания предложных конструкций в параллельном русско-чешском корпусе. В связи с этим имеет смысл рассмотреть более подробно особенности языков и самих конструкций.

Оба языка принадлежат славянской группе, но разным ее ветвям: русский — восточнославянской, чешский — западнославянской. Этим обусловлена большая степень схожести языков с точки зрения морфологии и синтаксиса, сведения о которых играют большую роль при разработке методики в нашем алгоритме.

2.4.1. Сопоставление синтаксиса языков

В связи с родственностью языков, при сопоставлении на уровне синтаксиса можно обнаружить много сходств. И русский, и чешский представляют ту группу, в которой наиболее типичной структурой является *SVO* (subject-verb-object). Структурные схемы простых предложений, заданные с точки зрения частей речи, совпадают, а элементы в составе сложных предложений также присоединяются друг к другу при помощи союзов или союзных слов. При анализе примеров из корпуса почти всегда заметен схожий способ организации элементов внутри синтаксически оформленной единицы. В качестве примера можно привести следующие два предложения из нашего корпуса, принцип составления которого мы описываем в разделе 3.2:

«*Ты ходишь в школу?*» — «*Chodíš do školy?*»

«*Я поднимался по крутым улочкам вдоль задних фасадов дворцов, я не торопился, я даже не знал, куда сверну на следующем перекрестке.*» — «*Stoupal jsem příkrými uličkami podél zadních traktů paláců, nespěchal jsem, nevěděl jsem, kudy se dám za příštím rohem.*»

Столь заметные сходства в синтаксисе позволяют нам говорить о том, что даже при художественном переводе мы часто можем найти переводной эквивалент для некоторой единицы примерно в той же части предложения, в которой она находится в исходном тексте. Путем умозрительного заключения нами было выявлено, что имеет смысл просматривать контекстное окно ±5 слов для поиска переводного эквивалента в предложении на целевом языке.

В обоих языках можно явно выделить такую синтаксическую единицу как словосочетание. Словосочетание можно определить как соединение знаменательных слов между собой либо при помощи соединительной связи, если части между собой равны, либо при помощи подчинительной связи, если мы можем явно выделить зависимое и главное слово.

При описании подчинительных связей внутри словосочетания обычно выделяют три вида связей: согласование, управление, примыкание. Подобные типы можно выделить и при анализе чешского корпуса. Зачастую типы связи в словосочетании в чешском сегменте корпуса соответствуют тем, которые мы наблюдаем у эквивалентных русскоязычных словосочетаний:

Согласование-управление: «*хороший друг*» — «*хорошего друга*» — «*хорошему другу*» = «*dobrý přítel*» — «*dobrého přítele*» — «*dobrému příteli*».

Управление: «*приглашать гостей*» — «*zvát hosty*»; «*заботиться о внуке*» — «*starat se o vnuka*»; «*забота о детях*» — «*starost o děti*»; «*страница книги*» — «*stránka knihy*»; «*похожа на сестру*» — «*podobná sestře*».

Примыкание: «*иду быстро*» — «*jdu rychle*».

Как мы можем увидеть, схожесть словосочетаний выражается не только синтаксически, но и морфологически: так, например, при согласовании-управлении зависимое слово, выраженное прилагательным в нашем примере, уподобляется по падежу, числу и роду главному слову, выраженному существительным.

В случае управления, мы также наблюдаем интересные закономерности: в первых двух словосочетаниях проявляется сильное глагольное управление. Синтаксическая валентность глагола может требовать после себя либо определенную падежную, либо определенную предложно-падежную форму. Как мы показываем в этой работе, проведя практический эксперимент, довольно часто конструкции с предлогом в русском языке переводятся конструкциями с предлогом на чешский. В случае сильной синтаксической

валентности глагол может несколько сужать значение предлога до какого-то конкретного; таким образом, при анализе некоторой предложной конструкции мы можем, например, из всего потенциала значений предлогов остановиться только на одном.

Существуют конструкции, которые мы можем рассматривать как частные случаи примыкания и управления: например, «*говорил час*» — «*mluvil hodinu*»; «*приехал в город*» — «*přijel do města*». Здесь реализованная связь является слабой в силу необязательности заполнения синтаксической валентности глагола.

2.4.2. Сопоставления морфологических структур языков

Морфология. В русском и чешском языке можно выделить примерно тот же набор частей речи, что говорит о схожести в функционировании слов с точки зрения синтаксических, морфологических и семантических свойств по определению, предложенному в «Лингвистическом энциклопедическом словаре» [Лингвистический энциклопедический словарь 1990].

Оба языка относятся к группе языков флективного типа. Это означает, что основным способ выражения грамматических форм и в чешском, и в русском служит словоизменение при помощи флексий.

В целом, части речи в чешском обладают почти такими же грамматическими категориями, что и в русском. Так, например, для существительных можно выделить словоизменительные категории падежа и числа и классифицирующую категорию рода. Система падежей в чешском почти полностью совпадает с русской, но, традиционно, чешская система включает в себя дополнительный звательный падеж, который в явном виде отсутствует в русском, хотя есть некоторый аналог, обращение, которое не фиксировано в падежной системе. К другим, более заметным различиям в морфологическом строе языков можно отнести систему склонений имен существительных, которая в чешском языке более сложная, чем в русском.

При рассмотрении морфологического строя обоих языков есть несколько аспектов, особенно важных для нас. К одному из них относится информация о переводе глаголов в форме среднего залога на чешский язык. Так, возвратный постфикс *-ся*, который участвует в образовании формы среднего залога в русском языке, в чешском передается при помощи синтаксически отделенной возвратной частицы «*se*». Стоит отметить, что не всегда русские глаголы с *-ся* переводятся с использованием частицы *se* на чешский язык, поэтому мы посчитали целесообразном включить проверку этого условия в наш алгоритм.

Помимо этого, как мы уже упоминали ранее в этой работе, в корпусной статистике наблюдаются различия в реализации причинной связи в словосочетаниях: в чешском языке подобная причинная связь может быть выражена беспредложной падежной формой («*vykřikla úlekem*»), но в русском языке скорее всего появится первообразный предлог «*от*» в сочетании с формой родительного падежа («*вскрикнула от испуга*»).

Приведенные характеристики языков позволяют сопоставлять параллельные корпусы с опорой на синтаксическую и морфологическую разметку.

2.5. Выявление предложных конструкций в русском и чешском языках

Схожести на уровне синтаксиса и морфологии приводят нас к тому, что предложные конструкции строятся по схожему принципу в обоих языках. Ранее мы уже рассмотрели подчинительные предложные связи, которые могут быть реализованы и в русском, и в чешском языках. Подобные подчинительные связи лежат в основе предложных конструкций.

Предлоги в чешском группируются в соответствии с теми же типологиями и классификации, что и предлоги в русском языке. Например,

можно выделить группы первообразных («*bez*», «*dle*», «*do*», «*k*», «*mezi*», «*na*» и т. д.) и производных предлогов («*díky*», «*za příčinou*», «*u příležitosti*», «*vlivem*» и т. д.). Вторичные, или производные, предлоги также мотивированы другими частями речи, которые в какой-то момент начинают функционировать в речи как предлоги, что в конце концов приводит к их грамматикализации.

Главной особенностью чешских предлогов является то, что они, также как и предлоги в русском языке, выражают и уточняют отношение между знаменательными словами, управляющим и зависимым словом в предложной конструкции. Некоторыми грамматистами выдвигается предположение, что зависимое слово, обычно существительное, управляет предлогом в предложной конструкции. Существует подход, согласно которому управляющее слово так же важно для уточнения значения предлога. Для нас в нашей работе важно то, что предлог никогда не выступает сам по себе, он всегда связан со своей конструкцией. По этой причине, мы считаем, что большую роль играют оба типа отношений: и отношение предлога с зависимым словом, и отношение предлога с управляемым словом.

Как и в русском, в чешском может возникнуть частичная синонимия в рамках предложной системы. Русскоязычная и чешская предложные системы обнаруживают большие сходства. Подтверждением этому может служить Таблица 1, демонстрирующая предлоги на русском и их "прямые" переводные эквиваленты на чешский, т. е. эквиваленты, совпадающие по внешней форме и частично по значению с предлогами из русской системы. В таблице также указаны падежи, с которыми могут сочетаться предлоги:

Первообразный предлог на русском	**Прямой перевод на чешский**
без(о): + Р.п.	*bez(e)*: + Р.п.
в(о): + Вин.п, Пр.п.	*v(e)*: + Вин.п., Пр.п.
для: + Р.п.	*pro*: + Вин.п

до: + Р.п.	*do*: + Р.п.
за: + Вин.п., Тв.п.	*za*: + Р.п., Вин.п., Тв.п.
из(о): + Р.п.	*z*: + Р.п.
из-за: + Р.п.	*z(e), zpoza*: + Р.п
из-под(о): + Р.п.	*zpod*: + Р.п.
к(о): + Дат.п.	*k(e), ku*: + Дат.п.
меж, между: + Р.п., Тв.п.	*mezi*: + Вин.п., Тв.п.
на: + Вин.п., Пр.п.	*na*: + Вин.п., Пр.п.
над(о): + Тв.п.	*nad(e)*: + Вин.п., Тв.п.
о: + Вин.п., Пр.п.; *об*: + Пр.п.	*o*: + Вин.п., Пр.п.
от(о): + Р.п.	*od(e)*: + Р.п.
перед(о), пред(о): + Тв.п.	*před(e)*: + Вин.п., Тв.п.
по: + Дат.п., Вин.п., Пр.п.	*po*: + Вин.п., Пр.п.
под(о): + Вин.п., Тв.п.	*pod(e)*: + Вин.п., Тв.п.
про: + Вин.п., Тв.п.	*pro*: + Вин.п.
против: + Р.п.	*proti*: + Дат.п.
при: + Пр.п.	*při*: + Пр.п.
с(о): + Р.п., Вин.п., Тв.п.	*s(e)*: + Р.п., Вин.п., Тв.п.
сквозь: + Вин.п.	*skrz(e)*: + Вин.п.
среди, средь, середь: + Р.п.	*uprostřed*: + Р.п.; *mezi*: + Вин.п., Тв.п.
у: + Р.п.	*u, vedle*: + Р.п.; *při* + Пр.п.
через(о), чрез(о): + Вин.п.;	*přes*: + Вин.п.;

Таблица 1. Предлоги на русском и их переводные эквиваленты.

Сходство предлогов, представленное нами в Таблице 1, зачастую носит формальный характер, проявляющийся в использовании одинаковых падежей или совпадению по внешней форме. С точки зрения семантики, сходство может быть лишь частичным: например, предлог «*в*» в русском языке насчитывает 23 значения, в то время как «*v*» в чешском — 16 [Захаров 2020]; это говорит о том, что отнюдь не всегда они могут быть переводными

эквивалентами друг для друга. Помимо этого, не стоит забывать о том, что предложные конструкции иногда могут переводиться беспредложно на чешский.

Межъязыковые соответствия всегда выражаются в конструкциях. В связи с этим, для целей лингвистической компаративистики важно работать с предложными конструкциями, чем мы и занимаемся.

2.6. Выводы

Предлог — служебная часть речи, которая с точки зрения синтаксического функционирования оформляет отношение между некоторыми действиями или объектами, которые в свою очередь выражены знаменательными частями речи. Эти знаменательные слова составляют минимальный контекст предлога и образуют с ним предложно-падежную конструкцию, где можно выделить главное и зависимое слово.

В русском языке предлоги можно рассматривать с точки зрения их происхождения (*первообразные*, *производные*), количества элементов, входящих в состав предлога (*простые*, *сложные*).

В отношении предлога неясным остается вопрос об их семантике: некоторые лингвисты выявляют у предлогов только грамматическое значение, в то время как другие призывают выделять и лексический компонент. С точки зрения функционального синтаксиса, значение предлога неразрывно от значения предложно-падежной синтаксемы, в состав которой он входит, а значит может быть реализовано только в контексте. Среди тех исследователей, которые поддерживают теорию о наличии у предлогов семантического компонента, можно услышать, что все предлоги обладают подобным значением, но оно по разному выражено у каждого отдельного предлога. В соответствии с этим, можно выделить группы немногозначных и многозначных предлогов: первую группу обычно составляют производные предлоги, в то время как вторую — первообразные. В нашей работе мы

придерживаемся той точки зрения, согласно которой у предлогов можно выделить семантический элемент. Мы используем векторные представления слов для отображения семантических характеристик предлогов и поиска их переводных эквивалентов.

Предложные конструкции в русском и чешском языках строятся по схожему принципу. Это обусловлено схожестью синтаксических и морфологических структур, которые проистекают из степени родства языков. В обоих языках можно выделить такую единицу, как словосочетание. Предложные также во многом обнаруживают большие сходства: они касаются как внешней формы, так и набора падежей, с которыми сочетаются предлоги.

Глава 3. Разработка методики и создание инструмента для выравнивания предложных конструкций

3.1. Описание методики выявления предложных конструкций

В данной работе для выравнивания предложных конструкций мы используем метод, опирающийся на морфологические, синтаксические и семантические маркеры словоформ. При разработке методики были учтены особенности обоих языков и закономерности, выявленные при анализе предложений из параллельного русско-чешского корпуса, собранного нами для проверки качества работы инструмента.

На первом этапе работы алгоритма мы ищем предложные конструкции в русском сегменте корпуса с опорой на грамматические и синтаксические маркеры слов в составе каждого предложения. Под предложными конструкциями в практической части мы понимаем предлог в сочетании с зависимым словом, выраженным знаменательной частью речи (существительным, местоимением-существительным, числительным) и стоящим в определенной падежной форме, а также главное слово, представленное знаменательной частью речи, чаще всего существительным или глаголом. В рамках практической части мы фокусируемся на выравнивании конструкций с простыми, или однословными, предлогами. После выявления предложной конструкции на русском мы анализируем чешский подкорпус и, используя векторные представления слов, грамматические и морфологические маркеры, находим подходящий переводной эквивалент для исходного сочетания на русском. В алгоритме мы стараемся учесть те случаи, когда переводной эквивалент является беспредложным.

Алгоритм был написан на языке программирования *Python*. Для разметки и определения границ предложных конструкций нами была выбрана

библиотека *UDpipe*, разрабатываемая в рамках проекта *Universal Dependencies* (https://universaldependencies.org/). При определении степени семантической схожести между словами были использованы предобученные векторные представления, распространяемые в рамках проекта *fasttext* (https://fasttext.cc/docs/en/aligned-vectors.html) [Joulin и др. 2018, Bojanowski и др. 2017]. Со сжатым принципом работы алгоритма можно ознакомиться на Рисунке 1.

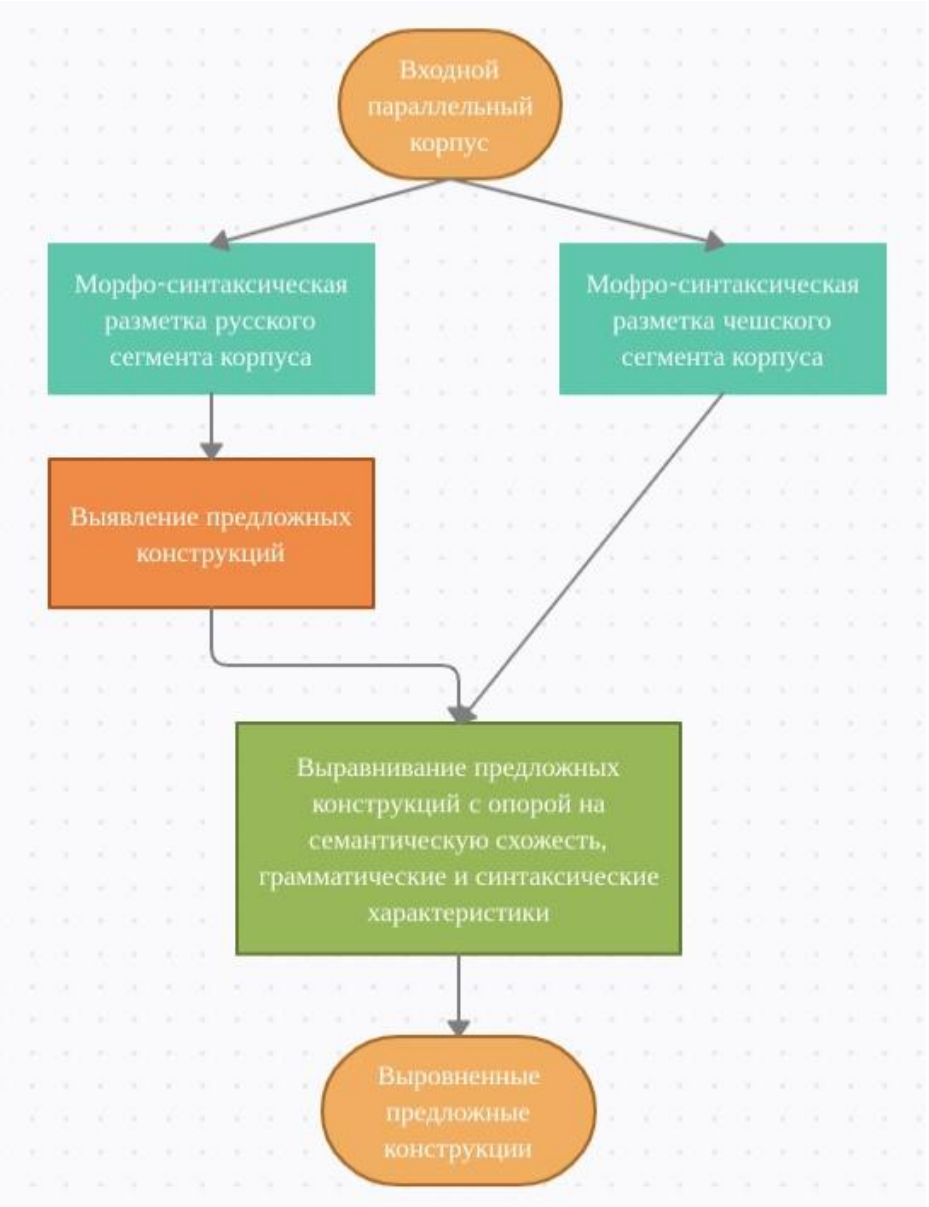


Рисунок 1. Схема, изображающая принцип работы созданного алгоритма.

Для проверки качества работы алгоритма нами был собран параллельный русско-чешский корпус, который содержит 200 предложений и 342 предложных конструкции. В состав корпуса входят предложения разной длины, из произведений художественного жанра, субтитры и новостные и экономические сводки. Все предложения были размечены вручную: сначала нами были выделены конструкции на русском языке, затем — соответствующие им словосочетания на чешском. Результаты ручной разметки были сопоставлены с результатами автоматической. Алгоритм был оценен при помощи таких метрик, как меры полноты, точности и F-мера.

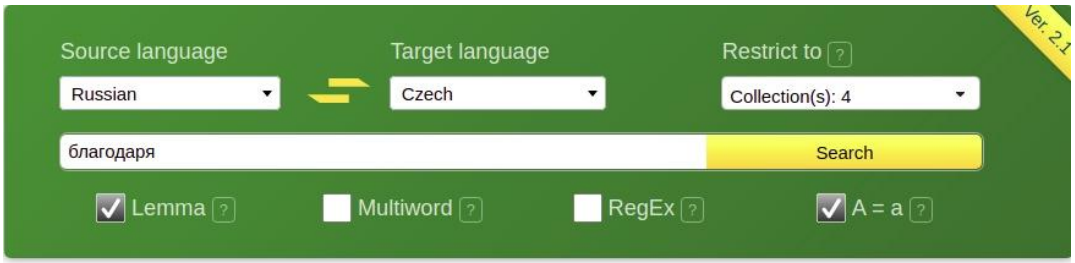
3.2. Сбор параллельного русско-чешского корпуса

Как мы уже упомянули ранее, материалом для нашего исследования и проверки качества работы алгоритма послужил собранный нами русско-чешский корпус. Данный корпус содержит параллельные предложения на русском и чешском. Главным основанием для включения того или иного предложения в состав корпуса стало наличие хотя бы одной предложной конструкции в русскоязычном предложении.

Все предложения были отобраны вручную из многоязычного параллельного корпуса InterCorp v.13 [Čermák, Rosen 2012], доступного в рамках проекта Чешского национального корпуса (ЧНК: https://www.korpus.cz/). Для просмотра корпуса были задействованы встроенные программы *KonText* и *Treq*. Первый инструмент позволил нам получить полные предложения, в состав которых входили указанные нами предлоги. *Treq* использовался для просмотра статистики частотности переводного эквивалента для того или иного русскоязычного предлога (см. Рисунок 2).

Единицы, взятые нами как основа для нашего корпуса, уже были грамотно выровнены на уровне предложений, что позволило нам без затруднений осуществить поиск конструкций и на русском, и на чешском

языках. В состав корпуса, как мы уже упомянули ранее, вошли предложения из художественной литературы, субтитров и новостей. Таким образом, в нашем корпусе представлены как образцы литературного, так и образцы делового и разговорного языков.



Frequency	Proportion	Russian	Czech
637	66.9	благодаря	díky
52	5.5	благодаря	kvůli
30	3.2	благодаря	z
29	3.0	благодаря	vzhledem
27	2.8	благодаря	prostřednictvím
21	2.2	благодаря	dík
21	2.2	благодаря	zásluha
12	1.3	благодаря	s
11	1.2	благодаря	proto
9	0.9	благодаря	důsledek
7	0.7	благодаря	skrz

Рисунок 2. Результат выдачи программы Treq для производного предлога «благодаря».

3.3. Разметка корпуса

Для того, чтобы изучить особенности русского и чешского языков и использовать эти особенности для выравнивания предложных конструкций, нам был необходим инструмент для разметки, который проводил бы качественное аннотирование по нескольким лингвистическим уровням и предусматривал бы универсальные наименования категорий для словоформ в обоих языках.

Для этой цели нами была выбрана библиотека *UDpipe*, разрабатываемая в рамках проекта *Universal Dependencies* (https://universaldependencies.org/). В библиотеке предусмотрены модули для проведения токенизации, лемматизации, морфологического и синтаксического анализа.

Для разметки русскоязычной части корпуса нами была использована модель *SynTagRus* [Straka, Straková 2017]. Именно *SynTagRus* показала наиболее точные результаты по сравнению с другими моделями, доступными для *UDpipe*: *Taiga* и *GSD* (Google Stanford Dependencies). В частности, у других моделей регулярно возникали проблемы с лемматизацией и соответственно определением частеречного тега для местоимений и глаголов. Так, например, в предложении «*Я поднимался по крутым улочкам, <...> не знал, куда сверну на следующем перекрестке*» для словоформы «*сверну*» модель корпуса *Taiga* определила лемму как «*сверна*» и приписала ей тег «*NOUN*», т. е. существительное.

Мы объясняем точность модели *SynTagRus* тем, что данная модель является наиболее крупной из представленных в *UDPipe* моделей корпусов: она насчитывает 61 889 предложений and 1 106 296 токенов. Для сравнения, три другие модели содержат как минимум в 6 раз меньше предложений и в 10 раз меньше токенов.

По аналогичному принципу нами была выбрана модель *Prague Dependency Treebank* (PDT) для разметки подкорпуса на чешском языке. PDT является одним из крупнейших корпусов чешского языка и размечает текст не только тегами универсальных грамматических категорий, но и тегами, специфичными для чешского языка.

Поскольку разметка инструментом *UDpipe* включает в себя множество языковых аспектов, для удобства ее хранения был разработан табличный формат *CoNLL-U* [Buchholz, Marsi 2006]. В таблице на каждый токен, или словоформу, отведена одна строка, состоящая из десяти полей, разделенных знаком табуляции. Поля содержат информацию о таких категориях, как порядковый номер словоформы в предложении, токен и его лемматизированная форма, часть речи, грамматические признаки, тип синтаксической связи с другим словом в предложении и т. д.

На Рисунке 3 представлен пример разметки для предложения из русскоязычной части корпуса, сохраненный в табличном формате *CoNLL-U*. На изображении представлены не все поля, но все равно виден общий способ организации данных.

```
# newdoc id = doc186
# sent_id = 1
# text = Утром я возвращаюсь в Париж.
1       Утром   утро    NOUN    _       Animacy=Inan|Case=Ins|Gender=Neut|Number=Sing   3
2       я       я       PRON    _       Case=Nom|Number=Sing|Person=1   3       nsubj   _
3       возвращаюсь     возвращаться    VERB    _       Aspect=Imp|Mood=Ind|Number=Sing|Pers
4       в       в       ADP     _       _       5       case    _       _
5       Париж   Париж   PROPN   _       Animacy=Inan|Case=Acc|Gender=Masc|Number=Sing   3
6       .       .       PUNCT   _       _       3       punct   _       SpacesAfter=\n
```

Рисунок 3. Пример разметки русскоязычного предложения инструментом *UDpipe*.

3.4. Алгоритм выявления русскоязычных предложных конструкций

В разделах 2.3, 2.4. и 2.5. мы уже затрагивали определение предложной конструкции и признаки, ее характеризующее. В разработанном нами алгоритме для выравнивания мы в первую очередь ставим перед собой задачу точного выявления предложных конструкций в русскоязычном сегменте собранного нами корпуса. Так, для выделения последовательности «хозяин»-предлог-«слуга» из текста на русском языке, мы опираемся на синтаксические и морфологические признаки, приписанные словоформам автоматическим инструментом *UDpipe*, а также на некоторые сведения о предложных конструкциях, рассмотренные нами в Главе 2.

Алгоритм выявления одной предложной конструкции может быть описан в 5 шагах:

1. Поиск предлога;
2. Определение кандидатов на главное и зависимое слова;
3. Дополнительная проверка найденных кандидатов на соответствие характеристикам, свойственным предложным конструкциям;
4. Добавление векторных представлений для слов;
5. Объединение отдельно стоящих элементов в конструкцию.

Рассмотрим каждых из этих этапов подробнее.

Выявление предлога. В разметке *UDpipe* предлог обозначен как *«ADP»*, сокращенно от *adposition*, и таким образом отнесен в одну категорию с послелогами. После того, как мы находим предлог, мы получаем доступ к сведениям о главном и зависимом словах, синтаксическая связь с которыми выражена при помощи особых тегов. Данные о зависимом и главном слове сохранены в отдельном поле в виде индекса, указывающего на порядковый номер слова-кандидата, а тип связи указан в соответствии со специальной системой, разработанной *UDpipe* и впервые представленной в работе Marie-Catherine de Marneffe и ее коллег (2014).

Для поиска предлога в предложении мы пробегаем по всем токенам в нем: каждый токен, частью речи которого является *«ADP*», начинает рассматриваться нами более серьезно, мы подбираем для него минимальный контекст и, если этот минимальный контекст действительно образует вместе с предлогом предложную конструкцию, выравниваем его вместе с сочетанием в чешском предложении.

Определение кандидатов на роли «хозяина» и «слуги» в предложной конструкции. Для обозначения связи между зависимым словом и предлогом в *UDpipe* предусмотрен тег *«case»*. В системе *UDpipe* данный маркер указывает на то, что предлог рассматривается как отдельный синтаксический элемент, указывающий на падеж, и который к тому же является зависимым от знаменательного слова, которое подобный синтаксический элемент предваряет или к которому он присоединен. Выявив при помощи типа синтаксической связи и индекса зависимое слово, мы можем от него определить кандидат на главное слово в составе конструкции. Для этого мы должны увидеть, что от зависимого слова до главного идет такой тип связи как *«obl»* или *«nmod»*. Первый тег указывает нам на то, что именное зависимое слово (например, существительное или местоимение-существительное) выступает как косвенное дополнение или

определение по отношению к глаголу или другому адвербиальному слову. Что касается *«nmod»*, он используется для указания на то, что зависимое слово является таковым для другого существительного, именного словосочетания или номинативной единицы языка и грамматически соответствует определению или дополнению, выраженному родительным падежом. Данные маркеры зачастую сопровождается тегом *«case»*, который как раз был рассмотрен нами выше.

Введение дополнительных ограничений на слова, составляющие предложную конструкцию. Несмотря на то, что типы синтаксических связей, описанные нами выше, достаточно четко ограничивают спектр категорий, слова которых могут функционировать в качестве зависимого и главного слов в предложной конструкции, в результате эксперимента было выявлено, что все же требуются дополнительные ограничения на части речи и теги. При ближайшем рассмотрении синтаксической функции предлога в разделе 2.1., мы увидели, что предлоги способны оформлять подчинение главному слову только падежных знаменательных слов. К этим категориям были отнесены существительные, местоимения-существительные, числительные и субстантивированные прилагательные. Чтобы исключить наименее релевантные случаи из нашей выборки, прежде чем окончательно объявить кандидатов как «хозяина» и «слугу» в конструкции, мы проверяем, не аннотированы ли они одними из следующих категорий и тегов: сочинительный союз (*«CCONJ»*), противительный союз (*«SCONJ»*), определяющее слово (например, артикль, квантификатор, или кванторное слово; *«DET»*), междометие (*«INTJ»*), частица (*«PART»*), знак пунктуации (*«PUNCT»*), символ (отличный от знаков пунктуации, например, эмодзи; *«SYM»*), предлог (*«ADP»*), наречие (*«ADV»*).

Помимо этого, при анализе соотношения предлогов и падежных форм можно заметить, что знаменательное слово достаточно редко стоит в форме именительного падежа. Обычно знаменательное слово представлено другими

падежами. В составе нашего корпуса не встретилось таких истинных случаев, но было некоторое количество ложных, когда слово в форме именительного падежа присоединяется к предлогу, который не может требовать после себя именительного падежа. В связи с этим, мы вводим дополнительное ограничение на этот аспект: если форма, связанная с предлогом, представлена именительным падежом, мы не рассматриваем эту конструкцию. Проверить то, что падежная форма представлена именительным падежом мы можем при обращении к набору грамматических характеристик словоформ. Таким образом, если мы видим среди списка *«CASE=Nom»*, мы не рассматриваем конструкцию.

Добавление векторных представлений для слов. После того, как мы убедились в правильности выбора главного и зависимого слова, мы добавляем для всех трех элементов будущей конструкции векторные представления (word embeddings), способные отображать семантические характеристики слов. Такие векторы нужны нам для дальнейшего поиска переводного эквивалента в чешском сегменте корпуса. Нами была выбрана выровненная векторная модель, обученная на Википедии и доступная в рамках проекта *fasttext* [Joulin и др. 2018, Bojanowski и др. 2017]. Ознакомиться со списком векторных моделей можно на официальном сайте проекта (https://fasttext.cc/docs/en/aligned-vectors.html).

Для каждого слова мы добавляем два вектора: один — для леммы, другой — для словоформы в том виде, в котором она представлена в тексте. Такой подход обусловлен тем, что векторная модель может не включать в себя определенную форму слова, но почти во всех случаях в ней представлена лемматизированная, исходная форма, степень схожести с которой будет определена для слов в предложении на чешском. Вектор для словоформы представляет запасной вариант на тот случай, если лемматизированная форма слова была определена неправильно и лемма не была найдена среди словаря модели *fasttext*. Если же ни лемма, ни

словоформа не представлены в словаре, слово получает значение 0 вместо вектора. Стоит отметить, что последний случай крайне редок.

Объединение элементов в предложную конструкцию. На последнем этапе определения русскоязычной предложной конструкции, все три элемента (главное слово, предлог, зависимое слово) объединяются в единую конструкцию. Затем мы переходим к поиску ее переводного эквивалента.

3.5. Алгоритм поиска переводных эквивалентов в чешском сегменте

На следующем этапе выравнивания мы переходим к поиску переводных эквивалентов в составе русско-чешского корпуса.

При просмотре корпуса и анализе результатов мы допускаем предположение о том, что знаменательные слова в конструкции на чешском будут представлены теми же частями речи. Мы создаем словарь, структура которого представлена ключами и их значениями: в качестве ключей выступают знаменательные части речи, в качестве их значений — списки потенциальных частеречных эквивалентов, заданных при помощи тегов *Universal Dependencies*. При сопоставлении подобных списков мы смотрим, какими частями речи представлены в собранном нами корпусе знаменательные слова: и управляющие, и управляемые. Обычно для каждой категории переводной эквивалент может быть только один, например, прилагательное всегда представлено прилагательным при переводе. Тем не менее, наблюдаются два исключения:

1. Переводным эквивалентом для нарицательного имени существительного может быть как нарицательное, так и собственное имя («*NOUN*»: «*NOUN*», «*PROPN*»).
2. Переводным эквивалентом для глагола может выступить как глагол, так и словоформа, отмеченная тегов прилагательного («*VERB*»: «*VERB*», «*ADJ*»). Такое допущение связано с

неточностями при разметке данных автоматическим инструментом *UDpipe*.

Поиск главного слова в предложной конструкции. Выявление переводного эквивалента начинается с того, что мы определяем, стоит ли главное слово в русскоязычной конструкции, если оно представлено глаголом, в форме среднего залога. Для этого мы смотрим на список грамматических признаков словоформы: если мы видим среди них «*Voice=Mid*», мы ищем в предложении на чешском частицу «*se*», которая должна быть синтаксически привязана к некоторому глаголу. Частица также будет включена нами в состав предложной конструкции.

Далее, по индексу синтаксически связанного с частицей слова мы получаем доступ к глаголу. Этот глагол будет служить нам кандидатом на место главного слова в переводном эквиваленте. Для того, чтобы определить, так ли это, мы рассчитываем степень его семантической схожести с управляющим словом в предложной конструкции на русском, т. е. сравниваем их векторные представления. В качестве меры сходства нами берется косинусный коэффициент:

$$cosine\ sim(x, y) = \frac{x \times y}{||x||\ ||y||} \quad (3)$$

Косинусная схожесть вычисляется в формуле для двух векторов x и y. $||x||$ и $||y||$ в Формуле 3 представляют собой евклидовы нормы для, соответственно, векторов $x = (x_1,..., x_n)$ и $y = (y_1,..., y_n)$. Если значение косинусной схожести для двух векторов равно 0, это указывает нам на то, что векторы ортогональны, или перпендикулярны, по отношению друг к другу на плоскости (т. е. угол между ними равен 90°) и не имеют общих точек. Чем ближе значение коэффициента к 1, тем меньше угол между векторами и больше их схожесть и направленность. Косинусную схожесть для двух векторов в *Python* можно рассчитать при помощи функции под названием *cosine()* из библиотеки *SciPy*, предназначенной для различных сложных

математических расчетов. Преимуществом библиотеки является повышенная скорость обработки данных.

Возвращаясь к степени схожести между главным словом в русскоязычной предложной конструкции и словом-кандидатом в чешском сочетании, мы отсеиваем наименее релевантные результаты следующим образом: если значение коэффициента больше 0.31, мы ищем синтаксему или зависимое слово для этого слова (принцип поиска синтаксемы и зависимого слова без предлога мы описываем в следующем подпункте), а затем записываем его вместе с синтаксемой или зависимым словом и суммарным значением всех трех или двух элементов получившейся конструкции в словарь; если значение меньше данного порога, мы не рассматриваем конструкцию дальше. Значение 0.31 было выбрано нами после ряда экспериментов: как оказалось на практике, именно единицы с величинами выше этого порога могут быть переводными кандидатами. Такой порог с одной стороны помогает отсеять наименее подходящие результаты, с другой стороны помогает выявить менее очевидный перевод конструкции: например, «*лежат на высотах*»: «*leží v horách*». Суммарное значение коэффициентов затем помогает нам выбрать наиболее вероятный эквивалент для предложной конструкции: в качестве переводного эквивалента на чешском мы берем конструкцию с максимальной суммарной величиной коэффициента по всем трем или двум элементам конструкции.

Если на выходе этого этапа мы получаем пустой словарь, это указывает нам на то, что главное слово не было найдено таким способом, либо потому, что управляющее слово в русскоязычной конструкции не было глаголом или не стояло в форме среднего залога, либо потому, что глагол в параллельном чешском предложении не связан синтаксически с частицей «*se*» или не обладал достаточной векторной схожестью с главным словом в русской конструкции.

В таком случае, мы рассматриваем другой подход. В языковой цепочке на чешском мы ищем все слова той же части речи, что и главное слово в словосочетании на русском; при этом, в силу того, что синтаксическая организация в русском и чешском зачастую совпадает, данное слово должно находиться в контекстном окне ±5 от индекса, соответствующего главному русскоязычному слову. Затем мы добавляем вектор для леммы слова-кандидата; далее, при помощи меры косинусной близости мы считаем схожесть слова-кандидата и управляющего слова в русской конструкции. Если данное значение выше 0.29, мы ищем синтаксему или только зависимое слово в случае беспредложной связи и добавляем в словарь полную конструкцию вместе с суммарным значением схожести, собранном из значений схожести для каждого отдельного слова с его переводными эквивалентами.

Если на данном этапе вместо значения коэффициента программа выдает нам *NaN* (т. е. *Not a value*), вероятнее всего, либо вектор для леммы слова-кандидата, либо вектор для леммы главного слова в русскоязычной конструкции не был найден в векторной модели *fasttext*. В таком случае, мы сравниваем векторы словоформ: если значение коэффициента схожести достигает хотя бы 0.3, мы ищем синтаксему или зависимое слово без предлога и также добавляем конструкцию в словарь.

Выявление предложно-падежной синтаксемы. Как только мы тем или иным способом выявили главное слово в переводном эквиваленте, мы можем приступить к поиску эквивалентов для предлога и падежной формы, представленной одной из знаменательных частей речи, чаще всего существительным или местоимением-существительным. Для этой цели нами было разработано два способа.

В первом случае, синтаксическую связность всех элементов предложной конструкции можно проследить по разметке автоматическим инструментов *UDpipe*. Мы отталкиваемся от поиска предлога, т. е. токена, маркированного

категорией «*ADP*»; как только мы находим предлог, в поле синтаксически связанного с ним элемента мы сможем увидеть слово, чью падежную форму предлог дополняет на синтаксическом уровне. Данное слово, если оно находится в списке частеречных эквивалентов для зависимого слова в русскоязычной конструкции, проверяется нами на предмет синтаксической связанности с уже найденным главным словом в конструкции на чешском. Для этого мы опять же смотрим на то поле разметки, где указан индекс зависимого или управляющего слова. Если между ними установлены синтаксические отношения, мы наконец вычисляем степень схожести векторов предлога и зависимого слова в потенциальной чешской конструкции с векторами предлога и зависимого слова в русскоязычной конструкции. Данное условие накладывается в связи с тем, что к главному слову могут быть присоединены несколько синтаксем. Если значение коэффициента схожести для обоих слова превышает 0.2, мы записываем в словарь предложно-падежную синтаксему с суммарным значением по обоим словам. На следующем шаге мы выбираем из словаря ту синтаксему, которая обладает наибольшим значением и объединяем ее с главным словом. Обычно в итоговом словаре у нас оказывается всего лишь одна синтаксема, но в некоторых случаях мы вынуждены выбирать из двух и более.

Если предложно-падежная синтаксема не была найдена таким способом, мы ищем синтаксему с достаточной степенью семантической схожести, но без прямой связи с главным словом в предложной конструкции. Идея такого подхода связана с тем, что при переводе перестановка лексем местами между собой может привести к тому, что теггер несколько иначе строит синтаксическое дерево предложения. Для иллюстрации подобного случая рассмотрим предложений на русском с параллельным ему предложением на чешском: «<...> *но Гурам Джохадзе — судьба его бережет! — успевает вырваться из-под обстрела, поворачивает вспять и благодаря своему могучему коню уносится вдоль берега по зарослям*»: «<...> *ale Gurama Džochadzeho osud uchrání, vyklouzne z palby, vrací se a díky svému neúnavnému*

koni prchá křovinami podél břehu». При посредстве автоматического инструмента для разметки *UDpipe* мы можем выделить три синтаксически связанные с глаголом «*уносится*» конструкции: «*уносится вдоль берега*», «*уносится по зарослям*», «*уносится благодаря коню*». В то же время, с опорой на аннотирование предложения на чешском с глаголом «*prchá*» можно найти следующие словосочетания: «*prchá díky koni*», «*prchá křovinami*». Предложно-падежная синтаксема «*podél břehu*», в силу своей присловной позиции к существительному «*křovinami*», присоединяется теггером к «*křovinami*» через тип связи *«nmod*». Таким образом, в предложении на чешском синтаксема служит как уточнение к существительному, а ее связь с глаголом никак не отображена. Для того, чтобы выровнять подобные случаи, мы точно так же, как при выявлении синтаксически связанной с глаголом синтаксемы, начинаем с поиска предлога и зависимого слова в предложной конструкции. Затем, мы вычисляем степень семантической близости их векторов с векторными представлениями предлога и зависимого слова в конструкции на русском: если значения коэффициента выше, соответственно, 0.4 и 0.3, мы записываем их в словарь и после окончания просмотра всего предложения выбираем синтаксему с максимальным суммарных значением коэффициентов ее частей.

Поиск зависимого слова без предлога. Если поиск синтаксемы не принес нам никаких результатов, вероятнее всего, зависимое слово присоединяется к главному без предлога, например, «*čtyři posluchači*» («*четверо из публики*»), «*plakala láskou*» («*плакала от любви*»). В таком случае, мы ищем в предложении на чешском все слова, которые эквивалентны по части речи зависимому слову в русскоязычной конструкции, и которые напрямую синтаксически присоединены тем или иным образом к уже определенному главному слову на чешском; мы снова просчитываем степень семантической близости между группами слов и принимаем за пороговое значение 0.2. Затем мы по тому же принципу, что и раньше, выбираем слово с наибольшим коэффициентом. Если и на этапом шаге у нас

нет никакого результата, вероятно, конструкция переводится однословно, например, в нашем корпусе встретился как минимум один такой случай: «*двое из них*»: «*dva*».

Поиск конструкции без учета схожести синтаксиса. Наконец, стоит упомянуть про последний подход к выравниванию предложных конструкций: он заключается в том, что мы ищем кандидат на место главного слова переводной конструкции в любом месте предложения, без учета потенциально схожей синтаксической структуры. Данный подход удовлетворительно показал себя в тех редких случаях, когда порядок слов в предложении на чешском сильно отличался от порядка слов в предложении на русском: обычно в качестве предложений с разной синтаксической организацией выступали фрагменты художественных текстов; это были сложные предложения, в состав которых входили придаточные части, причастные и деепричастные обороты и т. д.

Слово в произвольном месте текста также должно преодолеть порог схожести, чтобы мы включили его в список кандидатов: этим порогом является значение 0.3. Затем, по схеме, представленной нами выше, мы ищем в предложении предложно-падежную синтаксему или зависимое слово и выдаем результат пользователю.

Если переводной эквивалент для предложной конструкции не был обнаружен, т. е. выравнивание не произведено, пользователю выдается пустая строка.

3.6. Оценка качества работы алгоритма и анализ результатов

Для оценки качества работы алгоритма мы применили его для выравнивания собранного нами параллельного русско-чешского корпуса. Все предложения были заранее размечены нами вручную с целью последующего сопоставления результатов ручной разметки и разметки, проведенной нашим

алгоритмом. В составленном собрании текстов нами были встречены 342 русскоязычные конструкции с такими предлогами, как «*без*», «*благодаря*», «*в(о)*», «*вдоль*», «*вследствие*», «*для*», «*до*», «*за*», «*из*», «*из-за*», «*из-под*», «*к*», «*на*», «*о*», «*от*», «*перед*», «*по*», «*после*», «*посреди*», «*при*», «*про*», «*против*», «*с*», «*среди*», «*через*», «*у*». Для оценки качества работы мы используем такие метрики, как **полнота**, **точность** и **F-мера**.

Мера **точности** отражает число корректных выравниваний по отношению ко всем выравниваниям на корпусе:

$$precision = \frac{tp}{tp + fp} \quad (4)$$

Переменная tp в формуле означает количество правильно выявленных конструкций; fp — количество неверных результатов; сумма этих переменных дает нам все возможные выравнивания, которые были выявлены при автоматической обработке корпуса.

Мера **полноты**, напротив, отражает число корректных выравниваний по отношению ко всем потенциальным выравниваниям в составе корпуса, т.е. мы сравниваем правильно выровненные предложные конструкции одновременно и против неправильно выровненных, и против не выровненных вообще:

$$recall = \frac{tp}{tp + fn} \quad (5)$$

В данной формуле, fn соответствует количество всех потенциальных выравниваний, которые мы могли бы получить на примере рассматриваемого нами корпуса.

Наконец, **F-мера** выражает баланс между результатами выдачи метрик полноты и точности:

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \quad (6)$$

При ручной разметке корпуса возникла проблема с определением того, какую конструкцию можно считать верно выровненной, а какую — некорректно выровненной. В связи с этим, нами были отдельно оговорены следующие два случая:

1. Если сама предложная конструкция построена верно, т. е. включает в себя правильные главное слово, предлог и зависимое слово, но не включает в себя частицу «*se*», когда у главного слова в конструкции на русском не было выявлено формы среднего залога, мы даем такой конструкции 0. 5 балла;
2. Если синтаксема построена верно, но главное слово не выровнено верно вследствие того, что переводной эквивалент отсутствует в предложении на чешском языке, такая конструкция также получает 0. 5 балла.
3. В одном случае предложная конструкция была оценена нами на 0. 5 балла, потому что была выровнена концептуально правильно, хотя в чешском предложении не было точного лексического аналога существительному из русскоязычного предложения.

Во других случаях, выравнивание оценено либо 0, либо 1.

Оценка результатов в соответствии с представленными метриками приведена в Таблице 2:

	Точность	**Полнота**	**F-мера**
Наш алгоритм	75,8%	70,8%	73,2%

Таблица 2. Оценка алгоритма с точки зрения полноты, точности и F-меры.

Полные результаты выравнивания нашего корпуса представлены в Приложении 1, Приложении 2 и Приложении 3. В этом разделе мы постараемся выявить общую тенденцию: с какими предложениями алгоритм справился хорошо, а какие вызвали у него затруднения.

Как мы можем увидеть, проанализировав метрики, показатели выравнивания достаточно высокие. Алгоритм достаточно хорошо работает на предложных конструкциях, в состав которых входят однословные предлоги, и справляется с теми случаями, когда первообразный односоставный предлог переводится производным односоставным, и наоборот. Тем не менее, что ожидаемо, представленный нами алгоритм не захватывает производные предлоги, состоящие из двух и более элементов. Помимо этого, существуют проблемы с выравниванием, когда в предложении на чешском языке нет подходящих лексических единиц или другой синтаксис, например, вместо конструкции с предлогом идет придаточное предложение.

3.7. Выводы

В Главе 3 нами был создан алгоритм для выравнивания предложных конструкций в составленном нами параллельном русско-чешском корпусе. Алгоритм, предложенный нами, опирается на правила, примененные к синтаксическим деревьям, наборам морфологических категорий и степени семантической схожести между словами, выявленной при сопоставлении векторов словоформ.

Разработанная методика достаточно хорошо показала себя на корпусе; для оценки качества работы мы использовали такие метрики, как мера точности, мера полноты и F-мера. Показатель точности достиг 75,8%, что является качественным результатом.

Заключение

В данной выпускной квалификационной работе нами рассматривается проблема выравнивания предложно-падежных конструкций в параллельном русско-чешском корпусе. Благодаря изучению и разработке алгоритмов сопоставления словосочетаний и конструкций мы можем выявить закономерности и особенности перевода тех или иных сегментов оригинального текста на целевой язык.

В данной работе нами изучены теоретические особенности выравнивания и различных подходов к нему; рассмотрены границы предлога и предложной конструкции с точки зрения разных лингвистических уровней, а также изучены характерные особенности и закономерности при выделении предложных конструкций в параллельных русско-чешских корпусах.

В практической части мы представляем алгоритм для выравнивания подобных предложно-конструкций. Для проверки качества работы алгоритма нами был специально собран корпус, насчитывающий 200 предложений и 342 предложных конструкции на русском, для которых были найдены эквивалентные сочетания в чешском сегменте корпуса. Данные были размечены вручную; затем результат ручной разметки был сопоставлен с результатом выдачи разработанного нами алгоритма: показатель точности достиг 75,8%, что является достаточно хорошим показателем.

Алгоритм может успешно выравнивать предложные конструкции в параллельных русско-чешских корпусах. Тем не менее, есть дальнейшие возможности для развития: так, например, можно разработать правила для выравнивания предложных конструкций с составным предлогом, а также прописать больше правил для улучшения уже существующего алгоритма выделения сочетаний с односоставными предлогами. Помимо этого, можно попробовать применить алгоритм на других родственных языковых парах с

опорой на разметку инструментов *UDpipe* и выровненных векторных представлений слов, доступных в рамках проекта *fasttext*.

Список литературы

1. Азарова, И.В., Захаров, В.П., Москвина, А.Д. Семантическая структура русских предложно-падежных конструкций. // Компьютерная лингвистика и вычислительные онтологии, Т. 2, 2018. – С. 9-16. – [Электронный ресурс] URL: http://ojs.itmo.ru/index.php/CLCO/article/view/824 (дата обращения: 31 мая 2021).

2. Виноградов, В.В. Русский язык. – М., 1972.

3. Всеволодова, М.В., Кукушкина, О.В., Поликарпов, А.А. Русские предлоги и средства предложного типа. Материалы к функционально-грамматическому описанию реального употребления. Книга 1. Введение в объективную грамматику и лексикографию русских предложных единиц. – М.: Книжный дом "Либроком", 2014.

4. Захаров, В.П., Михайлова, В.Д. Контекстная грамматика предложных конструкций русского языка // Компьютерная лингвистика и вычислительные онтологии, Выпуск 1., 2017. – С. 57–71. – [Электронный ресурс] URL: https://ojs.itmo.ru/index.php/CLCO/article/view/842 (дата обращения: 31 мая 2021).

5. Захаров, В.П. О компьютерной онтологии русских предлогов. // Российская академическая лексикография: Современное состояние и перспективы. Сборник статей по материалам конференции. – Санкт-Петербург: Нестор-История, 2018.

6. Захаров, В.П. Comparative Corpus-driven Study of Prepositional Semantics. – Санкт-Петербург, 2020.

7. Золотова, Г.А. Синтаксический словарь: Репертуар элементарных единиц русского синтаксиса. // Изд. 4-е. – М.: Едиториал УРСС, 2011.

8. Малый академический словарь // РАН, Ин-т лингвистич. исследований. [под редакцией А.П. Евгеньевой] – Изд. 4-е. – М.: Рус. яз.; Полиграфресурсы, 1999.

9. Лингвистический энциклопедический словарь. // Институт языкознания АН СССР. [под редакцией В.Н. Ярцевой и др.] – М.: Советская энциклопедия, 1990.

10. Потемкин, С.Б., Кедрова, Г.Е. Выравнивание неразмеченного корпуса параллельных текстов. // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Междунар. конф. "Диалог-2008", Вып. 7 (14). – М.: РГГУ, 2008. – С. 431–436. – [Электронный ресурс] URL: http://www.dialog-21.ru/digests/dialog2008/materials/pdf/67.pdf (дата обращения: 31 мая 2021).

11. Потемкин, С.Б. Проблемы разработки параллельного корпуса переводов русской классики // Научно-информационный журнал Армия и общество, № 2. – М.: МГУ, 2012. –

С. 138-145. – [Электронный ресурс] URL: https://istina.msu.ru/publications/article/2791736/ (дата обращения: 31 мая 2021).

12. Русская грамматика [в 2 т.] // Акад. наук СССР, Ин-т рус. яз. [под редакцией Н.Ю. Шведовой (гл. ред.) и др.]. – М.: Наука, 1980.

13. Сичинава, Д.В. Параллельные тексты в составе национального корпуса русского языка: новые направления развития и результаты. // Труды Института русского языка РАН, № 6. – М.: Институт русского языка им. В.В. Виноградова РАН, 2015. – С. 194-235. – [Электронный ресурс] URL: http://ruslang.ru/doc/sitchinava/sitchinava-2015-parallel_corpus.pdf (дата обращения: 31 мая 2021).

14. Энциклопедический словарь Брокгауза и Ефрона. – М.: Терра, 2001.

15. Bahdanau, D., Cho, K., Bengio, Y. Neural machine translation by jointly learning to align and translate. // ICLR, 2014. – [Электронный ресурс] URL: https://arxiv.org/abs/1409.0473 (дата обращения: 31 мая 2021).

16. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T., Enriching Word Vectors with Subword Information. // Transactions of the Association for Computational Linguistics, Volume 5, 2017. – С. 135–146. – [Электронный ресурс] URL: https://www.aclweb.org/anthology/Q17-1010/ (дата обращения: 31 мая 2021).

17. Brown, P., Della Pietra, S., Della Pietra, V., Mercer, R. The mathematics of machine translation: Parameter estimation. // Computational Linguistics, Volume 19, Special Issue on Using Large Corpora: II, 1993. – С. 263-311. – [Электронный ресурс] URL: https://www.aclweb.org/anthology/J93-2003/ (дата обращения: 31 мая 2021).

18. Brown, P., Lai, J., Mercer, R. Aligning sentences in parallel corpora. // Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics, 1991. – С. 169-176.

19. Buchholz, S., Marsi, E. CoNLL-X shared task on Multilingual Dependency Parsing. // Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X). – New York City, USA.: Association for Computational Linguistics, 2006. – С. 149-164. – [Электронный ресурс] URL: https://www.aclweb.org/anthology/W06-2920/ (дата обращения: 31 мая 2020).

20. Čermák, F, Rosen, A. The Case of InterCorp, a multilingual parallel corpus. // International Journal of Corpus Linguistics, 2012 – С. 411-427. – [Электронный ресурс] URL: https://www.researchgate.net/publication/234118417_The_Case_of_InterCorp_a_multilingual_parallel_corpus (дата обращения: 31 мая 2021)

21. de Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., Manning, C.D. Universal Stanford Dependencies: A cross-linguistic typology. // Proceedings of

the Ninth International Conference on Language Resources and Evaluation (LREC'14). – Reykjavik, Iceland: European Language Resources Association (ELRA), 2014. – С. 4585–4592. – [Электронный ресурс] URL: https://www.aclweb.org/anthology/L14-1045/ (дата обращения: 31 мая 2021)

22. Dyer, C., Chahuneau, V., Smith, N. A Simple, Fast, and Effective Reparameterization of IBM Model 2. // Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. – Atlanta, Georgia: Association for Computational Linguistics, 2013. – С. 644–648. – [Электронный ресурс] URL: https://www.aclweb.org/anthology/N13-1073/ (дата обращения: 31 мая 2021).

23. Gale, W., Church, K. A Program for Aligning Sentences in Bilingual Corpora. // Computational Linguistics, Volume 19, Number 1, Special Issue on Using Large Corpora: I, 1993. – С. 75-102. – [Электронный ресурс] URL: https://www.aclweb.org/anthology/J93-1004/ (дата обращения: 31 мая 2021).

24. Joulin, A., Bojanowski, P., Mikolov, T., Jegou, H., Grave, E., Loss in Translation: Learning Bilingual Word Mapping with a Retrieval Criterion. // In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. – Berlin, Belgium: Association for Computational Linguistics, 2018. – С. 2979–2984. – [Электронный ресурс] URL: https://www.aclweb.org/anthology/D18-1330/ (дата обращения: 31 мая 2021).

25. Kay, M., Roscheisen, M. Text-Translation Alignment. // Computational Linguistics, Volume 19, Number 1, Special Issue on Using Large Corpora: I, 1993. – С. 121-142. – [Электронный ресурс] URL: https://www.aclweb.org/anthology/J93-1006/ (дата обращения: 31 мая 2021).

26. Ker, S.J., Chang, J.S. Aligning More Words with High Precision for Small Bilingual Corpora. // International Journal of Computational Linguistics & Chinese Language Processing, Volume 2, Number 2, August 1997. – С. 63-96. – [Электронный ресурс] URL: https://www.aclweb.org/anthology/O97-4004/ (дата обращения: 31 мая 2021).

27. Li, X., Li, G., Liu, L., Meng, M., Shi, S. On the Word Alignment from Neural Machine Translation // In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. – Florence, Italy: Association for Computational Linguistics, 2019. – С. 1293–1303. – [Электронный ресурс] URL: https://www.aclweb.org/anthology/P19-1124/ (дата обращения: 19 мая 2021).

28. Liu, L., Utiyama, M., Finch, A., Sumita, E. Neural machine translation with supervised attention. // Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. – Osaka, Japan: The COLING 2016 Organizing Committee, 2016.

– [Электронный ресурс] URL: https://www.aclweb.org/anthology/C16-1291/ (дата обращения: 31 мая 2021).

29. Ma, Y., Ozdowska, S., Sun, Y., Way, A. Improving Word Alignment Using Syntactic Dependencies. // Proceedings of the ACL-08: HLT Second Workshop on Syntax and Structure in Statistical Translation (SSST-2). – Columbus, Ohio: Association for Computational Linguistics, 2008. – С. 69–77. – [Электронный ресурс] URL: https://www.aclweb.org/anthology/W08-0409/ (дата обращения: 31 мая 2021)

30. Mi, H., Wang, Z., Ittycheriah, A. Supervised attentions for neural machine translation. // In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. – Austin, Texas: Association for Computational Linguistics, 2016. – С. 2283–2288. – [Электронный ресурс] URL: https://www.aclweb.org/anthology/D16-1249/ (дата обращения: 31 мая 2021).

31. Och, F., Ney, H. A systematic comparison of various statistical alignment models. // Computational Linguistics, Volume 29, Number 1, 2003. – С. 19-51. – [Электронный ресурс] URL: https://www.aclweb.org/anthology/J03-1002/ (дата обращения: 31 мая 2021).

32. Östling, R., Tiedemann, J. Efficient word alignment with Markov Chain Monte Carlo // The Prague Bulletin of Mathematical Linguistics, Volume 106, 2016. – С. 125–146. – [Электронный ресурс] URL: https://ufal.mff.cuni.cz/pbml/106/art-ostling-tiedemann.pdf (дата обращения: 31 мая 2021)

33. Sabet, M. J., Dufter, P., Yvon, F., Schütze, H. SimAlign: High Quality Word Alignments without Parallel Training Data using Static and Contextualized Embeddings. // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings. – Association for Computational Linguistics, 2020. – С. 1627–1643. – [Электронный ресурс] URL: https://www.aclweb.org/anthology/2020.findings-emnlp.147/ (дата обращения: 31 мая 2021).

34. Straka, M., Straková, J. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. // In Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. – Vancouver, Canada: Association for Computational Linguistics, 2017. – С. 88–99. – [Электронный ресурс] URL: https://www.aclweb.org/anthology/K17-3009/ (дата обращения: 31 мая 2021).

35. Pianta, E., Bentivogli, L. Knowledge Intensive Word Alignment with KNOWA. // COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics. – Geneva, Switzerland: COLING, 2004. – С. 1086–1092. – [Электронный ресурс] URL: https://www.aclweb.org/anthology/C04-1156/ (дата обращения: 31 мая 2021).

36. Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., Nagy, V. Parallel corpora for medium density languages. // In Proceedings of the RANLP 2005, 2005. – С. 590–596. – [Электронный ресурс] URL: https://www.researchgate.net/publication/282780901_Parallel_corpora_for_medium_density_languages (дата обращения: 31 мая 2021)

37. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł., Polosukhin, I. Attention is all you need. // In Advances in Neural Information Processing Systems, 2017 – С. 5998–6008. – [Электронный ресурс] URL: https://arxiv.org/abs/1706.03762 (дата обращения: 31 мая 2021).

38. Vogel, S., Ney, H., Tillmann, C. HMM-Based Word Alignment in Statistical Translation. // COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics, 1996. – С. 836-841. – [Электронный ресурс] URL: https://www.aclweb.org/anthology/C96-2141/ (дата обращения: 31 мая 2021).

39. Yamada, K., Knight, K. A Syntax-based Statistical Translation Model. // Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics. – Toulouse, France: Association for Computational Linguistics, 2001. – С. 836-841. – [Электронный ресурс] URL: https://www.aclweb.org/anthology/P01-1067/ (дата обращение: 31 мая 2021).

Приложение 1. Релевантные результаты выравнивания предложных конструкций

Предложная конструкция на русском	Предложная конструкция на чешском
четверо из публики	čtyři posluchači
умерли от кровоизлияния	zemřeli na krvácení
выжил благодаря тому	přežil díky tomu
устремление к совершенству	tíhnutí k dokonalosti
подозревая благодаря мечтам	tušil díky snům
утруждать при встрече	zatěžovat při setkání
захлебываясь в реке	topí se v řece
вырваться из-под обстрела	vyklouzne z palby
уносится благодаря коню	prchá díky koni
уносится вдоль берега	prchá podél břehu
уносится по зарослям	prchá křovinami
вычитал в газете	vyčetl z časopisů
лежат на высотах	leží v horách
несут в низины	stékají do nížin
все в природе	všecko v přírodě
спасся благодаря контейнеру	zachránil se díky kontejneru
контейнеру с мусором	kontejneru s odpadky
залечил благодаря заботе	uzdravil díky péči
существовал благодаря тому	existovalo díky tomu
вляпалась в историю	uvízla v příběhu
вляпалась благодаря особенностям	uvízla díky vlastnostem
завязались благодаря тамплиерам	navázal díky templářům
тамплиерам в баре	templářům v bistru
тамплиерам в эпоху	templářům v roce

анализ среди населения	analýza u populace
вакцинация от гриппа	očkování proti chřipce
полезна для защиты	úspěšné v ochraně
сочинял в эпоху	psal v roce
полезна от заражения	úspěšné proti chřipkou
ходишь в школу	chodíš do školy
возвращаюсь в париж	vrátím se do paříže
лекарства от болезни	léky proti chorobě
магазин на улице	obchod v ulici
играть на концерте	hrát na koncertu
ходили в кино	chodili do kina
ходили на прогулки	chodili na procházky
прогулки к морю	procházky k moři
прогулки в город	procházky do města
пошли в кино	šli do kina
пошли на детство	šli na dětství
показались на небе	ukázaly se na nebi
производилось из нее	vyráběla z té
сделал из рассеянности	udělal z roztržitosti
проснулся от холода	probudil se zimou
сделала из-за страха	udělala ze strachu
вскрикнула от испуга	vykřikla úlekem
черный от сажи	černý od sazí
проходит вследствие маневра	projde po manévru
достигают до млрд	dosahují až miliard
млрд в год	miliard ročně
недостатков в фундаменте	vadám v základech
выросла в кении	vyrostla v keni
следят вследствие связи	sledují kvůli vazbě

связи с англией	vazbě na anglii
возникла благодаря торговле	vznikla díky obchodu
удивлять из века	udivují po staletí
удивлять в век	udivují po staletí
жил в минуты	žil v minutách
жил на свете	žil na světě
было благодаря ему	existovalo díky jemu
окаменел от неожиданности	zkameněl překvapením
подал через минут	přiběhl za pár
начнем через минуту	začneme za chviličku
построена через сердце	vystavěná skrz srdce
умирайте от любви	umírejte láskou
плакала от любви	plakala láskou
любви к нему	láskou k němu
ничего на планете	nic na planetě
поводом для оптимизма	zdrojem optimismu
позвонил благодаря фотографиям	telefonoval díky fotografiím
позвонил после лет	telefonoval po letech
упорствовать в поведении	zraňovat chováním
двое из них	dva z nich
двое из них	dva
ушли по домам	odešli k rodinám
продолжала в моменты	šla v okamžicích
двигаться благодаря таланту	šla díky schopnostem
видел благодаря ему	viděl díky jemu
напоминает в отношении	připomíná v ohledu
то благодаря которому	taková díky kterému
бродил среди корпусов	obcházel vraky
отношении к правительству	názorem na vládu

стареют среди стран	udržely mezi státy
пережил за жизнь	prožil za let
пережил благодаря долготерпению	přežil díky trpělivosti
проводил до дома	doprovodil domů
отходит от дороги	odbočuje z silnice
трамвай с шапкой	tramvaj s pokrývkou
шапкой на крыше	pokrývkou na střeše
поехал на курсе	jel ve ročníku
поехал на раскопки	jel na vykopávky
поехал из-за этого	jel kvůli tomu
поднимался по улочкам	příkrými uličkami
поднимался вдоль фасадов	příkrými podél traktů
ждал из-за туч	čekal z mraků
поющего о любви	zpívající o lásce
держит за кулисами	drží za kulisami
палец на выключателе	prst na vypínači
работать без денег	pracovat bez peněz
ходит без меня	chodí beze mě
прибыли без приглашения	jsme bez ohlášení
летим без тебя	letíme bez tebe
останемся без горючего	zůstaneme bez hmot
приспосабливаться к нагрузкам	přizpůsobit se tlakům
развиваться благодаря поведению	vyvíjet se kvůli chování
находится в состоянии	nachází se ve stavu
находится благодаря дефициту	nachází se kvůli schodkům
смешалась с печалью	smísila se se smutkem
утонула в ней	utonula v něm
послышалось из-под стола	ozvalo se zpod stolu
стреляли в меня	stříleli na mě

стреляли из-под вагонов	stříleli zpod vagónů
вывести из-под удара	dostat z ohrožení
следил за мной	pozoroval mě
следил из-под листвы	pozoroval zpod listí
достали из-под подушки	vytáhli zpod polštáře
повоевал против французов	zaválčil proti francouzům
выступили против плана	říkám proti plánu
было благодаря ему	existovalo díky jemu
ничего против вас	nic proti vám
играем через месяц	hrajeme za měsíc
возникла благодаря торговле	vznikla díky obchodu
общается с помощью	komunikuje se s pomocí
общается с собой	komunikuje se se sebou
находит в себе	nachází v sobě
все на земле	všechno na zemi
совершается благодаря ему	děje se díky jemu
окружение в войне	obklíčení ve válce
стало благодаря подвижности	stalo se díky pohyblivosti
опирается на которую	umožňuje která
живет благодаря им	žije díky jim
живет в гармонии	žije v harmonii
гармонии с народом	harmonii s lidem
поверили с опозданием	uvěřili s zpožděním
существуют в россии	jsou v rusku
осознал спустя годы	uvědomil po letech
осознал в время	uvědomil v době
клялась в верности	zapřísahala se věrností
зашла благодаря обстоятельствам	stavila se díky příležitosti
отвоевали благодаря руководству	získali díky vedení

пользовались до войны	svítili před válkou
люди в городах	lidé ve městech
топили в деревнях	topili na venkově
пережила благодаря тому	přežila díky tomu
играла на скрипке	vyhrávala na housle
играла на площади	vyhrávala na nástupišti
играла в штрутхофе	vyhrávala v dachau
арии из вдовы	árie z vdovy
умирало в войны	umíralo v válkách
войны благодаря помощи	války díky pomoci
прогрессу в хирургии	pokrokům v chirurgii
видел благодаря ему	viděl díky jemu
напоминает в отношении	připomíná v ohledu
то благодаря которому	taková díky kterému
пробовал в жизни	pokoušel v životě
заронила в меня	roznítila ve mně
существует благодаря единству	drží díky úsilí
писал на досуге	psal ve chvílích
опубликованный в венеции	vyšel v benátkách
опубликованный в годах	vyšel v letech
второе из изданий	druhé z vydání
перемены в мире	změny na světě
происходят благодаря группе	probíhají díky skupině
ответить на вопрос	odpověděl na otázku
существую благодаря миру	existuji díky světu
существует благодаря мне	existuje díky mně
произошедших в годах	utrpěli v letech
распознал в тебе	viděl v tobě
попали благодаря таланту	dostal díky talentu

победил в соревновании	zvítězila ve přeboru
беженцы из уголков	emigrantů ze koutů
говорившие на идиш	hovořících jidiš
висел на стене	visela na stěně
смотрел на него	zůstal na ni
оттенок в голосе	tepla v hlase
обратился к ней	hovořil k ní
отобранной из скопища	vyzvedla z roje
оказалась благодаря звезде	octla se díky štěstí
оказалась в числе	octla se mezi skupinkou
сумела благодаря изобретательности	podařilo díky šikovnosti
обрела благодаря тишине	vyspat díky klidu
обрела за месяцы	vyspat za měsíce
избегли благодаря умению	unikaly díky umění
знала благодаря знакомству	zjistila díky známosti
знакомству с вандой	známosti s wandou
знала про селекцию	věděla o selekci
сел к столику	sedl k stolku
столику у рояля	stolku vedle klavíru
заказал у официантки	objednal u servírky
высилась на берегу	postaven na nábřeží
высилась у моста	postaven vedle mostu
висела у двери	visela vedle dveří
заглянул через плечо	nahlédl přes rameno
взял с собой	vzal s sebou
поднялась на метров	stoupla o metrů
переливалась через мост	přelévala se přes most
закрылся в кабине	zamkl se ve kabině
разговаривал сквозь дверь	mluvil skrz dveře

стоя на валуне	seděla na balvanu
выдавил сквозь зубы	procedil skrz zuby
скрывшийся под ней	zmizel pod ní
примеряла перед зеркалом	zkoušela před zrcadlem
полянка посреди дубов	palouk mezi duby
вытянула перед собой	natáhla před sebe
букет из перьев	kytici z pírek
равенство перед законом	rovnost před zákonem
товарищ по оружию	bratr ve zbrani
предстали перед судьями	předstoupili před soudce
встречаются перед заседанием	setkávají se před schůzkou
появился из-за ширмы	vynořil se zpoza paravánu
вышел из-за угла	vyšel zpoza rohu
направился к мосту	zamířil k mostu
выскочила из-за угла	přiběhla zpoza rohu
постучав в хижину	ťukal na dveře
видит сквозь землю	vidí skrz zem
стояли у трибуны	stáli vedle tribuny
слышна через полотно	slyšet přes plátno

Приложение 2. Релевантные результаты, в которых пропущен один элемент

Предложная конструкция на русском	Предложная конструкция на чешском	Корректная предложная конструкция
заживала благодаря уходу	hojilo díky péči	hojilo se díky péči
узнал о нем	dozvěděl o něm	dozvěděl se o něm
узнал благодаря книгам	dozvěděl díky knihám	dozvěděl se díky knihám
узнали благодаря постэлю	dozvěděli díky postelově	dozvěděli se díky postelově
узнали о тайне	dozvěděli o tajemství	dozvěděli se o tajemství
стемнело на улице	setmělo venku	setmělo se venku
попали благодаря таланту	dostal díky talentu	dostal se díky talentu
удалась вследствие недостатков	selhává vzhledem vadám	selhává vzhledem k vadám
соберёт через недели	sejde za týdny	sejde se za týdny
соберёт в париже	sejde v paříži	sejde se v paříži
вошла в историю	dostala do legendy	dostala se do legendy
вошла благодаря записям	dostala díky zápiskům	dostala se díky zápiskům
вынырнул из-за поворота	vynořila v ohbí	vynořila se v ohbí
попал в плен	dostal do zajetí	dostal se do zajetí
попал к голландцам	dostal k holanďanům	dostal se k holanďanům
следит за ветром	neohlíží po větru	neohlíží se po větru
должны благодаря которому	novém díky které	díky které
должны в столетии	novém v století	v století
вошла в историю	dostala do legendy	dostala se do legendy
представляет благодаря комментарию	poznamenala díky komentáři	díky komentáři
возможна благодаря свойству	geniální díky vlastnosti	díky vlastnosti
знаем благодаря описаниям	víme díky mužům	В предложении на чешском

		нет эквивалента слова «*описаниям*», подобная конструкция строится со словом «*людям*»
попали благодаря усилиям	dostali díky úsilí	dostali se díky úsilí
попали в сферу	dostali do péče	dostali se do péče
взлетели благодаря существованию	myslel díky existenci	díky existenci
поглядев на потолок	dívala na strop	dívala se na strop
выскочивший из-за угла	vyřítil zpoza rohu	vyřítil se zpoza rohu

Приложение 3. Нерелевантные результаты выравнивания

Предложная конструкция на русском	Найденный эквивалент на чешском	Корректный перевод
могу без них	tykat	nemohu bez nich
рассказывал о кошках	dozvěděl	příběhy o kočkách
отношения с якопо	styky	styky s jacopem
раз в год	-	ročně
играть в честь	hrát na koncertu	Нет лекс. эквивалента
темнело на улице	setmělo	setmělo se venku
проходит в острог	projde manévru	Нет лекс. эквивалента
затраты на здоровье	náklady v znečištění	zdravotní náklady
затраты вследствие загрязнения	náklady v znečištění	náklady v důsledku znečištění
достигают в ес	dosahují	dosahují v eu
следят за премьер-лигой	sledují keni	sledují premier league
продолжают благодаря простоте	udivují	udivují díky jednoduchost
сходите с ума	plačte	blázněte
сходите от радости	plačte	blázněte radostí
двое с орденами	dva fronťáci	dva s vyznamenáními
заявили среди консерваторов	činil podíl	Нет точного лекс. эквивалента

консерваторов о отношении	konzervativců	Нет точного лекс. эквивалента
уровень среди обладателей	lidmi vzděláním	mezi lidmi
обладателей среди женщин	-	mezi ženami
тропинка к посту	stezka	stezka na hlásku
курсе в марокко	-	Некорректная констр. на русском
сверну на перекрестке	-	dám se za příštím rohem
любви в держит	lásce drží	Некорректная констр. на русском
выглянули из-за хаты	-	vykoukli za chalupou
ходит на вечеринки	chodí na spousty	chodí na spousty večírků
увидел без него	všiml toho	všiml také
унесло без следа	nezanechala stopu	nezanechala nejmenší stopu
превзойдет в десятилетия	převýší počtu	převýší během deseti let
впечатляющему благодаря количеству	-	Некорректная констр. на русском
представляет благодаря воздействиям	nedokáže	Нет точного лекс. эквивалента
полагаются благодаря теорией	-	Некорректная констр. на русском
могут при издержках	-	Нет точного лекс. эквивалента
блеснувшая на миг	-	svitla na okamžik
пощады от меня	-	nečekejte ode slitování
уходит у него	měl	chvěla pod nohama

уходит из-под ног	měl	chvěla pod nohama
удается благодаря фонарю	viděl pomocí	viděl díky lucerně
зримы благодаря отражению	zřetelnější přítel	Нет точного лекс. эквивалента
отражению в душе	-	Нет точного лекс. эквивалента
поселились благодаря солженицыну	-	zabydlila se díky solženicynovi
поселились в словаре	-	zabydlila se ve slovníku
может благодаря им	-	může díky jim
отвешивать с точностью	-	odvažovat s přesností
войны от болезней	válkách	Некорректная констр. на русском
больше в раз	-	Нет точного лекс. эквивалента
больше в бою	-	Некорректная констр. на русском
звук с растеклись	-	Некорректная констр. на русском
растеклись по существованию	-	Нет точного лекс. эквивалента
проникновения в остальное		prostupující život
ключом на свете	paprsek zvuk	Некорректная констр. на русском
стали благодаря тебе	roznítila ve mně	Нет точного лекс. эквивалента
утонула в ней	nespadl	Нет точного лекс. эквивалента
утонула благодаря ушам	všiml uším	Нет точного лекс. эквивалента
ускользнуть от него	uniknout mu	uniknout mohl

четыре между удавалось	-	Некорректная констр. на русском
раз на свете	-	Нет точного лекс. эквивалента
победил благодаря игре	zvítězila	Некорректная констр. на русском
победил в квиддич	zvítězila	zvítězila ve famfrpálovém poháru
опережала благодаря знанию	-	Нет точного лекс. эквивалента
постигавших с трудом	-	Нет точного лекс. эквивалента
пришлось через время	musel	se musel později
стене под стеклом	stěně	stěně za sklem
взлетели на воздух	myslel ve stínu	Нет точного лекс. эквивалента
созданной из массы	vyletěli existenci	Нет точного лекс. эквивалента
созданной в тени	vyletěli díky existenci	vyrobeného ve stínu
одного из охранников	-	stráží
отъезда из кракова	-	Нет точного лекс. эквивалента
погружаться в грезы	-	nechat kolébat se vlnami snů
руководствовались в деятельности	zacházet jehlou	Нет точного лекс. эквивалента
боялась превыше всего	dovědělo co	děsila se nevýslovně
доносился до неё	dívala se	Нет точного лекс. эквивалента
небо за окном	oblohu lijavec	Нет точного лекс. эквивалента
льющиеся по стеклу	-	Нет точного лекс. эквивалента

запыхавшаяся с улыбкой	-	dechu s potěšeným úsměvem
выскочил из-за двери	vyklouznul malfoy	vyklouznul zpoza dveří
показались за ним	vyklouznul	Нет точного лекс. эквивалента
крэбб с гойлом	crabbem	crabbem a goylem
сшиб на землю	-	ho málem povalil
сшиб от радости	-	nadšením ho povalil
вышел из-за угла	vykročil	vykročil zpoza rohu
производит на поэтов	popisuje dojmy	Нет точного лекс. эквивалента