Санкт–Петербургский государственный университет

ЭГЕРВАРИ Кристиан Эрик

Выпускная квалификационная работа

Выравнивание русских предложных конструкций и их эквивалентов в агглютинативных языках при создании параллельных корпусов

Уровень образования: магистратура

Направление 45.04.02 «Лингвистика»

Основная образовательная программа ВМ.5805 «Компьютерная и прикладная лингвистика»

Профиль «Компьютерная лингвистика и интеллектуальные технологии»

Научный руководитель:

доцент, кафедра математической лингвистики

к.ф.н. Захаров Виктор Павлович

Рецензент:

зав. Информационно-библиотечным центром

Институт транспортных наук Венгрии

к.ф.н. Фюреди Михай

Санкт-Петербург

2021

Содержание

Введение

В дисциплине корпусной лингвистики в последние годы набирают большую популярность параллельные корпусы, которые позволяют проводить различные лингвистические исследования на основе многоязычного текстового материала. В публичном доступе имеются многочисленные параллельные корпусы, например, в составе Национального корпуса русского языка, Чешского национального корпуса, системы Sketch Engine и др., тем не менее, составление многоязычных параллельных корпусов является более сложной задачей, чем составление одноязычных корпусов. Кроме типичных задач, решаемых при составлении корпусов, таких как сбор текстовых данных, лингвистическая разметка и мета-разметка материала, также встают такие задачи, как выравнивание предложений и выравнивание словосочетаний, слов или морфем внутри предложений.

Проблема автоматического выравнивания параллельных текстов является одной из более важных задач современной компьютерной лингвистики, потому что высококачественное выполнение этой задачи является предварительным условием проведении большинства компаративных исследований: для изучения того, как те или иные лингвистические явления передаются разными языками, необходимо определить, какие конструкции в одном языке каким конструкциям другого соответствуют. Соответственно, для большинства многоязычных морфологических и синтаксических исследований, выравнивание – ахиллесова пята: от качества выравнивания зависит успех последующих шагов. Кроме теоретических задачах, выравнивание играет большую роль и во многих практических задач, например, в статистическом машинном переводе.

Для выравнивания параллельных текстов существует ряд распространённых методов, но эффективность и надёжность этих методов в большой

мере зависит от морфо-синтаксического сходства и близости изучаемых языков. Для пар языков из одной и той же семьи языков, как, например, немецкий и голландский, настоящие методы достигают удовлетворительной точности и полноты, но для более далёких друг от друга языков, как русский и венгерский, проблема выравнивания параллельных текстов до сих пор не решена.

Целью настоящей работы является создание и выравнивание двуязычного, русско-венгерского параллельного корпуса, на основе которого появляется возможность провести компаративное исследование русских предложно-падежных конструкций и эквивалентных им конструкций в венгерском языке. Для достижения данной цели необходимо решать следующие **задачи**:

- изучить особенности, способы и проблемы задачи выравнивания параллельных текстов;
- изучить морфо-синтаксические особенности русского и венгерского языка с точки зрения задачи выравнивания параллельных текстов;
- создать эффективный метод для выравнивания русских предложно-падежных конструкций и их эквивалентов в венгерском языке;
- оценить эффективность выработанного метода и сравнить её с результатами уже существующих методов;
- создать классификацию русских предложных конструкций и их эквивалентов в венгерском языке на основе выровненных текстов в параллельном корпусе.

Таким образом, **объектом** данной работы являются предложно-падежные конструкции в русском языке и способы передачи значения данных конструкций в венгерском, являющийся агглютинативным языком, в ко-

тором нет предлогов. **Предметом** работы является процедурный морфосинтаксический подход к русским предложно-падежным конструкциям и их аналогам в венгерском языке.

Тема предложно-падежных конструкций русского языка в контексте корпусной лингвистики является **актуальным**, но мало исследованным морфосинтаксическим вопросом, который в фокусе исследовательского внимания может выявлять до сих пор неизвестные закономерности русского языка и других языков, в нашем случае, венгерского языка. Изучение данной темы может содействовать появлению альтернативных решений многочисленных прикладных задач обработки естественного языка, в том числе, составления параллельных корпусов, выравнивания параллельных текстов, машинного перевода и т. д.

Однако, для изучения предложно-падежных конструкций русского языка и их эквивалентов в иностранных языках, необходимо создать подходящий, выравненный параллельный корпус, который позволяет проводить нужные исследования на основе высоко-качественного лингвистического материала. Для исследуемых языков, русского и венгерского, существует несколько параллельных корпусов (например, HunOr [33] и InterCorp [18]), но в них не выделены русские предложные конструкции, и даже если тексты выравнены, выравнивание является ненадёжным. По этим причинам для проведения данного исследования необходимо составление подходящего русско-венгерского параллельного корпуса. **Материалом** данной работы служат параллельные тексты на русском и венгерском языках в жанрах художественной, научной и популярной литературы и разговорной речи (субтитры фильмов и сериалов).

Научная новизна нашей работы заключается в выявлении соответствий между русскими предложными конструкциями и их эквивалентами в

венгерском языке. Подробное исследование для данной пары языков ещё не было проведено, поэтому данная работа может выявить до сих пор неизвестные аспекты исследуемых языков.

Теоретическая значимость данной работы состоит в применении сравнительных исследований на основе параллельных корпусов для изучения словоизменительных и словообразовательных характеристик принципиально разных языков с точки зрения морфологии и синтаксиса. Более того – создание классификации параллельных конструкций в разных языках может привести к развитию теории не только в области лингвистики, но и в области педагогики иностранных языков.

Практическая значимость заключается в создании и применения метода, основанного на теории графов для выравнивания русско-венгерских текстов при сохранении границ предложных конструкций и выделении их эквивалентов в венгерском языке. С созданием новых методов выявления и выравнивания лингвистических конструкций в разных языках даётся возможность дальше усовершенствовать методику проведения лингвистических исследований, кроме того, наш метод также может быть применен для практических задач, например, для автоматического машинного перевода.

Глава 1. Проблема выравнивания

1.1. Выравнивание в контексте обработки естественного языка

Выравнивание является предварительным этапом для многочисленных практических задач обработки естественного языка, как например, для машинного перевода, информационного поиска по текстам на различных языках, составления словарей, получения данных для обработки естественного языка, разрешение лексической неоднозначности и создания корпуса параллельных текстов [4; 12; 17]. Кроме практических применений, выравненные параллельные корпусы являются основой многих теоретических работ, входящих в дисциплину компаративной лингвистики: они могут быть использованы для расшифровки мёртвых языков, анализа разных вариантов (диалектов) одного языка, изучения исторических вариантов одного языка и анализа особенностей разных современных языков [22].

От качества выравнивания зависит успех следующих задач. Поэтому модуль для выравнивания часто интегрирован в программы для решения этих задач и, соответственно, высокая эффективность и производительность предполагается. Несмотря на это, в области выравнивания не так много инноваций в последнее время, как в других областях обработки естественного языка.

Самые важные работы о выравнивании были опубликованы в конце 1990-х и начале 2000-х годов: Brown, Della Pietra, Mercer (1993) опубликовали исследование, которое значительно повлияло на область статистического машинного перевода, и соответственно, на методы выравнивания параллельных текстов. Они первые применяли IBM-модели для решения задачи выравнивания, которые до сих пор являются релевантными и часто применяемыми [17]. Опираясь на теорию, изложенную авторами Brown, Della Pietra, Mercer, исследователи Och, Ney (2003) провели сравнительное исследование

различных IBM-моделей и создали популярную систему для статистического выравнивания (*Giza++*), которая до сих пор используется многими пользователями. В начале 2010-х годах, дальше развивая вышеописанные идеи и применяя IBM-модели, Dyer, Chahuneau, Smith (2013) создали новую систему выравнивания (*fast_align*), которая развивает производительность предыдущих методов. В отличие от предыдущих исследователей, Östling, Tiedemann (2016) создали систему выравнивания не на основе IBM-моделей, а на основе скрытых марковских цепей и байесовских моделей. Стоит обратить внимание на тот факт, что большинство упомянутых систем основаны на статистических моделях. Исключением является исследование, проведенное исследователями Jalili Sabet и др. (2020) о методе выравнивания на основе нейронных сетей и многоязычной модели mBert (*Simalign*). Создатели метода утверждают, что их метод способен соперничать со статистическими методами и в некоторых случаях даже способен достичь слегка большей эффективности, но при этом, производительность (performance) резко ухудшается.

Общим чертой вышеуказанных исследований и систем для выравнивания является отсутствие прогресса в эффективности после Och, Ney (2003): большинство исследователей после них развивают лишь производительность по сравнению с *Giza++*, а эффективность при этом не повышается, или, если повышается, то это улучшение происходит за счёт производительности. Это объясняется следующими факторами:

Во-первых, стремление исследователей создать универсальную систему, которая одинаково хорошо работает для любой пары языков и избегает специализированных методов и решений. К сожалению, достичь универсальности без учёта особенностей конкретных языков почти невозможно, так как особенности исследуемых языков всегда влияют на качество выравнивания. Соответственно, у универсальных систем выравнивания существует некий

потолок эффективности.

Во-вторых, развитие глубокого обучения с помощью нейронных сетей привлекает всё больше исследовательского внимания в области обработки естественного языка. Поскольку глубокое обучение не требует выполнения определённых этапов предварительной обработки, решение таких задач, как автоматический машинный перевод часто осуществляется без выравнивания текстов в параллельном корпусе или даже без использования параллельных текстов в качестве тренировочных данных [30]. По этим причинам некоторые исследователи считают выравнивание ненужным и устаревшим.

С другой стороны, проблема выравнивания до сих пор остаётся актуальным по многим причинам.

Во-первых, для решения специфических задач с параллельными текстами, как, например, объект данного исследования, необходимо чётко и надёжно выравнивать параллельные конструкции. Для таких задач применение глубокого обучения не даёт достаточного контроля и свободы для улучшения результатов: как правило, модели глубокого обучения решают с большой точностью самые частотные случаи, но не всегда дают возможность разобраться с исключениями и особенными случаями.

Во-вторых, в последние годы некоторые исследователи начали снова рассматривать выравнивание как потенциальный способ улучшение результатов автоматического машинного перевода: Xintong (2017) проводит выравнивание в системах машинного перевода, чтобы, по его словам, лучше понимать, каким образом работают модели машинного перевода по типу «черный ящик», где модели обучаются на основе тренировочных данных и ожидаемых результатов [27]. В некоторых приложениях машинного обучения также даётся возможность улучить результатов с использованием выровненных параллельных текстов в качестве тренировочного материала в целях обучения

моделей.

На данном основании, несмотря на генеральные тенденции последних лет, работа над улучшением методов выравнивания параллельных текстов непосредственно или опосредованно может повлиять на результаты многочисленных задач обработки естественного языка и, соответственно, работа над улучшением методов выравнивания является важной и актуальной задачей.

1.2. Определение выравнивания

При выравнивании слова, словосочетания или предложения исходного языка (дальше L1) автоматически сопоставляются их эквивалентам в целевом языке (дальше L2) в одном тексте [12].

Формально, можно определить выравнивание A в случае языков «L1» $S_{L1}^{I} = w_{L1}^{1}, w_{L1}^{2}, \ldots, w_{L1}^{i}$ и «L2» $S_{L2}^{J} = w_{L2}^{1}, w_{L2}^{2}, \ldots, w_{L2}^{j}$, как декартово произведение позиции выравниваемых сегментов текста , где S_{L1} – предложение на языке L1, S_{L2} – предложение на языке L2, i – количество лексем в S_{L1} и j – количество лексем в S_{L2} (см. формула 1) [26].

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

Формула 1 – самая обобщенное формальное описание проблемы выравнивания, которую можно применять для любой пары языков и получать более-менее приемлемые результаты. Некоторые исследователи, однако, добавляют дополнительные ограничения, с помощью которых появляется возможность повысить полноту результатов для определённых пар языков, при том, что для других языков при данных ограничениях она будет снижаться. Эти ограничения могут влиять на результаты в случае языков со свободным порядком слов.

Также существуют варианты решения, где применяются дополнительные лингвистические данные, как, например, частеречевая разметка, лемматизация, а также разбивка текста на синтагмы. С помощью этих данных можно улучить результаты для определённых языков, но модель потеряет универсальность – то есть, возможность применения инструмента для любой пары языков.

1.3. Уровни выравнивания

В зависимости от задачи и пар изучаемых языков, выравнивание может быть проведено на уровне:

- морфем;
- лексем (слов);
- синтагм (словосочетаний);
- и предложений.

1.3.1. Выравнивание на уровне морфем

Выравнивание на уровне морфем осуществляется между морфемами внутри слов в параллельных предложениях.

Этот подход применяется в тех случаях, когда один или оба языка имеют богатую морфологию, где морфемы внутри слов могут выражать не только *грамматическое*, но и *лексическое* значение, как, например, в случае агглютинативных языков.

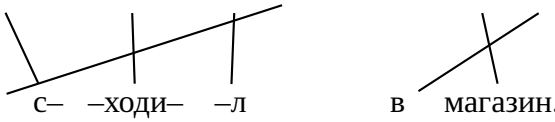
Выравнивание на уровне морфем необходимо для многих задач обработки полисинтетических языков, в которых элементы словосочетаний соединяются в единое целое без формальных показателей у каждого из них,

или для проведения специализированных морфологических исследований на основе одного или больше языков [14].

Надо отметить, что выравнивание на уровне морфем проводится значительно реже, чем выравнивание на других уровнях. Это, в основном, связанно со сложностью проведения морфологического анализа с учётом и выделением элементарных морфем: для проведения такого анализа необходимо установить границы морфем внутри слов, сделать соответствующую разметку и в случае выравнивания создать двуязычную классификацию морфем. Несмотря на то, что проведение такого исследования могло бы способствовать развитию систем машинного перевода, масштаб работы создаёт серьезные экономические барьеры для применения данного метода.

В примере 1 представлено предложение, в котором русский и венгерский глагол разделены на морфемы: *el–men–t–em с–ходи–л* – таким образом мы показываем связи между морфемами в словах. Применение такого подхода особенно выгодно тогда, когда, как в данном случае, не все морфемы L1 выравниваются с морфемами L2: в примере венгерское личное местоимение «én», которое по логике должно соответствовать русском личном местоимению «Я» отсутствует из-за того, что в глаголе «elmentem» морфема «em» выражает значение первого числа первого лица. Такой вид эллипсиса может вызывать сложности при выравнивании, но с применением данного метода мы можем избегать таких несоответствий.

(1) El– –men– –t– –em a bolt– –ba.



Я с– –ходи– –л в магазин.

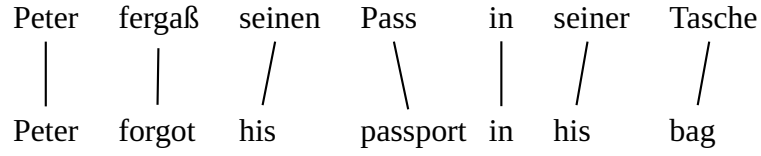
1.3.2. Выравнивание на уровне лексем

Выравнивание на уровне слов осуществляется между словами (лексемами) внутри параллельных предложений. При выравнивании на уровне слов к каждому слову в *L1* соотносится подходящее слово в *L2*. Здесь бывают *нуль-слова* – слова, у которых нет пар в другом языке. Также возможны такие случаи, где одному слову в *L1* соответствует несколько слов в *L2*.

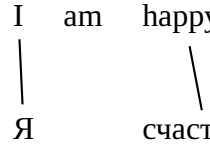
Данный метод является самым распространённым и часто-применяемым для выравнивания параллельных языков в большинстве практических применений и исследований. Он чаще всего применяется для языков с похожей морфологией, как, например, немецкий и голландский. В случае языков с богатой морфологией, как например *венгерский*, *финский* или *турецкий*, где существуют значительно больше суффиксов, чем в аналитических языках, данный метод может оказаться неподходящим в том случае, когда в другом языке в роли суффиксов выступают самостоятельные слова (аналитические языки), как например предлоги или частницы. Пример такой пары языков – *английский* и *венгерский*.

В примере 2 видно случай, где все слова выравнены в обоих языках, не встречаются нуль-слова и все слова линейно выравнены друг с другом. В примере 3 показывается случай, где не у каждого слова есть пара в другом языке. Соответственно «*am*» не выравнено – остаётся нуль-словом. Такое явление также называется **эллипсисом**.

(2)



(3)



1.3.3. Выравнивание на уровне словосочетаний

Выравнивание на уровне словосочетаний осуществляется между словосочетаниями в *L1* и в *L2* внутри параллельных предложений. В качестве выравниваемых словосочетаний может выступать любой синтаксически связанный сегмент слов в предложении. Данные сегменты могут содержать неразрывные или разрывные цепочки слов и могут состоять из одного или больше элементов. В большинстве случаев создаются сегменты из последовательных элементов, как, например, в случае **предложных конструкций** в индоевропейских языках. В данном случае, связи и соотношения между словами внутри словосочетаний не устанавливаются или устанавливаются на дальнейших этапах анализа.

Важно отметить, что не всегда имеется возможность выделять словосочетания в изучаемых языках: во многих случаях лексемы не сочетаются с другими лексемами, то есть, образуется словосочетание из одного слова. Это часто бывает у подлежащего или сказуемого предложения (см. пример 4), но для русского языка также характерно составное сказуемое (см. пример 5).

(4) Кот ест вкусную сметану
[S.подлежащее] [V.сказуемое] [obj.дополнение]

(5) Бушин был смел, честен в своих убеждениях
[S.подлежащее] [V.сказуемое] [adv.обстоятельство]

Главное преимущество метода состоит в том, что выравнивание на уровне словосочетаний требует меньше ресурсов, чем выравнивание на уровне слов или морфем, и дает очень высокую точность и полноту благодаря тому, что часть отличий в структуре выравниваемых языков «прячется» в выделенных словосочетаниях, которые воспринимаются как лингвистические единицы. Данный метод является самым оптимальным для некоторых из главных практических задач обработки естественного языка, как, например,

машинный перевод, выделение ключевых выражений, выделение именованных сущностей в параллельных текстах и т. д.

Недостатком в этом случае является то, что требуется предварительное выделение минимальных единиц выравнивания: словосочетаний. Процесс выделения словосочетаний (синтагм) в текстах также называется **чанкингом**. Задача чанкинга при выравнивании на уровне словосочетания отличается от одноязычного чанкинга в том, что выделенные словосочетаний должны соответствовать друг другу в изучаемых языках или должны по какой-то определенной логике выделяться в обоих языках.

Например, следующее предложение на английском языке *He is from Saint Petersburg* можно разделить, как *He*; *is*; *from Saint Petersburg* или *He*; *is*; *from*; *Saint Petersburg* – правильный вариант зависит от морфосинтаксических особенностей языка L2. В примерах 6 и 7 видно, что в случае английского и русского языков, даже у таких простых предложений есть несколько потенциально правильных способов выравнивания. В примере 6 выделяется целая предложная конструкция в обеих языках. В примере 7 предлог и именная группа выравнены отдельно. По идее, оба варианта правильны, тем не менее, в некоторых контекстах есть разница между разными способами выделения словосочетаний. В английском предложении не имеет значения, выделяется ли отдельно предлог или нет, а в русском предложении логичнее всю конструкцию выделять неразрывно из-за того, что предлог управляет последующим словам, как в примере 6.

(6) He is from Saint Petersburg

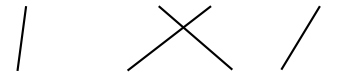
Он из Санкт-Петербурга

(7) He is from Saint Petersburg

Он из Санкт-Петербурга

Также проблемой при выравнивании на уровне словосочетаний является то, что из-за определённых особенностей порядка слов в некоторых языках синтаксически и морфологически связанные конструкции не всегда выступают в тексте последовательно. Примером данного явления являются отделяемые глагольные приставки в немецком языке (см. пример 8), где по правилам немецкого синтаксиса глагольная основа *«wache»* у глагола *«aufwachen»* должена быть во второй позиции, а приставка *«auf»* должна быть в конце простого предложения. Поскольку в английском переводе глагол вместе с приставкой находится в конце предложений, слова в двуязычном выравнивании соотносятся нелинейно, и образуется разрывная глагольная группа. Выравнивание подробных конструкций может вызывать затруднение для большинства инструментов.

(8) Ich wache nicht auf



I won’t wake up

1.3.4. Выравнивание на уровне предложений

При выравнивании на уровне предложений, предложения в *L1* выравниваются с предложениями в *L2* в параллельном тексте. Существует два подхода к выравниванию предложений. Первый подход, обеспечивающий суще-

ственно более высокую производительность, основан на длине предложений. Второй, более ресурсоёмкий подход, основан на лексических соответствий, устанавливаемых с помощью морфологических или семантических параметров. Выравнивание на уровне предложений не является непосредственно объектом данного исследования, но тем не менее является важным пререквизитом выравнивания на остальных уровнях анализа. Точность выравнивания на других уровнях зависит от точности выравнивания на уровне предложений, поскольку другие этапы выравнивания проводятся внутри параллельных предложений. В ходе данного исследования применяется гибридный подход, основанный на длине предложений и на лексической информации с помощью *hunalign* [29].

В таблице 1 показан простой пример выравнивания на уровне предложений на немецком и английском языках.

Таблица 1: Пример параллельных предложений после выравнивания

doch jetzt ist der Held gefallen .	but now the hero has fallen .
neue Modelle werden erprobt .	new models are being tested .
doch fehlen uns neue Ressourcen .	but we lack new resources .

1.4. Сложности при выравнивании

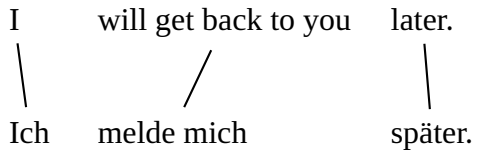
Как у всех задач обработки естественного языка, у выравнивания параллельных текстов также существуют некие ограничения. Данные ограничения связаны с органическом характером всех языков мира: можно наблюдать закономерности, правила и даже законы во всех языках, но нет правил без исключений. Правила и исключения разные во всех языках, поэтому создание универсального метода выравнивания невозможно.

Из-за разности языков, в процессе перевода слова, словосочетания и предложения могут разделяться, сливаться, удаляться, вставляться или менять последовательность.

Проблемы также возникают при вышеуказанном разделении задачи выравнивания на уровнях: даже в языках из одной семьи языков, уровни анализа могут смешиваться, поскольку не всегда существует буквального перевода тех или иных слов или словосочетаний. Это особенно часто вызывает сложности при выравнивании идиоматических выражений, фразеологизмов и жаргонизмов.

Пример 9 показывает выравнивания идиоматических выражений в немецком и английском языках. Как видно в примере, нельзя создать прямого выравнивания слов в этих предложений, и даже эквивалентные конструкции могут содержать разные по семантике слова при том, что предложения в контексте текста полностью эквивалентны: *melde mich* «сообщаю о себе», *will get back to you* «вернусь к вам». В данном случае самым подходящим способом выравнивания является гибридный метод – смесь выравнивания на уровне слов и словосочетаний.

(9)



Разрешить проблемы такого рода в языках из разных семьей языков еще сложнее. В случае русского и венгерского языка возникает ряд проблем из-за разных аспектов словоизменения и словообразования изучаемых языков. Данной теме просвещена глава 2 нашей работы.

1.5. Системы выравнивания параллельных текстов

Кроме лингвистических вопросов, относящихся к предмету данного исследования, необходимо уделять внимания техническим аспектам задачи выравнивания.

Для выравнивания параллельных текстов существуют два основных подхода: статистическое и нейронное выравнивание. Разные статистические методы существуют с конца 1990-ых годов и начала нового тысячелетия. Большинство статистических систем основано на IBM-моделях или на марковских моделях и проводит выравнивание на уровне слов. В последние годы также стали популярными системы выравнивания на основе нейронных сетей и глубокого обучения. Такие методы позволяют проводить выравнивание на любом уровне языка, но приобретение подходящих тренировочных данных вызывает затруднение у исследователей. Большинство популярных систем выравнивания применяют статистические методы, потому что системы нейронного выравнивания до сих пор являются экспериментальными.

Приведём примеры самых распространённых и часто-применяемых систем выравнивания параллельных текстов, которые соответствуют промышленным стандартам:

1. выравнивание на основе IBM-моделей с использованием *GIZA++* [26];
2. выравнивание на основе IBM-моделей с использованием *fast_align* [20];
3. выравнивание на основе марковских цепей Монте Карло с использованием *eflomal* [28];
4. выравнивание на основе нейронных сетей с использованием *Simalign* [31].

Дальше обсуждаются главные особенности статистических и нейронных моделей выравнивания.

1.6. Статистические модели выравнивания

Статистические модели выравнивания основаны на методах статистического машинного перевода. Och, Ney (2003) утверждают, что цель статистического машинного перевода – моделировать вероятность перехода $Pr(f_1^J|e_1^I)$, которая описывает отношение между строкой исходного языка f_1^J и строкой целевого языка e_1^I. В статистических моделях выравнивания $Pr(f_1^J, a_1^J|e_1^I)$, скрытое выравнивание a_1^J описывает соотношение между исходной позицией j и целевой позицией a_j. Отношение между моделями перехода и моделями выравнивания описано в формуле 2 [26].

$$Pr(f_1^J|e_1^I) = \sum_{a_1^J} Pr(f_1^J, a_1^J|e_1^I) \tag{2}$$

Выравнивание a_1^J может содержать выравнивание $a_j = 0$ с пустым словом (нуль-выравниванием) e_0, чтобы выражать слова исходного языка, не выравненные ни одним словом целевого языка.

В-общем, статистические модели во многом зависят от множества неизвестных параметров θ, которые получены из тренировочных данных. Och, Ney (2003) предлагают использовать модифицированную нотацию (см. формула 3) для выражения зависимости модели от неизвестных параметров.

$$Pr(f_1^J, a_1^J|e_1^I) = p_\theta(f_1^J, a_1^J|e_1^I) \tag{3}$$

Для определения неизвестного параметра θ, даётся параллельный тренировочный корпус, состоящий из S пар предложений $\{(f_s, e_s) : s = 1, \ldots, S\}$. Для каждой пары предложений (f_s, e_s), переменное выравнивание указано с $\mathbf{a} = a_1^J$. Неизвестные параметры θ определяются с максимализацией вероятности на основе тренировочного корпуса (см. формулу 4).

$$\hat{\theta} = \arg\max_{\theta} \prod_{s=1}^{S} \sum_{a} p_\theta(f_s, a|e_s) \tag{4}$$

Для описанных моделей обычно применяется EM-алгоритм [Dempster, Laird, Rubin, 1977] для нахождения оценок максимального правдоподобия параметров моделей, хотя это не во всех случаях обязательно [19].

Форма EM-алгоритма для нахождения наилучшего выравнивания $\hat{a}_1^J$ на паре предложений (f_1^J, e_1^I) называется выравниванием Витерби (см. формула 5).

$$\hat{a}_1^J = \arg\max_{a_1^J} p_{\hat{\theta}}(f_1^J, a_1^J|e_1^I) \tag{5}$$

1.6.1. Giza++

Одна из самых популярных и высокоэффективных систем статистического выравнивания – *Giza++*.

Система Giza++ основана на нескольких IBM-моделях, описанных Brown, Della Pietra и Mercer (1993) и Och и Nay (2000) и моделях на основе скрытых марковских цепей, описанной Vogel, Ney и Tillman (1996). Система является универсальной и проводит выравнивание на уровне слов.

Стандартная версия Giza++ применяет 5 итераций IBM-модели 1, 5 итераций модели HMM (скрытых марковских цепей) и 5 итераций IBM-модели 3 с выравниванием Витерби [26].

IBM-модель 1 использует равномерное распределение $p(i, j, I, J) = 1/(I + 1)$:

$$Pr(f_1^J, a_1^J|e_1 I) = \frac{p(J|I)}{(I+1)^J} \cdot \prod_{j=1}^{J} p(f_j|e_{a_j}) \tag{6}$$

Как мы уже упомянули выше, Giza++ также применяет стандартный HMM-модель для улучшения результатов на втором этапе работы алгоритма:

$$p(f_1^J|e_1^I) = p(J|I) \cdot \sum_{a_1^J} \prod_{j=1}^{J} [p(a_j|a_{j-1}, I) \cdot p(f_j|e_{a_j})] \tag{7}$$

Применяется ещё одна IBM-модель – модель 3:

$$p(B_i|B_{i-1}, e_i) = p(\phi_i|e_i)\phi_i! \prod_{j \in B_i} p(j|i, J) \tag{8}$$

Такая комбинация применённых моделей позволяет комбинировать некоторые особенности каждой модели. Например, HMM-модели хуже разбираются с нуль-словами и со свободным порядком слов, чем IBM-модели, зато выделяют более эффективно зависимые слова. Комбинация этих аспектов позволяет создать более универсальную программу.

1.6.2. fast_align

Система *fast_align* является непосредственным преемником Giza++, она создана с целью увеличения производительности Giza++. Инструмент *fast_align* использует IBM-модель 2 (см. формулу 9).

$$Pr(f_1^J, a_1^J|e_1^I) = \frac{p(J|I)}{(I+1)^J} \cdot \prod_{j=1}^{J} p(f_j|e_{a_j}) \tag{9}$$

Главное преимущество *fast_align* по сравнению с Giza++ состоит в том, что *fast_align* достигает повышенной производительности без потери качества выравнивания. В целях улучшения производительности не применяется так много разных моделей выравнивания, как в случае Giza++, но при этом точность и полнота заметно не ухудшается в случае популярных языков, как, например, английский, немецкий, испанский и т. д. *fast_align* достигает этого

улучшения с помощью тщательного подбора параметров модели. О подробностей метода пишется подробнее в работе [Dyer, Chahuneau, Smith, 2013]. Как и Giza++, fast_align является универсальной системой, где выравнивание проводится на уровне слов.

1.6.3. eflomal

В отличие от *Giza++* и *fast_align*, *eflomal* использует комбинацию совершенно других моделей. База системы *eflomal* – **Байесовская модель** в сочетании с моделью **HMM Монте Карло** и **семплированием по Гиббсу** (см. формулу 10).

$$P(a,\theta) = P(s,t,a,\theta,\alpha)\alpha \left(\prod_{k=1}^{K} \prod_{j=1}^{J^{(k)}} \theta_{s^{(k)}_{a^{(k)}_j}, t^{(k)}_j} \right) \cdot \left(\prod_{e=1}^{E} \prod_{f=1}^{F} \theta_{e,f}^{\alpha_f - 1} \right) \tag{10}$$

Данная система является похожей на предыдущие системы во многих аспектах: выравнивание проводится на уровне слов, метод является универсальным и т. д. Самая большая разница, опять же, в производительности. Очень интересно, что несмотря на то, что Giza++ является самым ранним из перечисленных систем, именно она достигает самой высшей точности и полноты. Последователи стараются улучшить в основном скорость метода, но значительно улучшить результаты не удаётся – вершина возможностей статистических универсальных методов была достигнута уже около 20 лет назад.

1.7. Нейронные модели выравнивания

Jalili Sabet и др. (2020) показывают, что нейронные модели выравнивания уже способны достигать похожих или даже лучших результатов по

сравнению со статистическими методами [31]. Некоторые модели достигают хороших результатов на основе одноязычного тренировочного материала, а другие используют модели, тренированные на параллельных текстах.

В данной работе мы изучаем систему *Simalign*, которая используют параллельные тексты на 104 языках с помощью языковой модели mBERT. *Simalign* применяет три разных метода для выравнивания, основанных на подобных матрицах: *Argmax*, *Itermax* и *Match*.

- **Argmax** простейший метод, который выравнивает самые частотные слова в подобной матрице.
- **Itermax** метод, предлагаемый авторами *Simalign*, которые применяют новый итеративный алгоритм для выравнивания слов в подобной матрице.
- **Match** метод теории графов, который определяет паросочетание на двудольном графе с применением венгерского алгоритма (также называется алгоритмом Куна – Манкреса) [24].

Jalili Sabet и др. (2020) сравнили эффективность (меры F1 и AER [1]) самых распространённых систем автоматического выравнивания со своей системой (Simalign) и пришли к выводу, что в некоторых датасетах для определённых пар языков их система достигает более высоких показателей, чем другие системы. Их результаты представлены в таблице 2.

Однако, несмотря на высокие показатели, представленные авторами исследования для английского и немецкого языков, для других пар языков статистические методы могут быть более эффективными. Более того, поскольку система Simalign основанная на машинном обучении, тренировоч-

[1] AER = 1.0 - F1

Таблица 2: Сравнение F1-меры разных методов выравнивания английского и немецкого языков (Jalili Sabet и др., 2020)

Система	F1	AER
Giza++	0.77	.23
fast_align	0.71	.29
eflomal	0.77	.23
Simalign - Argmax	0.87	.13

ные данные в большей мере влияют на результатов оценки, если оценка проводится на том же материале.

1.8. Выводы по главе 1

1. Под выравниванием подразумевается процесс установления соответствия морфем, слов, словосочетаний или предложений друг другу в двуязычных параллельных текстах.

2. Выравнивание проводится на уровне морфем, слов, словосочетаний и предложений.

3. Самые большие проблемы при выравнивании включают в себе проблему нуль-слов, проблему несоответствия уровней анализа, проблему порядка слов, проблему эллипсиса и проблему разных морфологических и синтаксических систем анализируемых языков.

4. Для выравнивание существуют разные системы: статистические, как, например, *Giza++*, *Eflomal* и *fast_align*, и основанные на машинном обучении, как *Simalign*.

5. Исследователи показывают, что в определённых контекстах *Simalign* может достигать похожих или даже лучших результатов, чем статистические системы выравнивания.

Глава 2. Описание изучаемых языков

2.1. Роль словоизменительных и словообразовательных особенностей языков в выравнивании

Для понимания ключевых отличий между венгерским и русским языками, необходимо ознакомится с базовыми словоизменительными и словообразовательными особенностями изучаемых языков. Венгерский и русский – очень разные языки по многим причинам, но самые большие отличия именно в словообразовании и словоизменении, поэтому необходимо изучать эти аспекты и подходить к данной проблеме с теоретической точки зрения. Именно такой подход гарантирует то, что в практической части работы удастся решенить сложности выравнивания изучаемых языков с непосредственным применением выводов данного теоретического обсуждения.

Кроме теоретических аспектов, в данной главе мы представляем схему русских предложно-падежных конструкций и их эквивалентов в венгерском языке, которая станет основой нашего метода выравнивания и непосредственно применяется для решения проблемы выравнивания.

2.2. Словоизменение и словообразование в русском языке

В русском языке, являющийся языком флективного типа, главным способом выражения грамматических форм является словоизменение с помощью изменения падежных окончаний (флексий) и изменения внутреннего строя слов. При флективном словоизменении изменения наблюдаются в конце слова (*книга – книги – книге – книгу – книгой*) или внутри слова (*собирать – соберу – сбор – собрать*), то есть, тождество слова не нарушается – не создаётся новое, отдельное слово, всего лишь модифицируется грамматическая форма исходного слова, как, например, в случае категории лица,

числа, времени и вида у глаголов (*объяснить – объясню – объяснят – объяснять – объясню – объясняют*), категории рода, числа и падежа у имён (*брат, братья, брату, братьям*). Словоизменение в случае имён существительных, прилагательных, числительных и местоимений называется склонением, а в случае глаголов – спряжением (хотя в более узком смысле склонением называют изменение имён по падежам и спряжением – образование личных форм глаголов) [6].

Вышеуказанные лингвистические процессы – самые употребляемые, но не единственные способы словоизменения в русском языке. Также имеются редко встречающиеся черты агглютинации, как, например в случае постфиксов «-ся / -сь», «-те» (*заниматься, извинись, извините*).

Кроме синтетических способов выражения грамматических значений слов, также наблюдаются аналитические способы, где словоизменение происходит с помощью отдельных слов, которые в некоторых контекстах выражают особенное грамматическое значение. В некотором случае аналитические конструкции могут выступать как единственно возможные, как, например формы «будущего сложного» времени глаголов несовершенного вида *буду говорить*, сослагательного наклонения *говорил бы*, выражение категорий существительных общего рода *умный сирота — умная сирота*, несклоняемых существительных *нет крепкого кофе* [11].

По нашему мнению, наличие и распространённость предложно-падежный конструкции тоже попадает в категорию аналитических признаков русского языка. Это утверждение обосновано тем, что

1. Предлог не является самостоятельной частью речи, но имеет грамматическое и возможно, лексическое значение (см. разд. 2.3);

2. Внутри предложной конструкции нельзя определить значение существительного без учёта соответствующего предлога, потому что зна-

чение падежной формы сильно зависит от предлога;

3. На взгляд большинства исследователей, внутри предложной конструкции наблюдается подчинительная связь между предлогом и последующим именам. Например, в случае конструкции *в доме* и *о доме*, несмотря на то, что подчинённое слово (слуга) в той же грамматической форме в обоих случаях, их значение не совпадает из-за предлога «в» и «о» (см. разд. 2.3).

Соответственно, в нашей работе предложно-падежная конструкция (или синтаксема) рассматривается как минимальная единица и не делится при выравнивания (см. разд. 1).

Реже, чем в агглютинативных языках, но в русском также наблюдается словообразование для порождения дериватов от однокоренных слов. В отличии от словоизменения, при словообразования порождаются новые, самостоятельные слова. Способы словообразования в русском включают в себе аффиксацию *ехать* $\rightarrow$ *выехать – заехать – приехать – уехать*, конверсию *рабочий человек* $\rightarrow$ *молодой рабочий*, сокращение *зонтик* $\rightarrow$ *зонт*, словосложение *везде + ходить* $\rightarrow$ *вездеход* и прочие [10].

Упоминание способов словообразования в русском может показаться излишней информацией в контексте нашей работы, но на самом деле это важная информация, которая может повлиять на качестве выравнивания. Например при субстантивации, прилагательные могут переходить в существительное, как в следующем случае: *в выходных дней* $\rightarrow$ *в выходных*. Здесь очень важно заметить и учесть, что при автоматическом выравнивании слово «выходные» именно в значении существительного, а не прилагательного. Если при выравнивании применяются морфологические данные, то необходимо снять лексическую неоднозначность между прилагательным «выходных» и субстантивированным существительным «выходных». В любом слу-

чае необходимо создать правила выравнивания, которые учитывают и решают такие проблемы. Проблемы также могут возникать при словосложении, например, слово «домофон» переводится на английский, как «door phone», поэтому прямое выравнивание в этом случае невозможно. При выравнивании с немецким и венгерским языками, где самый частотный способ словообразования – словосложение – таких проблем еще больше.

2.3. Предлоги и предложные конструкции в русском языке

Как мы уже упоминали выше, предлоги в русской грамматике занимают очень неоднозначную позицию: с одной стороны, все исследователи согласны в том, что предлог имеет грамматическое значение, а с другой стороны, имеются различные мнения о том, что можно говорить о и лексической значении у предлогов.

В классическом смысле понятие предлога в русском языке определяется следующим образом: «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» [13].

Существуют разные классификации предлогов в русском языке [3].

- Классификация по происхождению:
  - Первичные: *без (безо), в (во), до, для, за, из (изо), к, на, над, о (об, обо), от, по, под, пред (перед), при, про, с, у, чрез (через)*;
  - Наречные: *вблизи, вглубь, вдоль, возле, около и т. д.*;
  - Отымённые: *посредством, в роли, в зависимости от, путём, насчёт, по поводу, ввиду и т. д.*;

  - Глагольные: *благодаря, несмотря на, спустя и т. д.*.

- Классификация по структуре:
  - Простые: *в, с, к, у, над, на, перед, при и т. д.*;
  - Сложные: *из-под, из-за, по-над и т. д.*;
  - Составные: *несмотря на, в отличие от, в связи с и т. д.*.

- Классификация по употреблению с падежами:
  - с одним падежом:
    * винительный падеж: *про, через, сквозь*;
    * родительный падеж: *без, близ, вдоль, вместо, вне, возле, вокруг, до, для, у, ради и т. д.*;
    * дательный падеж: *к, благодаря, вопреки, согласно и т. д.*;
    * творительный падеж: *над, перед*;
    * предложный падеж: *при*;
  - с двумя падежами:
    * вин. и пред. п.: *в, на, о*;
    * вин. и твор. п.: *за, под*;
    * род. и твор. п.: *между*;
  - с тремя падежами:
    * вин., дат. и пред. п.: *по*;
    * род., вин. и твор. п.: *с*.

- Классификация по отношениям:
  - пространственные: *около, в, среди, на*;

– временные: *в течение*, *в продолжение*;

– причинные: *ввиду*, *благодаря*, *вследствие*, *в связи*, *из-за*;

– целевые: *для*;

– объектные: *о*;

– сравнительно-сопоставительные: *в отличии от*, *похоже на*;

– уступки: *несмотря на*;

– сопроводительные: *вместе с*.

В данной работе мы ориентируемся на все простые предлоги русского языка: *без, безо, благодаря, близ, в, вблизи, ввиду, вдогон, вдогонку, вдогоночку, вдоль, взамен, включая, вкруг, вместо, вне, внизу, внутри, внутрь, во, вовнутрь, возле, вокруг, вопреки, вослед, впереди, вперёд, вразрез, вроде, вслед, вследствие, встречу, выключая, выше, для, до, за, замест, заместо, из, изнутри, изо, именем, имени, исключая, к, касаемо, касательно, ко, кончая, кроме, кругом, меж, между, мимо, минус, на, наверху, навроде, навстречу, над, надо, назад, назади, накануне, наместо, наперекор, наперерез, наперехват, наподобие, наподобье, напротив, насупротив, насчёт, начиная, независимо, несмотря, ниже, о, об, обо, обок, обочь, около, окрест, окроме, окромя, округ, опосля, опричь, от, относительно, ото, перед, передо, по, поблизости, повдоль, поверх, под, подле, подо, подобно, позади, позадь, позднее, помимо, поперёд, поперёк, порядка, посверху, посереди, посередине, поперек, посерёдке, посередь, после, посреди, посредине, посредством, превыше, пред, предо, преж, прежде, при, про, промеж, промежду, против, противно, противу, путём, ради, с, сверх, сверху, свыше, середи, середь, сзади, силами, сквозь, скрозь, снаружи, снизу, со, согласно, сообразно, соответственно, со-*

размерно, *спереди*, *спустя*, *среди*, *средь*, *сродни*, *супротив*, *типа*, *у*, *через*, *чрез* [2].

Как мы уже упомянули, некоторые исследователи считают предлог словом без лексического значения [9]. А. Пешковский, А. Шахматов и другие утверждают, что у предлогов – только грамматическое значение, а нет лексического значения. В. Виноградов определяет предлог как служебное слово с грамматическим и лексическим значением [2]. Исходя из позиции В. Виноградова и Академической грамматики, В. Захаров поддерживают следующую позицию по поводу наличия значения у предлога: «… в любом случае предлог имеет лексическое значение, различна лишь степень его абстрактности <...> семантически „пустых“ предлогов не существует» [7]. Соответственно этому определению выделяется в данной работе формальное значение у всех предлогов, включая самые *грамматикализованные*.

Однако, **значение** предлога нельзя рассматривать без учёта падежной формой последующих слов или словосочетаний. Очевидно, что значение предлога зависит не только от самого предлога, а от сочетания предлога и падежной формы и от семантики зависимого слова.

Захаров, Михайлова (2017) рассматривают данное явление с точки зрения *функционального синтаксиса*, детальной системы Г. Золотовой для описания синтаксического механизма русского языка. В описавшей Г. Золотовой системе, базовой синтаксической единицей является *синтаксема*. **Синтаксема** – это минимальная неделимая семантико-синтаксическая единица русского языка, которая выступает одновременно как носитель элементарного смысла и как конструктивный компонент более сложных синтаксических конструкций, и которая характеризуется определенным набором синтаксических функций [8].

[2]Источник: `https://github.com/merionum/pphrase`

Различительные признаки синтаксемы:

1. категориально-семантическое значение слов, которые её образуют;
2. соответствующая морфологическая форма зависимого слова;
3. способность синтаксически реализоваться в определённых позициях.

Синтаксемы могут выступать *самостоятельно, в качестве компонента предложения* и *в качестве компонента словосочетания*.

В случае русских предложно-падежных конструкций, речь идет о синтаксемах, выступающих в качестве компонента словосочетания. Такие синтаксемы также называются **связанными**. Как мы раньше уже определили, значение предложно-падежной синтаксемы – результат **сочетания предлога с падежной формой**.

Это отлично иллюстрируется в случае пространственных предлогов. Предлог «в» в сочетании с предложными падежом *в городе, в доме, в стране, в парламенте* образует **локатив**, а в сочетании с винительным пажедом *в город, в дом, в страну, в парламент* – **директив**.

Опираясь на терминологии Золотовой, можно сказать, что *Предлог в сочетании с падежной формой образует предложно-падежную синтаксему — минимальную и нечленимую единицу синтаксиса* [7].

Исходя из этого определения следует предположить, что для того, чтобы всю предложную конструкцию изучать, полезно включать в неё и прилагательные (*в **красной** тачке*; *в **новом** доме*; *из-за **плохого** человека* и т. д.). Мы также включаем в предложную конструкцию слова, которыми управляет прилагательное (*в **очень красивом** городе*; *из **самого красивого** города*). Таким образом сохраняем целостность всей конструкции и последовательность входящих в неё слов. Данный подход является логичным для нашего

исследования, потому что он упрощает процесс выделения предложных конструкций с тем, что мы не должны разрывать порядок слов русского языка.

В данной работе рассматриваются предлоги и предложно-падежные конструкции (дальше предложные конструкции) именно с этой точки зрения, но не обязательно в полном согласии с классификацией и терминологией Г. Золотовой. Причины частичной несоответствия с классификацией, выдвинутой Г. Золотовой, объясняются в разделе 2.5.

2.4. Словообразование и словоизменение в венгерском языке

Венгерский язык, являющийся частью финно-угорской семьи языков, в отличии от русского, является языком агглютинативного типа. Это сразу ведёт к некоторым принципиальным положениям по сравнению с русским.

1. Агглютинативный характер языка даёт возможность словоизменения и словообразования с помощью накопления суффиксов (аффиксация). Таким образом к одному корню могут прикрепиться многочисленные словоизменительные и словообразовательные суффиксы (toldalékok), которые могут образовывать новые слова (*megold* гл. «решить», *megold**ás*** сущ. «решение», *megold**hatatlan*** прил. «нерешаемый») или изменять значение или грамматический строй без образования новых самостоятельных слов (словоизменение) (*néző* «зритель», *néző**k*** «зрители», *néző**t*** «зрителя», *néző**k-et*** «зрителей», *néző**k-kel*** «с зрителями», *néző**k-től*** «от зрителей»).

2. Склонение в венгерском языке осуществляется с помощью богатой системы падежей. Количество падежей в венгерском языке – 18 [16], в то время как их количество в русском языке варьируется от 6-и до 13-и [5]).

3. Падежи в венгерском выражают более специфичное значение, чем падежи в русском: их значение более похоже на значение русских предложных конструкций. Например в случае падежа **инессив**, слово *ház**ban***, где *ház* – корень и *-ban* – суффикс инессива, можно перевести на русский только с помощью предложной конструкции *в доме*. То есть, несмотря на то, что инессив переводится с использованием предложного падежа на русский, нельзя сказать, что предложный падеж – эквивалент инессива, потому что значение передаётся сочетанием предлога и падежной формы (предложной конструкцией). При этом некоторые падежи являются эквивалентами русских падежей без предлогов, например венгерский **датив** (*-nak / -nek*) – эквивалент русского дательного падежа без предлогов (*mondd meg a tanár**nak*** «расскажи преподавател**ю**»).

4. Кроме падежных суффиксов, также существуют так называемые беспадежные суффиксы, которые употребляются только со специфичными словами. Число беспадежных суффиксов венгерского языка – 9 [16]. Пример беспадежных суффиксов – **эссив-модалис** (*-ul / -ül*), который используется преимущественно с названиями языков. Это похоже на конструкцию в русском языке типа «по-русски» *orosz**ul***; «по-немецки» *német**ül***; «по-венгерски» *magyar**ul*** и т. д.

5. Вследствие богатой морфологии венгерского языка, в нем нет предлогов – значение предлогов выражается разными словоизменительными способами, в основном с помощью падежных суффиксов и послелогов (*utc**án*** «**на** улице»; *város **felett*** «**над** городом»);

6. В венгерском отсутствуют некоторые концепты, характерные для индоевропейских языков, как, например, категория рода. В отличие от

венгерского, в русском языке категория рода играет большую роль при склонении, определяя (среди прочих факторов), какое падежное окончание употребляется (*из город**а*** «város**ból**»; *из коробк**и*** «doboz**ból**»; *из желез**а*** «vas**ból**»; *из сливок* «tejszín**ből**»);

7. Важный аспект венгерского – наличие нескольких эквивалентных грамматических форм одной и той же морфемы как следствие сингармонизма (гармония гласных). Данное явление очень характерна для суффиксов, в том числе, суффиксов падежей, например, *tartó**n*** «в ёмкости»; *polc**on*** «на полке»; *kép**en*** «на картине»; *tükr**ön*** «на зеркале». Поэтому суффиксы данного падежа (суперессива) обозначаются следующим образом: *-n / -on / -en / -ön*, где все возможные формы суффикса указываются;

8. В именных группах венгерского языка прилагательные не склоняются (*a **nagy** városban* «в **большом** городе»; *a **nagy** házból* «из **большого** дома»; *a **nagy** bögrével* «с **большой** кружкой»);

9. Венгерский обладает своеобразной системой притяжательных суффиксов, примыкающих к именам (например *könyv**em*** «моя книга»; *könyv**ed*** «твоя книга»; *könyv**e*** «его/её книга» и т. д.).

10. Венгерский язык обладает и определённым *a / az* и неопределённым *egy* артиклем в отличии от русского, где вообще нет артиклей: *bementem **a** házba* «я зашёл в дом»; *nem találtam **az** ajtót* «я не нашёл дверь»; *bementem **egy** házba* «зашёл в какой-то дом».

Несмотря на большие отличия, некоторые аспекты венгерского языка очень знакомы носителям русского языка:

1. Венгерский язык обладает свободным порядком слов *olvasok egy könyvet* «читаю книгу», *egy könyvet olvasok* «книгу читаю»;

2. Глагольные приставки могут присоединятся к начале глаголов, изменяя или модифицируя значение глагола ***meg****jönni «****при****ехать»*, ***el****jönni «****от****ехать»*, или образуя совершенный вид *csinálni* «делать», ***meg****csinálni* «**с**делать»), но в отличие от русского, приставки могут отделяться от глагольной основы *csináld* ***meg*** «**с**делай», *nem oldottam* ***meg*** «не реш**и**л»);

3. В венгерском, как и в русском, существуют аналитические черты, как послелоги, которые могут выражать значение русских предложных конструкций аналитическим способом *szem* ***előtt*** «**перед** глазами», *a ház* ***mögött*** «**за** домом», *fej* ***felett*** «**над** головой») – с добавлением отдельного слова, которое изменяет значение, в данном случае, предыдущей лексемы. Важно заметить, что существительное перед послелогом не всегда изменяется по падежам – часто используется словарная форма слова в именительном падеже: *ház mögött, kép mellett, kertek alatt*.

До сих пор мы выявили два главных способа выражения значения русских предложных конструкций в венгерском языке: **аффиксацию** и **употребление аналитических черт языка с употреблением послелогов**. Но кроме этих способов существуют ещё много, менее регулярных эквивалентов: часто используется словосложение в значении предложных словосочетаний в русском *tejesdoboz* «коробка из-под молока», *lakáskulcs* «ключи от квартиры», *mosogatószer* «средство для мытья посуды». Выделение эквивалентов такого рода вызывает серьезные сложности при выравнивании, поэтому необходимо создать метод выравнивания изучаемых явлений, способный решать такие проблемы.

2.5. Способы классификации русских предложных конструкций и их эквивалентов в венгерском языке

Для классификации русских предложных конструкций и их эквивалентов в венгерском существуют разные попытки в области теоретической лингвистики и преподавания русского как иностранного. А. Шальга классифицирует русские предлоги на основе валентности и возможных падежных форм и по этим критериям соотносит русские предлоги с венгерскими эквивалентами. Однако он обращает внимание только на самые частотные, самые очевидные и самые обобщённые переводы. Также недостатком его классификации является то, что он указывает только падежные или послеложные способы передачи значении русских предложных конструкции. Более того, в научной литературе нет источников, которые указывают или изучают распределение употребления тех или иных эквивалентов при переводе.

Из-за отсутствия материала, подходящего для наших целей, было принято решение создания классификации, являющейся основой для некоторых дальнейших шагов нашего исследования в практический части работы (см. в главе 3).

Наша классификация русских предлогов и их эквивалентов в венгерском языке в большой мере отражает систему падежей в венгерском языке. Названия падежей, с некоторым исключением, тоже заимствованы из венгерских теоретических работ, где в основном используются названия латинского происхождения, как, например *элатив*, *аллатив*, *инессив*, *делатв*, *аблатив* и т. д. Данное решение объясняется тем, что наша классификация основана на морфологических и синтаксических данных и закономерностей изучаемых языков, в отличие от классификации предложных синтаксем, использованной Золотовой и Захаровым, которая по большей мере опирается на семантические аспекты предложных конструкций.

Главное преимущество нашего подхода в том, что таким образом даётся возможность обрабатывать текст без использования семантического словаря или без проведения семантического анализа двуязычного текста, поднимая таким образом производительность приложения и значительно упрощая решение проблем.

В таблице 3 представлена классификация русских параллельных конструкций с предлогами *в*, *на*, *из*, *с*, *до*, *от*, *к*, *у*, *за*, *для* и их аналогов в венгерском языке. В таблице указаны названия, теги, использованные в нашей программе, падежные вопросительные слова, модели порождения и примеры изучаемых конструкций:

Таблица 3: Русские предложные смнтаксемы и их эквиваленты в венгерском

тип	тег	язык	вопрос	модель	пример
инессив	Ine	рус.	в чём? в ком?	в + пред.	в доме, в комнате, в саду, в лесу
		венг.	miben? kiben?	-ban / -ben	házban, szobában, kertben, erdőben
иллатив	Ill	рус.	куда? во что?	в + вин.	в дом, в комнату, в сад, в лес
		венг.	hova? mibe?	-ba / -be	házba, szobába, kertbe, erdőbe
суперессив	Sup	рус.	на чём? на ком?	на + пред.	на крыше, на площади, на мосту, на территории
		венг.	min? kin?	-on / -ön	tetőn, téren, hídon, területen
сублатив	Sub	рус.	куда? на что?	на + вин.	на крышу, на площадь, на мост, на территорию
		венг.	hova? mire?	-ra / -re	tetőre, térre, hídra, területre

элатив	Ela	рус.	откуда? из чего?	из + род.	из дома, из Москвы, из России, из озера
		венг.	honnan? miből?	-ból / -ből	házból, Moszkvából, Oroszországból, tóból
делатив	Del	рус.	откуда? с чего?	с + род.	с крыши, с площади, с моста, с Закарпатья
		венг.	honnan? miről?	-ról / -ről	tetőről, térről, hídról, Kérpátaljáról
аблатив	Abl	рус.	откуда? от чего?	от + род.	от дома, от площади, от моста, от ворот
		венг.	honnan? mitől?	-tól / -től	háztól, tértől, hídtól, kaputól
терминатив	Ter	рус.	куда? до чего?	до + род.	до дома, до площади, до моста, до ворот
		венг.	hová? meddig?	-ig	házig, térig, hídig, kapuig
адессив	Ade	рус.	где?	у + род.	у дома, у здании, у родителей, у моста
		венг.	hol?	-nál / -nél	háznál, épületnél, szülőknél, hídnál
аллатив	All	рус.	куда? к чему?	к + дат.	к дому, к площади, к мосту, к воротам
		венг.	hová?	-hoz / -hez	házoz, térhez, hídhoz, kapuhoz
комитатив	Kom	рус.	с кем? с чем?	с + твор.	с домом, с Петром, с инструментом, с делом
		венг.	kivel? mivel?	-val / -vel	házzal, Péterrel, szerszámmal, dologgal
каузатив	Kau	рус.	за кем? за чем?	за + твор.	за хлебом, за молоком, за курицей, за вещами
		венг.	kiért? miért?	-ért	kenyérért, tejért, csirkéért, dolgokért

бенефактив	Ben	рус.	для кого? для чего?	для + род.	для него, для глаз, для разнообразия
		венг.	kiért? miért?	-ért	kedvéért, szemekért sokszinüségért

Данная классификация (см. табл. 3) является основой для ожидаемого распределения предложных конструкцией и их венгерских эквивалентов. Классификация основана на теоретических работ венгерских исследователей [15], на корпусных исследованиях и на лингвистической интуиции автора данной работы как носителя венгерского языка. В данной классификации не включены все русские предложные конструкции, и у них указаны не все возможные переводные эквиваленты. Дальнейшее развитие данной классификации рассматривается автором исследования.

Надо подчеркнуть, что поскольку цель создании классификации – лучше понять распределение русских предложных конструкций и их эквивалентов, классификация может меняться в будущем и мы никак не утверждаем, что она является цельной и единственно верной классификацией – она всего лишь первый шаг для понимания и изучения данного вопроса русской и венгерской языкознания.

2.6. Выводы по главе 2

1. Значение русских предложных конструкций выражается сочетанием предлога и падежной формы, при том некоторые предлоги в русском языке могут управлять несколькими падежными формами.

2. Эквиваленты русских предложных конструкций в венгерском языке выражаются аффиксацией, добавлением послелога к главному слову, а также некоторыми другими словообразовательными способами.

3. Предложные конструкции и их эквиваленты не всегда употребляются одинаково в обоих языках.

4. Несмотря на сложности учета различных лингвистических аспектов изучаемых языков, мы можем создать классификацию русских предложных конструкций и их эквивалентов в венгерском на основе самых частотных или самых очевидных соответствий, собранных теоретическими лингвистами или полученных путём корпусных исследований.

Глава 3. Создание системы выравнивания русских предложных конструкций и их эквивалентов в венгерском

3.1. Описание системы для выравнивания русских предложных конструкций и их эквивалентов в венгерском

В данной главе мы предлагаем новый метод для выравнивания русских предложных конструкций и их венгерских эквивалентов в параллельных текстах. Наш метод выравнивания проводится на уровне словосочетаний с учётом морфологических и синтаксических параметров, извлечённых из текстов. Метод создан для эффективной работы с русским и венгерским языками и в большой мере опирается на наши наблюдения о словообразовательных и словоизменительных процессах русского и венгерского языков, описанных в главе 2.

Поскольку главный фокус данной работы – проблема выравнивания русских предложных конструкций и их эквивалентов, метод выделяет и выравнивает только нужные конструкции: русские предложные конструкции и венгерские именные группы. В состав предложных конструкций входит предлог, существительное в падежной форме и другие зависимые от существительного части речи, такие как прилагательные, причастия и т. д. В нашей работе именными группами называются все конструкции венгерского языка, состоящие из существительного и зависимых от существительного частей речи, как артикли, прилагательные, числительные и послелоги.

Границы предложных конструкций и именных групп выделяются на основе проведенного морфосинтаксического анализа с использованием интерфейса *UDPipe* на языке программирования *Python*. Процесс предварительной обработки корпуса и выделения предложных конструкций и именных групп

описан в разделе 3.3.

Исходя из морфологических, синтаксических и лексических параметров, каждая выделенная группа взвешивается по определённым критериям: по морфологической близости, по позиционной близости и по длине словосочетаний. Дальше из выделенных предложных конструкций, именных групп и весов образуется граф, в котором узлы – предложные и именные группы, а ребра – связи между выделенными элементами в русском и венгерском тексте, которые устанавливаются с определением наибольших паросочетаний на графах с помощью алгоритма сжатия цветков (также называется алгоритмом Эдмондса нахождения наибольшего паросочетания в произвольных графах). Детали создания графа и решение проблемы выравнивания с помощью теории графов описаны в разделе 3.4.

После запуска алгоритма, результаты выравнивания сравниваются с составленным нами золотым стандартом и оцениваются по типичным лингвистическим мерам эффективности. Кроме качественных показателей (меры оценки), мы также собираем количественные показатели (статданные) запуска алгоритма: количество проанализированных пар предложений, выделенных предложных конструкций, выделенных потенциальных эквивалентов в венгерском и т. д. Все вышеописанные данные используется для понимания преимуществ и недостаток данного метода по сравнению с уже существующими методами. О составлении корпуса и сборе тестовых данных говорится в разделе 3.2 и о методике проведения оценки эффективности – в разделе 3.5.

Наше приложение представлено в виде блок-схемы ниже (см. рис. 1), где демонстрируется асинхронный характер обработки исходных текстов на двух языках и связи и взаимоотношения между процессами анализа.

Самым решительным преимуществом нашего подхода является способность метода эффективно выделять потенциальные эквиваленты пред-

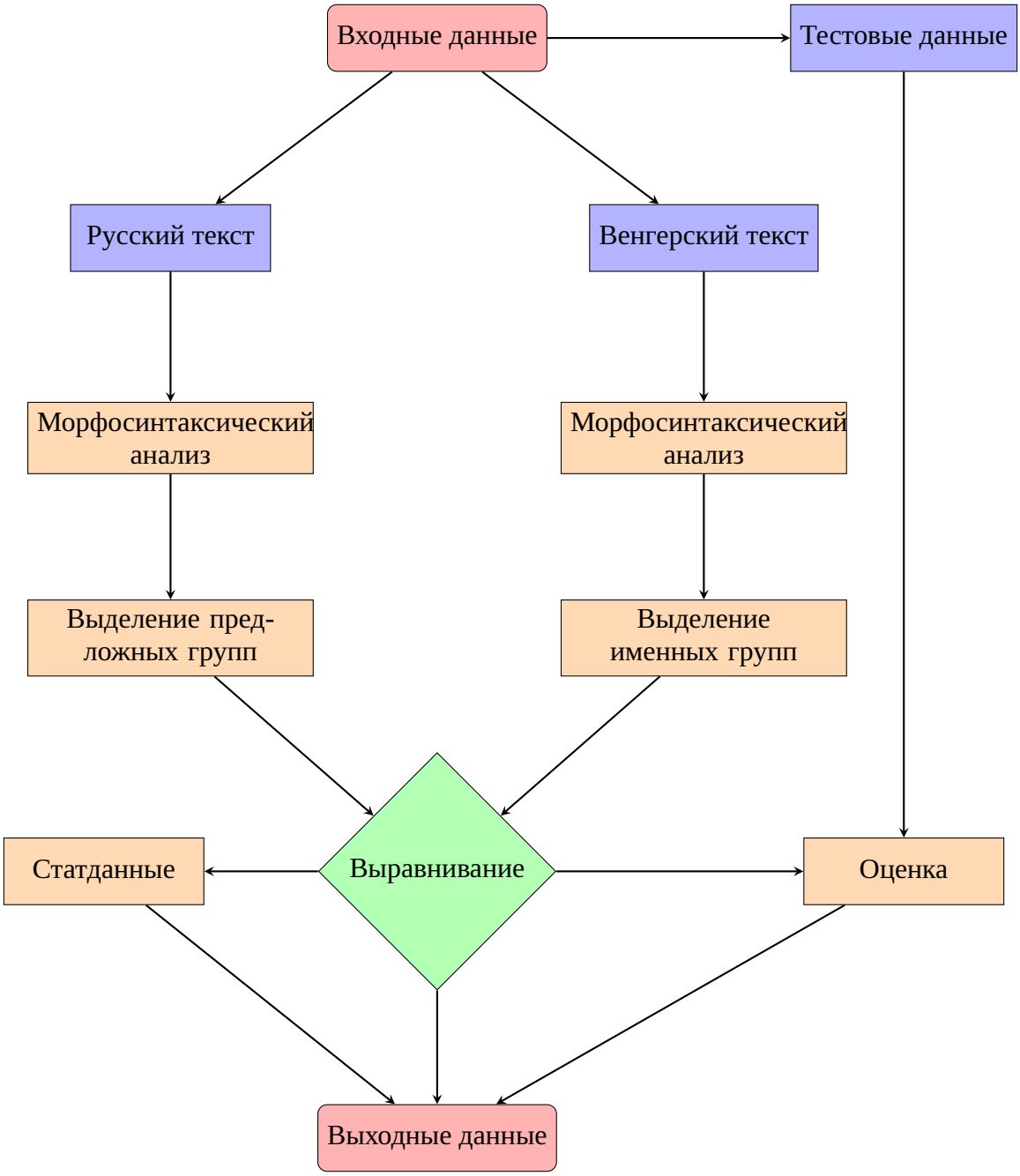


Рис. 1: Модель системы для выравнивания русских предложных конструкций и их эквивалентов в венгерском

ложных конструкций, несмотря на отличающиеся лингвистические особенности изучаемых языков и непосредственно включать в решение теоретическо-лингвистические предположения, которые в большинстве случаев значительно улучшают эффективность метода.

Самым большим недостатком нашего метода является то, что вследствие решения использовать морфосинтаксические параметры конкретных языков для выделения изучаемых конструкций, необходимо отдельно провести анализ и определить критерии выделения лингвистических конструкций в каждом изучаемом языке.

При дальнейшем развитии решения предполагается заменить модуль выделения предложных конструкций и именных групп на модели автоматического выделения словосочетаний, как, например, на модель, основанную на глубоком обучении. Таким образом расширится поле применения нашей системы на любую пару языков. Holmqvist (2010) представляет систему выравнивания на уровне словосочетаний на основе эвристики, где они автоматически выделяют словосочетания из текста на основе машинного обучения. Однако, на данный момент нет возможности создать такую систему для нашего исследования – это остаётся темой будущих работ в этом направлении.

Несмотря на вышеуказанное ограничение метода, он соответствует требованиям данного исследования и способен продвигать нашу работу и создать условия для изучения соответствия русских предложных конструкций и их эквивалентов в венгерском языке.

3.2. Сбор и составление русско-венгерского параллельного корпуса

Корпус, созданный в целях данного исследования, состоит из параллельных предложений на русском и венгерском языках, содержащих как ми-

нимум одну предложную конструкцию в русскоязычном предложении и потенциальные эквиваленты в венгерском предложении.

При создании корпуса все текстовые данные были извлечены из материалов проекта *InterCorp* [18]. Причиной использования параллельных корпусов *InterCorp* является тот факт, что данные корпусы хорошо обработаны, уже выравнены на уровне предложений и дают возможность изучать выравнивание предложных конструкций целонаправлено, изучая конструкции с отдельными предлогами с помощью языка поисковых запросов *CQL*. Соответственно, составленный нами параллельный русско-венгерский корпус является подмножеством корпуса *InterCorp v11*, который состоит из текстов художественной, научной и популярной литературы. В корпус также включены сегменты, извлеченных из субтитров для фильмов и сериалов, представляя таким образом элементы разговорной речи в нашей выборке.

На данный момент, наш корпус состоит из 500.000 параллельных предложений из разных текстов. Чтобы достичь репрезентативности, тексты были извлечены из основного корпуса случайным образом.

В целях создания тестовых данных для оценки работы алгоритма, русские предложные конструкции и их венгерские эквиваленты были вручную выделены из 200 параллельных предложений. Предложения тестового подкорпуса были выбраны случайным образом из выборки, содержащей русские предложные конструкции с предлогами *в, на, из, с, до, от, к, у, за, для*.

3.3. Предварительная обработка корпуса

В главе 2 уже упомянуты некоторые различаючиеся аспекты в морфологии и синтаксисе венгерского и русского языков, которые в какой-то мере влияют на процесс выравнивания между данными языками. Для того, чтобы изучить эти лингвистические проблемы с точки зрения самого выравнива-

ния и создать модель выравнивания, которая учитывает данную разницу и решает проблемы с большой эффективностью, необходимо провести предварительную обработку корпуса, с помощью которой снимаются возможные лексические, морфологические и синтаксические неоднозначности и создаются идеальные условия для применения алгоритмов выравнивания на дальнейших этапах анализа.

В процесс предварительной обработки корпуса входят следующие задачи для каждого из изучаемых языков:

1. Разбить текст на токены (токенизация);
2. Провести морфосинтаксический анализ с помощью системы *UDPipe*;
3. Выделить предложные конструкции на основе морфосинтаксических правил (только в русском);
4. Выделить именные группы на основе морфологических правил (только в венгерском).

Токенизация и морфосинтаксический анализ текста проводится с помощью системы UDPipe, основанной на машинном обучении [32]. Для анализа русскоязычной части корпуса применяется модель *syntagrus* и для венгерской части корпуса – модель Сегедского университета. При анализе создаётся комплексная разметка, состоящая из двух частей: морфологической и синтаксической. Морфологическая разметка в формате *CoNLL-U*[3] содержит частеречевые данные и специфичные данные для каждой части речи, например лицо, число, время и вид для глаголов или число, род и падеж для существительных. Синтаксическая часть содержит данные о синтаксической роли слов с точки зрения грамматики зависимостей (главные и зависимые слова)

[3]см. онлайн: `https://universaldependencies.org/format.html`

и отношений между словами в предложениях[4]. Для выделения предложных конструкций и именных групп используются данные о частях речи, падеже и синтаксической роли и связях существительных.

До этого момента точки шаги предварительной обработки совпадают для каждого из изучаемых языков, а далее при выделении предложных групп в русском и именных групп в венгерском применяются разные подходы из-за описанных в главе 2 отличий между изучаемыми языками. Самые существенные проблемы демонстрируются ниже с примерами.

При автоматическом выравнивании с традиционными инструментами выравнивания на уровне слов (как Giza++, FastAlign, и т. д.) самую большую проблему создаёт отсутствие предлогов в венгерском. Это объясняется тем, что в венгерском меньше лексем в одном предложении из-за агглютинативного характера языка и из-за эллипсиса, а в русском их больше из-за аналитических черт, в том числе, из-за наличия предлогв: *Я говорил об этом с Петром* «Beszéltünk erről Péterrel» (6 лексем в русском, 3 лексем в венгерском). Вследствие данных факторов возникает несоответствие в количестве лексем в предложении, создавая таким образом значительные сложности для статистических моделей выравнивания, которые работают на основе частотности, позиции и количества слов в предложений.

В примере 10 демонстрируется проблема: *встретился* ***с Петром*** «találkoztál **Péterrel**», то есть, два слова «с» и «Петром» должны быть выравнены с одним венгерским существительным «Péterrel».

[4]см. роли онлайн: `https://universaldependencies.org/u/dep/index.html`

(10) Ты когда встретился с Петром ?

Te mikor találkoztál Péterrel ?

(11) Ты когда встретился [с Петром] ?

Te mikor találkoztál Péterrel ?

Мы предлагаем решить эту проблему с применением подхода выравнивания на уровне словосочетаний для русскоязычной части корпуса, объединяя элементы русских предложных конструкций в одну логическую единицу при выравнивании. Самое главное преимущество этого шага заключается в том, что он даёт возможность взаимно однозначного выравнивания русских предложных конструкций с венгерскими эквивалентами (см. пример 11).

В дальнейшем для обозначения объединённых лингвистических единиц используются квадратные скобки вокруг предложной конструкций: *[от Кремля]*, *[из России]*, *[в европейских странах]*.

Как мы уже говорили в главе 2, прилагательные также воспринимаются частями предложных конструкций, поэтому, согласно нашим решениям объединять определённые элементы предложных конструкций на этапе предварительной обработки текста, прилагательные также были включены в конструкции: *в белом доме*, *в новой столице*, *из-за молодого неопытного сотрудника*, *от души*, *без очевидного логичного объяснения* (см. примеры 12-13).

(12) Пётр живёт [в красивом городе]

Péter [egy szép városban] él

(13) Президент ушёл [с поста] [без логичного объяснения]

[Az elnök] [logikus magyarázat nélkül] elhagyta posztját

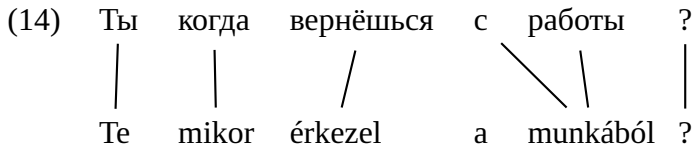
Согласно нашему решению прововить выравнивание на уровне словосочетаний, в венгерском языке были выделены именные группы, являющиеся потенциально эквивалентными русским предложным конструкциям. В примере 12 можно наблюдать, что именная группа *egy szép városban* «в красивом городе», состоящая из неопределённого артикля *egy*, прилагательного *szép* и существительного с падежным окончанием инессива выделена как именная группа в квадратных скобках.

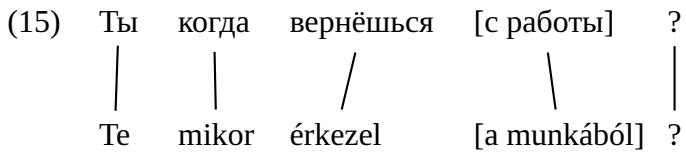
Артикли и прилагательные в венгерском языке также входят в именные группы: *egy nagy házban*, *kevés pénzből*, *teljes meggyőződésből*, *tiszta szívvel*, *egy szép városban*. В выделении именных групп помогает определённый порядок слов внутри групп: *артикль + прилагательные + существительное*, но артикли и прилагательные не обязательно присутствуют внутри каждой группы. Как видно выше, в венгерском артикли почти всегда занимают позицию предлогов русского – они находятся в самом начале именной группы, при их наличии, перед прилагательным и перед существительным[5], (***a*** *kertben* «в саду»; ***a*** *széken* «на стуле»).

Из-за совпадения позиции предлога в русском и артикля в венгерском при использовании способов выравнивания на уровне слов данная проблема может привести к тому, что инструменты, не учитывающие морфологию, выравнивают артикль с предлогом, как в примереах 14-15: «a» – определённый артикль венгерского занимает позицию русского предлога, а в этом случае даже длина не помогает традиционным инструментам найти правильное соответствие: большинство инструментов скорее всего выравнивает артикль «a» с предлогом «с», хотя «a» должно быть нуль-выравнено, поскольку у артикля нет эквивалента в русском языке. По данной причине, при предварительной обработке текста необходимо обращать внимание на проблему арти-

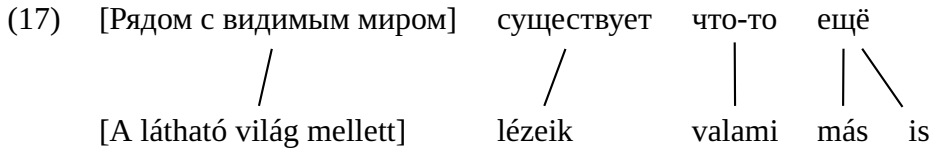
[5]исключением являются имена собственные, как в пр. 10

клей. Именно данной проблемой объясняется наше решение отнести артикли к именным группам: *a házban*, *a kertből*, *a városba*.





Как мы уже упомянули, в венгерском так же, как и в русском языке, имеются аналитические конструкции, которые могут выступать эквивалентами русских предложных конструкций. Послелоги могут изменять значение предыдущих существительных, поэтому необходимо их включать в именные группы и надёжно выделять их: *kertek alatt, ház felett, kuka mellett, szeretet nélkül, mindennel együtt* (см. пример 16-17).



После обработки корпуса, проведения морфосинтаксического анализа и определения критериев для выделения предложных конструкций в русском языке и именных групп в венгерском языке, мы можем выделить нужные конструкции с применением несколько простых правил. В случае русского языка шаблон предложной конструкций выглядит следующим образом: *предлог (1) + прилагательные / причастия (0+) + существительное (1)*. В нашей программы русские предложные конструкций выделяются следующим образом:

Algorithm 1 Алгоритм выделения русских предложных конструкций

ВХОД: Список токенов T

ВЫХОД: Список предложных конструкций P

function найти_предложные_конструкции(T) : P
    $P \leftarrow$ пустой список для выделенных предложных конструкций
    $p \leftarrow$ пустой список для токенов потенциальных словосочетаний
    for токен в T **do**
        if токен является предлогом **then**
            добавляем токен в p
        else if p не пустой **then**
            if токен является главным словам (head) для зависимого предлога $p[0]$ **then**
                добавляем токен в p
                создаем предложную конструкцию из содержания списка p и добавляем её в список P
                удаляем содержание списка p
            else
                добавляем токен в p
            end if
        end if
    end for
    return P
end function

То есть, мы опираемся не только на морфологические, но и на синтаксические параметры: после того, как встречается предлог, список элементов предложной конструкции дополняется, пока не встречается главное слово конструкции. Таким образом гарантируется, что все релевантные элементы входят в предложную конструкцию. На выходе получаем список выделенных предложных конструкций, готовых для выравнивания.

В случае венгерского языка шаблон именной группы выглядит следующим образом: *артикль (0-1) + прилагательное / причастие (0+) + суще-*

ствительное (1) + послелог (0-1). В нашей программе венгерские именные группы выделяются следующим образом:

Algorithm 2 Алгоритм выделения венгерских именных групп

ВХОД: Список токенов T

ВЫХОД: Список именных групп P

function найти_именные_группы(T) : P
    $P \leftarrow$ пустой список для выделенных предложных групп
    for токен в T **do**
        if токен имеет тег зависимого слова $nmod : obl$ или obl **then**
            $np \leftarrow$ *выделить_зависимые_слова*, T
            создаем именную группу из содержания списка np
            соотносим с именной группой подходящий тег на основе нашей классификации
            добавляем именную группу в список P
        end if
    end for
    return P
end function
// функция для выделения зависимых слов из предложения
// принимает главное слово в группе и список токенов
function выделить_зависимые_слова(tok, T) : np
    $np \leftarrow$ список элементов именной группы
    добавляем токен к концу np
    // выделяет послелог при наличии
    if следующий токен ($tok + 1$) является послелогом **then**
        добавляем следующий токен к концу np
    end if
    while предыдущий токен ($tok - 1$) является прилагательным **do**
        добавляем предыдущий токен к началу np
        $tok \leftarrow$ предыдущий токен
    end while
    while предыдущий токен ($tok - 1$) является числительным **do**
        добавляем предыдущий токен к началу np
        $tok \leftarrow$ предыдущий токен
    end while
    // выделяет артикли и другие детерминативы
    while предыдущий токен ($tok - 1$) является детерминативом **do**
        добавляем предыдущий токен к началу np
        $tok \leftarrow$ предыдущий токен
    end while

Для выделения венгерских именных групп мы опираемся больше на морфологические параметры: после выделения слова в роли адъюнкта ($obl, nmod:obl$), при их наличии, выделяются зависимые слова, в том числе, прилагательные, числительные, артикли и детерминативы. Для выделения послелога самый простой и надёжный способ – проверять их наличие после главного слова словосочетаний. В большинстве случаев эти шаги достаточны для выделения потенциальных эквивалентов русских предложных конструкций, но в некотором случае эквиваленты построены совсем по-другому, например, они выступают в роль субъекта или непосредственного объекта предложений. В этих случаях используются данные о позиции потенциальных эквивалентов в предложении и о длине словосочетаний для выделения нужных элементов.

3.4. Применение алгоритма сжатия цветков для выравнивания параллельных словосочетаний

После проведения предварительной обработки, морфосинтаксического анализа, выделения предложных конструкций и именных групп из корпуса и разметки выделенных конструкций, можно приступить к выравниванию выделенных конструкций.

Для самого эффективного решения проблемы выравнивания был выбран **алгоритм сжатия цветков** (также называется алгоритмом Эдмондса нахождения наибольшего паросочетания в произвольных графах), который был разработан Джеком Эдмондсом в 1961 году и опубликован в 1965 году [21]. Данный алгоритм способен решать проблему нахождения наибольшего паросочетания за полиномиальное время ($O(|E||V|^2)$).

Задача **нахождения наибольшего паросочетания** заключается в выборе как можно большего числа рёбер таким образом, чтобы ни одно выбран-

ное ребро не имело общей вершины ни с каким другим выбранным ребром.

Дан граф G с n вершинами $V = (v_1, v_2, v_3, \ldots, v_n)$, в котором требуется найти наибольшее паросочетание M так, что никакие два ребра v_j, v_k из выбранных не инцидентны друг другу (т.е. не имеют общих вершин). Простая цепочка $P = (v_1, v_2, \ldots, v_k)$ в G называется **чередующейся цепочкой**, если её рёбра попеременно не принадлежат M и содержатся в M. Чередующаяся цепочка P называется **увеличивающей**, если её первая и последняя вершины не принадлежат паросочетанию (также называются голыми). Иными словами, простая цепочка P является увеличивающей тогда и только тогда, когда вершина $v_1 \notin M$, ребро $(v_2, v_3) \in M$, ребро $(v_4, v_5) \in M, ...$, ребро $(v_{k-2}, v_{k-1}) \in M$, и вершина $v_k \notin M$. **Увеличение паросочетаний** вдоль пути P – это операция замены множества M на новое паросочетание $M_1 = M \oplus P = (M \setminus P) \cup (P \setminus M) M_1 = M \oplus P = (M \setminus P) \cup (P \setminus M)$ [25].

По теореме Бержа [1], паросочетание M является **наибольшим** тогда и только тогда, когда нет M-увеличивающего пути в G. Из этого следует, что паросочетание либо является наибольшим, либо его можно увеличить. Это означает, что начав с некоторого паросочетания, применяя алгоритм итеративно, можно вычислить наибольшее паросочетание путём увеличения текущего паросочетания с помощью увеличенного пути [25]. Формализация нахождения наибольшего паросочетания представлена в виде псевдокода ниже (см. алг. 3)[6]:

[6]Источник: `https://iq.opengenus.org/blossom-maximum-matching-algorithm/`

Algorithm 3 Алгоритм сжатия цветков

ВХОД: Граф G, начальное паросочетание M на G

ВЫХОД: наибольшее паросочетание $M*$ на G

1: **function** найти_наибольшее_паросочетание(G, M) : $M*$
2: $P \leftarrow$ *найти_увеличивающий_путь*(G, M)
3: **if** P не пустое **then**
4: **return** *найти_наибольшее_паросочетание*(G, увеличиваем M вдоль P)
5: **else**
6: **return** M
7: **end if**
8: **end function**

Для нахождения увеличивающего пути используется итеративный алгоритм, где фактически происходит сжатие цветков. **Цветок** B – это цикл в графе G, состоящий из $2k + 1$ рёбер, из которых в точности k принадлежат M и в котором есть вершина v (база) такая, что существует чередующаяся цепочка чётной длины (стебель) из v в голую вершину w. **Сжатие цветка** – это сжатие всего нечётного цикла в одну псевдовершину и, соответственно, все рёбра, инцидентные вершинам этого цикла, становятся инцидентными псевдовершине.

Для поиска увеличивающего пути используется дополнительная структура данных – лес F, множество индивидуальных деревьев, соответствующее порциям графа G. Лес F также может быть применён для поиска наибольших паросочетаний в двудольных графах. На каждой итерации алгоритм либо:

1. находит увеличивающий путь;
2. находит цветок и осуществляет рекурсию в сжатый граф;
3. делается вывод, что увеличивающего пути не существует.

Процедура построения просматривает вершины v и рёбра e графа G и инкрементально обновляет F соответствующим образом. Если v находится в дереве T леса, мы через $root(v)$ обозначим корень дерева T. Если и u, и v лежат в том же дереве T в F, через $distance(u, v)$ обозначим длину единственного пути из u в v в дереве T (см. алг. 4).

Algorithm 4 Нахождение увеличивающего пути

ВХОД: Граф G, паросочетание M в G

ВЫХОД: Увеличивающий путь P в G или пустой путь, если такого пути не найдено

1: **function** найти_увеличивающий_путь(G, M) : P
2: $F \leftarrow$ пустой лес
3: делаем все вершины и рёбра непомеченными в G, помечаем все рёбра M
4: **for** голой вершины v **do**
5: создаём дерево из одной вершины v и добавляем дерево в F
6: **end for**
7: **while** имеется непомеченная вершина v в F с чётным $distance(v, root(v))$ **do**
8: **while** существует непомеченное ребро e=v, w **do**
9: **if** w не в F **then**
10: // w входит в паросочетание, так что добавляем ребро, покрывающее e и w в F
11: $x \leftarrow$ сочетается с вершиной w в M
12: добавляем рёбра v,w и w,x в дерево для v
13: **else**
14: **if** $distance(w, root(w))$ чётно **then**
15: **if** $root(v) \neq root(w)$ **then**
16: // Сообщаем о увеличивающем пути в F $\cup\{e\}$.
17: $P \leftarrow$ путь $(root(v) \rightarrow \cdots \rightarrow v) \rightarrow (w \rightarrow \cdots \rightarrow root(w))$
18: **return** P
19: **end if**
20: **end if**
21: **end if**
22: помечаем ребро e
23: **end while**
24: помечаем вершину v
25: **end while**
26: **return** пустой путь
27: **end function**

Для решения проблемы выравнивания с помощью алгоритма сжатия

цветков можем внести некоторые изменения в выше представленный алгоритм. Во-первых, соответственно тому, что на графе представляются два вида вершин – русские предложные конструкции и венгерские именные группы, может быть реализован **двудольный граф**. В этом случае алгоритм сводится к стандартному алгоритму для паросочетаний в двудольных графах и решение упрощается. Во-вторых, появляется возможность извлекать из текстов лингвистическую информацию, позволяющую представлять определённую закономерность в графе. В этом случае встаёт вопрос о множестве M, которое даёт паросочетание с максимальным полным весом (мощность) $\sum w_1, w_2, \ldots, w_n$.

Соответственно, множество русских предложных конструкций U и множество венгерских именных групп V образуют взвешенный двудольный граф $G = (U, V, w(E))$, на котором вершины – предложные конструкции русского и именные группы венгерского языков и взвешенные ребра представляют потенциальные выравнивания.

На рисунке 2 указан взвешенный двудольный граф G, состоящий из множества русских предложных конструкций $U = r^1, r^2, \ldots, r^i$ и множества венгерских именных групп $V = h^1, h^2, \ldots, h^j$. Ребра на графе представляют *возможные* выравнивания с весами, но не все возможные рёбра представлены на графе ради читабельности. Высокий вес представляет собой большую вероятность правильного выравнивания, соответственно, низкий вес означает низкую вероятность правильного выравнивания.

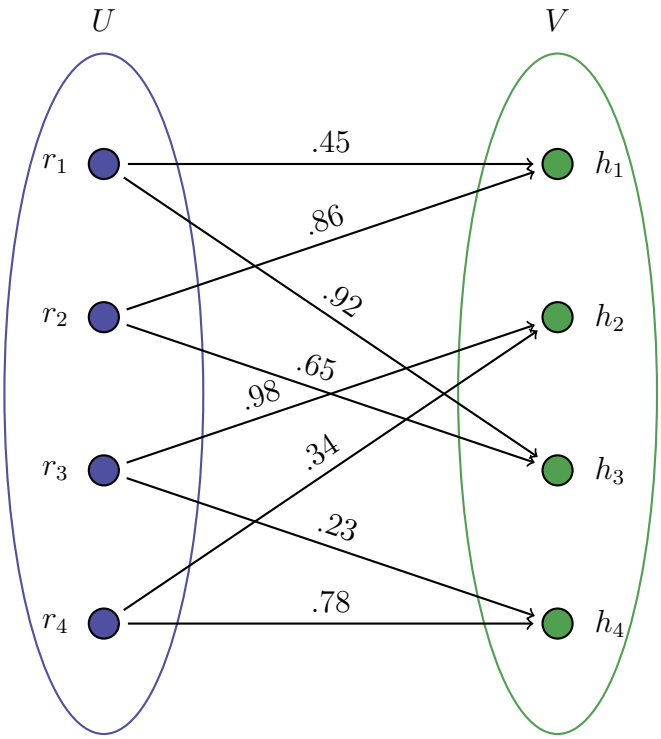


Рис. 2: Взвешенный двудольный граф

Весы, указывающие вероятность правильного выравнивания, состоят из трёх разных лингвистических компонентов, определённых на основе проведенного при предварительной обработке морфосинтаксического анализа:

1. Позиционный показатель близости параллельных словосочетаний;
2. Показатель близости по длине параллельных словосочетаний;
3. Морфологический показатель близости параллельных словосочетаний.

Первый компонент веса – **позиционный показатель близости**, извлеченный из параллельных предложений после выделения русских предложных конструкций и венгерских именных групп. Вес определяется путем вычета абсолютной величины разности индекса словосочетаний в русском и венгерском предложениях. Иными словами, те выравнивания, которые занимают ту же позицию в обоих предложениях получают повышенный вес.

Применение данного показателя обосновано тем, что в предложениях эквивалентные конструкции с большой вероятностью занимают ту же позицию, но в соседних позициях тоже могут быть эквиваленты, поэтому высокий вес даёт не только полное соответствие, а значения постепенно уменьшаются в зависимости от расстояния между позициями словосочетаний. Критерии и весы показаны на рисунке 3.

$$f(i,j) = \begin{cases} 12, & \text{if } |i-j| = 0 \\ 10, & \text{elif } |i-j| = 1 \\ 6, & \text{elif } |i-j| = 2 \\ 4, & \text{elif } |i-j| = 3 \\ 2, & \text{elif } |i-j| = 4 \\ 1, & \text{elif } |i-j| = 5 \\ 0, & \text{otherwise} \end{cases}$$

Рис. 3: Извлечение позиционного веса

Второй компонент веса – **показатель близости по длине параллельных словосочетаний**. Данный показатель определяется путем сравнения абсолютной величины разности длины русской предложной конструкций и венгерской именной группы. Длина – это количество слов в словосочетаний. Чтобы избежать возможного различия из-за отсутствия предлогов в венгерском, мы вычитаем единицу из разницы длин словосочетаний. Применение данного показателя обосновано тем, что эквивалентные конструкции с большой вероятностью состоят из примерно одинакового количества слов. По такой же логике, как у позиционного веса, значение показателя изменяется в зависимости от разности длины потенциальных эквивалентов. Критерии и весы показаны на рисунке 4.

Третий компонент веса – **морфологический показатель близости**, основанный на классификации соответствии между системой русских предложных конструкций и их самых вероятных эквивалентов в венгерском языке, представленной в разделе 2.5 главы 2. При высоком соответствии типов

$$f(n,m) = \begin{cases} 10, & \text{if } |n-m-1| = 0 \\ 4, & \text{elif } |n-m-1| = 1 \\ 2, & \text{elif } |n-m-1| = 2 \\ 1, & \text{elif } |n-m-1| = 3 \\ 0, \text{otherwise} \end{cases}$$

Рис. 4: Извлечение веса по длине словосочетаний

выделенных словосочетаний в русском и венгерском, морфологическим показателем веса назначено значение 16, при возможном соответствии – 8, в остальных случаях 0. Возможным соответствием являются все случаи, когда встречается кандидат второго уровня – то есть, редкое, но возможное выравнивание.

Морфологический показатель является самым значимым компонентом веса, потому что именно он может свидетельствовать о том, является ли выравнивание вероятным по нашей классификации. Однако, первый и второй компонент веса могут уверенно и точно указать на потенциально правильные выравнивания в случае отсутствия морфологического компонента.

Окончательная вероятность w – сумма вышеописанных трёх компонентов $w = \sum w1, w2, w3$. Вес определяется и назначается каждой возможной паре словосочетаний на графе $G = \forall x \in \{U \bigcup V\} \rightarrow w(u_j, v_k)$.

Lovász, Plummer (1986) утверждают, что можно решить взвешенную задачу о паросочетаниях на двудольных графах с помощью комбинаторного алгоритма, который использует невзвешенный алгоритм Эдмондса в качестве подпрограммы. Такой алгоритм реализован в библиотеке *networkx* на языке программирования *Python*. Мы применяем именно этот вариант алгоритма сжатия цветков.

3.5. Результаты анализа и проблемы метода

Благодаря нашему методу для выделения и выравнивания русских предложных конструкций и венгерских именных групп нам удалось достичь высоких показателей точности и полноты по сравнению с альтернативными методами.

Для оценки эффективности нашего метода для выравнивания были использованы меры **точность**, **полнота** и **F-мера**. Все эксперименты были проведены на оценочной части нашего корпуса, состоящей из 200 вручную выравенных параллельных предложений. В русскоязычной части оценочного корпуса около 392 предложных конструкций, в основном с предлогами *в, на, из, с, до, от, к, у, за, для*.

При оценке качества выравнивания вызывает сложности проблема выравнивания идиоматических выражений, свободных переводов, эллипсис и т. д. Это отрицательно влияет и на точность и на полноту. Также представляет проблему тот факт, что концепт «отношения между словами» (в случае выравнивания на уровне словосочетаний) является субъективным и не всегда существует четкая граница между правильным и не правильным выравниванием. Для обхода данной проблемы в оценочном корпусе были указаны ключевые слова с точки зрения выравнивания: в случае русского языка предлог и существительное, в случае венгерского языка существительное в падежной форме и, при наличии, послелог.

Точность – это доля правильных выравниваний относительно всех выравниваний при данном анализе:

$$prec = \frac{tp}{tp + fp} \tag{11}$$

где tp – правильные положительные (выделенные) выравнивания и fp –

неправильные положительные выравнивания.

Поскольку на точность не влияют не выровненные словосочетания, мы также рассматриваем полноту.

Полнота – это доля правильных выравниваний относительно всех потенциальных выравниваний в оценочном корпусе:

$$rec = \frac{tp}{tp + fn} \tag{12}$$

где tp – правильные положительные (выделенные) выравнивания и fn – неправильные отрицательные (не выделенные) выравнивания.

Надо отметить, что при изучении полноты мы сталкиваемся с тем, что не для каждой предложной конструкции можно найти эквивалента.

Полнота может неверно отражать качество выравнивания в экстремальных случаях, особенно при выравнивании художественной литературы. Для того, чтобы получить полную картину об эффективности выравнивания, мы также используем F-меру для оценки. **F-мера** представляет собой гармоническое среднее между точностью и полнотой:

$$F1 = \frac{2 \cdot prec \cdot rec}{prec + rec} \tag{13}$$

Согласно нашим ожиданиям, точность оказалась выше полноты. Высокая точность свидетельствует о том, что из выделенных конструкций наш метод эффективно выравнивает уверенные паросочетания, но не всегда находит эквивалент для каждой русской предложной конструкции в венгерском языке (см. таб. 4).

Таблица 4: Оценка точности, полноты и F-меры

	Точность	Полнота	F1
наш метод	81.3%	67.7%	73.9%
Giza++	32.2%	19.9%	24.6%
Simalign – Match	31.1%	27.4%	29.1%
Simalign – Itermax	31.1%	20.4%	24.5%
Simalign – Argmax	29.9%	14.0%	19.0%
fast_align	28.0%	12.4%	17.2%

Ниже демонстрируются примеры правильных результатов выделения предложных групп и их эквивалентов из нашего параллельного корпуса. В первой столбце указан русский предлог, во второй – класс (тег) выделенной русской предложной конструкции, в третьей – выделенная русская предложная конструкция и в четвертой – выделенная венгерская именная группа в пятой – класс (тег) выделенной венгерской именной группы. Перечень классов и тегов, использованной в таблице указан в таблице 3.

Таблица 5: Результаты выравнивания русских предложных конструкций и их эквивалентов в венгерском с указанием классов русской (клс1) и венгерской (клс2) конструкции

пред.	клс1	пред. гр.	им. гр.	клс2
в	Ine	в министерстве	a Minisztériumban	Ine
в	Ine	в Германии	Németországban	Ine
в	Ine	в Центральной Европе	a Közép-Európában	Ine
в	Ill	в порядок	rendbe	Ill
в	Ill	в воду	a vízbe	Ill
в	Ill	в свою клетку	a ketrecedbe	Ill
в	Ine	в конце	végén	Sup
в	Ine	в спящей конфигурации	ilyen alvós pozícióban	Ine
в	Ine	в антигосударственных кознях	elleni összeesküvéssel	Ins
в	Ine	В реальном коммунистическом мире	A valódi kommunista világban	Ine
в	Ine	в тумане	a ködben	Ine

в	Ine	в этом месяце	Ebben a hónapban	Ine
в	Ine	в глухих переулках	félreeső utcákban	Ine
в	Ill	в крепость	az erődbe	Ill
в	Ine	в своем типическом виде	tipikus formájában	Ine
в	Ine	в Гибралтаре	Gibraltárnál	Ade
в	Ine	в инвалидном кресле	tolószékben	Ine
в	Ine	в разной степени	különböző mértékben	Ine
в	Ine	в аэропорту	a repülőtéren	Sup
в	Ine	в твоей природе	a természetedre	Sub
в	Ill	в последний раз	utoljára	Sub
в	Ill	В первый снег	első hóesésében	Ine
на	Sup	на дне	fenekén	Sup
на	Sup	на двери	az ajtón	Sup
на	Sub	на битву	a csatába	Ill
на	Sub	на девчонку	egy lányra	Sub
на	Sup	на дороге	az úton	Sup
на	Sub	на минуту	egy percre	Sub
на	Sub	на запад	nyugatra	Sub
на	Sub	на диван	a heverőre	Sub
на	Sup	на территории	területén	Sup
на	Sup	на улице	utcán	Sup
на	Sub	на этот вопрос	a kérdésre	Sub
на	Sup	на факультете	A fakultáson	Sup
на	Sub	на него	Ford rá	Sub
на	Sup	на салате	salátánál	Ade
от	Abl	от автобуса	a busz elől	Nul
от	Abl	от солнца	a naptól	Abl
от	Abl	от него	tőle	Abl
от	Abl	от опасностей	a veszélytől	Abl
от	Abl	от вина	Bortól	Abl
от	Abl	от ужаса	a borzalomtól	Abl
от	Abl	от двери	Az ajtótól	Abl
от	Abl	от золотого пятнышка	a labdácskára	Sub

от	Abl	от него	belőle	Abl
от	Abl	от Коли	Koljától	Abl
от	Abl	от игры	a játéktól	Abl
с	Ins	с другим кавалером	más úrral	Ins
с	Nul	с удивительной важностию	csodálatos méltósággal	Ins
с	Del	с дома	egy házzal	Ins
с	Ins	с представителем	szószólójával	Ins
с	Ins	с мыслями	érveivel	Ins
с	Ins	с ватой	gumi füldugóval	Ins
с	Ins	с грустью	mélységes szomorúsággal	Ins
с	Ins	с бесконечной ненавистью	Határtalan gyűlölettel	Ins
к	All	к югу	keletre	Sub
к	All	к нам	hozzánk	Nul
к	All	к себе	magatokhoz	All
к	All	к стене	a falhoz	All
к	All	к горам	a hegyek felé	All
к	All	к ней	hozzá	All
у	Ade	у верхней ступеньки	a legfelső lépcsőn	Sup
о	Nul	об этом	erre	Sub
о	Nul	о болезни	betegségéről	Del
о	Nul	о том	arról	Del
из	Ela	из кукурузного початка	egy kukoricatorzsából	Ela
из	Ela	из двух девочек	a két leányból	Ela
за	Cau	за вещами	holmijáért	Cau
над	Nul	над ней	belőle	Abl
по	Dis	по небу	az égen	Sup

По выше представленным результатам видно, что в большинстве случаев (64% всех параллельных конструкций) класс русских предложный конструкций (клс1) совпадает с классом венгерской именной группы (клс2). Такой результат свидетельствует о том, что наша классификация, на основе которой были заданы параметры алгоритма, хорошо описывает самые ча-

стотные случаи соответствия изучаемых явлений.

Помимо этого, наши результаты также показывают, что в 32% случаев параллельные конструкции были выровнены не в соответствии с нашими ожиданиями, но их возможно было описать с нашими классами. Всего лишь в 4% случаев не нашлись соответствия в совокупности представленных нами классов.

Исходя из всех правильно выровненных параллельных конструкций, мы составили матрицу, в которой демонстрируется распределение переводных вариантов русских предложных конструкций в венгерском языке (см. табл. 6). Значение чисел в представленной матрице пропорционально соответствуют случаям, когда данный класс конструкций переводится на венгерский определённым образом.

Таблица 6: Матрица распределения переводных вариантов русских предложных конструкции в венгерском

	Abl	Ade	All	Cau	Del	Dis	Ela	Ill	Ine	Ins	Sub	Sup	др.
Abl	.76	-	-	-	-	-	-	-	.08	-	.08	-	.08
Ade	-	-	-	-	-	-	-	-	-	-	-	1.0	-
All	.58	-	-	-	-	-	-	-	-	-	.28	-	.14
Cau	-	-	-	1.0	-	-	-	-	-	-	-	-	-
Del	-	-	-	-	.77	-	-	-	-	.08	.15	-	-
Dis	-	-	-	-	-	-	-	-	-	-	-	1.0	-
Ela	-	-	-	-	-	-	1.0	-	-	-	-	-	-
Ill	-	-	-	-	-	-	-	.78	.1	-	.01	.01	-
Ine	-	.03	-	-	.03	-	-	-	.77	.03	.03	.01	-
Ins	-	-	-	-	-	-	-	-	-	.88	-	-	.12
Sub	-	-	-	-	-	-	-	.18	-	-	.82	-	-
Sup	-	.01	-	-	-	-	-	-	-	-	.01	.78	-

3.6. Выводы по главе 3

1. Мы создали специализированную систему выравнивания русских предложных конструкций и их эквивалентов в венгерском языке.

2. Выравнивание проводится на уровне словосочетания с использованием морфосинтаксических данных, извлечённых их текстов с помощью UDPipe – такой подход позволяет выделить и классифицировать изучаемые конструкции с большой эффективностью.

3. Метод основан на теории графов и применяет алгоритм сжатия цветков для выравнивания.

4. Данный метод может достичь значительно лучших результатов по сравнению с другими системами выравнивания благодаря решению проводить выравнивание на уровне словосочетаний и применять морфосинтаксические данные.

5. С помощью созданной нами системы мы проводили анализ распределения венгерских переводных эквивалентов русских предложных конструкций и подтвердили, что наша классификация хорошо описывает самые частотные соответствия изучаемых явлений.

Заключение

Данная работа посвящена изучению и раработке методов автоматического выравнивания русских предложно-падежных конструкций и их эквивалентов в венгерском языке.

В рамках настоящей работы были изучены особенности, проблемы и методы автоматического выравнивания параллельных текстов на разных уровнях и были представлены разные системы выравнивания – 3 на основе статистических и 1 на основе нейронных моделей выравнивания.

Были рассмотрены главные словоизменительные и словообразовательные процессы русского и венгерского языка с точки зрения задачи автоматического выравнивания параллельных текстов и была создана классификация соответствия русских предложных конструкций и их эквивалентов в венгерском языке.

Был создан русско-венгерский параллельный корпус, позволяющий изучать русские предложные конструкции и их эквиваленты в венгерском языке и был вручную размечен тестовый подкорпус, на основе которого можно оценить и сравнить разные методы выравнивания параллельных текстов. Корпус является подмножеством параллельных корпусов InterCorp, который содержит тексты литературного, научного и разговорного языка.

В рамках данной работы была создана система выравнивания на основе теории графов, которая с высокой эффективностью решает проблему выравнивания русских предложных конструкций и их эквивалентов в венгерском языке. Была проведена оценка метода и сравнение с уже существующими методами, которая показала его эффективность. Данная система была применена для вычета распределения венгерских переводных эквивалентов русских предложных конструкций и результаты были представлены в виде матрицы.

Таким образом, основные задачи настоящей работы были выполнены и

в результате проведенного исследования можно заключить, что созданный в рамках данной работы метод выравнивания выполняет задачу выравнивания русских предложных конструкций и их эквивалентов в венгерском языке с высокой эффективностью.

Наш метод выравнивания достигает 81.3 процент точности, то есть, по сравнению с уже существующими, универсальными системами выравнивания, данный специализированный для работы с русским и венгерским языками метод более чем 2 с половиной раза эффективнее альтернативных методов. Для получения таких высоких показателей были использованы морфосинтаксические параметры для выделения русских предложных конструкций и их эквивалентов в венгерском, был создан двудольный взвешенный граф из выделенных словосочетаний. Был применён алгоритм теории графов, алгоритм сжатия цветков для определения наибольших паросочетаний на графах.

Несмотря на высокие показатели оценки метода, у него также существует некое ограничение: данный метод был создан для работы с русским и венгерским языками, то есть, его применение за рамками изучения данных языков является ограниченным.

Планируется расширения метода для большего числа языков, для того чтобы его можно было бы использовать для решения ряда теоретических и практических задач компьютерной лингвистики, в том числе для изучения определённых конструкций в разных языках и для улучшения методов автоматического машинного перевода.

Список литературы

1. *Берж К.* Теория графов и её применение. — Москва : Издательство иностранной литературы, 1962.
2. *Виноградов В. В.* Русский язык. — Москва, Россия, 1972.
3. *Всеволодова М. В., Кукушкиеа О. В., Поликарпов А. А.* Русские предлоги и средства предложного типа. Т. 1. — Москва, 2018. — ISBN 978-5-9710-5505-1.
4. *Гарабик Р., Захаров В. П.* Параллельный русско-словацкий корпус // Труды международной конференции Корпусная лингвистика. — Санкт-Петербург, Россия, 2006. — С. 81—87.
5. *Зализняк А. А.* О понимании термина «падеж» в лингвистических описаниях // Проблемы грамматического моделирования. — Москва, Россия, 1973. — С. 53—87.
6. *Зализняк А. А.* Словоизменение // Большая российская энциклопедия. Т. 30. — Москва, Россия, 2015. — С. 445.
7. *Захаров В. П., Михайлова В. Д.* Контекстная грамматика предложных конструкций русского языка // Компьютерная лингвистика и вычислительные онтологии. — 2017. — С. 57—71.
8. *Золотова Г. А.* Синтаксический словарь: Репертуар элементарных единиц русского синтаксиса. — Москва, Россия : Эдиториал УРСС, 2006. — ISBN 5-354-01147-7.
9. *Коновалова Т. Е.* Специфика предлогов широкой семантики и её отражение в словах // Вопросы романо-германского языкознания. — Саратов, Россия : Издательство Саратовского университета, 1988. — С. 60—67.

10. *Кубрякова Е. С.* Словообразование // Лингвистический энциклопедический словарь. — Москва, Россия : Советская энциклопедия, 1990. — С. 467—469.

11. *Лопатин В. В., Улуханов И. С.* Восточнославянские языки. Русский язык // Языки мира. Славянские языки. — Москва, Россия : Academia, 2005. — С. 444—513. — ISBN 5-87444-216-2.

12. *Потемкин С. Б., Кедрова Г. Е.* Выравнивание неразмеченного корпуса параллельных текстов [Электронный ресурс]. — 2008. — URL: `http://www.dialog-21.ru/digests/dialog2008/materials/html/67.htm0` (дата обр. 03.12.2020).

13. Русская грамматика. — Москва, Россия, 1980.

14. *Скорик П. Я.* О соотношении агглютинации и инкоропорации // Морфологическая типология и проблема классификации языков. — 1965.

15. *Шальга А.* Венгерский язык в зеркале русского языка. — Будапешт, Венгрия : Танкёньвкиадо, 1984. — ISBN 963-17-7601-8.

16. *Antal L.* Egy új magyar nyelvtan felé. — Budapest, Hungary : Gondolat könyvkiadó, 1977. — С. 53—54. — ISBN 963-270-445-2.

17. *Brown P. F. Della Pietra S. A., Della Pietra V. J., Mercer R. L.* The mathematics of machine translation: Parameter estimation // Computational Linguistics. — 1993. — Т. 19(2). — С. 263—311.

18. *Čermák F., Rosen A.* The case of InterCorp, a multilingual parallel corpus // International Journal of Corpus Linguistics. — 2012. — Т. 17(3). — С. 411—427.

19. *Dempster A. P., Laird N. M., Rubin D. B.* Maximum Likelihood from Incomplete Data via the EM Algorithm // Journal of the Royal Statistical Society. Series B. — 1977. — T. 39 (1). — C. 1—38.

20. *Dyer C., Chahuneau V., Smith N. A.* A Simple, Fast, and Effective Reparameterization of IBM Model 2. — 2013. — URL: https://www.cs.cmu.edu/~ark/cdyer/fast_valign.pdf (дата обр. 26.01.2021).

21. *Edmonds J.* Maximum matching and a polyhedron with 0,1-vertices // Journal of Research of the National Bureau of Standards Section B. — 1965. — T. 69. — C. 125—130.

22. *Harris B.* Bi-text, a new concept in translation theory // Language Monthly (UK). — 1988. — T. 54. — C. 8—10.

23. *Holmqvist M.* Heuristic word alignment with parallel phrases // LREC 2010, Seventh International Conference on Language Resources and Evaluation. — Linköping, Sweden, 2010.

24. *Kuhn H. W.* The Hungarian Method for the assignment problem // Naval Research Logistics Quarterly. — 1955. — T. 2. — C. 83—97.

25. *Lovász L., Plummer M. D.* Matching Theory. — Budapest, Hungary : Akadémiai Kiadó, 1986. — C. 358—379. — ISBN 0 444 87916 1.

26. *Och F. J., Ney H.* A Systematic Comparison of Various Statistical Alignment Models // Computational Linguistics. — 2003. — T. 29, № 1. — C. 19—51.

27. On the Word Alignment from Neural Machine Translation [Электронный ресурс] / L. Xintong [и др.]. — 2017. — URL: https://www.aclweb.org/anthology/P19-1124.pdf (дата обр. 21.02.2021).

28. *Östling R., Tiedemann J.* Efficient word alignment with Markov Chain Monte Carlo // Prague Bulletin of Mathematical Linguistics. — 2016. — Т. 106. — С. 125—146.

29. Parallel corpora for medium density languages / D. Varga [и др.] // Proceedings of the RANLP 2005. — 2005. — С. 590—596.

30. *Sennrich R., Haddow B., Birch A.* Improving Neural Machine Translation Models with Monolingual Data // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Т. 1. — Berlin, Germany : Association for Computational Linguistics, 08.2016. — С. 86—96. — DOI: 10.18653/v1/P16-1009. — URL: https://www.aclweb.org/anthology/P16-1009.

31. SimAlign: High Quality Word Alignments without Parallel Training Data using Static and Contextualized Embeddings / M. Jalili Sabet [и др.] // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings. — Association for Computational Linguistics, 2020. — С. 1627—1643.

32. *Straka M.* UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task // Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. — Brussels, Belgium : Association for Computational Linguistics, 10.2018. — С. 197—207. — DOI: 10.18653/v1/K18-2020. — URL: https://www.aclweb.org/anthology/K18-2020.

33. *Szabó M. K., Vincze V., Nagy I.* HunOr: A Hungarian–Russian Parallel Corpus [Электронный ресурс]. — URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/262%5C_Paper.pdf (дата обр. 26.12.2020).