Автоматизация проверки текста на корейском языке на соответствие грамматическим правилам

А.Е. Елфимова, И.Н. Полякова

Аннотация **- В последнее время все большее внимание привлекают страны Азии, в частности Южная Корея. В результате в обществе набирает популярность изучение корейского языка. Но не все готовы ходить на курсы или заниматься индивидуально с преподавателем. В связи с этим стали быстро развиваться приложения для самостоятельного изучения иностранных языков. Для корейского языка таких приложений пока очень мало, и инструментарий у них скудный.**

В данной работе сделана попытка разработать и реализовать программы, осуществляющие автоматическую проверку текста на корейском языке на соответствие грамматическим правилам. С помощью такой программы любой желающий сможет самостоятельно проверить свой текст и исправить в нем возможные ошибки.

Для автоматизации проверки текстов на корейском языке на соответствие грамматическим правилам было решено построить модель грамматики языка, выбрать оптимальный работающий морфологический анализатор, а также подобрать наиболее подходящий для анализа текстов на корейском языке алгоритм синтаксического анализа, модифицировать его, адаптировать для корректного синтаксического анализа корейского языка по построенной грамматике и реализовать. Полученная программа позволяет анализировать текст на корейском языке и автоматизировать его проверку на соответствие грамматическим правилам.

В дальнейшем, для улучшения результата можно расширять набор базовых грамматических правил - чем больше правил, тем лучше программа будет справляться с более сложными предложениями. Расширяя модель, можно будет получить большее покрытие грамматических правил и особенностей корейского языка. Также можно попробовать улучшить результат морфологического анализа путем комбинированного использования различных морфологических анализаторов. Кроме того, одним из возможных направлений развития полученного анализатора является распознавание типов ошибок в тексте. Программа будет более дружелюбной для пользователей, если будет выводить сообщение о том, какой тип ошибки был допущен и как ее можно исправить.

Ключевые слова **- автоматический анализ текста, алгоритм Эрли, грамматика, корейский язык, морфологический анализ, синтаксический анализ.**

I. ВВЕДЕНИЕ

В настоящее время особую популярность набирает изучение иностранных языков. По данным ВЦИОМ [1] по состоянию на 2019 год растет не только доля людей, которые считают нужным изучать иностранный язык, но и доля людей, которые действительно начали изучать иностранный язык.

В частности, растет спрос на изучение корейского языка. Южная Корея – это одна из быстро развивающихся стран, их культура быстро распространяется среди молодежи. В результате молодые люди стремятся выучить их язык, чтобы смотреть фильмы и сериалы в оригинале, ездить в Южную Корею и общаться с ее жителями. В отчете за 2023 год от Duolingo – популярная онлайн платформа для изучения иностранных языков – корейский язык занимает шестое место по популярности [2], а в отчете ассоциации современного языка Америки [3] говорится о том, что студенты все больше и больше записываются на изучение корейского языка, в то время как европейские языки частично теряют свою аудиторию.

Благодаря такому спросу на изучение языков распространились приложения для самостоятельного освоения иностранного языка. Данные приложения предлагают и карточки со словами, и упражнения на запоминание грамматики, и прослушивание правильного произношения. Кроме прочего, особенно востребованы системы для автоматической проверки текста. Например, специальная система для проверки текстов используется на международном экзамене по английскому – TOEFL. Несмотря на это, для корейского языка еще нет достаточно хорошего варианта, который бы освободил учителей от ручной проверки работ.

Задача автоматизации проверки письменных текстов существует давно. Еще в 1999 была опубликована статья [4], описывающая соответствующую систему. Эта система была спроектирована на основе латентно-семантического анализа. Тем не менее, в корейском языке много различных нюансов, из-за которых еще нет систем автоматической проверки текста, которые по качеству сравнились бы с аналогичными системами, например, для английского языка.

Для решения задачи написания подобной системы нужен хороший синтаксический анализатор. Существуют различные алгоритмы

синтаксического анализа [5]. Их достаточно легко применять к языкам программирования, так как в них строгие правила, но для естественного языка нужно сначала эти правила формализовать и выразить в виде некоторой модели.

В корейском языке довольно строгие грамматические правила. Например, в предложении слова идут в определенном порядке, падежи, союзы записываются в виде окончаний слова. По таким правилам можно составить модель грамматики, которая будет использоваться для синтаксического анализа. А чтобы выделить в словах основы и окончания, нужно применить морфологический анализ. Для корейского языка существуют несколько хороших морфологических анализаторов.

Получается, что для автоматизации проверки текстов на корейском языке на соответствие грамматическим правилам нужно построить модель грамматики языка, выбрать морфологический анализатор, а также выбрать, модифицировать (адаптировать к корейскому языку) и реализовать алгоритм синтаксического анализа.

II. МОДЕЛЬ ГРАММАТИКИ КОРЕЙСКОГО ЯЗЫКА

A. Грамматические правила корейского языка

В корейском языке определена достаточно строгая последовательность слов в предложении. Конечно, некоторые члены предложения могут быть в разных местах предложения, но это вполне возможно показать в модели грамматики. Для сравнения, в русском языке допустимы варианты как подлежащее + сказуемое, так и сказуемое + подлежащее. В корейском же сказуемое расположено всегда в конце предложения, остальное рассматривается как ошибочное написание. Для такого языка намного проще построить модель грамматики, которая покроет как можно больше грамматических правил.

Еще одной важной особенностью можно считать окончания слов в корейском языке. Для обозначения положения слова в предложении используются окончания. По ним можно понять, что является подлежащим, что сказуемым, что дополнением и т.д. Это означает, что в нем для каждого предложения можно построить одно дерево разбора. В противовес, в русском языке существует некоторая неоднозначность. Например, мать любит дочь. Это предложение синтаксически возможно разобрать несколькими способами: подлежащим может быть как «мать», так и «дочь». К счастью, в корейском таких неоднозначность нет благодаря различным окончаниям у членов предложения.

Таким образом, предложение «Я открыл дверь» в корейском языке выглядело бы следующим образом:

«Я는 дверь을 открыл다»,

а именно «나는 문을 열었습니다». Из этого примера видно, что, во-первых, слова в предложении стоят в определенном порядке, и он не совпадает с постановкой слов в русском языке, и во-вторых, члены предложения обозначены своими окончаниями.

Аналогично обозначаются союзы в предложениях. Например:

Я спал, затем учил корейский - 저는 자고 한국어를 공부했어요.

На русском это предложение выглядело бы следующим образом:

Я는 спал고, корейский를 учил다.

Окончание у слова «спал» в этот раз не является окончанием сказуемого, а является окончанием-союзом, который и придает значение «затем».

Пример сложноподчиненного предложения на корейском языке:

버스가 출발하자마자 사람들은 움직였어요

Как только автобус поехал, люди двинулись.

В первом слове окончание 가, которое означает, что это подлежащее, но не определенное. В русском языке сложно подобрать аналогию, но в английском языке используются артикли перед словами. В этом предложении в английском переводе стоял бы неопределенный артикль «a». Во втором слове окончание 자마자 – это означает союз «как только». Союзы в корейском языке могут быть как отдельными словами, так и окончаниями. Окончание 은 в третьем слове тоже означает, что это подлежащее, но определенное, в английском переводе предложения стоял бы артикль «the». Наконец, последнее слово, как было написано выше, - сказуемое и заканчивается на 였어요.

Как видим, наличие окончаний сильно облегчает построение простой модели грамматики корейского языка. К сожалению, не у каждого слова в предложении есть окончание. И эту важную особенность корейского языкат также необходимо учитывать при построении грамматики.

B. Построение формальной грамматики

Для реализации синтаксического анализатора для текстов на корейском языке была рассмотрена и использована теория формальных языков и грамматик, а также классификация грамматик и языков по Хомскому [6, 7]. Для решаемой задачи подходят распознающие грамматики, так как на вход программа получает предложение для проверки, а на выходе выдает, правильно ли оно.

Основная идея заключается в том, чтобы в модели грамматики использовать нетерминалы как члены предложения, а терминалы – часть речи, полученные из тега после морфологического анализа. Для построения грамматики по такому принципу важно изучить теорию синтаксического

анализа естественного языка. Основой предложения являются главные члены – подлежащее и сказуемое. Кроме того, в предложении могут быть второстепенные члены: дополнение, определение и обстоятельство.

Также предложения бывают простыми и сложными. Это определяется количеством основ в предложении. В сложных предложениях есть союзы, которые соединяют части предложения. Все это верно как для русского, так и для корейского языка.

В результате изучения корейской грамматики были собраны следующие основные члены предложения: подлежащее, сказуемое, наречие, дополнение, обстоятельство времени или места, определение. Кроме того, в предложении могу быть союзы как для перечисления, так и для соединения частей предложения.

Очевидно, что модельная версия формальной грамматики не сможет покрыть все грамматические правила корейского языка. По этой причине для решения задачи рассматривается лишь такое подмножество множества правил, чтобы алгоритм работал с большей частью предложений, использующих наиболее распространенные конструкции языка.

Формальная грамматика строилась в соответствии с правилами корейского языка, взятыми из классических учебников [8]. В полученной грамматике использовались следующие обозначения: S – начальный символ, T – обстоятельство времени, P – обстоятельство места, O - дополнение, V – сказуемое; t, p, o, v– обозначают окончания, которые соответствуют данной позиции, n – подлежащее, a – определение, поэтому стоит перед разными членами предложения, c – союз, может быть связкой между двумя словами, a может – между двумя частями предложения, d – наречие, поэтому находится только перед глаголом. Для понимания идеи приведен пример ранней версии грамматики G:

S -> anT | nT | T
T -> atP | tP | P
P -> apO | pO | O
O -> aoV | oV | V
V -> cS | dv | v

Данная грамматика покрывает небольшую часть грамматики корейского языка, которую изучают на первых занятиях. Итоговая грамматика содержит больше правил, в ней учитывается как основа слова, так и окончание. Разделение на основу слова и окончание возможно благодаря морфологическому анализу. Разработанная грамматика является контекстно-свободной [9]. Данная модель была выбрана в силу удобства, она достаточно интуитивно понятна. Члены предложения рассматриваются в той последовательности, в которой они могли бы находиться в предложении. Так как в конце точно должно быть сказуемое, то пустое предложение недопустимо.

III. МОРФОЛОГИЧЕСКИЙ АНАЛИЗАТОР ДЛЯ КОРЕЙСКОГО ЯЗЫКА

Для достижения поставленной цели будем использовать библиотеку KoNLPy [10]. Это библиотека для обработки текстов на естественном языке, разработанная специально для корейского языка. Для морфологического разбора данная библиотека предлагает несколько вариантов: Hannanum, Kkma, Komoran, Mecab, Okt.

Hannanum [11] был разработан еще в 1999 в Корейском институте передовых технологий. С тех пор этот анализатор еще дорабатывали. Этот процессор сначала разбивает предложение на слова, находит основы слов (стемминг), определяет все возможные варианты морфологического разбора, по результатам наиболее вероятного варианта проставляет словам теги с информацией о части речи. Одной из основных проблем этого анализатора считается устаревший словарь. По этой причине он плохо работает со сленгом, например.

Kkma [12] был разработан Сеульским национальным университетом. В отличие от предыдущего, данный анализатор устойчив к отсутствию пробелом между словами. Все возможные морфемы генерируются с помощью динамического программирования, подходящие варианты сортируются в порядке целесообразности. Для улучшения анализа используются скрытые Марковские модели. Для ускорения процесса используется заранее скомпилированный словарь. Поиск подходящих морфем осуществляется с помощью метода проверки смежных условий.

Komoran отличается от других морфологических анализаторов своей простотой. Словарь для данного анализатора представлен в виде текстового файла, который в любое время можно прочитать и отредактировать под свои нужды.

В Mecab [13] морфемы определяются с помощью модели условного случайного поля и алгоритма Витерби.

Okt – морфологический анализатор, который разрабатывался для работы с текстами из социальных сетей. В данном анализаторе тегов меньше, чем в остальных анализаторах. Например, в остальных анализатор несколько тегом для существительных, чтобы различать собственные имена, нарицательные и прочие возможные существительные. В Okt любое существительно определяется просто тегом «Noun».

Чтобы выбрать наиболее подходящий морфологический анализатор, сравним результаты их работы на разных предложениях.

아버지가방에들어가신다. – Отец заходит в комнату.

Hannanum: 아버지가방에들어가 – N, 이 – J, 시ㄴ다 - E

Kkma: 아버지 – NNG, 가방 – NNG, 에 – JKM, 들어가 – VV, 시 – EPH, ㄴ다 – EFN

Komoran: 아버지가방에들어가신다 - NNP

Mecab: 아버지 – NNG, 가 – JKS, 방 – NNG, 에 – JKB, 들어가 – VV, 신다 - EP+EC

Okt: 아버지 – Noun, 가방 – Noun, 에 – Josa, 들어가신 – Verb, 다 – Eomi

В этом примере подробно расписаны результаты работы каждого морфологического процессора. Видно, что второй, третий и четвертый расставляют теги по единой системе. В этих тегах не только часть речи, но еще и некоторые уточнения. Например, NNP – это имя собственное, а NNG обычное существительное. В первом и последнем анализаторах теги означают только часть речи.

나는 밥을 먹는다 и 하늘을 나는 자동차 – «Я ем рис» и «Летающая машина»

В этих двух предложениях слово 나 выступает в роли разных частей речи. В первом предложении оно означает «я». Второе предложение можно дословно перевести как «Небо летать машина», в нем 나 означает летать. Этот пример важен, чтобы указать на то, что все анализаторы, кроме Kkma, во втором предложении обозначили 나 как существительное. И только Kkma обозначила его как глагол.

Также важно отметить, что Kkma и Okt лучше остальных анализаторов справились со сленгом. Интересно отметить, что Okt изначально разрабатывался как морфопроцессор для социальных сетей.

По результатам тестирования для реализации автоматической проверки текста на корейском языке был выбран морфопроцессор Kkma. В данном морфологическом анализаторе больше всего тегов, а также заметно лучше скорость работы, по сравнению с остальными. Как уже уточнялось ранее, система разрабатывается для студентов или людей, для которых корейский язык не является родным, поэтому сленговые предложения не будут рассматриваться как правильные. Намного важнее, чтобы анализатор правильно расставил теги для последующего синтаксического анализа.

IV. ОБЗОР АЛГОРИТМОВ СИНТАКСИЧЕСКОГО АНАЛИЗА

Пусть дана некоторая контекстно-свободная грамматика G с начальным символом S, L(G) – язык порождающей грамматики, и цепочка w принадлежит этому языку [14]. Если занумеровать правила этой грамматики, то левым разбором будем называть последовательность правил, примененных при левом выводе цепочки w из S. Аналогично определяется правый разбор.

Алгоритмы синтаксического анализа применяются в различных сферах и уже давно исследуются. За это время были разработаны самые разные подходы. Самые простые и известные – это нисходящий и восходящий разборы. Алгоритмы достаточно интуитивно понятны, и их реализация не доставляет больших проблем. Самым главным недостатком данных алгоритмов является число шагов, необходимых для разбора. Однако есть несколько подходов к ускорению разборов [15]. Данные алгоритмы называются методами синтаксического разбора с возвратами.

В отличие от методов синтаксического анализа с возвратами, табличные методы применимы ко всем контекстно-свободным грамматикам, так как любая такая грамматика приводима к нормальной форме Хомского. К табличным методам относится алгоритм Кока-Янгера-Касами [16]. В результате алгоритма строится таблица, по которой можно получить все левые разборы цепочки w [17]. Однако в общем случае нельзя получить все левые разборы входной цепочки за менее чем экспоненциальное время, так как у нее может быть экспоненциальное число разборов.

Другим табличным методом является алгоритм Эрли [18]. Как и предыдущий алгоритм, он требуют времени n^3 и емкости n^2, где n – длина входной цепочки. Алгоритм Эрли во всех отношениях так же хорош, как и алгоритм Кока-Янгера-Касами, а для многих грамматик даже лучше. Для однозначных грамматик время квадратичное, а с некоторыми модификациями можно добиться и линейных времени и емкости.

Идея алгоритма Эрли заключается в том, чтобы построить всевозможные нетерминалы из подстрок входной цепочки. Каждый символ считывается по порядку слева направо, и на каждой позиции формируется список всех ситуаций - частично завершенных продукций грамматики, из которых могут быть выведены полученные префиксы входной цепочки его части.

В алгоритме используются 6 правил [15]:

1. Если S -> α – правило из P, включить [S -> * α, 0] в I0.
2. Если [B -> γ*, 0] принадлежит I0, включить в I0 ситуацию [A -> α B* β, 0] для всех [A -> α *B β, 0], уже принадлежащих I0.

3. Допустим, что [A -> α *B β, 0] принадлежит I0. Для каждого правила из P вида B -> γ включить в I0 ситуацию [B -> γ *, 0].
4. Для каждой ситуации [B -> α *a β, i] из Ij-1, для которой a = aj, включить в Ij ситуацию [B -> α a* β, i].
5. Пусть [A -> α *, i] принадлежит Ij. Искать в Ij ситуации вида [B -> α *A β, k]. Для каждой из них включить в Ij ситуацию [B -> α A* β, k].
6. Пусть [A -> α *B β, i] принадлежит Ij. Для каждого B -> γ из P включить в Ij ситуацию [B -> * γ, j].

Сначала применяется первое, чтобы начать формировать первое множество ситуаций. Затем применяются второе и третье, пока это возможно. После того, как все ситуации для нулевой позиции получены, применяется четвертое правило для начала формирования следующего списка ситуаций. Аналогично, пятое и шестое правила применяются, пока это возможно. Эти шаги повторяются для всех позиций.

Чтобы понять, принадлежит ли входная цепочка языку, проверяются правила в последнем множестве ситуаций. Если там есть правило вида [S -> γ *, 0], то есть в котором левый нетерминал является начальным, символ «звездочка» в конце правила, а позиция равна 0, то это значит, что входная цепочка принадлежит языку, иначе – не принадлежит.

Алгоритм Эрли далеко не идеален. У него есть два основных недостатка. Самый главный из них – это ошибки при обработке правил нулевой длины. Кроме того, при правосторонней рекурсии приходится применять алгоритм дважды. В связи с этими недостатками, Ахо и Ульман в 1972 году описали вариант алгоритма Эрли с возможностью работать с правилами нулевой длины [19]. К сожалению, такой подход сильно утяжелил алгоритм, необходимые ресурсы на его работу значительно увеличились. В последующее время также разрабатывались способы исправления проблем алгоритма, чтобы он работал и с нулевыми правилами, и с правой рекурсией, и быстрее.

V. РАЗРАБОТКА СИНТАКСИЧЕСКОГО АНАЛИЗАТОРА

A. Реализация алгоритма Эрли

Для реализации задачи синтаксического анализа текстов на корейском языке были использованы техники объектно-ориентированного программирования [20]. Класс Earley [21] (реализует синтаксический анализ) написан на языке Python [22]. Работа программы построена таким образом, что на вход подается предложение для проверки, далее поочередно применяются функции класса Early, и на выходе получается результат проверки. Функции класса реализуют правила алгоритма Эрли.

Первоначальная реализация алгоритма была проверена на тестовых предложениях английского и русского языков и соответствующим им грамматикам, взятых из различных источников. Этот этап был важен для отладки самого алгоритма в случае возникновения проблем. На тестовых входных данных было проверено, что синтаксический анализ на основе реализованного алгоритма работает корректно, все тесты были успешно пройдены. Таким образом, можно было приступить к этапу адаптации алгоритма для корейского языка.

B. Модификация алгоритма для синтаксического анализа корейского языка

В первую очередь для работы синтаксического анализатора была написана структура с правилами вывода из построенной модели грамматики корейского языка. Полученная структура несколько отличается от самой формальной модели грамматики. В корейском языке у одного члена предложения могут быть разные окончания. Это никак не влияет на модель, но влияет на реализацию синтаксического анализатора. Окончание у слов может быть разным по двум причинам. Во-первых, оно зависит от написания слова, к которому присоединяется, а именно от последней буквы в слове. Во-вторых, окончания также передают степень уважения в предложении. Обращение к другу и к руководителю будут явно выражены в речи человека.

Например, предложение
«Я всегда вижусь со своим учителем» в обращении к другу выглядит следующим образом:

나는 나의 선생님을 항상 봐, но при обращении к руководителю окончание глагола, то есть последнего слова в предложении, должно быть в уважительной форме:

저는 저의 선생님을 항상 봅니다 .

Из примера видно, что поменялось окончание глагола на более формальное: было «봐», стало «봅니다», хотя смысл слова не изменился. И также использовался более уважительный вариант местоимения «я».

При реализации проверяющей системы важно сделать ее как можно более гибкой, учесть как можно больше правил, случаев. Терминалы в программе реализованы как множества строк – окончаний. Такой вариант программы (см. Рис.1) был тщательно отлажен и проверен.

Было замечено, что одного алгоритма Эрли недостаточно, чтобы работать с корейским языком – алгоритм работает правильно, предложения анализируются в соответствии с грамматикой, но не хватает дополнительных проверок, соответствующих особенностям языка. Таким образом, были ложно интерпретированы неверные

предложения, и их было больше, чем хотелось бы. Так было принято решение расширить возможности класса Early.

К реализованному алгоритму Эрли были добавлены вспомогательные функции. Они направлены на работу с особыми случаями, которые не обрабатываются при морфологическом и синтаксическом анализе. Например, если вежливость глагола не соответствует вежливости местоимения.

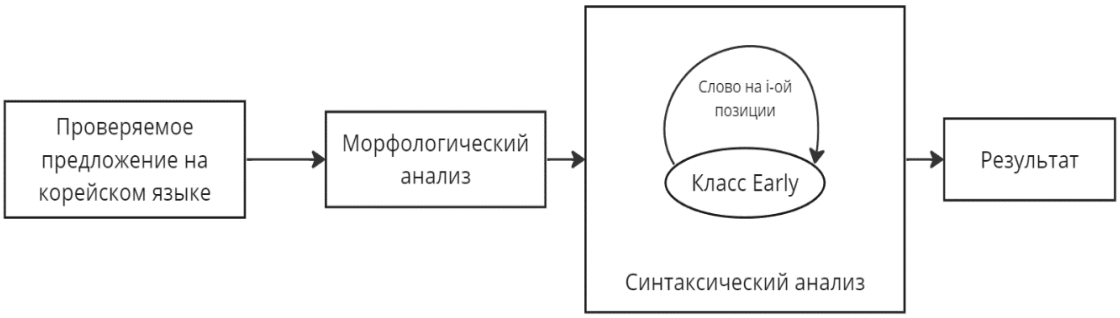

Рисунок 1. Схема работы реализованного синтаксического анализатора

Формальная грамматика является необходимым скелетом предложения, последовательность главных членов предложения не может быть изменена. Однако между ними могут находиться другие описательные слова. Это означает, что программа может работать с достаточно сложными предложениями, потому что не анализирует каждое слово. С некоторыми потерями в точности получается более эффективный анализ предложений.

Таким образом, были добавлены функции, необходимые для работы с корейским языком. Полученная модификация заметно улучшила работу алгоритма. Реализованная программа работает по схеме, изображенной на рис. 2.

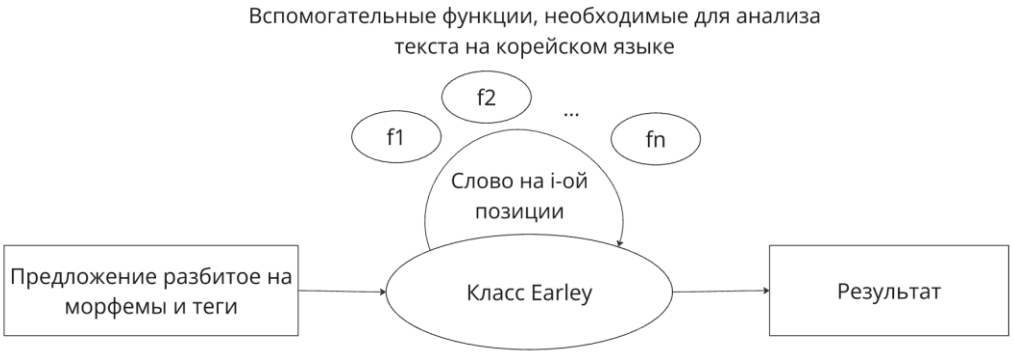

Рисунок 2. Схема работы реализованного синтаксического анализатора

На вход подается предложение для проверки, которое прошло морфологический анализ и было разбито на морфемы и теги. Класс Earley последовательно применяет функции – шаги алгоритма Эрли, но теперь в необходимых местах применяются вспомогательные функции, которые прямо влияют на будущий результат. После всех шагов вызывается завершающая функция, которая проверяет, соответствует ли предложение грамматике корейского языка.

VI. ИСПОЛЬЗОВАНИЕ РЕАЛИЗОВАННОГО АЛГОРИТМА ДЛЯ ПРОВЕРКИ КОРЕЙСКИХ ТЕКСТОВ

Реализованная программа была сначала проверена на тестовых предложениях из учебников по грамматике корейского языка [23, 24]. Они подбирались на такие правила, которые программа распознает, то есть по тем грамматическим правилам, на основе которых разрабатывалась модель. Целью этих тестов было проверить, что программа распознает правильные предложения,

которые соответствуют формальной грамматике и не пропускает неправильные варианты использования выбранных правил. Был получен ожидаемый результат в 100% (см. Рис. 3). Данная диаграмма говорит о том, что в результате работы программы все правильные предложения были отмечены правильными и все неправильные предложения были отмечены неправильными. Таким образом, система работает ровно так, как и ожидалось.

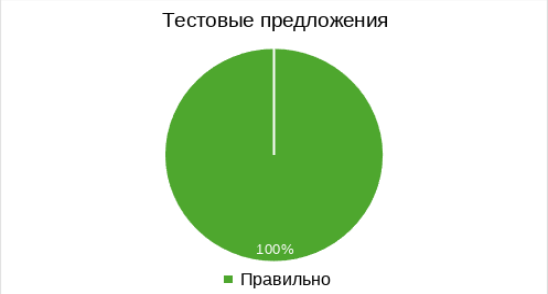

Рисунок 3. Результат работы синтаксического анализатора на тестовых предложениях.

Затем проверка проводилась на текстах из классической литературы [25,26,27], где ожидались довольно высокие показатели правильности соответствия грамматике. Были подобраны различные предложения, независимо от того, подходят они под смоделированную формальную грамматику или нет.

Рисунок 4. Результат работы синтаксического анализатора на текстах из художественной литературы.

На полученной диаграмме (см. Рис. 4) видно, что большую часть предложений система смогла правильно оценить. Однако есть небольшой процент случаев, с которыми она не справилась. Эти предложения отмечены неправильными по двум причинам. Первая заключается в том, что предложения включали в себя некое исключение, которое не было обработано вспомогательными функциями, а также не попало в формальную грамматику корейского языка. Вторая причина– писатель мог осознанно написать грамматически неправильное предложения. Это могла быть фраза персонажа, который по каким-то причинам говорит неправильно, или это мог быть литературный прием.

Таким образом, полученная статистика по художественной литературе позволяет заметить, что реализованный синтаксический анализатор достаточно хорошо справляется с поставленной задачей, однако есть некоторый небольшой процент ошибочно неправильных предложений.

Также была собрана группа предложений из социальных сетей или комментариев пользователей на различных сайтах [28, 29, 30]. Принято считать, что в интернете люди намного меньше следят за грамматикой. Такой вариант предложений был выбран с целью проверить, насколько хорошо система различает неправильные предложения. Кроме того, главное отличие предложений из социальных сетей от художественной литературы – это наличие сленга. Результаты проверки приведены на рис. 5.

Рисунок 5. Результат работы синтаксического анализатора на текстах из социальных сетей.

На диаграмме видно, что в социальных сетях заметно больше грамматически неправильных предложений. Однако результат заметно лучше, чем ожидалось. Из этого следует, что корейцы достаточно хорошо следят за своим языком. А также в реализованной программе допустим неформальный стиль, без него данный результат был бы совершенно другим. В процент неправильных предложений в основном входят все те же исключения, описанные ранее.

Таким образом, была собрана статистика по соблюдению грамматических правил в разных сферах использования корейского языка, а также была проверена работоспособность реализованного синтаксического анализатора.

VII. ЗАКЛЮЧЕНИЕ

Построенная грамматика и реализованный синтаксический анализатор позволяют автоматизировать проверку текста на корейском языке на соответствие грамматическим правилам. Студенты, изучающие корейский, или любые желающие проверить себя могут самостоятельно проанализировать свой текст и исправить возможные ошибки.

Для реализации было использовано ограниченное количество грамматических правил корейского языка. Данные правила составляют некоторый базис, они изучаются на уроках корейского языка для иностранных учащихся. В дальнейшем, для улучшения результата можно расширять набор грамматических правил - чем больше правил, тем лучше программа будет справляться с более сложными предложениями. Расширяя модель, можно будет получить большее покрытие грамматических правил и особенностей корейского языка.

Морфологический анализ предложений тоже можно улучшить. Минимизировать вероятность ошибки морфопроцессора можно с помощью комбинирования различных морфологических анализаторов, сравнения их результатов и использование того варианта, который встретился у большинства.

Кроме того, одним из возможных направлений развития полученного анализатора является распознавание типов ошибок в тексте. Программа будет более дружелюбной для пользователей, если будет выводить сообщение о том, какой тип ошибки допущен и как ее можно исправить.

БЛАГОДАРНОСТИ

Работа выполнена в рамках НИР “Математическое и программное обеспечение перспективных систем обработки символьной информации с элементами искусственного интеллекта“, проводимой на кафедре алгоритмических языков факультета вычислительной математики и кибернетики Московского государственного университета имени М.В.Ломоносова.

БИБЛИОГРАФИЯ

[1] ВЦИОМ [электронный ресурс] – Электрон.дан. – URL: https://wciom.ru/analytical-reviews/analiticheskii-obzor/inostrannyj-yazyk-perspektivnaya-investicziya (дата обращения 05.02.2025)
[2] Duolingo [электронный ресурс] – Электрон.дан. - URL: https://blog.duolingo.com/2023-duolingo-language-report/ (дата обращения 03.11.2024)
[3] Modern Language Association [электронный ресурс] – Электрон.дан. – URL: https://www.mla.org/Resources/Guidelines-and-Data/Reports-and-Professional-Guidelines/Enrollments-in-Languages-Other-Than-English-in-United-States-Institutions-of-Higher-Education (дата обращения 05.10.2024)
[4] Foltz, P.W., Laham, D. and Landauer, T.K. (1999). The intelligent essay assessor: Applications to educational technology. Interactive Multimedia Electronic Journal of Computer-Enhanced Learning (IMEJ) 1(2), 939–944.
[5] Компиляторы: принципы, технологии и инструментарий / Ахо А. В., Лам М. С., Сети Р., Ульман Д. Д. – М.: Издательский дом «Вильямс», 2008. – 1128 с.
[6] Хомский Н. Синтаксические структуры // Новое в лингвистике. – 1957. Вып. 2. – С. 412–537.
[7] Волкова И.А., Вылиток А.А., Руденко Т.В. Формальные грамматики и языки. Элементы теории трансляции: Учебное пособие для студентов II курса (издание третье, переработанное и дополненное). – М.: Издательский отдел факультета ВМиК МГУ им. М.В.Ломоносова (лицензия ИД № 05899 от 24.09.2001), 2009 – 115 с.
[8] Ahn. J., Lee K., Han H. Korean Grammar in Use Beginning to Early Intermediate. – Darakwon, 2010. – 345 с.
[9] Кук Д., Бейз Г. Компьютерная математика. – М.: Наука. Физматлит, 1990. – 385 с.
[10] KoNLPy documentation [электронный ресурс] – Электрон.дан. – URL: https://konlpy.org/en/latest/ (дата обращения 15.09.2024)
[11] Park, E.L.; Cho, S. KoNLPy: Korean Natural Language Processing in Python - In Proceedings of the Annual Conference on Human and Language Technology, Chuncheon, Republic of Korea, 10–11 October 2014; - 133–136 с.
[12] Lee, D.; Yeon, J.; Hwang, I.; Lee, S. KKMA : A Tool for Utilizing Sejong Corpus Based on Relational Database – 2010 - 1046–1050 с.
[13] Wumaier, A.; Yibulayin, T.; Kadeer, Z.; Tian, S. Conditional Random Fields Combined FSM Stemming Method for Uyghur - In Proceedings of the 2009 2nd IEEE International Conference on Computer Science and Information Technology, ICCSIT - 2009, Beijing, China, 8–11 August 2009 - 295–299 с.
[14] Хопкрофт Д., Мотвани Р., Ульман Д. Д. Введение в теорию автоматов, языков и вычислений. – М.: Издательский дом «Вильямс», 2008. — 528 с.
[15] Ахо А., Ульман Д. Теория синтаксического анализа, перевода и компиляции. – М.: изд-во «Мир», 1978. – 612 с.
[16] Younger D., Daniel H. Recognition and parsing of context-free languages in time n3 // Information and Control. – 1967. Т. 10, вып. 2. – С. 189–202.
[17] Теория и реализация языков программирования / Серебряков В.А., Галочкин М.П., Гочар Д.Р., Фуругян М.Г. – М.: МЗ Пресс, 2006. – 358 с.
[18] Earley J. An efficient context-free parsing algorithm // Communications of the ACM. – 1970. Т. 13, вып. 2. – С. 94–102.
[19] Parsing: a timeline // Ocean of Awareness blog [Электронный ресурс]. – 2014 – URL: http://jeffreykegler.github.io/Ocean-of-Awareness-blog/individual/2014/09/chron.html (дата обращения 05.12.2024)
[20] Л.Н. Кузина, И.Н.Полякова. Объектно-ориентированное программирование. Учебно-методический комплекс. 2013.
[21] Gamma E. Design Patterns: Elements of Reusable Object-Oriented Software. – Addison-Wesley Professional, 1994.– 395 с.
[22] Python documentation [Электронный ресурс] - Электрон.дан. – URL: https://docs.python.org/3/index.html (дата обращения 08.11.2024)
[23] Бакланова М.А., Хохлова Е.А., Ю Чжо Ен. Корейский язык. Базовый курс: Учебное пособие. – М.: изд-во ВШЭ, 2021. – 360 с.
[24] Ан Чинмён, Ли Кёна, Хан Хуён. Грамматика корейского языка для начинающих. – М.: АСТ, 2021. – 384 с.
[25] Ли Ёндо. Дракон Раджа. – изд-во «Золотая ветвь», 2012.
[26] Хён Чжингон. Удачный день. – М.: АСТ, 2019. – 224 с.
[27] Чон Ючжон. Семилетняя ночь. – М.: АСТ, 2021. – 528 с.
[28] Samsung. [Электронный ресурс]. – Электрон.дан. – URL: https://www.samsung.com/global/galaxy/galaxy-z-flip4. (дата обращения 21.09.2024)
[29] Daum. [Электронный ресурс]. – Электрон.дан. – URL: https://www.daum.net. (дата обращения 03.02.2025)
[30] Naver News [электронный ресурс] – https://news.naver.com. (дата обращения: 05.01.2025)

Статья получена 10 марта 2025.
Елфимова А.Е., магистрант, МГУ имени М.В. Ломоносова (email: elfimova509@yandex.ru)

Полякова И.Н., МГУ имени М.В. Ломоносова (email: polyakova@cs.msu.ru)

Automated grammar checking of text in Korean

Anna Elfimova, Irina Polyakova

***Abstract* - Recently, more and more attention has been attracted to Asian countries, in particular South Korea. As a result, learning the Korean language is becoming more popular in society. However, not everyone is ready to attend language learning courses or hire a tutor. In this regard, applications for independent study of foreign languages began to develop rapidly. There are very few such applications for the Korean language, and their tools are poor.**

In this paper, an attempt was made to develop and implement a program that automatically checks text in Korean for compliance with grammar rules. With the help of such program, anyone can independently check their text and correct possible errors in it.

To automate the verification of texts in Korean for compliance with grammar rules, it was decided to build a model of the grammar of the language, select the most suitable morphological analyzer, and also select the most suitable algorithm for syntactic parsing of texts. In addition, syntactic parser should be modified to work with Korean language. Thus, it was decided to adapt it for the correct syntactic analysis of the Korean language according to the constructed grammar and implement it. The resulting program allows you to analyze a text in Korean and automate its verification for compliance with grammar rules. With the help of such program, students studying Korean or anyone who wants to check themselves can independently study their text and correct possible errors. To improve the program in the future, the set of grammar rules can be expanded. The more rules, the better the program will cope with more complex sentences. By expanding the model, you can make the program work with a larger number of grammar rules and peculiars of Korean language. Also, you can try to improve the result of morphological analysis by combining the use of morphological analyzers. In addition, one of the possible ways to improve and develop the resulting analyzer is to detect typical errors in the text. The program will be more user-friendly if it also displays a message about where the error was made, what type of error, and how it can be corrected.

***Keywords* - automatic text analysis , Earley algorithm, grammar, Korean language, morphological analysis, syntactic analysis.**

REFERENCES

[1] RPORC [electronic resource] – Electronic data – URL: https://wciom.ru/analytical-reviews/analiticheskii-obzor/inostrannyj-yazyk-perspektivnaya-investicziya (Date of access: 05.02.2025)

[2] Duolingo [electronic resource] – Electronic data - URL: https://blog.duolingo.com/2023-duolingo-language-report/ (Date of access: 03.11.2024)

[3] Modern Language Association [electronic resource] – Electronic data – URL: https://www.mla.org/Resources/Guidelines-and-Data/Reports-and-Professional-Guidelines/Enrollments-in-Languages-Other-Than-English-in-United-States-Institutions-of-Higher-Education (Date of access: 05.10.2024)

[4] Foltz, P.W., Laham, D. and Landauer, T.K. (1999). The intelligent essay assessor: Applications to educational technology. Interactive Multimedia Electronic Journal of Computer-Enhanced Learning (IMEJ) 1(2), 939–944 p.

[5] Aho. A. V., Lam M. S., Sethi R., Ulman J. D. Compilers: Principles, Techniques, and Tools – Publisher «Pearson Education, Inc», 2006. – 1128 p.

[6] Chomsky N. Syntactic Structures – Publisher «Mouton & Co». – 1957 – 157 p.

[7] Volkova I.A., Vylitok A.A., Rudenko T.V. Formal Grammars and Languages. Elements of Translation Theory: A Textbook for 2nd-Year Students. – M.: Publisher «MAX Press», 2009 – 114 p.

[8] Ahn. J., Lee K., Han H. Korean Grammar in Use Beginning to Early Intermediate. – Darakwon, 2010. – 345 p.

[9] Cook D. J., Bays G. Computer mathematics. – M.: Publisher «Nauka», 1990. – 385 p.

[10] KoNLPy documentation [electronic resource] – Electronic data – URL: https://konlpy.org/en/latest/ (Date of access: 15.09.2024)

[11] Park, E.L.; Cho, S. KoNLPy: Korean Natural Language Processing in Python - In Proceedings of the Annual Conference on Human and Language Technology, Chuncheon, Republic of Korea, 10–11 October 2014; - 133–136 p.

[12] Lee, D.; Yeon, J.; Hwang, I.; Lee, S. KKMA : A Tool for Utilizing Sejong Corpus Based on Relational Database – 2010 - 1046–1050 p.

[13] Wumaier, A.; Yibulayin, T.; Kadeer, Z.; Tian, S. Conditional Random Fields Combined FSM Stemming Method for Uyghur - In Proceedings of the 2009 2nd IEEE International Conference on Computer Science and Information Technology, ICCSIT - 2009, Beijing, China, 8–11 August 2009 - 295–299 c.

[14] Hopcroft J., Motwani R., Ullman J.D. Introduction to Automata Theory, Languages, and Computation. – Publisher: «Addison-Wesley», 2001. — 521 p.

[15] Aho A., Ulman J.D. The Theory of Parsing, Translation, and Compiling. – Publisher « Prentice-Hall, Inc. », 1972. – 2051 p.

[16] Younger D., Daniel H. Recognition and parsing of context-free languages in time n3 // Information and Control. – 1967. T. 10,vol. 2. – p. 189–202.

[17] Serebryakov V., Galochkin V., Gochar D., Furugyan M. Theory and implementation of programming languages - M.: Publisher: «MZ Press», 2006. – 358 p.

[18] Earley J. An efficient context-free parsing algorithm // Communications of the ACM. – 1970. T. 13, vol. 2. – p. 94–102.

[19] Parsing: a timeline // Ocean of Awareness blog [electronic resource]. – 2014 – URL: http://jeffreykegler.github.io/Ocean-of-Awareness-blog/individual/2014/09/chron.html (Date of access: 05.12.2024)

[20] Kuzina L.N., Polyakova I.N. Object-oriented programming - 2013.

[21] Gamma E. Design Patterns: Elements of Reusable Object-Oriented Software. – Addison-Wesley Professional, 1994.– 395 c.

[22] Python documentation [electronic resource] - Electronic data – URL: https://docs.python.org/3/index.html (Date of access: 08.11.2024)

[23] Baklanova M.A., Khokhlova E.A., Yu Zho Yeon. Korean Language Basic Course: Study Guide. – M.: Publisher: HSE, 2021. – 360 p.

[24] Ahn Chinmyeon, Lee Kyunga, Han Huyeon. Korean Grammar for Beginners. – M.: Publisher AST, 2021. – 384 p.

[25] Lee Yeondo. Dragon Raja. – Publisher: «Golden Branch», 2012.

[26] Hyun Jing-gon. Lucky Day. – M.: Publisher: «AST», 2019. – 224 p.

[27] Jeong Yu-jeong. Seven Years' Night. – M.: Publisher: «AST», 2021. – 528 p.

[28] Samsung. [electronic resource]. – Electronic data – URL: https://www.samsung.com/global/galaxy/galaxy-z-flip4. (Date of access: 21.09.2024)

[29] Daum. [electronic resource]. – Electronic data – URL: https://www.daum.net. (Date of access: 03.02.2025)

[30] Naver News [electronic resource] – Electronic data –URL: https://news.naver.com. (Date of access: 05.01.2025)