МРНТИ 20.23.25, 16.31.21, 28.23.39. https://doi.org/10.26577/JMMCS-2019-3-m8

Графематический анализ текста на казахском языке

Шарипбай А.А., Евразийский национальный университет имени Л.Н. Гумилева,
г. Нур–Султан, Казахстан, E-mail: sharalt@mail.ru
Ниязова Р.С., Евразийский национальный университет имени Л.Н. Гумилева,
г. Нур–Султан, Казахстан, E-mail: rozamgul@list.ru
Туребаева Р.Д., Евразийский национальный университет имени Л.Н. Гумилева,
г. Нур–Султан, Казахстан, E–mail:58stud@mail.ru
Разахова Б.Ш., Евразийский национальный университет имени Л.Н. Гумилева,
г. Нур–Султан, Казахстан, E–mail:utalina@mail.ru
Зулхажав А., Евразийский национальный университет имени Л.Н. Гумилева,
г. Нур–Султан, Казахстан, E-mail: altinbekpin@gmail.com
Елибаева Г.К., Евразийский национальный университет имени Л.Н. Гумилева,
г. Нур–Султан, Казахстан, E-mail: gaziza_y@mail.ru

В данной работе рассматривается графематический анализ текста на казахском языке, являющаяся одним из основных этапов автоматической обработки текстов. Графематический анализ показывает местоположение автоматического анализа текста. Описаны различные классы графематических дескрипторов для описания графем, такие как главные и альтернативные графематические дескрипторы. Приведены какие задачи решаются при автоматическом анализе текста. В данной работе представлены графематические дескрипторы, задачи графематичекого анализа, приводятся алгоритм разделения текста на предложения и описывает графематический анализатор казахского языка. Также описан алгоритм деления текста на предложения, где ключевой задачей графематического анализа является правильный поиск границ слов и предложений. В данной статье приведены примеры вспомогательных примитив, также приведен некоторые замечаний относительно аббревиатур, сокращений, перечислений, определений и фрагментов. В статье также приведены какие задачи должны решать графематический анализ, рассматриваются дескрипторы, связанные к макросинтаксическому анализу. Приведены примеры основных графематических дескриптор. А также приведены примеры макросинтаксических дескрипторов. Все алгоритмы, описанные в данной работе были реализованы на Python.

Ключевые слова: графематический анализатор, графематические дескрипторы, автоматическая обработка текста, графема, графематический анализ.

Қазақ тіліндегі мәтінді графематикалық талдау

Шәріпбай А.Ә., Л.Н. Гумилев атындағы Еуразия ұлттық университеті,
Нұр–Сұлтан қ., Қазақстан, E-mail: sharalt@mail.ru
Ниязова Р.С., Л.Н. Гумилев атындағы Еуразия ұлттық университеті,
Нұр–Сұлтан қ., Қазақстан, E-mail: rozamgul@list.ru
Туребаева Р.Д., Л.Н. Гумилев атындағы Еуразия ұлттық университеті,
Нұр–Сұлтан қ., Қазақстан, E–mail: 58stud@mail.ru
Разахова Б.Ш., Л.Н. Гумилев атындағы Еуразия ұлттық университеті,
Нұр–Сұлтан қ., Қазақстан, E–mail: utalina@mail.ru
Зулхажав А., Л.Н. Гумилев атындағы Еуразия ұлттық университеті,
Нұр–Сұлтан қ., Қазақстан, E-mail: altinbekpin@gmail.com
Елибаева Г.К., Л.Н. Гумилев атындағы Еуразия ұлттық университеті,
Нұр–Сұлтан қ., Қазақстан, E-mail: gaziza_y@mail.ru

Бұл жұмыста мәтінді автоматты өңдеудің негізгі кезеңі болатын қазақ тіліндегі мәтінді графематикалық талдау қарастырылады. Графемді сипаттау үшін әр түрлі класстық дескрипторлар қолданылады. Мәтінді автоматты талдау кезінде қандай есептер шешілетіндігі

көрсетілген. Мәтінде сөйлемдерге бөлудің алгоритмі де сипатталған, онда графематикалық талдау кезеңіндегі негізгі мәселе - сөз бен сөйлем шекараларын дұрыс іздеу. Сонымен қатар бірқатар түсініктемелер берілген және де қосымша мысалдар келтірілген. Графематикалық талдаудың мәтінді автоматты талдаудағы орны көрсетіледі. Графематикалық талдаудың есептері көрсетіледі, графематикалық дескрипторлар анықталады, мәтінді сөйлемдерге бөлу алгоритмі келтіріледі және қазақ тілінің графематикалық талдауышы сипатталады. Сондай-ақ, графематикалық талдау қандай міндеттерді шешуі керек екендігі, макро синтаксистік талдауға байланысты дескрипторлар қарастырылған. Негізгі графиктік дескрипторлардың мысалдары келтірілген. Сонымен қатар макро синтаксистік дескрипторлардың мысалдары келтірілген.Жұмыс барысында сипатталған барлық алгоритмдер Python программасында орындалған.

Түйін сөздер: графематикалық талдауыш, графематикалық дескрипторлар, автоматты мәтінді өңдеу, графема, графематикалық талдау.

Graphematic analysis of Kazakh language text

Sharipbay A., L.N. Gumilyov Eurasian National University, Nur–Sultan, Kazakhstan,
E-mail: sharalt@mail.ru
Niyazova R., L.N. Gumilyov Eurasian National University, Nur–Sultan, Kazakhstan,
E-mail: rozamgul@list.ru
Turebayeva R., L.N.Gumilyov Eurasian National University, Nur–Sultan, Kazakhstan,
E–mail: 58stud@mail.ru
Razakhova B., L.N. Gumilyov Eurasian National University, Nur–Sultan, Kazakhstan,
E–mail: utalina@mail.ru
Zulkhazhav A., L.N. Gumilyov Eurasian National University, Nur–Sultan, Kazakhstan,
E-mail: altinbekpin@gmail.com
Yelibayeva G., L.N. Gumilyov Eurasian National University, Nur–Sultan, Kazakhstan,
E-mail: gaziza_y@mail.ru

In this paper, a grammatical analysis of the text in the Kazakh language is considered, which is one of the main stages of the automatic processing of texts. Grammatical analysis shows the location of the automatic analysis of the text. Various classes of grammar descriptors for describing grammar are described, such as main and alternative graphematic descriptors. What tasks are presented are solved by the automatic analysis of the text. This work presents the grammatical descriptors, tasks of the grammatical analysis, provides an algorithm for the separation of the text on the sentences and describes the grammatical analyzer of the Kazakh language. Also described is the algorithm for dividing text into sentences, where the key task of a grammatical analysis is the correct search for word and sentence borders. This article gives examples of auxiliary primitives, as well as some notes on abbreviations, abbreviations, enumerations, definitions, and fragments. The article also presents what tasks should be solved by grammatical analysis; descriptors related to macro syntactic analysis are considered. Examples of basic graphical descriptors are given. And also examples of macro syntactic descriptors are given. All algorithms described in this work were implemented in Python.
Key words: graphematic analyzer, graphematic descriptors, automatic text processing, grapheme, graphematic analysis.

1 Введение

В настоящее время происходит интеллектуализация цифровых информационных ресурсов на основе автоматической обработки (анализа и синтеза) текста на естественном языке. При этом анализом текста считается извлечением грамматической и семантической инфор-мации из текста, а синтезом текста – генерация текста с заданными характеристиками по за-данному алгоритму в соответствии с описанием естественного языка.

Нас интересует только автоматический анализ текста, который состоит из нескольких этапов, и они выполняются последовательно. Результаты предыдущих этапов используются на каждом этапе, при этом может происходить снятие неоднозначности, возникшая на предыдущем этапе. Ошибки в каждой фазе влияют на последующие этапы. Выход из преды-дущего этапа - это вход в следующий этап. Выделяются следующие этапы[1-4]:

- Графематический анализ;
- Морфологический анализ;
- Синтаксический анализ;
- Семантический анализ.

Автоматическая анализ текста позволит решать следующие задачи:

1. Анализ текста и информационный поиск:

- Поиск текстовой информации – анализ текста запроса;
- Поиск близких текстов (документов) – анализ двух текстов на близость;
- Автоматический (машинный) перевод с исходного одного языка на целевой язык – анализ текста исходного языка;
- Извлечение информации из текста – анализ исходного текста;
- Системы вопросов–ответов – анализ вопроса;
- Автоматическое резюмирование – анализ исходного текста;
- Автоматическое реферирование;
- Автоматическое аннотирование;
- Кластеризация и классификация документов – анализ текстов документов на схожесть;
- Контентный анализ авторизованного текста – определение характеристик текста и автора;
- Эмоциональный (Сентимент) анализ текста – определение эмоциональной окраски текста.

2. Синтез текстов:

- Поиск близких текстов (документов) – синтез текста о заимствований и плагиате;
- Автоматический (машинный) перевод с исходного языка на целевой язык – синтез текста целевого языка;
- Извлечение информации из текста – перевод текста к структурированной информации и её перенос в базу данных;
- Вопросно–ответные системы – синтез ответа;
- Автоматическое резюмирование – синтез текста резюме;
- Кластеризация и классификация документов – синтез текста об отнесении документов к предопределенному классу;
- Контентный анализ авторизованного текста – построение психолингвистического портрета автора;
- Эмоциональный (Сентимент) анализ текста – синтез текста с выводами на эмоциональной окраски текста.

3. Устное взаимодействие с компъютером:

- Распознавание речи – автоматическое преобразование устной речи в письменную речь.
- Синтез речи – автоматическое преобразование письменноой речи в устную речь.

2 Графематический анализ

На этапе графематического анализа выявляются все графемы текста и размечаются они с помощью графематических дескрипторов. При этом графемой считается минимальная смыслоразличительная единица графической системы (письменности) языка, имеющая в качестве своего значения фонему, слога, морфему, слова или другого неделимого лингвистического содержания.

Графематический анализ – начальный этап автоматической обработки текста на естественном языке, в котором выявляются и описываются его графемы с помощью графематических дескрипторов. Графематический анализ должен решать следующие задачи [2-5]:

- выделение в тексте заголовков, примечаний и абзацев;
- выделение в абзаце предложений (нахождение границ предложений);
- выделение устойчивых оборотов, не имеющих словоизменительных вариантов;
- выделение токенов (цепочек символов, образующих слова, чисел, формул, собственных имён, аббревиатур, единиц измерения и др.), ограниченных с двух сторон разделителями в предложении казахского языка;
- выделение устойчивых фраз (тұрақты сөз тіркестері) казахского языка;
- поиск географических наименований;
- поиск наименований страны, государства и населенных пунктов;
- поиск почтовых адресов;
- поиск адресов сайтов, ссылок на интернет ресурсы, имен файлов;
- поиск электронных адресов;
- поиск телефонных кодов (кодов зон, стран, регионов, населенных пунктов) и номеров;
- поиск ФИО (фамилия, имя, отчество), когда имя и отчество написаны инициалами;
- поиск сокращений и аббревиатур и определение их расшифровок;
- поиск литералов типа true, false, null и др.;
- поиск чисел, представленных в различной форме и системе счисления;
- поиск слов и устойчивых оборотов;
- сборка слов, написанных в разрядку;
- определениеь чередования верхнего и нижнего регистров.

3 Графематические дескрипторы

В языке графемы могут быть разных типов [2,16]:

Омофоничные. Это графемы, передающие один и тот же звук.

Аллографы - разные начертания одной и той же буквы. Вспомогательные - графемы, включающие знаки препинания и знаки диакритики.

Лигатуры - слитные написания графем.

Монографы - графемы, состоящие из одной буквы.

Графемные комплексы - единицы графики, состоящие из двух и более графем (например, английские ch, sh, augh и др.).

Для описания графем используются различные классы графематических дескрипторов: *главные дескрипторы, альтернативные дескрипторы, макросинтаксические де-*

скрипторы, контекстные графематические дескрипторы [17]. Каждый класс графематического дескриптора описывает текст с определенной точки зрения. На основе множества графематических дескрипторов можно построить аннотированный корпус текстов, удобный для дальнейшего разрешения задач ААТ [1-7]. Главные графематические дескрипторы определяют спецификации выделяемых структурных единиц текста [11-14]. Примеры таких дескрипторов приведены в таблице 1.

Альтернативные графематические дескрипторы служат для уточнения спецификаций выделяемых единиц. Примеры альтернативных графематических дескрипторов, которые записаны прописными и строчными буквами приведены в Таблице 2.

Следует различать дескрипторы макросинтаксического графематического анализа и де-скрипторы контекстного графематического анализа. Дескрипторы макросинтаксического графематического анализа выделяют условные предложения (УП), которыми могут быть структурные единицы такие, как заголовки, примечания, абзацы и т.д. Дескрипторы кон-текстного графематического анализа ставятся в зависимости от контекста строки, учитыва-ются номера не только из текущей строки, но номера строки, которые находятся выше и ни-же от неё [17].

В таблице 3 показаны примеры дескрипторов макросинтаксического графематического анализа.

В таблице 4 показаны примеры дескрипторов контекстного графематического анализа.

4 Алгоритм деления текста на предложения

Основная проблема на этапе графематического анализа состоит в корректном поиске границ слов и предложений. Например, в казахском языке разделителем предложений не всегда является ’.’ – точка и ’:’ – двоеточие, а разделителем слов не всегда является ’-’ – тире. Например, *25.04.2019, 19:45, Әл-Фараби, үп-үлкен (прибольшой)*. Рассмотрим алгоритм деления текста на предложения. На вход этого алгоритма подают-ся два кода StartPos и EndPos, обозначающие начальную и конечные строки текста, соответственно. Сначало ищется конец предложения, затем – начало предложения. Один из вариантов алгоритма деления текста на предложения приведен в [3]. Ниже предлагается некоторая модификация этого алгоритма, основанная на следующих утверждениях:

1. Начало первого предложения совпадает с началом текста, конец последнего предло-жения – с концом текста;
2. Начало предложения всегда является прописная буква;
3. Абзац не бывает меньше предложения;
4. Предложение не может состоять только из цифр и специальных знаков.
5. Конец предложения заканчивается знаком препинания ‘.’, ‘?’, ‘!’, ‘;’, ‘:’, ‘–’ или ‘...’.
6. Конец последнего предложения будет концом текста.

Замечание: Знаки ‘.’, ‘;’, ‘:’ и ‘–’ используются не только для определения конца предложения, но и для представления перечислений, определений, сокращений и фрагментов, что потребует использовать отдельное правило.

Определим вспомогательный примитив *IsSentEndMark*. На вход подаем номер строки, а на выходе получим истина, если эта строка содержит знак ‘.’, ‘?’, ‘!’, ‘;’, ‘:’ или ‘...’.

Таблица 1: Главные графематические дескрипторы

Название графемы	Описание графемы	Примеры графемы
	Начало предложения – Start of sentence	
ANC	Буквенно-цифровой комплекс – Alphanumeric chain. Присваивается цепочкам, состоящим из букв и цифр.	A125БС3, SIP970W
CBr	Закрывающая скобка – Closing bracket. Присваивается одиночному знаку закрывающей (правой) скобки.	’) ’, ’] ’, ’} ’, ’> ’, ”’, ’» ’, ””
Del	Разделитель – Separator. Присваивается цепочкам из одинаковых знаков разделителя.	’␣’, ’+’, ’–’, ’×’, ’*’, ’/’, ’=’, ’≠’, ’≈’, ’<’, ’≤’, ’>’, ’≥’, ’∨’, ’∧’, ’%’, ’→’
DPUN	Последовательность одинаковых символов, длина которой больше 20.	
ELI	Признак конца строки – End of line indication.	\0
ES	Конец предложения – End of sentence	
FLE	Иностранная лексема – Foreign lexeme. Присваивается цепочкам из латиницы.	England, London, Richard, hi, thanks
INu	Целое число – Integer number. Присваивается цепочкам из цифр 0,1,2,3,4,5,6,7,8,9	2019
KLE	Казахская лексема – Kazakh lexeme. Присваивается цепочкам из кириллицы.	Казахстан, Алматы, Абай, сәлем, рахмет
KSPH	Казахская устойчивая фраза – Kazakh sustainable phrase. Присваивается цепочкам из кириллицы.	Бас имеу, Ат ізін салмау, Екі езуі екі құлағына жету
LST	Строка пробелов или табуляций – Line of spaces or tabs. Присваивается к строке, которая не содержит ни одного символа.	’␣ ␣ ␣ ␣ ␣ ␣ ␣ ␣’
OBr	Открывающая скобка – Opening bracket. Присваивается одиночному знаку открывающей (левой) скобки.	’(’, ’[’, ’{ ’, ’< ’, ’> ’, ’‘ ’, ’« ’, “‘
Par	Скобки – Parentheses. Присваивается цепочкам внутри спаренных скобок.	’(’, ’) ’, ’[’, ’] ’, ’{ ’, ’} ’, ’< ’, ’> ’, ’‘ ’, ’ ’ ’, ’« ’, ’» ’, ’“ ’, ””
Pg	Абзац – Paragraph	
Pg Sym	Символ параграфа – Paragraph symbol.	§
PLP	Последовательность одинаковых символов, длина которой больше 1.	

Определим вспомогательный примитив *IsSentenceEndSeq*. На вход подаем номер строки. Примитив возвращает истину в двух случаях: - Если функция *IsSentEndMark*

Таблица 2: Альтернативные графематические дескрипторы

Название графемы	Описание графемы	Примеры графемы
EMSYM	Нулевой символ, но не пробел.	
FCLE	Все символы английской лексемы большие буквы – – All symbols of the English lexeme capital letters. Альтернатива FLE.	MOTHER
FCSLE	В английской лексеме первый символ большая буква, а остальные малые буквы – The first symbol in the English lexeme is capital letter, and the rest are small letters. Альтернатива FLE.	Mother
FSLE	Все символы английской лексемы малые буквы – All symbols of the English lexeme small letters. Альтернатива FLE.	mother
HYP	Дефис – Hyphen. Разновидность дескриптора Del.	
KCLE	Все символы казахской лексемы большие буквы – All symbols of the Kazakh lexeme capital letters. Альтернатива KLE.	АНА
KCSLE	В казахской лексеме первый символ большая буква, а остальные малые буквы – The first symbol in the Kazakh lexeme is capital letter, and the rest are small letters. Альтернатива KLE.	Ана
KSLE	Все символы казахской лексемы малые буквы – All symbols of the Kazakh lexeme small letters. Альтернатива KLE.	ана
RCLE	Все символы русской лексемы большие буквы – All of the Russian lexeme capital letters. Альтернатива RLE.	МАТЬ
RCSLE	В русской лексеме первый символ большая буква, а остальные малые буквы – The first symbol in the Russian lexeme is capital letter, and the rest are small letters. Альтернатива RLE.	Мать
RSLE	Все символы русской лексемы малые буквы – All symbols of the Russian lexeme small letters. Альтернатива RLE.	мать

верна для этой строки и непосредственно справа нет закрывающей кавычки (если предложение заключено в кавычки, то закрывающая кавычка входит в его состав);

- Если функция *IsSentEndMark* верна для строки, стоящей непосредственно слева от строки, заканчивающейся закрывающей кавычкой.

Программа поиска предложений на основании этих случаев будут описываться так:

1. Пусть i – номер строки между *StartPos* и *EndPos*.

2. Если на строке i стоит разметка начала абзаца, то нужно дойти до конца преды-

Таблица 3: Дескрипторы макросинтаксического графематического анализа.

Название графемы	Описание графемы
CS	Ставится на конце простого УП
CS?	Ставится на конце УП, тип которого не определен
CS_AUX	Ставится на конце УП, заключенного в скобки
CS_PRNT	Указывает конец УП, заканчивающегося на двоеточие
DOC	Указывает нулевую строку
HDNG	Указывает конец заголовка

дущего абзаца, пройдяназад все пробелы и длинные разделители (PLP).

3. Если в конце абзаца стоит строка, удовлетворяющая *IsSentEndSeq*, то нужно поставить SENT_END этой строке, иначе нужно поставить SENT_END на конец предыдущего абзаца.

4. Если на строке i стоит макросинтаксический дескриптор УП, то нужно сделать как в пункте 2 и учесть, что дескриптор УП ставится на конце абзаца.

5. Если до начала текущего предложения стояла открывающая скобка или кавычка, и те-кущая строка указывает на слово до соответствующей закрывающей скобки или кавычки, тогда нужно поставить SENT_END на закрывающую скобку (кавычку), а i сместить на ближайшее после закрывающей кавычки (скобки) слово.

6. Если текущая строка не стоит внутри графематических групп (FIO1–FIO2 и т.д.) и для неё верна функция *IsSentEndSeq*, то нужно пройти все знаки препинания от текущей строки. Знак препинания не должен стоять в самом начале строки, если он заканчивает предложения. Началом нового предложения считается найденное первое слово от текущей строки.

Надо знать, что иногда фрагменты выделяются с одной стороны или с обеих сторон.

Некоторые организации кавычки включаются в состав предложения и разрабатывают свои алгоритмы выделения предложений, которые похожи на такой:

1. Передвинуть признак окончания предложения после закрывающей кавычки, если та-ковая существует;

2. Убрать признак окончания предложения в следующих случаях:

- Если предыдущее слово является известным сокращением, использование которого не предполагается в конце предложения, например ‘проф.’, ‘г.’, ‘ул.’, ‘д.’

- Если предыдущее слово является известным сокращением, но за которым не следует слово с заглавной буквы, например: ‘т.д.’, ‘т.е.’.

3. Убрать признак окончания предложения после ‘!’ и ‘?’ в случая, если за ними следуют слова без заглавной буквы.

Таблица 4: Дескрипторов контекстного графематического анализа

Название графемы	Назначение графемы	Примеры графемы
ABB1	Указывает начало сокращения	и т.п.
ABB2	Указывает конец сокращения	
BEG	Указывает начало текста и ставится на нулевой строке в таблице, которая используется как служебная, при этом содержимое её первого столбца не входит в рассматриваемый текст	
BUL	Указывает начало пункта перечисления	
EA	Указывает адрес сайта или электронной почты	www.zerdet.kz qazaq@mail.ru
EndPos	Указывает последнюю строку входного текста	
EOP	Указывает конец фразы, которым считается только ‘;’.	
EXPR1	Указывает начало оборота	‘келе жатыр’
EXPR2	Указывает конец оборота	
FAM1	Ставится на начале фамилия с инициалами	‘Уркер М.Н.’
FAM2	Ставится на конце фамилия с инциалами	
FILE1	Указывает начало имени файла	C:\test.txt
FILE2	Указывает конец имени файла	
INDENT	Указывает начало абзаца	
KEY1	Указывает начало цепочки обозначений клавиш	ctrl–f
KEY2	Указывает конец цепочки обозначений клавиш	
NAM?	Указывает часть собственного имени и присваивается лексеме, начинающейся с большой буквы и не имеющей перед собой символа конца предложения.	
SENT_END	Конец предложения	
SENT_START	Начало предложения	
StartPos	Первая строка входного текста	

5 Графематический анализатор

Программа, которая осуществляет графематический анализ называется графематическим анализатором. Его выход будет использоватьс Морфологическим и Синтаксиче-

ским анализаторами.

Графематический анализатор должен:

- иметь возможность поиска текстовых конструкций с использованием шаблонов;

- иметь возможность добавления новых шаблонов для поиска;

- иметь возможность представления результатов в формате JSON (текстовый формат обмена данными, основанный на JavaScript) для обмена данными с другими модулями анализа текста;

- иметь возможность интегрироваться с морфологическим анализатором;

- работать без предварительного обучения на правильно обработанной экспертами набора текстов;

- затрачивать малое количество времени, памяти и других ресурсов в своей работе.

Программа должна работать в разных операционных системах и представляться в виде библиотеки исходного кода. Разработка должна осуществляться на основе свободных и бес-платных программного обеспечения и системы управления базами данных.

Входным данным графематического анализатора является текст на естественном языке, состоящий из цепочки единиц (графем) в системе кодировки Unicode, а выходным данным – графематическая таблица (5 таблица), которая состоит из 2 столбцов: в первом столбце будут выделенные единицы входного текста, а во втором столбце – графематические дескрипторы (теги), характеризующие эту единицу. Например, из текста ‘Асан пришел’ будет построена следующая таблица.

Таблица 5: Графематическая таблица

Выделенные единицы	Графематические дескрипторы
Асан	RCSLE NAM?
Пришел	RSLE SENT_END

В первый столбец всегда помещается часть входного текста. Если входные символы яв-ляются последовательностью из мягких разделителей (пробел, табуляция, возврат каретки), тогда используются другие символы, номера которых включены в описание на языке Uni-Turk [5-10].

Все выделенные единицы текста проходят проверку в соответствующих словарях (сло-варе словоформ, словаре терминов, словаре аббревиатур, словаре сокращений и т.д.). Одна-ко, случаи, когда все 100% слов текста представлены в используемых системой словарях, крайне редки. В тексте могут встречаться неологизмы, устаревшие слова, искаженные слова.

В случаях, когда встреченное системой слово отсутствует в используемых ей словарях, система строит гипотезы о том, является ли данное слово искаженным. Для этого учитывается, что графематические слова имеют неслучайную структуру – они построены из определенных полиграмм. Состав и количество полиграмм определяются рядом факторов, важнейшими из которых являются: ономастическая система данного языка, правила орфографии, принятая система обозначения звуков на письме и др. Диагностика искажений в словах основывается на предположении, что среди полиграмм,

входящих в искаженное слово, есть такая, которая не вложится в стохастическое дерево, и её частота будет достаточно низка[14].

Стохастическое дерево (дерево вероятностей) – это граф, в узлах которого записаны элементы моделируемого объекта, а в дугах – вероятности перехода от предыдущей цепочки узлов к следующему узлу[19]. В нашем случае, моделируемый объект – это текст на казах-ском языке. Соответственно, в узлах графа будут находиться графемы, а в дугах – вероят-ность следования данной графемы за предыдущей частью полиграммы.

6 Заключение

В этой работе особое внимание уделяется на выделение устойчивых фраз казахского языка, которые сильно отличаются от фраз других естественных языков, а выделение остальных единиц текста на казахском языке в основном совпадает с выделением аналогич-ных единиц текстов на других алфавитных естественных языках. Следует отметить, для кор-ректного выделения таких фраз необходимо использовать словарь устойчивых фраз казах-ского языка. В противном случае придется применить технологию глубоко обучения, основанную на нейронных сетях.

7 Благодарность

Работа выполнена при поддержке грантового финансирования научно-технических программ и проектов Министерством науки и образования Республики Казахстан (грант № AP05132249, 2018-2020 годы)

Список литературы

[1] *Jackson, P., Mouliner, I.* Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization: John Benjamins Publishing Co.– 2002. – 237 p.

[2] *Автоматическая обработка текста.* [Электр.ресурс]. – 2003. – URL: http://aot.ru/docs/graphan.html (дата обращения: 25.07.2019)

[3] *Первушин А.* Модуль графематического анализа в системе обработки русскоязычных текстов [Электр.ресурс]. – 2003. – URL: https://cyberleninka.ru/article/n/modul-grafematicheskogo-analiza-v-sisteme-obrabotki-russkoyazychnyh-tekstov (дата обращения: 02.08.2019)

[4] *Графема - это ...* Виды и особенности графем [Электр.ресурс]. – 2018. – URL: https://fb.ru/article/432209/grafema—eto-vidyi-i-osobennosti-grafem (дата обращения: 25.07.2019)

[5] *Шәріпбай А.Ә., Гатиатуллин А.Р., Ергеш Б.Ж., Қажымұхан Д.А.* Разработка единого метаязыка морфологии тюркских языков // Вестник КазНУ. Серия математика, механика, информатика. – Алматы. – 2018. – N. 4(100). – С.78–87.

[6] *Yelibayeva G., Mukanova A., Sharipbay A., Zulkhazhav A., Yergesh B., Bekmanova G.* Metalanguage and Knowledgebase for Kazakh Morphology // Lecture Notes in Computer Science. – 2019. No. 11619. – P. 717–730.

[7] *Sharipbay A., Mukanova A., Yergesh B., Zhetkenbay L., Zulkhazhav A., Yelibayeva G.* Ontology modeling of morphological rules of the Kazakh and Turkish languages // Abstract of the VI international conference «Modern problems of applied mathematics and information technology - al-Khorezmiy 2018». – Tashkent, Uzbekistan. – 2018. – P. 51-52.

[8] *Zhetkenbay L., Sharipbay A., Bekmanova G., Kamanur U.* Ontological modeling of morphological rules for the adjectives in Kazakh and Turkish languages // Journal of Theoretical and Applied Information Technology. – 2016. – Vol. 91. No.2. – P. 257- 263.

[9] *Bekmanova G., Sharipbay A., Altnbek G., Adalı E., Zhetkenbay L., Kamanur U., Zulkhazhav A.* The uniform morphological analyzer for the Kazakh and Turkish languages. // Proceedings of the Sixth International Conference on Analysis of Images, Social Networks and Texts (AIST 2017), Moscow, Russia, July 2017. –P. 20-30.

[10] *Жеткенбай Л., Шарипбай А., Бекманова Г., Қажымұқан Д., Каманур У.* Сравнение морфологических правил глагола казахского и турецкого языков. // Вестник. Алматы: Казахский национальный университет им. аль-Фараби. – 2018.4(100).–С. 42-51.

[11] *Garside, R., Leech G. and Sampson G. (eds).* The CLAWS Word-tagging System // The Computational Analysis of English: A Corpus-based Approach. – London: Longman. – 1987.

[12] *Jurafsky D., James H.* Speech and Language Processing. // An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. – 2nd Edition. – Prentice-Hall. –2009. – 988 p.

[13] *Nitin I., Fred J. D.* Handbook of Natural Language Processing. – 2nd ed. – Chapman & Hall/CRC. – 2010.– 704 p.

[14] *Dunaev A.A.* Research system for analyzing natural language texts https://www.iis.nsk.su/files/articles/sbor_kas_13_dunaev.pdf

[15] *Berg K.* Identifying graphematic units: vowel and consonant letters. // Writ. Lang. Lit. 15. – 2012. P.26–45. 10.1075/wll.15.1.02ber;

[16] *Eisenberg P.* Uber die Autonomie der graphematischen analyse. // in Probleme der Geschriebenen Sprache, eds Nerius D., Augst G., editors. Berlin: Akademie Verlag . – 1988. P. 25–35.

[17] *Aronoff M.* Morphological stems. what William of Ockham really said. Word Struct. 5. – 2012. P. 28–51. 10.3366/word.2012.0018

[18] *Frost R., Katz L.* The reading process is different for different orthographies. The orthographic depth hypothesis, in Orthography, Phonology, Morphology and Meaning, eds Frost R., Katz L., editors. Amsterdam/London: North Holland. – 1992, P.67–84.

[19] *Saenger P.* Space Between Words. The Origins of Silent Reading. Stanford, CA: Stanford University Press. – 1997.

References

[1] Jackson P., Mouliner I., Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization: John Benjamins Publishing Co., 2002. – 237 p.

[2] Avtomaticheskaya obrabotka teksta [Automatic text processing], accessed July 25, 2019, http://aot.ru/docs/graphan.html

[3] Pervushin A., Modul grafematicheskogo analiza v sisteme obrabotki russkoyazyichnyih tekstov [Graphematic analysis module in the system for processing Russian-language texts], accessed August 2, 2019, https://cyberleninka.ru/article/n/modul-grafematicheskogo-analiza-v-sisteme-obrabotki-russkoyazychnyh-tekstov

[4] Grafema - eto... Vidyi i osobennosti grafem [A grapheme is ... Types and features of graphemes], accessed July 25, 2019, https://fb.ru/article/432209/grafema—eto-vidyi-i-osobennosti-grafem

[5] Sharipbay A., Gatiatullin A., Yergesh B., Kazhymukhan D., "Development of an unified meta language of the turkic languages morphology ", *Journal of Mathematics, Mechanics and Computer Science.* –Almaty, 2018. – N. 4(100). – C.78–87.

[6] Yelibayeva G., Mukanova A., Sharipbay A., Zulkhazhav A., Yergesh B., Bekmanova G., "Metalanguage and Knowledgebase for Kazakh Morphology", *Lecture Notes in Computer Science.* – 2019. No. 11619. – P. 717–730.

[7] Sharipbay A., Mukanova A., Yergesh B., Zhetkenbay L., Zulkhazhav A., Yelibayeva G., "Ontology modeling of morphological rules of the Kazakh and Turkish languages", *Abstract of the VI international conference «Modern problems of applied mathematics and information technology - al-Khorezmiy 2018».* – Tashkent, Uzbekistan. – 2018. – P. 51-52.

[8] Zhetkenbay L., Sharipbay A., Bekmanova G., Kamanur U., "Ontological modeling of morphological rules for the adjectives in Kazakh and Turkish languages", *Journal of Theoretical and Applied Information Technology.* – 2016. – Vol. 91. No.2. – P. 257- 263. [

[9] Bekmanova G., Sharipbay A., Altnbek G., Adalı E., Zhetkenbay L., Kamanur U., Zulkhazhav A., "The uniform morphological analyzer for the Kazakh and Turkish languages", *Proceedings of the Sixth International Conference on Analysis of Images, Social Networks and Texts* (AIST 2017), Moscow, Russia, July 2017. –P. 20-30.

[10] Zhetkenbay L., Sharipbay A.A., Bekmanova G.T., Kazhymukhan D., Kamanur U., "Comparison of the morphological rules of the Kazakh and Turkish languages", *Journal of Mathematics, Mechanics and Computer Science.* –Almaty, 2018. – N. 4(100). – P. 42-51.

[11] Garside, R., Leech G. and Sampson G. (eds)., "The CLAWS Word-tagging System", The Computational Analysis of English: A Corpus-based Approach. – London: Longman. – 1987.

[12] Jurafsky D, James H. M., "Speech and Language Processing", An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. – 2nd Edition. – Prentice-Hall. –2009. – 988 p.

[13] Nitin I., Fred J. D., Handbook of Natural Language Processing. – 2nd ed. – Chapman and Hall/CRC. – 2010.– 704 p.

[14] Dunaev A.A., Research system for analyzing natural language texts https://www.iis.nsk.su/files/articles/sbor_ka_13_dunaev.pdf

[15] Berg K., Identifying graphematic units: vowel and consonant letters. Writ. Lang. Lit. 15. – 2012. P.26–45. 10.1075/wll.15.1.02ber;

[16] Eisenberg P., "Uber die Autonomie der graphematischen analyse", *in Probleme der Geschriebenen Sprache,* eds Nerius D., Augst G., editors. Berlin: Akademie Verlag . – 1988. P. 25–35.

[17] Aronoff M. "Morphological stems: what William of Ockham really said. Word Struct. 5. – 2012. P. 28–51. 10.3366/word.2012.0018

[18] Frost R., Katz L., "The reading process is different for different orthographies: the orthographic depth hypothesis", in Orthography, Phonology, Morphology and Meaning, eds Frost R., Katz L., (Amsterdam/London: North Holland). – 1992, P.67–84.

[19] Saenger P., Space Between Words. The Origins of Silent Reading. Stanford, CA: Stanford University Press. – 1997.