Министерство образования и науки Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего образования
«Южно-Уральский государственный университет
(национальный исследовательский университет)»
Институт лингвистики и международных коммуникаций
Кафедра лингвистики и перевода

ДОПУСТИТЬ К ЗАЩИТЕ
Заведующий кафедрой,
д.филол.н., доцент
_____/Т.Н. Хомутова/

АВТОМАТИЗАЦИЯ ИЗВЛЕЧЕНИЯ ЛЕКСИКОГРАФИЧЕСКОЙ ИНФОРМАЦИИ ИЗ ТЕКСТОВ (НА МАТЕРИАЛЕ КИТАЙСКОГО ЯЗЫКА)

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА
ЮУрГУ – 450303.2018.286.ВКР

Руководитель, д.филол.н., проф.
/С.О. Шереметьева/
«___»_____2018г.

Автор
студент группы ЛМ-437
_____/И.А. Рожин/
«____»_____2018г.

Нормоконтролер,
к. филол. н., доцент
_____/О.И. Бабина/
«____»_____2018г.

Работа защищена с оценкой
_____
«____»_____2018г.

Челябинск

2018

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Автоматическое извлечение информации из текстов на естественном языке является одной из важных проблем в области автоматической обработки естественного языка, решение которой позволит повысить эффективность использования информационных ресурсов, хранящихся в виде электронных текстовых документов [21]. Востребованность в эффективных методах для решения данной проблемы возрастает, если речь идёт о документах на иностранных языках, отличных от русского, английского, немецкого и других алфавитных языков. Если для европейских языков существует достаточно большое количество методов автоматического извлечения ценной информации из текстов, списков текстов, корпусов, то для изолирующих языков, например, китайского, корейского, японского, набор методов решения данной проблемы обработки естественного языка, на сегодняшний день, не является удовлетворительным. Особенно если речь идёт о китайском языке.

В двадцать первом веке китайский язык получает большое распространение по всему миру, благодаря политике, которую проводит Китай. Из этого следует вывод, что и количество информации, хранимой и передаваемой на китайском языке, с каждым днём увеличивается и необходимость в её обработке возрастает.

Группа Стэнфордского университета по обработке естественного языка активно занимается проблемой анализа китайского языка средствами ЭВМ. Такие прикладные лингвисты как, Стивен Берд, Эдвард Лопер и Эван Клейн, создатели пакетов библиотек и программ для обработки естественного языка. Нельзя забывать и об энтузиастах, которые выкладывают свои программы для свободного использования на сайте GitHub, крупнейшем веб-сервисе для хостинга IT-проектов и их совместной разработки. Стоит отметить, что данный список людей, занимающихся анализом китайского языка, не является исчерпывающим.

Актуальность обработки китайского языка растёт, о чём свидетельствует привлечение первых двух вышеупомянутых больших групп учёных-лингвистов.

Актуальность дипломной работы заключается в необходимости нахождения метода, способного решать проблемы в области обработки китайского языка, а именно в извлечении именных групп.

Объект исследования дипломной работы – список текстов на китайском языке по лингвострановедческой тематике.

Предмет исследования дипломной работы – именные группы китайского языка.

Одной из задач автоматического извлечения информации является автоматическое формирование словарей предметной области [21].

Лексикография (наука о создании, изучении и использовании словарей) включает в себя как теоретические знания, так и практические исследования: теоретическая часть лексикографии включает в себя теорию и историю создания словарей, в свою очередь, практическая занимается непосредственным созданием словарей и сбором первичного словарного материала [8].

Компьютерная лексикография представлена совокупностью методов и программных средств обработки текстовой информации для создания словарей [17]. В рамках компьютерной лексикографии разрабатываются компьютерные технологии составления и эксплуатации словарей. Специальные программы, базы данных, компьютерные картотеки, программы обработки текста позволяют в автоматическом режиме формировать словарные статьи, хранить словарную информацию и обрабатывать её [35].

Из выше изложенного – актуальности проблемы обработки китайского языка, тезисов из теории по лексикографии, объекту и предмету исследования – формируется цель дипломной работы.

Цель исследования – написание программы для автоматического создания словаря именных групп китайского языка.

Для достижения цели дипломной работы нами были поставлены следующие ***задачи***:

1. Сбор и изучение теоретических знаний по лексикографии, компьютерной лексикографии, именным группам и автоматическому извлечению информации из текстов на естественном языке.

2. Создание списка текстов из учебного пособия по страноведенью Китая.

3. Поиск метода для извлечения именных групп из текста на китайском языке.

4. Создание программы на языке программирования python по извлечению именных групп и автоматическому формированию словаря предметной области.

В нашем исследовании мы не ставим помимо основной цели, второстепенную – разрешить все проблемы, с которыми столкнёмся во время работы. Одной из таких проблем стала проблема определения частей речи в тексте на китайском языке. Данная проблема является решаемой за короткий промежуток времени, но после определения частей речи, встаёт вопрос, о неоднозначности лексики, что является также проблемой. Так как иероглиф может являться сразу несколькими частями речи, например, иероглиф 家 [jiā] может быть, как существительным семья, так и служебным словом: счётным словом для зданий. Для решения этих проблем мы использовали стороннюю программу SegmentAnt для тегирования и сегментации текста [28].

Методы, которые были использованы в дипломной работе:

1. Метод непосредственно составляющих;
2. Корпусный анализ;
3. Метод лингвистического моделирования.

Научная новизна дипломной работы заключается, во-первых, в том, что метод для извлечения именных групп разрабатываемый Шереметьевой С.О. ранее использовался только на европейских языках. Во-вторых,

предварительная обработка текста в методе, предлагаемом Шереметьевой С.О, заключается в построение списка n-грамм, что в данной дипломной работе не будет использовано. Мы предлагаем вместо построения n-грамм поочерёдный отбор кандидатов в именную группу. Отбор начинается с первого иероглифа и заканчивается на иероглифе, тег которого не входит в состав частей речи, которые могут использоваться в именной группе.

Результаты исследования данной дипломной работы носят, как теоретическую, так и практическую значимость.

Теоретическая значимость дипломной работы заключается в том, что процесс построения программы для автоматического создания словаря предметной области полностью описан во второй главе дипломной работы, что в свою очередь может послужить в обучающих и исследовательских целях.

Практическая значимость непосредственно заключается в созданной программе, которая находится в свободном доступе на сайте GitHub. Данная программа может быть использована в виде функции по извлечению именных групп для программ более широкой направленности, нацеленные на автоматическую обработку китайского языка.

Структура дипломной работы обусловлена объектом, предметом, целью и задачами исследования. Работа состоит из следующих разделов:

1. Введение раскрывает актуальность темы, определяет объект, предмет, цель, задачи и методы исследования, раскрывает теоретическую и практическую значимость работы.

2. В первой главе рассматривается теория по лексикографии, составлению словарей, словарная статья и её состав, компьютерная лексикография, составление автоматических словарей, различие между автоматическим словарём и лексиконом, методы извлечения вокабуляра.

3. Во второй главе описывается создание программы: алгоритм, метод извлечения именных групп, использованные функции языка

программирования python, представлены примеры работы программы как в письменном виде, так и в виде рисунков.

4. В заключении подводятся итоги исследования: была ли достигнута поставленная цель, выполнены ли все задачи, подтверждается или опровергается гипотеза, формируется окончательный вывод по изучаемой теме.

5. В приложении находится вспомогательный материал: скриншот кода программы, список тегов частей речи, которые могут входить в именную группу в китайском языке, список именных групп, составленных вручную и список именных групп, извлечённых автоматически.

ГЛАВА 1 ЛЕКСИКОГРАФИЯ – НАУКА О СОЗДАНИИ И ИЗУЧЕНИИ СЛОВАРЕЙ

1.1Понятие лексикография

Как написано в трудах Щербы Л.В., лексикография с древнегреческого обозначает lexikon – словарь и grapho – пишу [23]. В настоящее время существует множество взглядов на науку лексикографию, в работах отечественных лексикографов можно встретить следующие определения.

Так, Пустошило Е.П. в книге «Лексикология. Фразеолгоия. Лексикография» определяет лексикографию как раздела языкознания, занимающийся изучением теории и практики составления словарей [14]. В это же время, Морковкин В.В. определяет лексикографию как область филологической и инженерно-филологической деятельности, состоящей в создании словарей и других произведений словарного типа, а также в осмыслении всей суммы, относящейся к этой проблеме [12].

По мнению Дубчинского В.В., которое он высказал в своей книге «Лексикография русского языка» теоретическая часть лексикографии занимается непосредственно теорией и историей создания словарей, а практическая лексикография занимается созданием словарей и сбором первичного словарного материала [8]. В свою очередь, Пустошило Е.П. даёт более развёрнутое определение для каждого направления лексикографии. По её мнению, практическая лексикография возникла намного раньше теоретической, т.к. составление словарей относится к древнейшему виду лингвистической деятельности. Теоретическая лексикография, по мнению Пустошило Е.П., охватывает комплекс проблем, связанных с разработкой макроструктуры и микроструктуры словаря.

Под макроструктурой словаря, со слов Елены Петровны, следует понимать: отбор лексики, объём и характер словника и принцип расположения материала. Также, по её мнению, под микроструктурой мы понимаем следующие: структура словарной статьи, типы словарных

определений, соотношения разных видов информации о слове, описание общей типологии словарей и создание словарей новых типов [14].

Также Пустошило Е.П. выделяет следующие проблемы, решаемые практической лексикографией:

1. Описание и нормализация языка.
2. Обеспечение межязыкового общения.
3. Обучение языку (родному и иностранному).
4. Научное изучение лексики языка [14].

Мнения многих лексикографов о науке лексикографии схожи, так Щерба Л.В. считал: «Лексикография стремится найти наиболее оптимальные и допустимые для восприятия способы словарного представления всей совокупности знаний о языке» [23, 24]. Советский и российский лингвист, лексикограф Сороколетов Ф.П. видел науку лексикографию так: «Лексикография представляет слово в совокупности всех его свойств, поэтому словарь оказывается не только уникальным и незаменимым пособием по языку, но и важнейшим инструментом научных исследований» [16].

В 1985 году Ступин Л.П в своём учебном пособии «Лексикография английского языка» отмечал, что статьи на слово лексикография нет в таких современных энциклопедиях, как энциклопедии «Британика» и «Американа». Ученый объясняет данный факт отсутствием единогласного мнения по поводу определения лексикографии как науки. Часто приводятся высказывания Виганда Г.Э. о лексикографии, как о не науке: «Лексикография – это не наука, не искусство, не отрасль лингвистики, не прикладная лингвистика», «Лексикография никогда не была, не является и, скорее всего, не станет наукой» [17].

Из истории лексикографии нам известно, что в XIX веке в России лексикография получила большое развитие, появились словари разных типов:

1. Исторические словари.

2. Словари иностранных слов.
3. Двуязычные словари.
4. Толковые словари.

Большинство мнений лексикографов сходятся на том, что наибольшее значение для развития русской лексикографии имели следующие словари:

1. «Словарь церковнославянского и русского языка» изданный в 1847 году.

2. «Толковый словарь живого великорусского языка». Этот словарь был написан Далем В.И. и издан в 1847 году. В последующем, он был дополнен и исправлен Бодуэном де Куртенэ И.А. и перевыпущен в 1903 году.

3. «Словарь русского языка» под редакцией Грота Я.И. изданный в 1895. Последующие издание словаря продолжал Шахматов А.А. уже по принципам ненормативного словаря – тезауруса в 1907 году.

По утверждениям, найденным в исторических справочниках, в СССР лексикография превратилась в ведущую отрасль прикладной лингвистики. Такое становление было обусловлено необходимостью фиксировать русский и другие языки страны на современном этапе. Необходимо было закрепить языковые нормы для многих языков и создать двуязычные словари для народов СССР. Стоит отметить, что в советской лексикографии были применены многие решения, к которым зарубежная лексикография пришла позднее, например, указание на зависимость значения глагола от семантики его актантов в толковых словарях русского языка.

В настоящее время, как считает Морковкин В.В., новым стимулом для развития теоретической лексикографии является разработка учебных словарей и использование компьютерной техники в лексикографической практике [12]. Здесь стоит отметить, что с его мнение схожи мнения и других лексикографов: Герда А.С., Зализняк А.А. и другие.

1.2 Словарь как основной лексикографический ресурс

Сороколетов Ф.П. в книге «История русской лексикографии» писал: «Лингвистическая наука двадцать первого века стремится воплотить в

словарной форме все аспекты полученных знаний, поэтому в новейших словарях объектом описания становятся не только слова, но и иные языковые единицы» [16].

Объектом описания словаря – заголовочной единицей – является лексическая единица языка. В книге «Лексикография русского языка» автор пишет, что в качестве заголовочной единицы могут выступать следующие единицы языка: морфемы, слова и словосочетания [8]. Для описание данной заголовочной единицы учёными-лингвистами, лексикографами выделяются следующие критерии или, если говорить более точно, то характеристики языковой единицы: фонетическая; грамматическая; сочетаемость описываемой лексической единицы с другими единицами языка; словообразование; этимологическая справка; иллюстрации; пометы лексикографического характера; энциклопедическая информация; отсылки на другие статьи в самом словаре, так и на другие источники информации; примечания [12-14].

Этот список можно дополнить исследованием Дубчинского В.В. По его мнению, для описания заголовочной единицы используется ещё и семантизация лексической единицы. Это словосочетание автор понимает, как термин, который вбирает в себя знания о толкование, дефиниции и переводном эквиваленте слов [8].

В учебниках по введению в языкознание и лексикографии можно найти дополнительную информацию о языковых единицах, которые могут выступить в роли заголовочной единицы словаря. Это такие единицы как: фонема и словоформа.

В книге Мельчук И.А. «Опыт теории лингвистических моделей «Смысл ⇔ Текст»» автор выделяет лексическую функцию слова, как один из критериев, по которому можно охарактеризовать заголовочную единицу [11].

На ряду с выше представленными критериями для характеристики лексической единицы, среди учёных-лингвистов есть мнение, что дополнительно заголовочную единицу также можно охарактеризовать по

следующим критерия: прагматический компонент [1], однозначность или многозначность слов, их сфера употребления, прямое или переносное значение слова.

Обобщая выше приведённые знания, можно составить следующий список критериев, по которому лексикограф может охарактеризовать заголовочную единицу в словаре:

1. Фонетическая характеристика включает в себя указание на произношение лексической единицы, например, транскрипцию, знак ударения, тон, интонация.

2. Грамматическая характеристика даёт информацию об основных морфологических свойствах лексической единицы.

3. Семантическая характеристика даёт отличительную характеристику слова, которая служит для различения его значения от значений других слов.

4. Этимологическая характеристика указывает на происхождении слов.

5. Синтаксическая сочетаемость даёт набор синтаксических связей одной лексической единицы с другими лексическими единицами, с которыми первая может употребляться.

6. Прагматический компонент – это область исследований в семиотике и языкознании, в которой изучается функционирование языковых знаков в речи.

7. Многозначность слова, полисемия. Наличие у слова более чем одного значения, т.е. способность одного слова передавать различную информацию о предметах и явлениях внеязыковой действительности.

8. Лексическая функция – это зависимость, которая связывает слово с его лексическими коррелятами.

1.2.1 Классификация словарей и их цели

В русской лексикографии накоплен значительный опыт создания словарей и справочников разных типов. Тип словаря определяется информацией о слове, которая является для данного справочника основной. Различают два типа словарей. Это филологические словари, содержащие

знания о языке, и энциклопедические справочники, содержащие знания о мире [22].

История лексикографии знает не одну классификацию словарей. В русской лексикографии фундаментальной классификацией является труд Щербы Л.В. В своей работе «Опыт общей теории лексикографии» Лев Владимирович выделил следующую классификации словарей, в основе которой лежат шесть противоположений:

1. Словарь академического типа противопоставляется словарю-справочнику. Словарь академического типа является нормативным, описывающим лексическую систему данного языка: в нём не должно быть фактов, противоречащих современному употреблению. В противоположность академическим словарям словари-справочники могут содержать сведения о более широком круге слов, выходящих за границы нормативного литературного языка.

2. Энциклопедический словарь противопоставляется общему словарю. Энциклопедические словари описывают вещь, реалии. В сою очередь, лингвистических словарей описывают слова.

3. Тезаурус противопоставляется обычному (толковому или переводному) словарю. Тезаурусами считаются словари, в которых приводятся все слова, встретившиеся в данном языке хотя бы один раз.

4. Обычный (толковый или переводной) словарь противопоставляется идеологическому или идеографическому словарю. В идеологическом словаре слова должны идти по порядку.

5. Толковый словарь противопоставляется переводному словарю.

6. Исторический словарь противопоставляется неисторическому словарю [27].

Из числа современных лингвистов, лексикографов, российский лексикограф Морковкин В.В. представил собственную классификации словарей:

1. Лексические словари: орфографические, орфоэпические, толковые, переводные, лингвострановедческие, словоизменительные и словари морфемной структуры слов: словари синонимов и антонимов.

2. Словари сочетаемости, именного и глагольного управления, словосочетаний.

3. Фразеологические словари: фразеологизмов, пословиц и поговорок, крылатых слов, клише и речевых формул.

4. Аспектные и полиаспектные словари.

5. Диахронические и синхронические словари.

6. По характеру расположения языкового материала: формально упорядоченные (алфавитные: прямые и обратные) и содержательно упорядоченные (идеографические, гнездовые).

7. Инкорпорированные словари.

8. Книгопечатные и компьютерные словари [12].

В совместной работе Шереметьевой С.О. и других авторов «К вопросу об электронных ресурсах профессиональной лексики» авторы разделяют электронные словари на два основных вида: электронные копии бумажных словарей и электронные словари с пользовательским интерфейсом, основанные на базе данных или знаний разной глубины в предметной области [25].

На сайте Санкт-Петербургского государственного университета представлена классификация различающая следующие виды словарей:

1. Орфоэпические словари хранят сведения о произношении, ударении и образовании грамматических форм каждого слова, включенного в словник. В словарях этого типа записаны произношение по норме русского литературного языка для каждой единицы словника [22].

В зависимости от объема слов в словнике, такие словари могут быть предназначены как специалистам в какой-либо области знаний, так и более широкому кругу читателей. Например, «Орфоэпический словарь русского языка» под редакцией Аванесова Р.И. является самым известным словарем

данного типа и рассчитан на специалистов в области лингвистики и людей, работающих в сфере коммуникации с общественностью.

2. Этимологические словари содержат справки о происхождении слов, из какого языка пришли некоторые слова в нашу речь. Объектом описания этимологического словаря служит заимствованная лексика. В словарях этого типа указывается написание заимствованного слова на языке, откуда данное слово было заимствованно; указывается первоначальное звучание; значение, которое слово имело в языке-источнике. Полнота информации о слове меняется в зависимости круга читателей на который ориентируется автор, лексикограф составляя словарь [22].

Известные этимологические словари русского языка являются: «Этимологический словарь русского языка» Цыганенко Г.П., «Краткий этимологический словарь русского языка» Иванова В.В., Шанской Т.В. и Шанского Н.М.

3. Словари общего типа – обычные толковые и двуязычные (переводные) словари [22].

Говоря о словарях общего типа, учёные-лингвисты имеют в виду словари разной степени полноты, в которых отражена народная и литературная лексика. К словарям этого типа относятся следующие словари: «Словарь русского языка в четырёх томах» Ушакова Д.Н., «Толковый словарь русского языка» Ожегова С.И.

4. Словари трудностей (правильностей) хранят в себе информацию: о написании отдельных слов; произношении или выборе места ударения в слове или его словоформе; словоупотребление, которому соответствует языковая ситуация; грамматическая атрибуция слова: использование правильной формы падежа и числа в зависимости от речевой ситуации [22].

Как пример данного типа словарей можно указать «Словарь трудностей русского произношения» Каленчук М.Л., Касаткина Р.Ф.

5. В научных классификациях словарей термин «полный словарь» обозначает тип издания, содержащий исчерпывающий состав тех слоев и

разрядов лексики, которые служат объектом описания данного справочника [22].

Так, например, «Полный филологический словарь русского языка» Орлова А.И. включает подробное объяснение всех отличий разговорной речи от письменной, замену всех иноязычных слов, вошедших в состав русского языка, чисто русскими словами. По такому принципу, словарем полного типа можно считать «Орфографический словарь русского языка» под редакцией Лопатина В.В.

6. В словарях неологизмов описываются слова, значения слов и словосочетаний, появившиеся в определенный период времени [22].

Как показывают исследования, развитые языки активно пополняются новыми словами благодаря компьютерным технологий, активно покорившим наш мир. Яркий пример такого типа словарей – «Толковый словарь русского языка начала XXI века: Актуальная лексика» под редакцией Скляревской Г.Н.

7. Грамматические словари содержат сведения о формальных, словоизменительных и синтаксических, свойствах слова. Порядок слов в таких словарях может быть, как прямым, когда слова располагаются в алфавитном порядке, так и в обратном, когда слова располагаются в алфавитном порядке, начиная с последней буквы слова. Обратный порядок позволяет представить читателям словообразовательные свойства слова [22].

Одним из фундаментальных словарей этого типа является «Грамматический словарь русского языка. Словоизменение».

8. Во фразеологических словарях заголовочной статьёй являются фразеологические единицы. Эти единицы являются одним из наиболее закрытых разрядов лексики [22].

В качестве наиболее значимых словарей фразеологизмов можно привести: «Фразеологический словарь русского языка» под редакцией Молоткова А.И. и «Французско-русский фразеологический словарь» Гака В.Г. и других авторов.

9. Отраслевые справочники хранят языковые единицы, ограниченные в сфере словоупотребления. Отраслевые словари принято разделять на два типа: словари, в которых толкуются значения слов и энциклопедические справочники, в которых описаны знания о мире [22].

Как пример, для словарей первого типа, можно привести «Толковый словарь избранных медицинских терминов. Эпонимы и образные выражения» под редакцией Л. П. Чурилова Л.П., Колобов А.А. и Строев Ю.И. Примером для второго типа является «Военно-морской словарь» под редакцией Чернавин В.Н.

10. Орфографический словарь предназначен указать нормативное, правильное написание слов [22].

Наиболее полным словарём общего типа на сегодняшний день является «Русский орфографический словарь» под редакцией Лопатина В.В.

11. Словообразовательные словари содержат сведения о морфемном членении слова, его словообразовательной структуре. В таких справочниках представлена информация о структуре слова и тех элементах, из которых это слово состоит. В словообразовательных словарях, в большинстве случаев, слова собраны по корневым гнёздам.

Блумфельд в своих трудах писал, что в компьютерной лингвистике лексикон – это информация о лексемах, упорядоченных в виде списка [Bloomfield, 1933]. Например, лексема *словарь* в лексиконе представляет собой граф, который охватывает все словоформы, которые могут существовать в языке *словарь*, *словарём*, *словарю*. Как ссчитают иностранные учёные-лингвисты, в ряде случаев, в лексему включаются синтаксические и прагматические знания о слове [34-36].

В словаре, как и в лексиконе основная форма слова соотносится со всеми словоформами, но, например, существительное *игрок* не будет соотносится со глаголом *играть*, т.к. *игрок* является производной слова, а не словоформой. В свою очередь, в лексиконе производная *игрок* будет одной из вершин графа [30].

Лексикон направлен на компьютер и имеет разное представление в виде таблиц, графов, отдельных файлов: следовательно, имеет удобочитаемую для ЭВМ форму, в свою очередь совершенно непонятное для человека.

Когда речь идет о программировании, лексикон – это группа слов, которые используются для создания языка программирования [32]. Также их называют зарезервированными словами. К ним относятся, например, for, or, in, if, while являются лишь несколькими примерами слов, найденных во многих языках программирования.

1.2.2 Источники лексикографической информации

Говоря о лексикографических источниках, многие учёные имеют ввиду прежде всего словари, которые должны быть использованы в качестве материала для составления других словарей. При этом материал словарей, являющихся источником, может быть использован не обязательно в том виде, в каком он представлен [13, 14].

В своей статье Щитова О.Г. «Лексикографические источники изучения функциональной эквивалентности иноязычных новаций в русском языке начала двадцать первого века» пишет, что в группу лексикографических источников, констатирующих состояние лексико-семантической системы, входят словари любого типа [29].

Учёные-лингвисты опираются не только на языковой материал, извлечённый из разного рода словарей, но и на материал различных типов корпусов, выбор которых зависит, прежде всего, от темы исследования и поставленных целей.

Одним из ведущих направлений прикладной лингвистики является корпусная лингвистика, занимающаяся разработкой общих принципов построения и применения лингвистических корпусов с использованием информационных технологий [32]. Стройков С.А. под лингвистическим или языковым корпусом текстов понимает большой, представленный в электронном виде, унифицированный, структурированный, размеченный,

филологически компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач [19].

В качестве базы для корпуса в основном используются тексты, которые представляют язык во всех его проявлениях [23]. Классифицировать корпусы можно по разным признакам: цель создания корпуса, тип языковых данных, динамичность, тип разметки, объем текстов, параллельность, аннотирование [22].

Среди множества определений корпуса учёные-лингвисты выделяют его главные свойства:

1. Электронный – в современном понимании корпус должен быть в электронном виде.

2. Репрезентативный – должен хорошо представлять объект, который моделирует.

3. Размеченный – главное отличие корпуса от коллекции текстов

4. Прагматически ориентированный – должен быть создан под определенную задачу [22].

В качестве примеров электронных корпусов, доступных в открытом доступе в сети интернет, стоит привести:

1. Национальный корпус русского языка, доступный для поиска текстов в электронном формате. Также доступный для поиска исторических, церковнославянских, древнерусских (XI – XIV века) и среднерусских (XV – начало XVIII века) текстов.

2. Британский национальный корпус из ста миллионов слов, содержащий образцы письменного и разговорного британского английского языка из широкого круга источников. Корпус охватывает британский английский конца двадцатого века, представленный широким разнообразием жанров, и задуман как образец типичного разговорного и письменного британского английского языка того времени.

1.2.3 Этапы построения словарей

Реализация любого проекта, связанного со словами естественного языка, предполагает поиск и сбор громадной, постоянно увеличивающейся, а иногда и теряющей свою актуальность информации.

В статье Щитовой О.Г. «Лексикографические источники изучения функциональной эквивалентности иноязычных новаций в русском языке начала двадцать первого века» автором выделяются следующий критерии для отбора словарного материала. С её слов мы понимаем, что основу словаря составляет общеупотребительная лексика, обозначающая понятия из различных сфер жизни [29].

Обобщая труды многих лингвистов, следует сказать, что родственные, однокоренные слова обычно представлены в словаре не всегда полно. Более полно словообразовательное гнездо оформляется в тех случаях, когда значения словообразовательных аффиксов неочевидны, неоднозначны или, когда обозначаемые этими словами понятия чрезвычайно важны. Тем не менее, если все входящие в гнездо слова обладают развитой системой значений либо их значения далеко разошлись друг от друга; при этом каждое слово описывается в отдельной словарной статье [28].

Из-за ограниченного объёма словаря в его состав не включаются следующие группы слов, найденные нами на образовательных сайтах сети интернет: полностью или частично названия животных; названия растений; названия химических элементов и соединений; воинские звания за исключением исторических, таких, как вахмистр, штабс-капитан; сложные слова, если их значение является простой суммой значений составляющих их корней и приставок, также не включаются в словарь, например, геоботаника, антикоммунизм, деблокировать [22].

В статьях, расположенные на сайте Санкт-Петербургского государственного университета, находится информация о «разном» порядок слов в словарях. Порядок слов в таких словарях может быть, как прямым, когда слова располагаются в алфавитном порядке, так и в обратном, когда

слова располагаются в алфавитном порядке, начиная с последней буквы слова [22]. Одним из фундаментальных словарей этого типа является «Грамматический словарь русского языка. Словоизменение» Зализняка А.А.

В своём интервью учёный-лингвист Дмитриев Д.Н. поделился своим мнением о концепции создания словаря. С его слов мы понимаем, что для создания словаря в бумажном виде необходимо:

Во-первых, необходимо составить словник и дать толкование каждой языковой единице.

Во-вторых, провести более полную проверку собранной лексики, то есть создать тезаурус.

В-третьих, составление подробных словарных статей из собранного в тезаурус материала, что и является заключительным этапом в написание бумажного словаря [8].

Как пишет в своей книге Нелюбин Л.Л. «Перевод и прикладная лингвистика» электронный словарь – это словарь упорядоченный, относительно конечный массив лингвистической информации, представленный в виде списка, таблицы или перечня, удобного для размещения в памяти ЭВМ и снабжённого программами автоматической обработки и пополнения [13].

В свою очередь зарубежные учёные-лингвисты отмечают, что термин электронный словарь может быть использован для обозначения любого справочного материала, хранящегося в электронном виде и предоставляющего информацию о написании, значении или использовании слов [34-36]. В статье Стройкова С.А. «Основные понятия лингвистической концепции электронного лексикографического гипертекста» написано, что для автоматического словаря, чаще всего источниками выступают корпуса текстов [19].

Стоит отметить, что объектом описания автоматического словаря – заголовочной единицей – является лексическая единица языка также, как и в классическом словаре. Следовательно, в качестве заголовочной единицы

могут выступать следующие единицы языка: фонемы, морфемы, слова, словоформы и словосочетания [8, 54].

По мнению Беляевой Л.Н. использование ЭВМ в лексикографии предполагает ряд задач, сравнительно схожих с классическими задачами лексикографии:

1. Решение задач отбора лексических единиц, извлечения информации о лексической единице из ориентированного массива текстов.

2. Создание, редактирование словаря, его последующих изданий.

3. Создание и ведения терминологических баз данных и онтологий.

4. Исследование лексического состава и динамики лексического спектра конкретного языка [4].

Авторы работы «К вопросу об электронных ресурсах профессиональной лексики» Шереметьева С.О. и другие предлагают следующие этапы по созданию автоматического словаря.

На первом этапе определяется предметная область, цель и пользователи словаря.

На втором этапе определяется модель знаний: типы, объем и формализм представления лингвистической информации в словарной статье.

Третий этап состоит в определении источника и состава вокабуляра с лингвистической информацией в соответствии с моделью знаний.

На четвёртом этапе определяются эквиваленты для каждой единицы вокабуляра с лингвистической информацией о переводных эквивалентах в соответствии с моделью знаний.

Пятый этап предполагает спецификацию и программную реализацию словаря.

Шестой этап – введение знаний, полученных в результате выполнения третьего и четвёртого этапов в электронную словарную оболочку через интерфейс лексикографа [25].

Агаповоа Н.А. и Картофелева Н.Ф. в совместной работе «О принципах создания электронного словаря лингвокультурологического типа: к

постановке проблемы» описали следующие этапы по созданию электронного словаря:

1. Сбор материала, который выступит в качестве основы словаря.

2. Написание словарных статей.

3. Разработка общей концепции словаря

4. Проектирование архитектуры компьютерного приложения с несколькими вариантами поиска и группировки материала, выбор средств разработки.

5. Составление программистом технического задания в соответствии с разработанной концепцией.

6. Завершение разработки электронного ресурса: создание дружелюбного интерфейса и удобной навигации [1].

1.3 Компьютерная лексикография

Одной из задач автоматического извлечения информации является автоматическое формирование словарей предметной области [25].

Лексикография (науки о создании, изучении и использовании словарей) включает в себя как теоретические знания, так и практические исследования: теоретическая часть лексикографии включает в себя теорию и историю создания словарей, в свою очередь, практическая занимается непосредственным создании словарей и сбором первичного словарного материала [8].

Под термином компьютерная лексикография, со слов, Федосова Ю.В. в данной дипломной работе мы понимаем, что компьютерная лексикография представлена совокупностью методов и программных средств обработки текстовой информации для создания словарей [30].

В тоже время, под термином компьютерная лексикография Чернышева М.И. и Филиппович А.Ю. понимают, что это прикладная научная дисциплина в языкознании, которая изучает методы использования компьютерной техники для составления словарей [31].

По мнению Чепика Е.Ю., целью компьютерной лексикографии является разработка компьютерных технологий, составление и эксплуатация словарей. В качестве примера автор приводи, базы данных; компьютерные картотеки; программы обработки текста, позволяют в автоматическом режиме извлекать и формировать словарные статьи, хранить словарную информацию и обрабатывать её [23].

Из истории нам известно, что компьютерная лексикография возникла как отдельная дисциплина в прикладной лингвистике с появлением машиночитаемых словарей, начиная с создания Джоном Олни карманного словаря в компаниеи Merriam-Webster в 1960 годах. В 1987 году Берд, Кальцолари и Чодоров разработали вычислительные инструменты для анализа текста. Данная программа решала проблему неоднозначности слов. Первоначально электронные словари имели такую же форму записи, как обычные словари, и исследователям приходилось тратить много времени для интерпретации такой формы записи, например, чтобы определить, к какой части речи относится определённое слово. С развитием технологий издатели решили отделить базу данных электронного словаря от того, как он выглядит при печати [29].

Сегодня компьютерная лексикография наиболее известна благодаря созданию и применению WordNet. По мнению Кашеварова И.С., высказанному в статье «Электронный словарь как новый этап в развитии лексикографии» компьютерная лексикография – это научная дисциплина представлена совокупностью методов и программных средств обработки текстовой информации для создания словарей.

Как пример ресурсов компьютерной лексикографии, можно привести российскую компанию-разработчика ABBYY. Наиболее известные продукты – программа для распознавания текстов ABBYY FineReader, система потокового ввода данных ABBYY FlexiCapture и электронные словари ABBYY Lingvo. Специалисты этой компании к возможностям электронного словаря относят:

1. Возможность вывода содержания словарной статьи в форме, необходимой пользователю, включая возможность частичного вывода по разным критериям, например, части речи.

2. Использование для доступа к содержанию различных лингвистических технологий, таких как морфологический и синтаксический анализ, распознавание и синтез звука [32].

К иностранным аналогам можно отнести такого гиганта как переводчик от компании Google. Google Translate это веб-служба компании Google, предназначенная для автоматического перевода части текста или веб-страницы на другой язык. В марте 2017 года Google полностью перевела свой онлайн словарь на двигатель, построенный как нейросеть, для более качественного перевода.

1.4 Автоматизация построения вокабуляра словаря

По мнению зарубежных учёных, именная группа – это словосочетание, в котором имя существительное является вершиной, то есть главным словом, определяющим характеристику всей составляющей. В современных синтаксических теориях принято считать, что даже если имя не содержит зависимых, оно всё равно является именной группой (состоящей из одного слова) [34-36].

Учёные-лингвисты выделяют два основных подхода для извлечения именных групп из текста рациональный и эмпирический. Рациональный подход заключается в составлении шаблонов для идентификации именных групп в тексте на определённо языке. Эмпирический в свою очередь даёт возможность основываться на данных текста и не требует больших затрат для разработки базы знаний.

Эффективным способом выделения именных групп является гибридный подход, разрабатываемый Шереметьевой С.О. и основанный на применении баз знаний стоп-слов, которые не могут использоваться в начале, середине или конце именной группы [2, 24-25].

В совместной работе Вэнь Чжана и других авторов «A Study on Multi-word Extraction from Chinese Documents» подробно описан принцип извлечения именных групп из текстов на китайском языке. Вэнь Чжан и другие авторы применили в своей работе метод n-грамм для предварительного обработанного текста, размеченного по частям речи.

Во-первых, из предварительного размеченный текст по частям речи формируется список n-грамм.

Во-вторых, используя список всевозможных шаблонов, каждый кандидат проходи проверку на то, соответствие фразе по выбранной проблематике. За исключением, если фраза является одним словом [39].

Бабина О.И. и Тамгина Е.С. полагают, что при коллективной работе над лексиконом следует ввести этап верификации переводов, целью которой будет унификация предлагаемых для вхождений лексикона переводов. Предлагаемая ими процедура верификации перевода для компонента именной группы, встречающегося в составе нескольких многокомпонентных именных групп и допускающего множественную манифестацию на языке перевода, включает следующие этапы:

1. Отбор из двуязычного лексикона именных групп с общим лексическим компонентом и соответствующими переводами.

2. Построение иерархии именных групп (от более длинных к более коротким).

3. Формирование перечня кандидатов для перевода общего лексического компонента.

4. Проверка отобранных кандидатов и, при необходимости, корректировка переводов в лексиконе.

Этап отбора включает поиск именных вхождений лексикона, содержащих целевую подстроку на языке оригинала. Все вхождения, удовлетворяющие этому условию, и соответствующие им переводы заносятся во множество верифицируемых именных групп.

На основе парадигматического подхода программа лемматизации имитирует работу лексикографа, который извлекает и упорядочивает слова, относящиеся к одной лексеме. Из совместной работы Ляшевской О.Н и других нам известно, что процедура автоматического сведения парадигм на первом этапе разделяет словоформу на псевдооснову и псевдоокончание, затем по всему массиву подсчитывается количество повторений каждой квазиосновы и каждого квазиокончания.

В совместной работе Шереметьевой С.О. и Осминина П.Г., авторами были обозначены следующие методы извлечения ключевых слов, в нашем случае вокабуляра. Наиболее простой статистический метод извлечения ключевых слов предполагает построение множества кандидатов ключевых слов путём ранжирования всех словоформ или лексем документа по частоте. Фильтрация заключается в отборе в качестве ключевых определённого количества наиболее частотных лексем.

Гибридные методики, в которых статистические методы обработки документов дополняются одной или несколькими лингвистическими процедурами (морфологическим, синтаксическим, или семантическим анализами) и лингвистическими базами знаний различной глубины. Гибридные методы извлечения ключевых слов из документа, также как и статистические, могут требовать или не требовать корпуса текстов [25].

Обобщая вышесказанное можно сделать вывод, что сбор вокабуляра для словаря является нетривиальным делом несмотря на количество продолженных методов по извлечению лингвистических знаний из текстов. Мнения авторов всех предложенных методов были приняты нами во внимание и наиболее подходящим для нашего исследования методом является метод Шереметьевой С.О., основанный на применении баз знаний стоп-слов.

Выводы по главе 1

На основании собранного и рассмотренного теоретического материала в первой главе нами была установлена противоречивость различных точек

зрения, что требует осмотрительного подхода к выполнению дальнейшего исследования и учёта особенностей объекта исследования дипломной работы – списков текстов на китайском языке по лексикографической тематике и предмета исследования дипломной работы – именных групп китайского языка.

На основании проведённого анализа над теоретическим материалом по составлению словарей, словарной статье и её составу, составлению автоматических словарей был сделан следующий обобщающий список о словарях в целом (микроструктуре и макроструктуре):

Во-первых, в словарь могут входить различные языковые единицы, такие как: фонемы, морфемы, слова, словоформы и словосочетания.

Во-вторых, заголовочная единица словаря может быть охарактеризована по следующим пунктам: фонетическая характеристика; грамматическая характеристика; семантическая характеристика; этимологическая характеристика, указывающая на происхождение слов; синтаксическая сочетаемость одной лексической единицы с другими лексическими единицами, с которыми первая может употребляться; прагматический компонент; полисемия и лексическая функция.

В-третьих, алгоритм составления словаря должен состоять из следующих этапов: определения предметной области, цели и пользователей словаря; определения модели знаний; разработки концепции словаря и создания словаря на языке программирования.

ГЛАВА 2 РАЗРАБОТКА МОДЕЛИ ДЛЯ АВТОМАТИЧЕСКОГО СОЗДАНИЯ КИТАЙСКОГО СЛОВАРЯ ИМЕННЫХ ГРУПП

2.1 Мотивация для создания программы по формированию словарей именных групп на китайском языке по предметной области

Страноведение важно в культурном аспекте, тексты по страноведческой тематике являются основным содержателем сведений страны изучаемого язык. Существующих словарей недостаточно, чтобы обеспечить вниманием нужды студентов и преподавателей при прочтении текстов по страноведению. Словари с однокомпонентной лексикой являются превалирующим, нежели с многокомпонентной лексикой, а вот именно многокомпонентная лексика составляет основной словарный запас студента, только что начавшего изучать иностранный язык.

Необходимость в создании программы для автоматического формирования словарей предметной области на китайском языке была обусловлена следующим фактором – автоматизацией процесса создания словаря. Чтение и перевод текста, объёмом, примерно, один лист формата А4, на китайском языке занимают достаточное количество времени и сил у студента. Для последующего отбора и записи словарного материала также необходимо время, а именно этот этап работы с текстом может быть автоматизирован: автоматизация повторяющихся действий – извлечение и систематизация.

Опираясь на теорию, собранную и проанализированную в первой главе дипломной работы, мы можем определить, что цель, модель знаний и пользователей словаря.

Цель создания словаря: словарь создаётся в учебных целях в области лингвистического страноведения.

Модель знаний словаря: словарь будет состоять из многокомпонентных лексем, именных групп китайского языка, т.к. именные группы являются наиболее частотными компонентами текста.

Из цели создания словаря следует вывод о непосредственно прямых пользователях словаря.

Пользователи словаря – студенты и преподаватели.

2.2 Общая характеристика программы

Алгоритм работы программы следующий, на вход поступает протегированный текст на китайском языке, на выходе пронумерованный список именных группы (Рисунок 1).

Рисунок 1 – Блок схема алгоритма извлечения именных групп

Для создания программы по автоматическому формированию словарей именных групп китайского языка, в первую очередь, необходимо понять, что есть именная группа в китайском языке.

По мнению зарубежных учёных, именная группа – это словосочетание, в котором имя существительное является вершиной, то есть главным словом, определяющим характеристику всей составляющей. В современных синтаксических теориях принято считать, что даже если имя не содержит зависимых, оно всё равно является именной группой (состоящей из одного слова) [34-36].

Следующая задача – создание корпуса текстов для проведения исследований путём использования google translator, в частности его функции компьютерное зрение.

Принцип отбора именных групп следующий. Есть список, тех частей, которые могут входить в именную группу и только эти части речи автоматически извлекаются в именную группу. Следовательно, показателем границы, конца именной группы будут служить все остальные части речи, не входящие в список частей речи для именной группы.

Например, представим, что у нас есть список частей речи, которые могут входить в именную группу. В предложении «中国位于亚洲大的东部、太平洋的西岸。» нам необходимой найти все именные группы, мы берём первый иероглиф и смотрим является ли он той частью речи, которая может быть в именной группе. «中国» – это существительное, обозначающее географический объект, следовательно, данный иероглиф может быть в именной группе. Берём на проверку следующий иероглиф «位于». Это глагол, следовательно, мы не включаем его в именную группу и данный иероглиф служит показателем того, что именная группа до него закончилась. Из этого следует вывод, что первая именная группа, найденная в предложении – «中国 ».

Берём на проверку следующий иероглиф «亚洲». Этот иероглиф является существительным, следовательно, он удовлетворяет критерию отбора в кандидаты в именную группу. Берём на проверку следующий иероглиф «大陆», так как он тоже удовлетворяет критериям отбора в именную группу, следует объединить его с иероглифом «亚洲», так как они стоят вместе в тексте, следовательно, являются одной именной группой. Берём на проверку следующий иероглиф «的», так как он тоже удовлетворяет критериям отбора в именную группу, следует объединить его с иероглифами или уже кандидатом в именную группу «亚洲大陆», так как они стоят вместе в тексте,

следовательно, являются одной именной группой. Берём на проверку следующий иероглиф «东部», так как он тоже удовлетворяет критериям отбора в именную группу, следует объединить его с кандидатом в именную группу «亚洲大陆的», так как они стоят вместе в тексте, следовательно, являются одной именной группой.

Следующий символ предложения – точка, данный маркер не относится к списку объектов, которые могул употребляться в именной группе, следовательно, мы нашли конец именной группы «亚洲大陆的东部».

Таким образов в данном предложении есть две именные группы: «中国» и «亚洲大陆的东部».

Обобщая вышеизложенное, работу программы можно разбить на этапы, через которые проходит анализ текста:

1. Синтаксический анализ.

На данном этапе отбираются кандидаты в именные группы по методу, разработанному Шереметьевой С.О. Работа по этому методу основана на применении баз знаний стоп-слов, которые не могут использоваться в начале, середине или конце именной группы [24-25]. Данный метод был нами, авторами этой дипломной работы, дополнен. Суть нашего подхода заключается в следующем, каждый кандидат, а кандидатом в именную может быть и один иероглиф, если его тег не в списке стоп-слов, тогда кандидат отбирается в список именных групп. Пройдя первый этап анализа, список кандидатов отправляется на второй этап – фильтрацию.

2. Фильтрация кандидатов в именную группу.

Фильтрация кандидатов в именную группу происходит следующим образом, обусловленным тематикой текстов. Если кандидат состоит из одного иероглифа, и тег этого иероглифа не обозначает географического места, как например, города, реки, озёра, страны, то данный кандидат не включается в именную группу. Пройдя второй этап анализа, список кандидатов отправляется на третий этап – создание словаря.

3. Сохранение результатов работы.

Результаты работы программы, список именных групп, сохраняется в файле с названием 中文字典 [Zhōngwén zìdiǎn], что в переводе на русский язык означает словарь китайского языка. Микроструктура словаря представлена следующем образом, заголовочная единица – именная группа – не имеет какой-либо характеристики, за исключение нумерации. Нумерация в словаре совпадает с последовательностью этих именных групп в тексте.

2.3 Использование сторонней программы SegmentAnt для морфологического анализа текста

В нашей работе, как мы уже упомянули выше, мы не ставим среди основной цели, второстепенную – разрешить все проблемы, с которыми столкнёмся во время работы. Одной из таких проблем стала проблема определения частей речи в тексте на китайском языке. Научные исследования по разрешению лексической многозначности находятся в поле зрения прикладной и компьютерной лингвистики достаточно давно и имеют многолетнюю историю. С течением лет количество предложенных решений и их эффективность неуклонно росли до тех пор, пока эффективность не достигла определённого уровня сравнительно-эффективных показателей точности для определённого спектра слов и типов многозначностей. Полного решения задача пока не получила, поскольку на пути успешного решения стоит много проблем, напрямую связанных с языковыми особенностями человеческой речи.

Так как иероглиф может являться сразу несколькими частями речи, например, иероглиф 家 [jiā] может быть, как существительным семья, так и счётным словом для зданий. Для решения этих проблем мы использовали стороннею программу SegmentAnt для тегирования и сегментации текста [37]. Принцип работы программы SegmentAnt показан на рисунке под номером один (Рисунок 2).

Рисунок 2 – Принцип работы программы SegmentAnt

2.4 Сбор списка текстов и их обработка в программе SegmentAnt

Корпус для данной работы был собран путём использования google translator, в частности его функции компьютерное зрение. Компьютерное зрение – это междисциплинарная область, которая занимается проблемой того, как компьютеры могут получать и понимания информацию с цифровых изображений или видео [36].

При обработке текстов вручную нами были выделены следующие формальные признаки. Значимая лексическая информация содержится в основном в именных группах. В состав именной группы в китайском языке могут входить следующие части речи: существительное и прилагательное.

1. Существительные собственные обозначают отдельных лиц, единичные предметы и явления, например, 北京 [Běijīng] – Пекин, 黄河 [Huánghé] – Жёлтая река. Существительные нарицательные обозначают однородные лица, однородные предметы и явления, например, 雪 [xuě] – снег, 山 [shān] – гора. Существительные – единицы измерения называют меры, например, 里 [lǐ] – мера длинны и предметы мер длинны 盘 [pán] – чашка [3].

2. Прилагательное в китайском языке обозначает качественный признак предмета, может выполнять функцию определения или сказуемого. Обозначая относительный признак предмета, прилагательное обычно выполняет функцию определения [3].

Для нахождения именной группы необходимо знать какой частью речи является каждый иероглиф. Проблема неоднозначности частей речи может быть решена набором тегов. Тег – это тип метаданных, приписываемый и характеризующий объект [20]. В лингвистике частеречная разметка (Part of Speech Tagging) – это процесс определения частей речи и грамматических характеристик слов в тексте (корпусе) с приписыванием им соответствующих тегов [20, 33]. Разметка по частям речи в китайском языке является не тривиальной проблемой т.к. иероглифы в предложение идут друг за другом без пробелов. Следовательно, на первом этапе необходимо определить границы слов в предложении. Сегментация текста – это процесс разделения письменного текста на значимые единицы, такие как слова, предложения или заголовки [33].

В данной работе мы воспользовались сторонней программой по сегментации и тегированию иероглифических текстов. SegmentAnt продукт Энтони Лоуренс [33].

Список тегов частей речи китайского языка был взят с сайта [20] т.к. ни в инструкции к SegmentAnt, ни на официальном сайте [33] не было данного списка. Нами было проведено сравнение, сопоставление тегов из двух источников, что послужило базой для создания списка тегов.

После обработки текста в программе SegmentAnt, необходимо составить список всех тегов и разбить его на группы, которыми в последующем будет оперировать программа. Общий список тегов получился из двадцати четырёх маркеров, из данного списка будут выделены теги иероглифов стоп-слов. В результате мы получим список стоп-слов.

Теги знаков пунктуации по причине их не востребованности удаляются из текста. В группе тегов для именной группы все теги т.к. они относятся к

одному классу целесообразней, в нашем случае, объединяются под тегом «n»: «f» и «s» – части света, «nr» – имена собственные, «nt» – темпоральные существительные, «nz» – другие имена собственные, «l» – существительные места и «t» – аббревиатуры. Тег вспомогательных частиц «uj» заменяется на тег «u» (Рисунок 3).

Данное решение позволит нам сократить базу лингвистических знаний о тегах частей речи, что в свою очередь будет полезно для алгоритма работы программы; для работы непосредственно компьютера, т.к. его оперативная память не будет задействована при обработке тегов частей речи, относящихся к оному классу объектов. Стоит отметить, что данное решение не появляет на точность извлечения именных групп, т.к. заменённые теги обладают всей необходимой информацией об иероглифе.

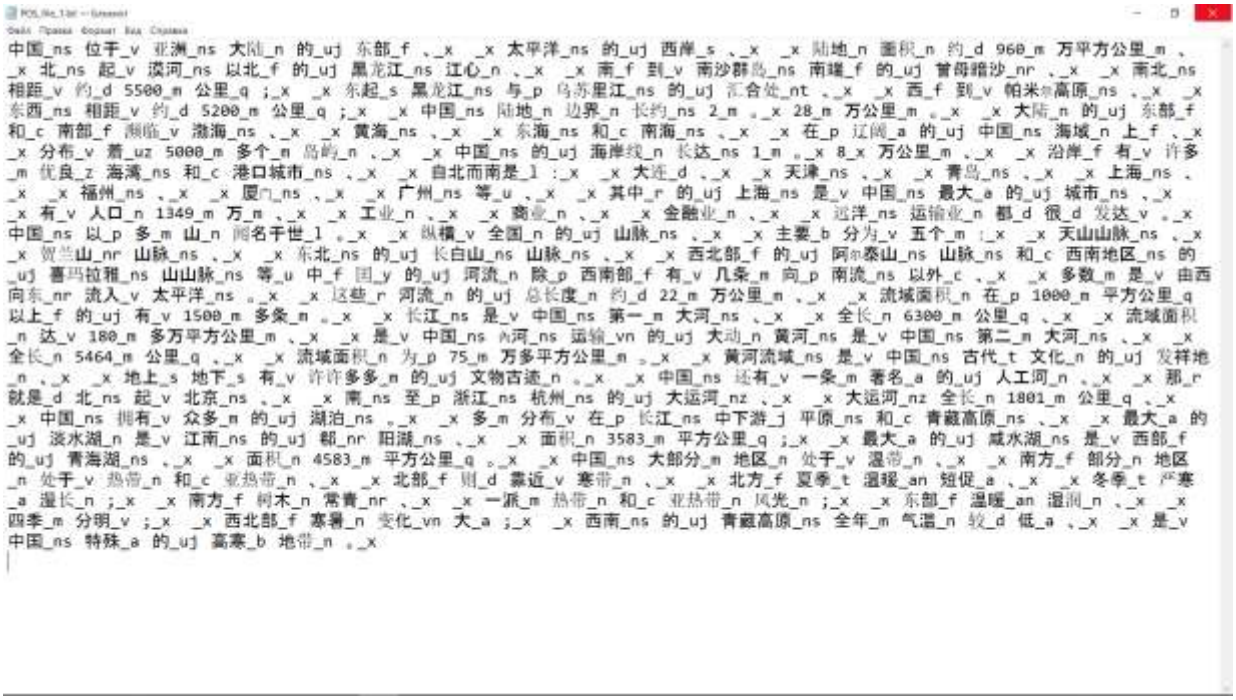


Рисунок 3 – Текст с изменёнными тегами речи

Решение по замене некоторых тегов, удалению и объединению обусловлено следующими факторами.

Во-первых, в данной работе нет необходимости использовать теги с расширенной информацией о иероглифах т.к. нам не нужна информация о синтаксических отношениях.

Во-вторых, данное решение упрощает процесс перенесения информации о метаданных на язык программирования, что в свою очередь экономит память, а последнее не тормозит процесс работы программы.

2.5 Процесс извлечение кандидатов в именные группы

Процесс извлечение именных групп в нашей работе происходит следующим образом, Программа осуществляет проход по элементам предложения, если взятый на поверку иероглиф, а именно, его тег соответствует одному из тегов списка для номинативной группы: иероглиф считается кандидатом. Если взятый на поверку иероглиф, его тег не соответствует одному из тегов списка для номинативной группы, то есть находится в списке стоп-слов, в таком случае, иероглиф пропускается программой.

В качестве примера рассмотрим предложение «中国_ns 位于_v 亚洲_ns 大陆_n 的_u 东部_n、太平洋_ns 的_u 西岸_n 、陆地_n 面积_n 约_d、北_ns 起_v 漠河_ns 以北_n 的_u 黑龙江_ns 江心_n?».

Программа берёт имя существительное, начиная с начала предложения, и записывает его в отдельную переменную. Этот процесс происходит до тех пор, пока тег отдельно взятого иероглифа соответствует списку тегов номинативной группы.

Первый в кандидаты иероглиф будет «中国_ns» и по логике программы «中国_ns» является номинативной группой т.к. следующий тег иероглифа «位于_v» относится к списку стоп-слов.

Второй кандидат будет состоять из следующих иероглифов «亚洲_ns 大陆_n 的_u 东部_n». Логика действий программы в таком случае, программа путём конкатенации, если тег иероглифа соответствующий, создаёт строку: в переменную будут складываться иероглифы «亚洲_ns» + «大陆_n» + «的_u» + «东部_n» до момента несоответствия тегов. Для второй именной группы индикатором конца служи знак пунктуации «、», после его нахождения

программа прекратит конкатенацию и запишет сформированную строку как кандидата.

Третий кандидат будет формироваться по такой же логике, как и второй. Если тег иероглифа соответствующий, создаёт строку: в переменную будут складываться иероглифы «太平洋_ns» + «的_u» + «西岸_n» до момента несоответствия тегов. Для второй именной группы индикатором конца служи знак пунктуации «、», после его нахождения программа прекратит конкатенацию и запишет сформированную строку как кандидата.

Третий кандидат – «太平洋_ns 的_u 西岸_n».

Четвёртый кандидат будет формироваться по такой же логике, как второй и третий кандидаты. – «陆地_n 面积_n». Индикатором конца для четвёртого кандидата будет служить тег иероглифа «约_d» т.к. тег не относится к списку тегов именной группы.

Пятая кандидат будет состоять из одного иероглифа «北_ns» т.к. следующий тег иероглифа «起_v» относится к списку тегов глагольной группы.

Последний, шестой кандидат будет состоять из всех иероглифов начиная от последнего несоответствующего тега до конца предложения, то есть до знака пунктуации «?», шестой кандидат – «漠河_ns 以北_n 的_u 黑龙江_ns 江心_n». Список кандидатов в именную группу представлен на рисунке под номером два (Рисунок 4).

Рисунок 4 – Список кандидатов в именную группу

Как было упомянуто выше, если кандидат состоит из одного иероглифа, и тег этого иероглифа не обозначает географического места, как например, города, реки, озёра, страны, то данный кандидат не включается в именную группу (Рисунок 5).

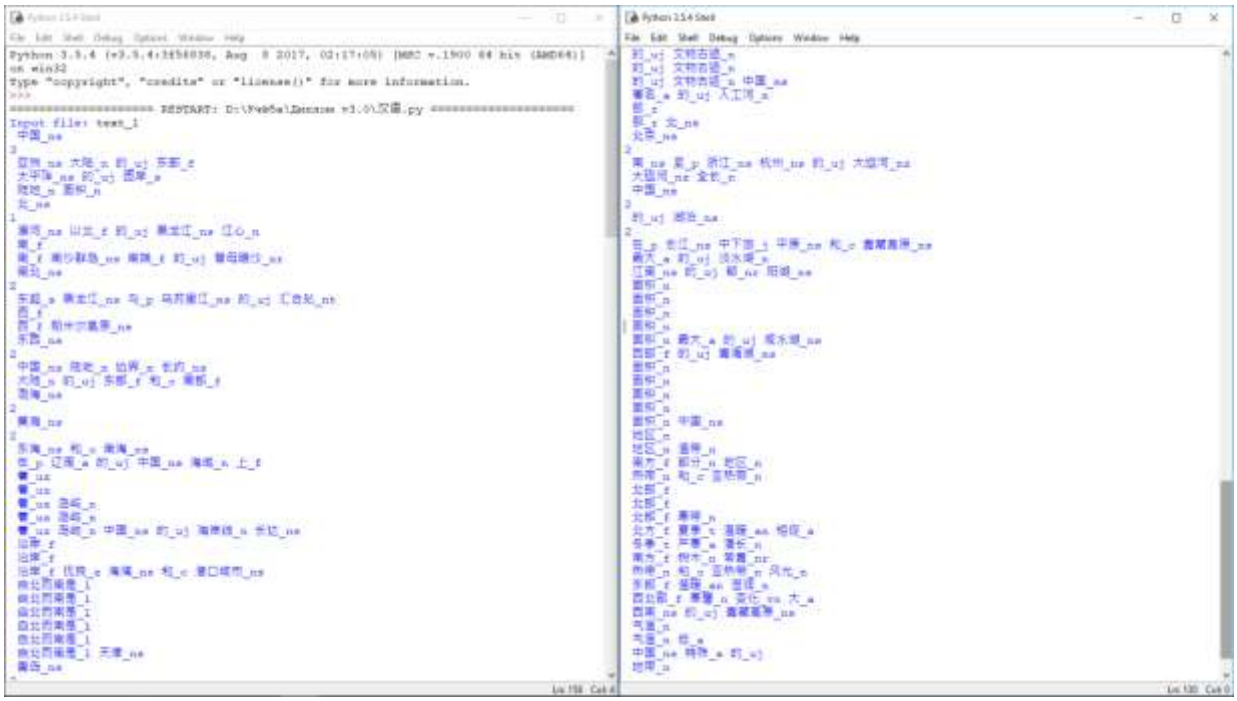


Рисунок 5 – Список кандидатов в именную группу без односоставных иероглифов

Лингвистические правила, которые мы создали и использовали в процессе извлечения кандидатов в именную группу:

1. Если взятый на поверку иероглиф, его тег не соответствует одному из тегов списка для номинативной группы, то есть находится в списке стоп-слов, в таком случае, иероглиф пропускается программой.

2. Если взятый на поверку иероглиф, его тег соответствует одному из тегов списка для номинативной группы, то есть не находится в списке стоп-слов, в таком случае, иероглиф значится как кандидат в именную группу.

3. Если взятый на поверку иероглиф, его тег соответствует одному из тегов списка для номинативной группы, то есть не находится в списке стоп-слов, и за ним стоит ещё один иероглиф, тег которого не находится в списке стоп-слов, тогда путём конкатенации создаётся кандидат в именную группу.

После обработки текста программа сформировала список из шестидесяти восьми кандидатов в именные группы, но не каждый кандидат в нашем понимание может стать именной группой, особенно если речь идёт о группах с одним или двумя иероглифами. Как было упомянуто выше, если кандидат состоит из одного иероглифа, и тег этого иероглифа не обозначает географического места, как например, города, реки, озёра, страны, то данный кандидат не включается в именную группу.

2.6 Процесс фильтрации кандидатов в именные группы

После обработки текста программа сформировала список из шестидесяти восьми кандидатов в именные группы, но не каждый кандидат в нашем понимание может стать именной группой, особенно если речь идёт о группах с одним или двумя иероглифами. Было принято решение не включать в именную группу тех кандидатов, которые состоят из одного и из двух иероглифов. В последнем случае, условие выполняется, если тег кандидата не сигнализирует о том, что это имя собственное, а если быть точными, то имя собственное для географических объектов. Например, удовлетворяющий

критерию отбора тег – «中国_ns», не удовлетворяющие критерию отбора теги – «地区_n», «低_a», «那_r» и «着_u» (Рисунок 6).

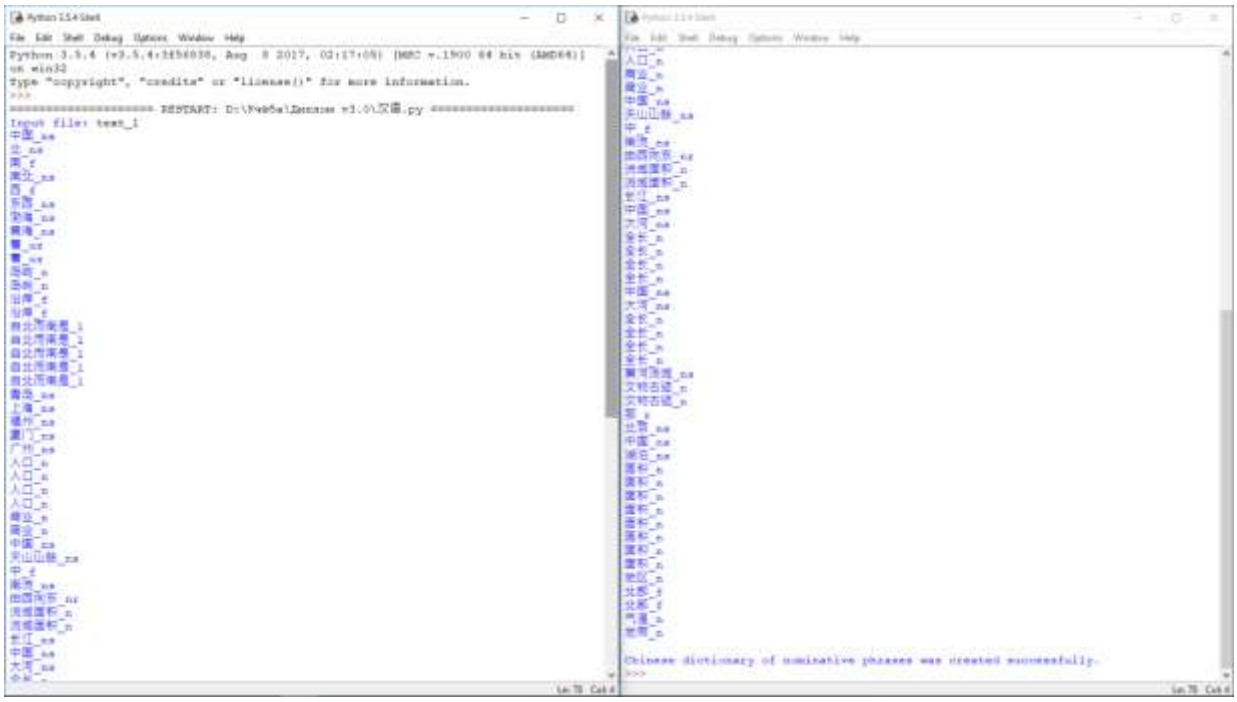


Рисунок 6 – Кандидаты в именную группу, состоящие из одного иероглифа

Затем каждый кандидат, если он состоит более чем из одного иероглифа, проверяется на наличие стоп-слов в начале и в конце. Например, из кандидата «向_p 南流_ns 以外_c» из начала и из конца будут удалены иероглифы с тегами, обозначающими предлог «向_p» и союз «以外_c». В итоге мы получаем именную группу из одного иероглифа, «南流_ns», который удовлетворяет первому критерию отбора – тег иероглифа соответствует тегу имени собственному для географических объектов (Рисунок 7).

Рисунок 7 – Кандидаты в именную группу, обозначающие географические объекты

Лингвистические правила, которые мы создали и использовали в процессе фильтрации кандидатов в именную группу:

1. Если у кандидата вначале стоит иероглиф, если быть точными то, тег с которого не может начинаться именная группа, необходимо удалить этот иероглиф.

2. Если у кандидата в конце стоит иероглиф, если быть точными то, тег на который не может заканчиваться именная группа, необходимо удалить этот иероглиф.

3. Если кандидат после отсечения иероглифов состоит из одного иероглиф и тег этого иероглифа не равен тегу «ns», тег, который обозначает географические объекты, тогда необходимо удалить кандидата.

Итог фильтрации – список именных групп входного текста на китайском языке. После окончания фильтрации, результат сохраняется в файл с названием 中文字典 [Zhōngwén zìdiǎn], что в переводе на русский язык означает словарь китайского языка (Рисунок 8).

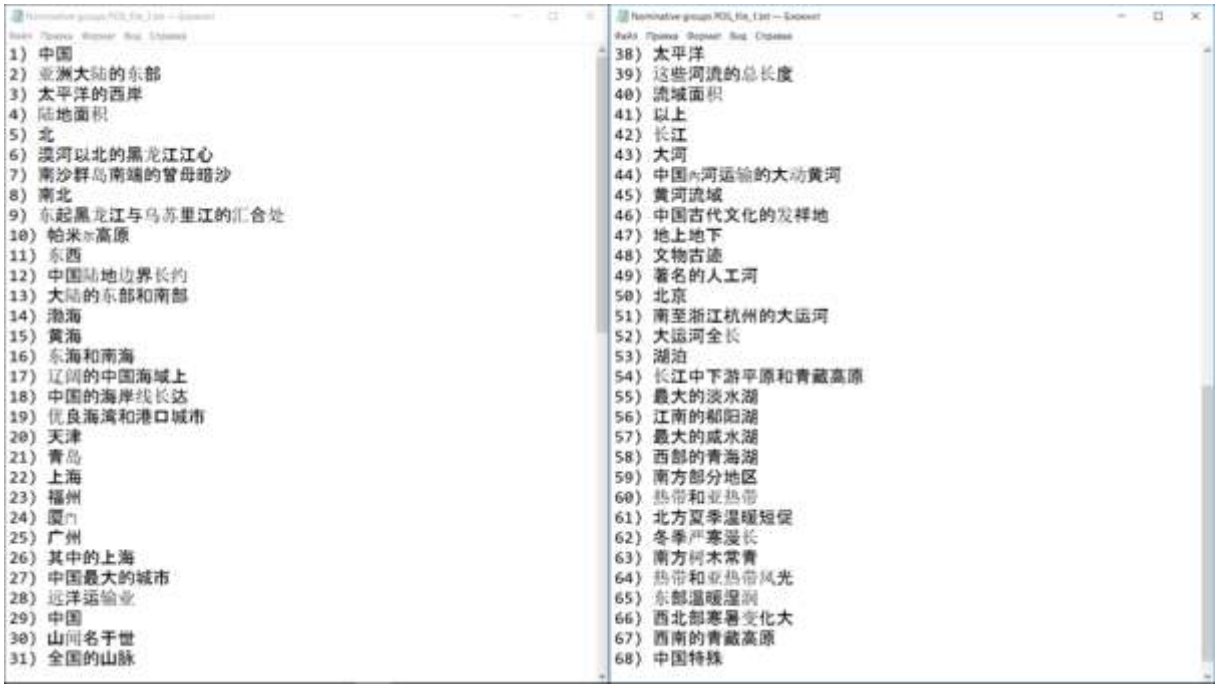


Рисунок 8 – Словарь именных групп

Вывод по главе 2

На основании материала, изложенного нами во второй главе дипломной работы можно сделать вывод о проделанной работе. Использование утилит, созданных лингвистами прикладниками, облегчает работу с языком, как филологам, лингвистам, так и самим прикладным лингвистам, что было ярко показано в нашей работе.

Использование сторонней программы SegmentAnt помогло нам в достижении поставленной цели и облегчило ряд задач.

Как мы упоминали раньше, формальный алгоритм построения программы, который мы предлагаем, позволит использовать данную утилиту как основу для программ более узкой направленности в обработке текстов на китайском языке. Это представляется возможным благодаря тому, что алгоритм, предложенный нами в дипломной работе, может быть использован на текстах с другой тематикой, т.к. работа утилиты будет осуществляется эффективно при наличии программы для определения частей речи, покрывающей большую часть иероглифов китайского языка.

ЗАКЛЮЧЕНИЕ

Автоматическое извлечение информации из текстов на естественном языке будет является одной из важных проблем в области автоматической обработки естественного языка, решение которой позволит повысить эффективность использования информационных ресурсов, хранящихся в виде электронных текстовых документов [21].

Компьютерная лексикография представлена совокупностью методов и программных средств обработки текстовой информации для создания словарей [17]. В рамках компьютерной лексикографии разрабатываются компьютерные технологии составления и эксплуатации словарей. Специальные программы, базы данных, компьютерные картотеки, программы обработки текста, позволяют в автоматическом режиме формировать словарные статьи, хранить словарную информацию и обрабатывать её [35]. Так наша программа не стала исключением. Мы уверенны, что формальный алгоритм построения программы, который мы предлагаем, позволит использовать данную утилиту как основу для программ более узкой направленности в обработке текстов на китайском языке. Это представляется возможным благодаря тому, что алгоритм, предложенный нами в дипломной работе, может быть использован на текстах с другой тематикой, т.к. работа утилиты будет осуществляется эффективно при наличии программы для определения частей речи, покрывающей большую часть иероглифов китайского языка.

Возвращаясь к цели дипломной работы и её задачам, можно сказать, что ***цель исследования*** – написание программы для автоматического создания словаря именных групп китайского языка – была достигнута.

Задачи, поставленные для достижения цели дипломной работы (сбор и изучение теоретических знаний по лексикографии, компьютерной лексикографии, именным группам и автоматическому извлечению информации из текстов на естественном языке; создание списка текстов из учебного пособия по страноведению Китая; определить, что есть именная

группа в китайском языке; найти метод для извлечения именных групп из текста на китайском языке; создать программы на языке программирования python по извлечению именных групп и автоматическому формированию словаря предметной области), были выполнены.

Полнота извлечения именных групп программой была проверена сопоставлением двух списков именных групп: первый список был составлен нами вручную, а второй список был составлен программой. Результата этого сопоставления показывает, что именная группа находится в тексте с точностью 80%, но это при учёте того, что программы была нацелена на тексты с лингвострановедческой тематикой.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Агапова, Н.А. О принципах создания электронного словаря лингвокультурологического типа: к постановке проблемы [Текст] / Н.А. Агапова, Н.Ф. Картофелева // Вестн. Том. Гос. Ун-та. – 2014. – № 386. – С. 6–10.

2. Бабина, О.И. Извлечение именных групп из корпуса текстов на испанском языке [Текст] / О.И. Бабина, Т.Ю Мыларщикова // Вестник ЮУрГУ. Серия: лингвистика. – 2011. – № 22. – С. 47–53.

3. Бессмертный, И.А. Статистический метод извлечения терминов из китайских текстов без сегментации фраз [Текст] / И.А. Бессмертный, Юй Чуцяо Ма Пенюй // Научно-технический вестник информационных технологий, механики и оптики. – 2016. – № 6. – С. 1096–1102.

4. Беляева, Л.Н. Потенциал автоматизированной лексикографии и прикладная лингвистика [Текст] / Л.Н. Беляева // Известия РГПУ им. А.И. Герцена. – 2010. – № 134. – С. 186–216.

5. Виноградов, В.В. Основные типы лексических значений слова, «Вопросы языкознания» [Текст] / В.В. Виноградов. – М.: Просвещение, 1953. – 125 с.

6. Горелов, В.И. Теоретическая грамматика китайского языка: учеб. Пособие для студентов пед. ин-тов по спец. «Иностр. яз.» [Текст] / В.И. Горелов. – М.: Просвещение, 1989. – 318 с.

7. Демина, Н.А. Страноведение: учебное пособие [Текст] / Н.А. Демина, Чжу Канцзи. – М.: Вост. Лит., 2004. – 351 с.

8. Дубичинский, В.В. Лексикография русского языка: учеб. пособие [Текст] / В.В. Дубчинский. – М.: Наука: Флинта, 2008. – 432 с.

9. Конкатенация [Электронный ресурс]. – Режим доступа: https://ru.wikipedia.org/wiki/Конкатенация, свободный. – Загл. с экрана. – (Дата обращения: 25.03.2018).

10. Лексический разбор слова [Электронный ресурс]. – Режим доступа: http://megabook.ru/article/лексический%20разбор%20слова, свободный – Загл. с экрана. – (Дата обращения: 25.03.2018).

11. Мельчук, И.А. Опыт теории лингвистических моделей «Смысл ⇔ Текст» [Текст] / И.А. Мельчук. – М.: Наука, 1974. – 314 с.

12. Морковкин, В.В. О всеохватном лексикографическом представлении лексического ядра русского языка [Текст] / В.В. Морковкин // Вестн. Том. гос. ун-та. Филология. – 2011. – № 3 – С. 129–135.

13. Нелюбин, Л.Л. Перевод и прикладная лингвистика. [Текст] / Л.Л. Нелюбин. – М.: Высшая школа, 1983. – 208 с.

13. Пустошило, Е.П. Лексикология. Фразеолгоия. Лексикография: учебно-методический комплекс по русскому языку для студентов педагогических специальностей [Текст] / Е.П. Пустошило. – Гродно: ГРГУ им. Я. Купалы, 2011. – 181 с.

14. Сегментация в лингвистике [Электронный ресурс]. – Режим доступа: https://ru.wikipedia.org/wiki/Сегментация_(лингвистика), свободный. – Загл. с экрана. – (Дата обращения: 20.03.2018).

15. Сороколетов, Ф.П. История русской лексикографии [Текст] / Ф.П. Сороколетов. – СПб.: Наука, 2001. – 616 с.

16. Ступин, Л.П. Лексикография английского языка: учеб. пособие [Текст] / Л.П. Ступин. – М.: Высшая школа, 1985. – 36 с.

17. Стоп-слова [Электронный ресурс]. – Режим доступа: http://datalytics.ru/all/spisok-stop-slov-yandeks-direkta, свободный. – Загл. с экрана. – (Дата обращения: 09.02.2018).

18. Стройков, С.А. Основные понятия лингвистической концепции электронного лексикографического гипертекста [Текст] / С.А. Стройков // Известия самарского научного центра ран. 2010. – С. 808–811.

19. Список тегов для китайского языка [Электронный ресурс]. – Режим доступа: https://www.ltp-cloud.com/intro/en/#cws_how, свободный. – Загл. с экрана. – (Дата обращения: 25.03.2018).

20. Словарь Ожегова [Электронный ресурс]. – Режим доступа: https://slovarozhegova.ru/, свободный. – Загл. с экрана. – (Дата обращения: 25.03.2018).

21. Типы словарей [Электронный ресурс]. – Режим доступа: http://rus-gos.spbu.ru/index.php/dictionary/type, свободный – Загл. с экрана. – (Дата обращения: 25.03.2018).

22. Чепик, Е.Ю. Компьютерная лексикография как одно из направлений современной прикладной лингвистики [Текст] / Е.Ю. Чепик // Ученые записки таврического национального университета им В.И. Вернадского. – 2006. – № 3-4. – С. 274–279.

23. Шереметьева, С.О. Интерактивное реферирование, ориентированное на машинный перевод [Текст] / С.О. Шереметьева // Вестник ЮУрГУ. Серия: Лингвистика. – 2014. – № 1. – С. 89–92.

24. Шереметьева, С.О. К вопросу об электронных ресурсах профессиональной лексики [Текст] / С.О. Шереметьева, П.Г. Осминин, Е.С. Щербаков // Вестник ЮУрГУ. Серия: Лингвистика. – 2014. – №1. – С. 57 – 63.

25. Шереметьева, С.О. Методы и модели автоматического извлечения ключевых слов [Текст] / С.О. Шереметьева, П.Г. Осминин, // Вестник ЮУрГУ. Серия: Лингвистика. – 2015. – №1. С. 76 – 81.

26. Шмелев, Д.Н. Проблемы семантического анализа лексики [Текст] / Д.Н. Шмелев. – М.: Комкнига, 2006. – 149 с.

27. Щерба, Л.В. Опыт общей теории лексикографии языковая система и речевая деятельность [Текст] / Л.В. Щерба. Л.: Наука, 1974. – 428 с.

28. Щерба, Л.В. Языковая система и речевая деятельность [Текст] / Л.В Щерба. – М.: Наука, 1974. – 304 с.

29. Щитова, О.Г. Лексикографические источники изучения функциональной эквивалентности иноязычных новаций в русском языке начала XXI в [Текст] / О.Г. Щитова // Вестн. Том. Гос. Ун-та, 2012. – №355. – С. 113–118.

30. Федосов, Ю.В. Электронный научно-образовательный журнал / Ю.В. Федосов – Москва: ВГПУ «Грани познания», 2010. – С. 93–98.

31. Филиппович, Ю.Ф. Историческая компьютерная лексикография - terra incognita в компьютерном мире / Ю.Ф. Филиппович, М.Т. Чернышева. – М.:Гнозис, 1999. – С. 56–67.

32. Abbyy [Электронный ресурс]. – Режим доступа: https://www.abbyy.com/ru-ru/science/technologies/lexicography, свободный. – Загл. с экрана. – (Дата обращения: 20.03.2018).

33. Antconc [Электронный ресурс]. – Режим доступа: http://www.laurenceanthony.net/software.html, свободный. – Загл. с экрана. – (Дата обращения: 20.03.2018).

34. Hilary, N. Electronic dictionaries in second language vocabulary comprehension and acquisition: the state of the art [Текст] / N. Hilary. – Stuttgart, Germany, 2000. – P. 839–847.

35. Huang, T. Computer Vision: Evolution and Promise [Текст] / T. Huang // University of Illinois at Urbana-Champaign, 1996. – P. 225–228.

36. Joachims, T. Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms [Текст] / T. Joachims // Kluwer Academic Publishers, 2002. – P. 199–205.

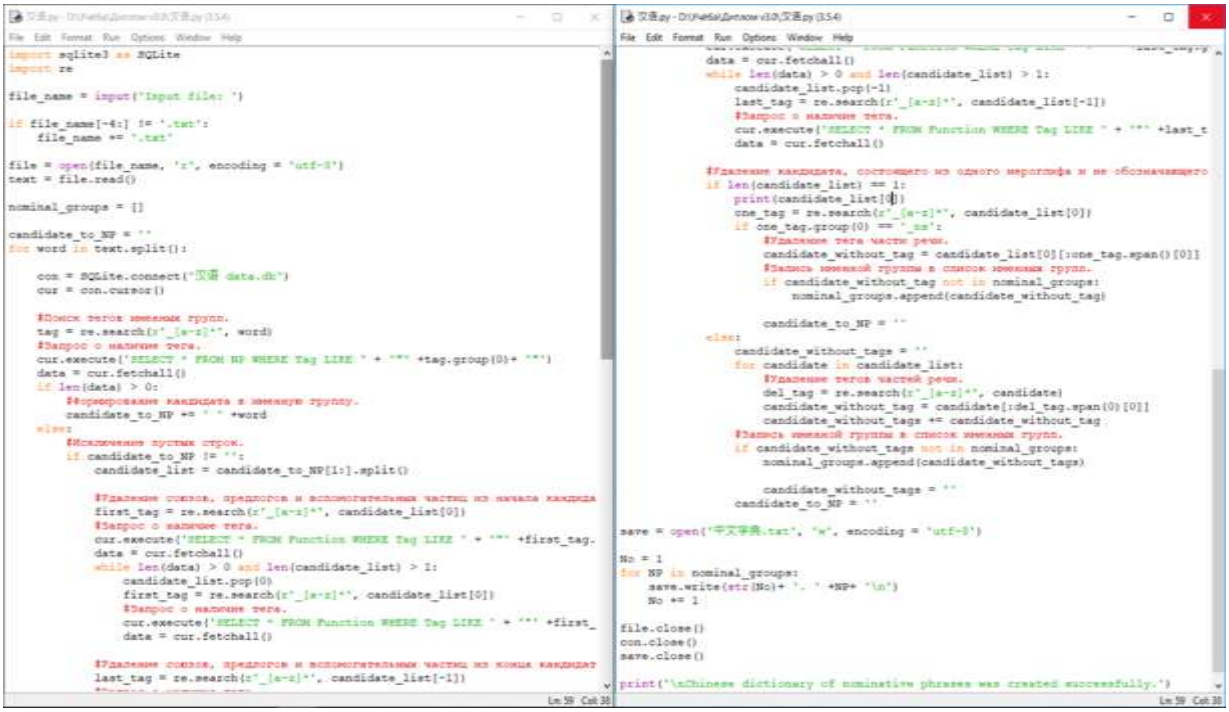
37. POS tagging [Электронный ресурс]. – Режим доступа: https://ru.wikipedia.org/wiki/Частеречная_разметка, свободный – Загл. с экрана. – (Дата обращения: 25.03.2018).

38. Sheremetyeva, S. On extracting multiword np terminology for MT [Текст] / S. Sheremetyeva // Proceedings of the thirteen conference of european association of machine translation (EAMT-2009). – Barcelona, Spain, 2009. – P. 205–212.

39. Wen, Z. A Study on Multi-Word Extraction from Chinese Documents [Текст] / Wen Zhang, Taketoshi Yoshida, Xijin Tang // School of Knowledge Science, Japan Advanced Institute of Science and Technology. – Japan, 2017. – P. 42–53.

ПРИЛОЖЕНИЕ 1

Код программы на языке программирования python по извлечению именных групп из лингвострановедческих текстов на китайском языке и формированию словаря.



```
import sqlite3 as SQLite
import re

file_name = input('Input file: ')

if file_name[-4:] != '.txt':
    file_name += '.txt'

file = open(file_name, 'r', encoding = 'utf-8')
text = file.read()

nominal_groups = []

candidate_to_NP = ''
for word in text.split():

    con = SQLite.connect("汉语 data.db")
    cur = con.cursor()

    #Поиск тегов именных групп.
    tag = re.search(r'_[a-z]*', word)
    #Запрос о наличие тега.
    cur.execute('SELECT * FROM NP WHERE Tag LIKE ' + '"' +tag.group(0)+ '"')
    data = cur.fetchall()
    if len(data) > 0:
        #Формирование кандидата в именную группу.
        candidate_to_NP += ' ' +word
    else:
        #Исключение пустых строк.
        if candidate_to_NP != '':
            candidate_list = candidate_to_NP[1:].split()

            #Удаление союзов, предлогов и вспомогательных частиц из начала кандида
            first_tag = re.search(r'_[a-z]*', candidate_list[0])
            #Запрос о наличие тега.
            cur.execute('SELECT * FROM Function WHERE Tag LIKE ' + '"' +first_tag.
            data = cur.fetchall()
            while len(data) > 0 and len(candidate_list) > 1:
                candidate_list.pop(0)
                first_tag = re.search(r'_[a-z]*', candidate_list[0])
                #Запрос о наличие тега.
                cur.execute('SELECT * FROM Function WHERE Tag LIKE ' + '"' +first_
                data = cur.fetchall()

            #Удаление союзов, предлогов и вспомогательных частиц из конца кандидат
            last_tag = re.search(r'_[a-z]*', candidate_list[-1])
```

```
            data = cur.fetchall()
            while len(data) > 0 and len(candidate_list) > 1:
                candidate_list.pop(-1)
                last_tag = re.search(r'_[a-z]*', candidate_list[-1])
                #Запрос о наличие тега.
                cur.execute('SELECT * FROM Function WHERE Tag LIKE ' + '"' +last_t
                data = cur.fetchall()

            #Удаление кандидата, состоящего из одного иероглифа и не обозначающего
            if len(candidate_list) == 1:
                print(candidate_list[0])
                one_tag = re.search(r'_[a-z]*', candidate_list[0])
                if one_tag.group(0) == '_nr':
                    #Удаление тега части речи.
                    candidate_without_tag = candidate_list[0][:one_tag.span()[0]]
                    #Запись именной группы в список именных групп.
                    if candidate_without_tag not in nominal_groups:
                        nominal_groups.append(candidate_without_tag)

                candidate_to_NP = ''
            else:
                candidate_without_tags = ''
                for candidate in candidate_list:
                    #Удаление тегов частей речи.
                    del_tag = re.search(r'_[a-z]*', candidate)
                    candidate_without_tag = candidate[:del_tag.span(0)[0]]
                    candidate_without_tags += candidate_without_tag
                #Запись именной группы в список именных групп.
                if candidate_without_tags not in nominal_groups:
                    nominal_groups.append(candidate_without_tags)

                candidate_without_tags = ''
                candidate_to_NP = ''

save = open('中文字典.txt', 'w', encoding = 'utf-8')

No = 1
for NP in nominal_groups:
    save.write(str(No)+ '. ' +NP+ '\n')
    No += 1

file.close()
con.close()
save.close()

print('\nChinese dictionary of nominative phrases was created successfully.')
```

ПРИЛОЖЕНИЕ 2

Список тегов частей речи для программы SegmentAnt, которые могут быть использованы в именной группе:

n – general noun

f – world sides noun

ns – other proper noun

u – adjective

a – adjective

nr – geographic noun

j – abbreviation

p – preposition

r – pronoun

ПРИЛОЖЕНИЕ 3

Список именных группы, извлечённых вручную, для первого текста, пример работы с которым представлен в дипломной работе.

中国, 亚洲大陆的东部, 太平洋的西岸, 陆地面积, 漠河以北的黑龙江江心, 南北, 西帕米尔, 原, 东西, 中国陆地边界长约, 大陆的东部和南部, 渤海, 黄海, 东海和南海, 辽阔的中国海域上, 岛屿中国的海岸线长达, 沿岸优良海湾和港口城市, 自北而南是天津, 青岛, 上海, 福州, 厦门, 广州, 其中的上海, 人口工业, 商业金融业, 远洋运输业, 山闻名于世, 全国的山脉, 天山山脉, 贺兰山山脉, 东北的长白山山脉, 中的河流除西南部, 南流, 由西向东太平洋, 这些河流的总长度, 流域面积在以上, 长江, 大河, 全长流域面积, 中国内河运输的大动黄河, 黄河流域, 中国古代文化的发祥地, 著名的人工河, 北京, 南至浙江杭州的大运河, 大运河全长, 湖泊, 长江中下游平原和青藏高原, 江南的郗阳湖, 面积最大的咸水湖, 西部的青海湖, 面积中国, 地区温带, 南方部分地区, 热带和亚热带, 北部寒带, 北方夏季温暖短促, 冬季严寒漫长, 南方树木常青, 热带和亚热带风光, 东部温暖湿润, 西北部寒暑变化大, 西南的青藏高原, 气温低, 中国特殊

ПРИЛОЖЕНИЕ 4

Список именных группы, извлечённых автоматически, для первого текста, пример работы с которым представлен в дипломной работе.

中国, 亚洲大陆的东部, 太平洋的西岸, 陆地面积, 北, 漠河以北的黑龙江江心, 南南沙群岛南端的曾母暗沙, 南北, 东起黑龙江与乌苏里江的汇合处, 西帕米尔高原, 东西, 中国陆地边界长约, 大陆的东部和南部, 渤海, 黄海, 东海和南海, 辽阔的中国海域上, 岛屿中国的海岸线长达, 沿岸优良海湾和港口城市, 自北而南是天津, 青岛, 上海, 福州, 厦门, 广州, 其中的上海, 中国最大的城市, 人口工业, 商业金融业, 远洋运输业, 中国, 山闻名于世, 全国的山脉, 天山山脉, 贺兰山山脉, 东北的长白山山脉, 西北部的阿尔泰山山脉和西南地区的喜玛拉雅山山脉, 中的河流除西南部, 南流, 由西向东太平洋, 这些河流的总长度, 流域面积在以上, 长江, 大河, 全长流域面积, 中国内河运输的大动黄河, 黄河流域, 中国古代文化的发祥地, 地上地下, 文物古迹中国, 著名的人工河, 那北, 北京, 南至浙江杭州的大运河, 大运河全长, 湖泊, 长江中下游平原和青藏高原, 最大的淡水湖, 江南的郗阳湖, 面积最大的咸水湖, 西部的青海湖, 面积中国, 地区温带, 南方部分地区, 热带和亚热带, 北部寒带, 北方夏季温暖短促, 冬季严寒漫长, 南方树, 常青, 热带和亚热带风光, 东部温暖湿润, 西北部寒暑变化大, 西南的青藏高原, 气温低, 中国特殊