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 ИМЕННЫХ СЛОВОСОЧЕТАНИЙ (НА МАТЕРИАЛЕ ИСПАНСКОГО ЯЗЫКА)

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА
ЮУрГУ – 45.03.03.2018.286.ВКР

Руководитель, к.филол.н., доцент
_____ /О.И. Бабина/
«____» _____ 2018 г.

Автор
студент группы ЛМ-436
_____ /Д.Д. Сарасек /
«____» _____ 2018 г.

Нормоконтролер,
к.филол.н., доцент
__________ /О.И. Бабина/
«____» _____ 2018 г.

Работа защищена с оценкой
_____
«____» _____ 2018 г.

Челябинск
2018

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

В эпоху технического прогресса активно растет количество текстовой информации. Современные компьютерные технологии позволяют ускорить процесс её обработки, сделать его качественным и удобным для пользователя. Выделение именных словосочетаний является одной из значимых составляющих частичного анализа текста. Оно является необходимым при автоматическом выявлении фактов, анализе медицинской и технической документации, при извлечении информации об отношениях.

Актуальность данного исследования заключается в необходимости создания автоматизированных шаблонов для поиска именных словосочетаний в текстовых корпусах.

Объектом исследования выступают именные словосочетания.

Предметом исследования является автоматизация процессов извлечения именных словосочетаний из корпуса текста.

Цель настоящего исследования заключается в разработке автоматизированных поисковых шаблонов для выявления именных словосочетаний, их тестирование и анализ устойчивости в корпусе фармацевтических рецептов на испанском языке.

Для достижения поставленной цели необходимо решить следующие **задачи**:

1) Рассмотреть понятие «именное словосочетание»;

2) Выделить критерии и лингвистические особенности для составления правил;

3) Разработать модель извлечения именных словосочетаний на испанском языке;

4) Составить базу знаний, включающую шаблоны для извлечения именных словосочетаний из корпуса на испанском языке;

5) Апробировать составленные шаблоны на экспериментальном корпусе, проанализировать их точность и полноту отбора на основе составленной базы знаний.

Для решения поставленных задач были использованы следующие **методы** исследования:

– описательный с использованием приемов обобщения анализируемого материала;

– сплошной выборки;

– корпусный анализ;

– моделирование;

– инструментальный;

– экспериментальный.

Теоретико-методологической базой исследования послужили работы отечественных и зарубежных лингвистов, педагогов, посвященных:

– проблеме определения термина «словосочетания» (В.Н. Ярцевой, Ф.Ф. Фортунатова, В.В. Виноградова, Н.С. Валгина, Д.Э. Розенталь, Н.Ю. Шведова);

– компьютерной лингвистике и анализу текста (Б.Ю.Городецкий, В.Ю. Захаров, Э.С. Клышинский).

Достоверность и обоснованность результатов исследования обеспечивается:

– использованием адекватных методов исследования;

– результатами эксперемента.

Научная новизна исследования обусловлена тем, что в нем разработаны шаблоны для автоматизированного поиска именных словосочетаний в испанском языке на базе операционной системы UNIX и утилиты Grep, а также предложен метод последующего анализа

извлеченных конструкций для улучшения точности выявления именных словосочетаний в постобработанном материале на основе корпуса.

Теоретическая значимость исследования заключается в том, что модель расширяет теорию прикладной лингвистики в области извлечения информации из текста при проведении корпусных исследований.

Практическая ценность исследования заключается в том, что разработанная база знаний может быть использована при создании реальных систем по обработке текста; представленная модель найдет применение при автоматизация извлечения именных словосочетаний для процедуры составления специализированных словарей и баз данных; результаты могут быть применены при обучении студентов в таких областях как корпусная и прикладная лингвистика.

Апробация и внедрение результатов работы проводилось на экспериментальном корпусе.

Цель и задачи исследования определили его **структуру и объем**. Данная работа состоит из введения, двух глав, заключения библиографического списка и 4 приложений.

Во **введении** дается обоснование актуальности и выбора темы исследования, определяются объект, предмет, цель, задачи и методы исследования, а также его научная новизна, теоретическая и практическая значимость; формулируются основные положения, выносимые на защиту.

Основная часть исследования, представленная двумя главами, посвящена последовательному решению поставленных задач.

Первая глава состоит из двух разделов, включающих в себя девять подразделов и посвящена трактовке понятий «слова» и «словосочетание», словообразовательным моделям в испанском языке, семантическому единству и сочетаемости двух слов.

В результате рассмотрения теоретических основ тестирования лингвистически ориентированных электронных учебных ресурсов в

выводах по первой главе сформулированы теоретические принципы, положенные в основу исследования.

Во **второй главе** приведены правила и поисковые шаблоны на инструменте Grep для выявления именных словосочетаний в испанском фармацевтическом корпусе, описан механизм работы составленной базы знаний, проведен анализ устойчивости выявленных именных словосочетаний и предложен метод их постобработки.

В заключении подводятся основные итоги проведенного исследования, формулируются общие выводы, намечаются перспективы дальнейшего исследования в этой области.

Библиографический список представлен 50 наименованиями.

В качестве приложений включена лексическая база знаний использованная в эксперименте, таблица с регулярными выражениями, а также графически представленные правила алгоритмов для построения поисковых шаблонов в терминах регулярных выражений.

ГЛАВА 1 ИМЕННОЕ СЛОВОСОЧЕТАНИЕ В ИСПАНСКОМ ЯЗЫКЕ

1.1 Слово

1.1.1 Понятие «Слово»

Во все времена перед лингвистами стояла проблема формального определения понятия «слово», что порождало дискуссии и неоднозначные взгляды на его природу. Л.В.Щерба писал: «В самом деле, что такое «слово»? Мне думается, что в разных языках это будет по-разному. Из этого, собственно, следует, что понятия «слово» вообще не существует».

Ю.С. Маслов, считает что «слово», является неопределенной единицей, как с точки зрения структурного аспекта и формальных признаков, так и с точки зрения смыслового содержания, как в пределах одного языка, так и при проведении сравнительного анализа разных естественных языков [1]. В.В. Виноградов, также указывал на недостаток прочных теоретических основ в современной грамматике, в отсутствии определения или точного описания основных грамматических понятий, особенно понятий слова и предложения [2].

В различные исторические эпохи, определение понятия «слова» менялось. Так, в 19 веке, А.Г. Нурен определял слово как: «независимая морфема (un morphème indépendant), которую наше языковое чутье воспринимает как целое по звуку и значению, так что она или ощущается неразложимой на более мелкие морфемы (например, здесь, почти, там), или – в случае, если это можно сделать, – она воспринимается независимо от значения этих более мелких, составляющих ее морфем» [3]. Для Э. Сепира слово есть один из мельчайших вполне самодовлеющих кусочков изолированного «смысла», к которому сводится предложение» [4].

Б.Т. Ганеев определяет слово, как минимально значащую единицу языка, которая может быть предложением или членом предложения [5].

В словаре В.Н. Ярцевой, понятие слово, трактуется как – основная структурно-семантическая единица языка, служащая для именования предметов и их свойств, явлений, отношений действительности, обладающая совокупностью семантических, фонетических и грамматических признаков, специфичных для каждого языка [6].

Т.Ф. Ефремова трактует слово, как единицу речи, представляющую собою звуковое выражение отдельного предмета [7]. Название понятия в отличие от самого понятия. По Д.Н. Ушакову, *слово* – единица речи, представляющая сою звуковое выражение отдельного предмета мысли [8].

Трудность в определении понятия «слово», сподвигла многих лингвистов отказаться от введения данного понятия и рассматривать вместо него термины с условным более узким значением (напр.: словема, лексема, вокабула, словоформа, лексико-семантический вариант).

А.И. Смирницкий писал: «В одних языках... слова выделяются более или менее четкими фонетическими признаками (ударение, сингармонизм, законы конца слова и пр.); в других, напротив, фонетические признаки слова совпадают с тем, что мы находим у других образований (например, у морфем или, напротив, целых словосочетаний). Все многообразие особенностей отдельных языков может, однако, нисколько не препятствовать определению «слова вообще», поскольку в этом многообразии выделяются и общие черты, выступающие как наиболее существенные признаки слова, при всех возможных отклонениях от типичных случаев» [9].

В.Н. Ярцева выделяет следующие характерные признаки для вычленения слова:

1) Цельность (воспроизводимость слова в его фонетическом и морфологическом единстве, типологическая черта флективных языков; цельность флективного слова поддерживается единым ударением,

фузионной связью основы и аффикса, несамостоятельностью основы и фонетическим отличием аффиксов от служебных слов);

2) Выделимость (наличие морфологического оформления);

3) Свободная воспроизводимость в речи [10].

В истории языкознания было выдвинуто свыше семидесяти различных критериев определения слова, в их основе лежали фонетические, грамматические, структурные, фонетические, синтаксические, семантические и системные принципы [11].

В представлении американского философа Ч.У. Морриса, значение слова составляют три базовых компонента (прагматический, семантический, синтаксический), каждый из которых специфичен и обладает неразрывной связью с другими.

Н.М. Шанский, считает, что в определении слова стоит отразить, его наиболее существенные признаки. По его мнению, основными признаками слова как лингвистической единицы в целом являются: фонетическая оформленность, семантическая валентность, непроницаемость, недвуударность, лексико-грамматическая отнесенность, постоянство звучания и значения, воспроизводимость, цельность и единнооформленность, преимущественное употребление в сочетаниях слов, изолируемость, номинативность, фразеологичность [12].

В.В. Виноградов, отмечает, что существуют слова, которые являются только морфемами, и морфемы, которые иногда являются словами. Слово может выражать и единичное понятие, конкретное, абстрактное, и общую идею отношения (напр.: предлоги *от*, *об* или союз *и*), и законченную мысль (например, афоризм Козьмы Пруткова: *«Бди!»*) [13]. Однако В.В. Виноградов подмечает глубокую разницу между словами и морфемами, так как лишь слово свободно перемещаться в пределах предложения, а морфемы, входящие в состав слова – неподвижны.

1.1.2 Лексическое и грамматическое значения слова

Содержательная (внутренняя) сторона слова представляет собой сложный, многогранный феномен. Традиционно в языкознании выделяется два значения слова: лексическое и грамматическое. В истории языка грамматическое и лексическое значение органически связаны и подвержены влиянию друг друга. Изучение грамматического строя языка не возможно без учета взаимодействия грамматических и лексических значений.

В словаре В.Н. Ярцевой под *лексическим значением* понимается содержание слова, отображающее в сознании и закрепляющее в нём представление о предмете, свойстве, процесс, явлений, а под *грамматическим значением* – обобщённое, отвлечённое языковое значение, присущее ряду слов, словоформ, синтаксических конструкций и находящее в языке своё регулярное (стандартное) выражение [14]. Грамматические значения слова характеризуются своей не универсальностью и образуют четкий структурированный класс. А.А.Зализняк, понимал под *грамматическим значением* – значение, выражение которого обязательно для всех словоформ данного класса лексем [15].

Д.Н. Шмелев, пишет, что собственное лексическое значение слова определяется не только его непосредственным «предметным содержанием», но и семантической соотнесенностью с рядом других слов [16]. Лексическое значение не изменяется во всех грамматических формах слова, в том числе и аналитических. Оно принадлежит к не определенной словоформе, а лексеме в целом. Исследованием природы лексического значения слова занимаются лексическая семасиология и лексикология.

В лингвистике нет четкого взгляда на определение лексического значения у служебных слов. Ю.С. Маслов считает, что служебные слова функционируют в предложении как выразители тех или иных грамматических значений отдельных слов и тех или иных смысловых и формальных связей между словами; грамматическое значение в их содержании представляется основным, если не единственным в своем роде.

И.А. Стернин выделяет лексическое значение (закрепленное словом отражения языковой реальности) и структурное-языковое значение (информация о признаках слова как функциональной единицы языка, то есть отражение в значении языковой действительности) [17].

Д.Н. Шмелев в лексическом значении выделяет *денотативный макрокомпонент* (основной компонент, указывающий на свойства, признаки предмета номинации; передает коммуникативно значимую информацию) и *коннотативный макрокомпонент* (выражает эмоционально-оценочное отношение говорящего к денотату слова, несет дополнительную информацию) [18].

Концептуальное значение слова отображает денотат (класс денотатов). В состав лексического значения слова входят ядро и коннотации (эмоциональные, стилистические, экспрессивные добавочные элементы), придающие свойственную слову эмоциональную окраску.

Отнесенность лексического значения характеризуют:

1) предметная отнесенность слова (отношение к денотату);

2) понятийная отнесенность (отношение к категориям логики);

3) значимость (отношение к концептуальным и коннотативным значениям других слов).

1.1.3 Двусторонняя сущность слова

Слово является сложной двусторонней единицей языка. Ему присуще план выражения (форма) и план содержания (значение) (Рисунок 1.1).

Ю.С. Маслов определяет *план выражения* как звуковую материальную сторону слова, воспринимаемую слухом (на письме – буквенное обозначение), а *план содержания* как заключенную в слове мысль, передающую ту или иную информацию и те или иные сопровождающие эту информацию эмоциональные моменты.

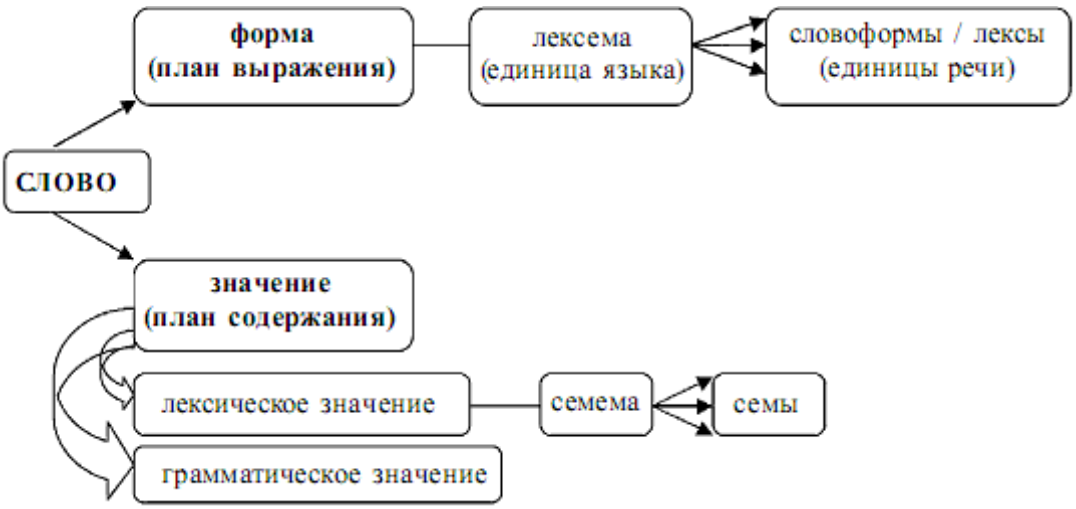


Рисунок 1.1 – Схема представления слова как двусторонней единицы языка

Многие лингвисты именуют план выражения – лексемой (абстрактной единицей, представляющей слово в совокупности всех его форм, и значений). В речи лексема воспроизводится в определенных словоформах или лексах (единицы речи). Лексемы могут быть представлены, как в одной словоформе (кофе, визави), так и в нескольких (стол, столом, стулу и т.д.).

План содержания слова составляют его лексическое и грамматическое значение. Лексическому значению свойственно выражение в слове того или иного явления действительности, а также конкретность и индивидуальность. Грамматическое значение определяет принадлежность слова к определенной части речи. На уровне лингвистического анализа план содержания слова также именуют «семемой» (компоненты значения слова). Семема – высшая единица плана содержания, которая включает в себя саму сему или их совокупность (напр.: «соседка» состоит из сем: человек + женский пол + живущий по соседству).

В семантических полях встречается два вида сем: интегрирующие (способные объединяться в одну группу) и дифференциальные семы (отличающие члены семантического поля друг от друга).

Грамматическое значение отражается морфологически.

Рисунок 1.2 – Схема описания структуры слова стул

1.1.4 Мотивация слова

Каждому предмету или явлению в реальности присущи некие признаки, но при именовании выбирается лишь один из них – самый заметный и не обязательно существенный, который в дальнейшем и представляет его в целом (напр. кукушка – по характерному крику; el corte (порез) – по длине раны). *Мотивация слова* – выражение в слове одного или нескольких признаков называемого предмета, используемое в качестве названия данного предмета в целом (Ю.С. Маслов). Е.А. Земская под *мотивацией* понимает отношения между морфологической и фонетической структурой слова с одной стороны, и его значением с другой.

Выделяют три типа мотивации:

1) *Фонетическая мотивация* имеет место при естественной связи между значением слова и его звучанием (cuckoo, splash, purr, buzz, bubble);

2) *Морфологическая мотивация* строится по имеющимся в естественном языке моделям из существующих морфем (extranumerario: extra – сверх, numerar – штат, io – признак (сверхштатный); superabundancia: super – лишние, abundancia – обилие (изобилие)). Морфологически мотивированное

слово сформировано из компонентов, несущих в себе определенное значение. Проведя морфологический анализ испанского слова electroencefalografista, можно выделить в нём следующие компоненты: electro (корень) + encefalo (корень) + graf (корень) + ista (суффикс имени существительного);

3) *Семантическая мотивация* опирается на связи между первичным и вторичным значением слова, изменяются значение и функция слова (sable – соболь (зверь), sable – соболь (мех), azafata – служанка королевы, помощница, azafata – стюардесса). Слова, основанные на семантической мотивации, часто носят образный характер.

Глубина мотивации может быть различной: ярко выраженной (белошапка, старорусский), стертой (волейбол, жимолость), утраченной (дом, парашют).

1.1.5 Словообразование в испанском языке

В лингвистическом энциклопедическом словаре Виктории Николаевны Ярцевой, понятие «словообразование» имеет двойственный характер, и трактуется как:

1) Образование слов, называемых производными и сложными, обычно на базе однокорневых слов по существующим в языке образцам и моделям с помощью аффиксации, словосложения, конверсии и других формальных средств;

2) Раздел языкознания, изучающий все аспекты создания, функционирования, строения и классификации производных и сложных слов.

Словообразование, обеспечивая процесс номинации и его результаты, играет важную роль в классификационно–познавательной деятельности человека и выступает как одно из основных средств пополнения словарного состава языка, а также установления связей между отдельными частями речи.

Морфологический способ словообразования, подразумевающий в себе деривацию (с использованием средств аффиксации – суффиксации и префиксации) и словосложение, является ведущим способом словообразования в испанском языке.

Деривационный способ словообразования включает в себя:

1) Суффиксацию:

1.1) *Именная.* Производство имен существительных и прилагательных посредством суффиксации (пр.: ganch**uelo**, perr**una**, polvor**íenta**). В испанском языке именная деривационная модель доминирует над глагольной и крайне разнообразна.

1.2) *Глагольная* (пр.: act**uar**, tard**ecer,** diplom**ar**). Является ограниченной. Включает в себя не только суффиксы, но и инфинитивные флексии (–ar, –er, –ir).

2) Префиксацию (пр.: **ante**sala, **sin**sabor, **des**hacer). В испанском языке насчитывается 12 наиболее продуктивных префиксов сходных по своей форме с предлогами (предложные префиксы). Остальные же имеют иноязычное происхождение, в основном греческое и латинское. Префиксация распространяется на все части речи и не имеет ограничения в употреблении. Испанский характеризуется активным употреблением префиксойдов (**tele**puente, **mini**televisor, **mini**bolsa) [19].

Приоритетными для испанского языка считаются аффиксальные деривации, в которых происходит добавление аффиксов к лексеме (*in–constante* «непостоянный», *transporta–ble* «переносной»). Не аффиксальные деривации встречаются реже, но всё же присущи данному типу языка. Их формирование происходит на основе опущения некоторых морфологических составляющих лексемы (*retén* «стоп», «застава» образовано от *retener* «удерживать»). Похожим на не аффиксальную деривацию, можно назвать морфологический процесс, в котором глагольная лексема превращается в существительное лишь при добавлении безударной гласной к корню слова (конверсия). Эта «добавочная» гласная

может совпадать с сильной гласной (например как в словах guard–a–r «охранять» – guard–a «охранник») или наоборот изменяться на родственный гласный звук: atrac–a–r «атаковать» – в atrac–o «грабеж»; empuj–a–r «нажимать» – empuj–e «давление». В некоторых случаях одинаковый отглагольный корень может выстраивать парадигму с тремя различными типами гласных: costar «стоить» – cost–e, cost–o, cost–a(s) «стоимость». Во всех испанских диалектах наиболее распространен аффиксальный способ деривации.

При ***словосложении*** в испанском языке принимают участие различные части речи. Сочетание основ происходит 3 способами:

1) При образовании сложного слова, слова не изменяются: madre + selva = madreselva (жимолость), sordo + mudo = sordomudo (глухонемой);

2) Если основа второго слова начинается с гласной – первое слово не теряет конечную гласную: tela + araña = telaraña – паутина;

3) Последняя буква первой основы заменяется на i (в таких случаях часто первый компонент оканчивается на гласную, а второй начинается с согласной): verde + negro = verdinegro – темно-зеленый; pelo + corto = pelicorto – короткошерстный.

В образовании сложных слов, состоящих из двух основ, могут участвовать следующие части речи:

1) Имя существительное + имя существительное: madre + perla = madreperla (жемчужница), coche + cama = cochecama (вагон-кровать);

2) Имя прилагательное + имя прилагательное: claro + oscuro = claroscuro (светотень), agre + duice = agriduice (кисло-сладкий);

3) Имя существительное + имя прилагательное: vino + agre = vinagre (уксус); tío + vivo = tiovivo (карусель);

4) Имя прилагательное + имя существительное: salvo + conducto = salvoconducto (пропуск, залог), media + noche=medianoche (полночь);

5) Глагол + Глагол: gana + pierde = ganapierde (поддавки);

6) Глагол + Имя существительное: limpia + botas = limpiabotas (чистильщик обуви), rasca + cielos = rascacielos (небоскреб);

7) Местоимение + глагол: cual + quiera = cualquiera (любой);

8) Наречие + Имя прилагательное: mal + contento = malcontento (горестный).

В.В. Виноградов отдельно выделяет *именное соположение*. Такие устойчивые лексические единицы образуются по единой формальной модели существительное + существительное и единой семантико-грамматической схеме определяемое + определяющее (пр.: cuento chino – небылица, pájaro carpintero – дятел).

В испанском языке наблюдается явление *синапсии* (исп. Sinapsia). В терминологии Е. Бенвенисте термин рассматривается, как соединение состоящее из нескольких синтаксически коррелированных лексем, где детерминированный элемент несёт в себе главную смысловую нагрузку, и каждая лексема сохраняет свою первоначальную единозначность (пр.: goma de mascar, golpe de estado, traje de luces).

1.2 Словосочетание

1.2.1 Понятие словосочетание

В первом подходе отечественная лингвистическая школа рассматривает синтаксическую группу как отдельную синтаксическую единицу. В понимании Ф.Ф.Фортунатова словосочетание представляет «то целое по значению, которое образуется сочетанием одного полного слова с другим полным словом» [20]. А.А. Шахматов трактует понятие словосочетание как «такое соединение слов, которое образует грамматическое единство, обнаруживаемое зависимостью одних из этих слов от других» [21]. Классик отечественной лингвистики В.В. Виноградов рассматривает словосочетания как образования, в которых слово и определяющая его форма слова воедино соединены синтаксической связью [22]. В целом отечественные лингвисты рассматривают словосочетание как непредикативную единицу с

характерным синтаксическим отношением подчинения одного элемента другим («пластиковая тарелка», но не «кружка и тарелка»).

В рамках второго подхода словосочетание рассматривается как непредикативная единица с синтаксической связью среди других возможных синтаксических отношений между элементами [23].

Третий подход является достаточно весомым, так как рассматривает словосочетание (и целую синтаксическую группу) как соединение, в составе которого обнаруживается более одного слова [24]. Ф.Ф. Фортунатов и М.Н. Петерсон опираясь на морфологические свойства слова понимали под словосочетанием сочетание двух полнозначных слов независимо от их структуры и без учета семантических и синтаксических особенностей. Предложение является разновидностью словосочетания, с единственным признаком – интонацией.

Соположение и взаимная обусловленность данных слов в контексте объясняет их общность. Обусловленность укладывается в ряд критериев выделения синтаксических групп, среди которых самостоятельность (способность к топикализации, парцелляции, фрагментированию), неразделимость и т.п.

Согласно «Словарю лингвистических терминов Д.Э. Розенталя», в целом *словосочетание* – синтаксическая конструкция, образуемая соединением двух или более знаменательных слов на основе подчинительной грамматической связи [25].

Существует три типа подчинительной грамматической связи:

1) Согласование. Ярцева определяет согласование как один из видов подчинительной связи компонентов словосочетания, при котором в зависимом слове повторяются граммемы или часть граммем главенствующего слова [26]. В случае изменения вершинного слова, изменения затрагивают и все зависимые компоненты. В языках с развитой флективной системой данный тип связи широко используется для выражения атрибутивных отношений в субстантивных словосочетаниях

(рус.: «зелёный лес», «зелёная трава», «зелёное дерево»; нем. kalter Wein, kalte Milch, kaltes Wasser).

Я. Г. Тестелец выделил 4 базисных типа согласования в русском языке:

1) Согласование полных прилагательных (включая причастия и прилагательные-местоимения) с существительным в числе и падеже: узкая дорога, узкую дорогу, узкие дороги;

2) Согласование сказуемого с подлежащим в числе: мама приедет, мы приедем, они приедут; милая девушка, милые девушки;

3) Согласование полного прилагательного с существительным в роде: узкая река – узкий мост;

4) Согласование сказуемого с подлежащим в лице (у глаголов в настоящем и будущем времени) и в роде (у глаголов в прошедшем времени и кратких прилагательных): я прибуду – ты прибудешь; папа пришел – мама пришла; она добрая – он добр [27].

В классификации Золотовой приводятся два основных типа согласования:

1) Полное. Зависимое слово способно принимать все свои грамматические формы: ранним утром (согласование в падеже, числе и роде); отметки выставлены (согласование по числу); первые дни (согласование в числе и падеже);

2) Неполное. Зависимое слово не всегда уподобляется стержневому: озеро Кисегач (нет согласования по роду); на озере Кисегач (согласование по роду и падежу). Данный тип характерен для сочетаний с аппозитивными отношениями: река Лена, доцент Петрова. Крайне редко неполное согласование встречается в словосочетаниях с атрибутивными отношениями: умелая токарь, молодая врач.

2) Управление. Г.А.Золотова определяет управление как тип связи, при котором наличие у слова активной валентности требует определенного грамматического оформления соответствующей зависимой группы [28].

И.А. Мельчук выделил следующую классификацию случаев управления для русского языка, где под управление попадают:

1) Синтаксические актанты глагола, существительного и прилагательного: Афанасий (им. п.) посылает телеграмму (вин. п.) друзьям (дат. п.); мне и Насте (дат. п.) не осталось супа (род. п.);

2) Именная группа-дополнение предлога: на полу (предл. п.); к офису (дат. п.);

3) Глагол при союзе «чтобы» в сослагательном наклонении: чтобы увиделись;

4) Существительное с числительным в именительном или винительном падеже – две табуретки (род. п.), восемь табуреток (род. п.);

5) Прилагательное в роли присказуемостного имени при некоторых глаголах: знал его молодым (твор. п.), нашел ее здоровой (твор. п.);

6) Именная группа-актант при морфологически выраженной сравнительной степени прилагательного: слабее лося (род. п.).

Я.Г. Тестелец отмечает, что И.А. Мельчук не учитывает предложно падежные (забрать у собаки) и союзные (спросить что случилось: что + придаточное) средства выражения. Согласно теории И.А. Мельчука управление – случай морфологической зависимости между словоформами, под которое не попадают падежные и союзные конструкции [29]. Я.Г. Тестелец считает, что данный подход противоречит традиции и необходимо включить вышеизложенные средства выражения в управление.

3) Примыкание. Согласно В.А. Белошапковой примыканием называется тип связи, «который выражается не изменением формы зависимого компонента словосочетания, а лишь местоположением и зависимой грамматической функцией» [30]. При примыкании зависимая слово или группа не обладает никаким выраженным морфологическим признаком: сладко спать, юбка цвета неба, любить бегать, озеро Байкал, яйцо вкрутую и т.п.. П.А. Лекант различает сильное (при употреблении неизменяемых слов при глаголах, требующих информативно-

восполняющего слова: находиться вдали, относиться безрассудно, думать впопыхах) и слабое примыкание (остальные случаи) [31].

При примыкании к стержневому слову способны примыкать:

1) Наречия (слишком старый, говорить по-русски, свернуть налево);

2) Неизменяемые прилагательные (юбка хаки, соль экстра);

3) Прилагательные и наречия в сравнительной степени (дети помладше, быть глупее, идти медленнее);

4) Инфинитивы (привычка пить, люблю бегать, приехал работать);

5) Деепричастия (ест стоя, молча крича, идет прихрамывая);

6) Связь между именем существительным и некоторыми формами притяжательных местоимений его, её, их [32]. Грамматически данные формы схожи с неизменяемыми прилагательными из-за отсутствия форм падежа и не участии категорий рода и числа при выражении связи (ср.: их/его/её машина, кровать, еда и т.п.).

В языках со слабой развитой морфологией, синтаксические отношения могут быть выражены линейным расположением группы слов в цепочке. Такая особенность характерна для английского и китайского языков.

1.2.2 Смысловые отношения в словосочетании

Е.С. Скобликова также рассматривает смысловые отношения (грамматическое значение словосочетания), которые возникают в словосочетании между его компонентами:

1) Атрибутивные – в словосочетаниях со стержневым словом именем существительным, где роль зависимого компонента может выполнять: прилагательное (раннее утро), причастие (построенный магазин), местоимение (наш сад), порядковое числительное (второй номер), (бес)предложными формами существительного (платье из шелка, газета с объявлениями), наречие (сосед сверху), инфинитив (необходимость выспаться). Е.С. Скобликова четко разграничивает атрибутивные и аппозитивные отношения, которые возникают между существительным и приложением к нему. В аппозитивных отношений четко выражен денотат и

два соотнесенных с ним понятия (студент Петров, газета «Ночной Петербург»), а атрибутивных же представлен предмет и его признак.

2) Объектные отношения возникают в глагольных словосочетаниях, где зависимым словом может выступать: существительное в предложной или беспредложной форме (интересоваться книгами), местоимение (знать его) или числительное (забыть троих) или адъективных словосочетаниях (ласковый с матерью, гордый наградой).

3) Субъектные отношения выделяются, когда зависимое слово обозначает производителя действия или носителя состояния (приезд матери, освобожденный армией).

4) Обстоятельственные (адвербальные) – это отношения, при которых действие, состояние, признак определяются со стороны своего качества или условий его проявления (слишком тусклый, гулять на пруду) [33].

5) Комплетивные (восполняющие) – зависимое слово восполняет информативную недостаточность стержневого слова, является обязательным смысловым дополнением (три окна, прослыть мудрецом) [34].

Отношения между компонентами СС зависят не только от принадлежности их к той или другой части речи, но и от лексических значений как главного слова, так и подчиненного [35].

М.Я. Блох выделил четыре основных типа синтагматических связей между словами в их синтаксических объединениях:

1) *Сочинение* – последовательно эквипотентные связи. Компоненты эквипотентных словосочетаний равны по синтаксическому рангу. Синтаксические связи между компонентами таких словосочетаний могут осуществляться с помощью специальных сочинительных союзов или без каких-либо связующих слов (напр.: pobre pero honesto, mal y peligroso);

2) *Подчинение* – последовательно доминационные связи – одно слово в широком смысле определяет, или модифицирует другое. Главный

компонент в таком словосочетании называется ядром, а определяющего его слово адъюнктом (напр.: chica linda, basset bueno);

3) *Предикация* – взаимодоминационные связи. Сказуемое подчиняет подлежащее, называя событие предикации, некое действие, состояние или признак; в трансформациях номинализации отглагольный трансформ занимает позицию ядра в словосочетании, а подлежащее становится адъюнктом (напр.: él resolvió – su solución);

4) *Присоединение (кумулятивные)* – внутреннее присоединение, внутренняя кумуляция – объединения слов, в которых последующий компонент, хотя и присоединяется к предшествующему компоненту с помощью сочинительного союза, не равнозначен ему по характеру номинации (напр.: llego con retraso, pero; o en vías de estarlo) [36].

1.2.3 Именное словосочетание

Каждую синтаксическую группу можно выделить в определенный класс, базируясь на принадлежности стержневого слова к определенной части речи. П.А. Лекант, опираясь на Д.Э. Розенталя, определяет *именное словосочетание* – группа слов, вершину которых определяет имя. П.А. Лекант предложил свою классификацию именных словосочетаний:

1) Субстантивные (с существительным: нарицательным, собственным, субстантивированным), имеют определительное (атрибутивное) содержание («предмет – признак») (рус.: суровая зима, Елена прекрасная, свет, апрельский день; исп.: mujer fureosa, belleza loca). В словосочетаниях с отвлеченными существительными проявляются объектные, пространственные и причинные отношения (встреча на реке Москве,

опоздание из-за пробки);

2) Адъективные (с прилагательным) выражают качественную или количественную характеристику признака (очень приятный, по-детски плаксивый), обстоятельственные значения (готовый сражаться, удобный для работы), объектные (полный до краев, бедная влагой);

3) Прономинальные (с местоимением – личное, указательное, неопределенное, определенное, вопросительное) (рус.: кто-то из них, она красивая; исп.: nadie más bella, alguien dormido);

4) Нумеративные (с числительным) стержневым словом всегда выступает числительное в форме именительного или винительного падежей (рус.: два стула, три числа; исп.: dos pajaros, tres meses) [37].

Стандартная конструкция именной группы в испанском языке, состоит из двух элементов (но может включать и более): стержневое слово + зависимый компонент. Детерминант является не обязательным элементом субстантивного словосочетания, в некоторых случаях он может быть опущен.

В роли детерминанта выступают: определенные артикли (el, la, los, la); неопределенные артикли (un, una, unos, unas); указательные местоимения (este, ese, aquel); притяжательные местоимения (mi, tu, su, nuestro); порядковые числительные (tercero, quinto, séptimo); количественные числительные (dos, cinco, veinte); неопределенно-количественные числительные (varios, algunos, muchos); вопросительные местоимения (qué, cuánto); относительные местоимения (cuyo, cuya, cuyos, cuyas).

В испанском языке в адъективном словосочетании, к прилагательному (стержневому слову) в качестве зависимого компонента могут также присоединяться определительные наречия:

1) Количественные наречия (muy, tan, bastante, extremadamente);

2) Образа и способа действия (cuán y qué);

3) Качественные наречия (un poco).

Часто дополняет основную конструкцию предложная синтагма (пр.: cansado *de ti*, seguro *de todo*, lleno *con su actitud*, triste *por la música*).

В роли зависимого компонента в именном словосочетании могут выступать:

1) Существительное (рус.: поездка по городу; исп.: **capital riesgo);**

2) Прилагательные (рус.: интересный сюжет, исп.: **sector puntero);**

3) Местоимения (рус.: мой рюкзак; исп.: **gato mío);**

4) Порядковые числительные (рус.: четвертый курс; исп.: **día cuatro);**

5) Причастие (рус.: проверенный тест; исп.: ventana cerrada);

6) Наречия (рус.: яйца всмятку; исп.: muy rapido);

7) Инфинитив: (рус.: желание спать, умение петь).

1.2.4 Семантическая близость и сочетаемость слов в словосочетании

При семантической сочетаемости учитывается способность слова сочетаться с другими группами слов, при этом связанных с ним некой общностью смысла. В 1998 году В.Г. Гак отметил, что: «Основной закон сочетания слов сводится к тому, что для того, чтобы два слова составили правильное сочетание, они должны иметь, помимо специфических сем, одну общую сему» [38]. Французские лингвисты обозначали эту сему «классемой». Она может быть выражена любой семой и не всегда заключать в себе первостепенное значение суждения.

В лингвистике принято считать закон семантического согласования В.Г. Гака классическим для определения характеристики сочетаемости. Отечественный лингвист В.В. Морковкин считает, что семантическая сочетаемость обязательно должна содержать в себе указание на сему, которая должна присутствовать в значении всех слов, заполняющих соответствующую синтактико-семантическую позицию [39].

До сих пор открыт вопрос о дифференциации семантической и лексической сочетаемости. Ю.Д. Апресян описал, что в случае семантической сочетаемости ограничение на сочетаемость задаются указанием на семантический признак, а в случае лексической сочетаемости – только списком слов, с которыми может сочетаться данное слов (лексическая сочетаемость – оказать услугу, но не оказать заботу; семантическая сочетаемость – птицы вылетели из гнезда) [40].

При семантическом согласовании все компоненты сочетания не должны иметь противоречащих сем, иначе происходит нарушение языковой нормы, либо переосмысление одного из компонентов суждения.

Дистрибутивная семантика занимается вычислением степени семантической близости между лингвистическими единицами на основании их дистрибуции (на изучении окружения отдельных единиц в тексте) в больших массивах данных [41].

Выводы по главе 1

Отечественные лингвисты по-разному определяет понятие «слово». Природа «слова» остается малоизученной и актуальной.

Слово обладает лексическим значением и грамматическим значением. Исследованием природы лексического значения слова занимаются лексическая семасиология и лексикология. Грамматическое значение выражается морфологически.

Слово – двусторонне, обладает планом выражения и планом содержания. Оно семантически, фонетически и морфологически мотивированно.

Деривационный и способ словосложения в испанском языке являются ведущими способами формирования слов.

Изучение научных работ: Ф.Ф. Фортунатова, В.Н. Ярцевой, Г.А. Золотовой, В.В. Виноградова, по проблеме определения термина «словосочетание» выявило противоречия в его трактовке, а также неоднозначность подходов к его выявлению в контексте.

Словосочетание основано на одном из трех (согласование, управление, примыкание) видов подчинительной грамматической связи. В нём могут возникать смысловые отношения между компонентами: атрибутивные, объектные, субъектные, обстоятельственные и комплетивные.

Отношения между компонентами СС зависят не только от принадлежности их к той или другой части речи, но и от лексических значений как главного слова, так и подчиненного.

По П.А. Леканту выделяются четыре типа именных словосочетаний, в зависимости от принадлежности вершинного слова к определенной части речи: субстантивные, адъективные, прономинальные и нумеративные.

ГЛАВА 2 АВТОМАТИЗАЦИЯ ИЗВЛЕЧЕНИЯ ИМЕННЫХ СЛОВОСОЧЕТАНИЙ В ИСПАНСКОМ КОРПУСЕ

2.1 Теория конечных автоматов и регулярные выражения

Конечный автомат – абстрактный автомат без выходного потока, число возможных состояний которого конечно. Результат работы автомата определяется по его конечному состоянию [42].

Существует два типа автоматов:

1) Детерминированный конечный автомат (ДКА) (англ. deterministic finite automaton (DFA)) – набор из пяти элементов

($\sum$, Q, s Є Q, T C Q, ∂ : Q X $\sum$ → Q), где $\sum$ – алфавит (англ. alphabet), Q – множество состояний (англ. finite set of states), s – начальное (стартовое) состояние (англ. start state), T – множество допускающих состояний (англ. set of accept states),∂ – функция переходов (англ. transition function);

2) Недетерминированный конечный автомат (НКА) является обобщением детерминированного.

Для представления конечного автомата составляют расширенную таблицу переходов (таблицу значений функции переходов δ, первая строка которой соответствует начальному состоянию, а заключительные состояния помечены единицами в дополнительном столбце) (Рисунок 2.1) или диаграмму переходов (ориентированный граф, вершины которого – состояния автомата, а дуги помечены элементами алфавита) (см. Рисунок 2.1).

На автомат можно смотреть как на физическое устройство, состоящее из устройства управления и входной ленты.

Конечные автоматы широко используются на практике, например в синтаксических, лексических анализаторах, и тестировании программного обеспечения на основе моделей.

Для работы с ДКА активно используют регулярные выражения (формальный язык поиска и осуществления манипуляций с подстроками в

тексте, основанный на использовании метасимволов (символов-джокеров, англ. wildcard characters) [43].

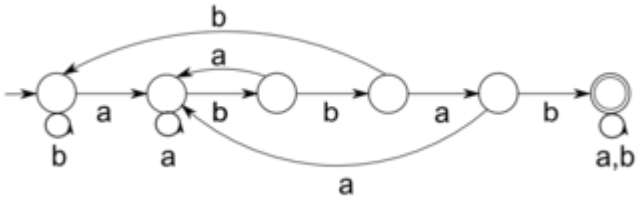


Рисунок 2.1 – Автомат для поиска образца в тексте для строки abbab

	a	b	F
q_0	q_0	q_1	0
q_1	q_2	q_3	0
q_2	q_2	q_3	1
q_3	q_3	q_3	0

а

	a	b	F
q_0	q_0	q_1	0
q_1	q_2	q_3	0
q_2	q_2	q_3	1
q_3	q_3	q_3	0
q_4	q_4	q_4	0

б

	a	b	F
q_0	q_0	q_1	0
q_1	q_2	q_3	0
q_2	q_2	q_4	1
q_3	q_3	q_4	0
q_4	q_3	q_4	0

в

Рисунок 2.2 – Расширенные таблицы переходов автоматов A1 (а), A2 (б), A3 (в)

Сейчас регулярные выражения используются многими текстовыми редакторами и утилитами для поиска и изменения текста на основе выбранных правил. Многие языки программирования уже поддерживают регулярные выражения для работы со строками. Например, Perl и Tcl имеют встроенный в их синтаксис механизм обработки регулярных выражений. Набор утилит (включая редактор sed и фильтр grep), поставляемых в дистрибутивах Unix, одним из первых способствовал популяризации понятия регулярных выражений [44].

Регулярные выражения используются для сжатого описания некоторого множества строк с помощью шаблонов, без необходимости перечисления всех элементов этого множества. При составлении шаблонов используется специальный синтаксис, поддерживающий обычно такие операции как перечисление, группировка, квантификация. Синтаксис регулярных выражений вместе с метасимволами приведен в приложении 1.

2.2 Морфологический анализ и частеречная разметка

Морфологический анализ стал особой формой лингвистического исследования лишь с 40-х годов нашего века. Значимость морфологического анализа заключена в определении морфологической структуры слова, т. е. его строение, описанное в терминах данного уровня. В отличие от морфемного анализа, целью которого является вычленение морфем, составляющих данное слово, и их структурная и функциональная классификация, в задачи морфологического анализа входит вся область изучения структуры слова в ее грамматическом аспекте [45].

Основной задачей морфологического анализа является выявление по данной словоформе её первоначальную нормальную форму, от которой и было произведено данное слово, а также указание набора параметров. В результате анализа одной заданной словоформе может быть сопоставлено несколько таких пар.

Под морфологическим параметром понимают такие особенности как род, число, склонение, время, краткость формы прилагательного и другие признаки, свойственные анализируемому языку [46].

С развитием автоматической обработки текста морфологическая разметка приобретает широкую популярность при работе с корпусами текстов и большими массивами данных. На её фоне появляется частеречная разметка, которая также носит название POS tagging или part-of-speech tagging. Теперь перед частеречной разметкой стоит задача не только в определении части речи и грамматических характеристик слов в тексте или корпусе, но и присвоение им свойственных тегов. POS tagging считается одним из первых этапов компьютерного анализа текста, который используется при составлении морфологических анализаторов.

Морфологическая разметка составляет основу для последующих этапов лексического анализа – семантического и синтаксического.

Каждая автоматизированная частеречная разметка должна обязательно базироваться на определенных правилах и тегах, которые хранятся в

схематичном виде. Каждому тегу прилагается его описание (расшифровка). Как правило размер и количество тегов может варьироваться в зависимости от того какую задачу выполняет морфологический анализатор, и что на выходе хочет получить пользователь. В последнее время из-за активно развивающейся корпусной лингвистики наметилась тенденция к сокращению количества морфологических тегов (помет) для ускорения анализа текстовых массивов и выполнения строго структурированных задач [47].

2.3 Методы и способы автоматизации извлечения именных словосочетаний

Из-за значительных затрат при разработке, полный синтаксический и семантический анализ текста по-прежнему недоступен для большинства исследователей. Если этап полного анализа текста пройден, то встаёт задача извлечения целевой информации из результатов анализа. В системах автоматического анализа текстов предпочтение отдается подходам, основанным на частичном синтаксическом анализе, который позволяет решать довольно большой спектр практических задач по извлечению и поиску информации [48].

Из-за своей гибкости работы модуля анализа текста, частичный синтаксический анализ позволяет в полной мере обработать входной текст, следуя конкретной задачи исследователя. Выделение именных групп (NP-chunking) является одной из значимых составляющих частичного анализа текста. Выделение именных словосочетаний является необходимым при автоматическом выявлении фактов, анализе медицинской и технической документации, при извлечении информации об отношениях [49].

В испанском языке именное словосочетание представляет сложный объект изучения. Есть необходимость в разработке корректного локального понимания входного текста ЭВМ для распознавания именной группы из-за таких лингвистических особенностей, как стирание склонений имен существительных и прилагательных, согласование по грамматическим

признакам, а также унификации множественного числа посредством окончания –s в испанском языке.

В настоящее время выделяется 2 основных подхода для извлечения информации из текстов, в частности именных словосочетаний:

1) *Рационалистический подход или инженерный* (rule-based) – базируется на составлении шаблонов с учетом лингвистических особенностей именных групп в обрабатываемом тексте, содержит правила, основывается на регулярных выражениях. Инженерный подход опирается на тот факт, что извлекаемая информация употребляется в рамках определённых языковых конструкций (напр.: название города пишется с большой буквы и нередко предваряется словами город, гор. или г.). Подобная лингвистическая информация обычно вручную описывается в виде формальных шаблонов распознаваемых конструкций и правил их обработки. Затем правила применяются IE-системой к анализируемому тексту: в нем ищутся описанные шаблонами фрагменты, из которых извлекается искомая информация.

2) *Машинное обучение* (machine learning) – подход, основанный на самообучающейся системе. Несмотря на свою прозрачность, выходные данные сложно поддаются лингвистической интерпретации и практически не учитывают лингвистические особенности именной группы. Машинное обучение включает в себя: методы обучения с учителем (supervised), методы обучения без учителя (unsupervised), методы частичного обучения с учителем (bootstrapping).

Чаще всего применяется обучение с учителем, которое подразумевает построение математической и программной модели, которая умеет отличать искомые данные от всех остальных. Построение такого машинного классификатора (т.е. обучение модели) происходит на специально размеченном вручную текстовом корпусе (обучающей выборке), в котором значимым объектам, их атрибутам, отношениям, фактам приписаны соответствующие метки. Метки кодируют признаки для распознавания этих

данных. Для вышеприведенного примера для извлечения названия города в качестве признаков могут выступать: регистр (верхний) первой буквы слова, конкретные слова, стоящие перед ним (город, город-курорт, город-музей, город-герой, гор. или г.), а также признаки последующих слов (для выявления многословных названий, таких как Нижний Тагил).

В последнее время появляется все больше *гибридных методов*, сочетающих в себе достоинство рационалистического и инженерного подхода.

Гибридный подход С.О. Шереметьева опирается на базе знаний стоп-слов, учитывающих их позицию при обработке входного текста. В базу включены особые словоформы с установленными правилами на запрет их расположения в начале, середине или конце именной группы.

Первые IE-системы были построены в рамках инженерного подхода, наиболее известной из них была AutoSlog. Среди первых отечественных разработок стоит упомянуть семейство мультиязычных систем извлечения информации из деловых текстов OntosMiner, которые обеспечивали переход от неструктурированной информации к ее семантическому представлению в формате онтологий предметных областей, заложенных в систему (бизнес-события, судебная тематика и полицейские отчёты). Разработка прикладных IE-систем является сложным и трудоемким процессом, существенную помощь в котором могут оказать инструментальные системы, включающие стандартные модули анализа текста и даже средства сборки и отладки приложений.

Инструментальные системы, предназначенные для разработки приложений в рамках инженерного подхода, имеют обычно встроенный формальный язык для задания лингвистических правил и шаблонов — с их помощью стандартные программные модули настраиваются на решение конкретной прикладной задачи.

2.4 Корпус и инструменты для машинного анализа

Для машинного анализа был собран корпус фармацевтических рецептов с испанского сайта www.doctoralia.es. Предварительно корпус был очищен от лишних элементов (html теги, нумерация страниц, ссылки на вопросы и форум).

В постобработонном виде корпус содержит 1538045 слов, объемом 7,4 мб в формате .txt.

Для более продуктивного анализа и оценки выходных результатов, автоматический поиск именных словосочетаний был разбит на несколько этапов с учетом части речи вершинного слова: нумеративные, прономинальные, адъективные и субстантивные.

Перед процедурой создания шаблонов на языке регулярных выражений были составлены правила для каждой группы (см. Приложение 2).

Количество правил, как и количество шаблонов – бесконечно. Они строятся под определенные задачи и обладают вариативностью.

Обработка корпуса выполнялась в операционной системе Ubuntu 14.04 LTS, с поддержкой графической оболочки UNIX. UNIX содержит встроенную командный интерпретатор (терминал), позволяющий работать с открытыми исходными кодами и запускать командные скрипты, что позволяет произвести быстрый автоматический анализ текста при работе с массивными базами данных.

Для автоматического поиска использовалась утилита командной строки «Grep», которая полностью отвечает заданному регулярному выражению и выводит строки, содержащие заданный поисковый элемент. Инструмент Grep позволяет реализовать подход основанный на правилах с использованием регулярных выражений. Для выделения именных групп применялись как морфологические, так и грамматические признаки вершинной части речи словосочетания, способные охарактеризовать искомый элемент в целом. В нескольких случаях поисковый шаблон включал в себя детерминанты и морфологические признаки зависимого

слова, что позволило создать наиболее точный поисковый шаблон, базирующийся на признаках составляющих его словоформ.

Для оценки качества работы предложенной модели по извлечению именных словосочетаний и составленной базы знаний были использованы следующие метрики: *точность* (Precision) как количество правильных ответов, делённое на количество всех найденных ответов и *полнота* (Recall) — как количество правильных ответов, делённое на общее число правильных ответов.

2.5 Автоматизированный поиск нумеративных словосочетаний

Для эксперимента мы взяли количественные от 2 до 30. Первые 15 были представлены в своей полной форме, остальные имеют в составе одинаковый элемент dieci– и veinti– (кроме числа 20 veinte). Согласно правилам испанской грамматики, если имя числительное возглавляет словосочетание, то зависимое слово или группа слов строго следует за ним.

Шаблон: «\<**(dos|tres|cuatro|cinco|siete|ocho|nueve|diez|once|doce|trece|catorce|quince|veinte)**\>\s\w*|\<**(dieci*|veinti*)**\w*\s\w*». На выходе из поискового шаблона было получено 575 образований. Из них двухкомпонентными цельнооформленными словосочетаниями являются: 501 (см. рисунок 2.3).

Кроме них поисковый шаблон выдает конструкции с числительным и последующим предлогом или союзом, что указывает на незаконченность синтаксических отношений (явное начало именной группы) и на неточность построения поискового шаблона (см. таблицу 2.1).

Необходимо изменить поисковый шаблон с учетом всех исключений. Мы прибегнем к контексту для анализа структуры именной группы, в дальнейшем это поможет создать максимально точный поисковый шаблон.

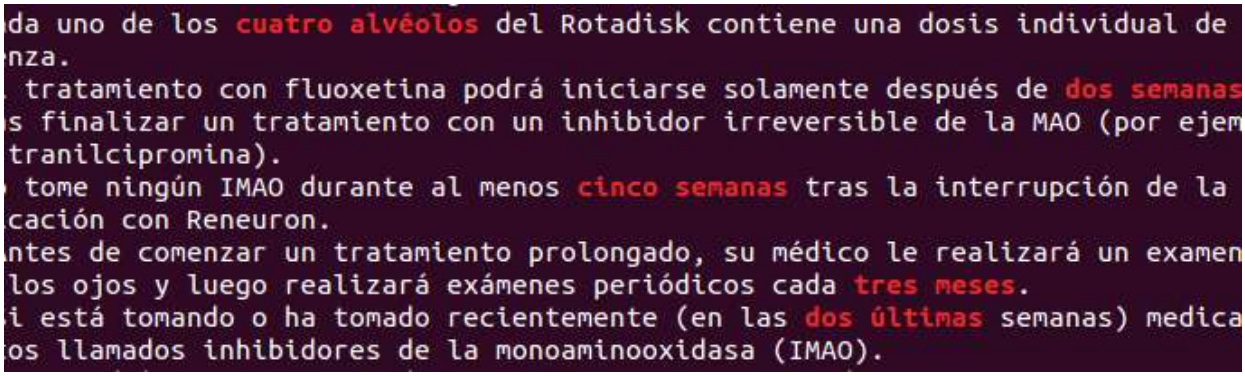


Рисунок 2.3 – Фрагмент текста для поиска нумеративной ИГ

Таблица 2.1 – Примеры предложных конструкций с числительным

tres **o**	dos por	dos **al**
dos a	Nueve **en**	dos sin

Конструкции с сочинительными союзом «o» и предлогом: dos o **más mucosas**; tres o **más años;** dos o **tres días;** dos o tres **días de tratamiento.** Наличие предлога «de» указывает на то, что элемент является составляющим компонентом более крупной именной группы и его следует рассматривать как отдельную смысловую единицу текста, его включение в поисковый шаблон не обязательно, но возможно.

Шаблон для поиска узкой ИГ с союзом «o»: **«\<(dos|tres|cuatro|cinco|siete|ocho|nueve|diez|once|doce|trece|catorce|quince|veinte)\>\so\s\w*\s\w*|\<(dieci*|veinti*)\so\s\w*\s\w*»**. Найдено: 44 ИГ - среди них такие словосочетания как: tres o más bebidas, dos o tres tomas, dos o más mucosas .

Шаблон для поиска расширенной ИГ с предлогом «de» и союзом «o»: **«\<(dos|tres|cuatro|cinco|siete|ocho|nueve|diez|once|doce|trece|catorce|quince)\>\so\s\w*\s\w*\sde{0,1}\s\w*|\<(dieci*|veinti*)\so\s\w*\s\w*\sde{0,1}\s\w*».**
В корпусе имеется лишь одно словосочетание с расширенной ИГ: dos o tres días de tratamiento (два или три дня лечения).

Конструкции с предлогами «por», «a», «al», «en»: dos **por codeína** cuya frecuencia; máximo dos **al día**; nueve **en niños**, dos **sin problemas**. Структура

именной группы с предлогами включает в себя один или более зависимых элементов, также встречаются расширенные именные группы.

Шаблон для поиска узкой ИГ: «**\<(dos|tres|cuatro|cinco|siete|ocho|nueve|diez|once|doce|trece|catorce|quince|veinte)\>\s(sin|al|en|por)\s\w***». Найдено 4 словосочетания данного типа (см. рисунок 2.4).

```
whiterabbit@whiterabbit-TravelMate-5744Z:~$ grep -E -o '\<(dos|tres|cuatro|cinco
|siete|ocho|nueve|diez|once|doce|trece|catorce|quince|veinte)\>\s(sin|al|en|por)
\s\w*' result.txt
dos por codeína
dos sin problemas
nueve en niños
dos al día
```

Рисунок 2.4 – Шаблонный поиск ИГ с предлогами «por», «a», «al», «en»

Предложные словосочетания с предлогом «a»: tres a cinco días de tratamiento, dos a cuatro semanas, dos a siete dias. Особенность контекста фармацевтического корпуса заключается в том, что предлог «a» требует за собой постановку ещё одного количественного числительного. Данную особенность необходимо учитывать при составлении поискового шаблона.

Шаблон для поиска ИГ с предлогом «a»: **«\<(dos|tres|cuatro|cinco|siete|ocho|nueve|diez|once|doce|trece|catorce|quince|veinte)\>\sa\s\w*\s\w*».**

Всего обнаружено 8 словосочетаний данного типа, 2 из которых повторяются несколько раз на протяжении всего текста (см. рисунок 2.5).

```
whiterabbit@whiterabbit-TravelMate-5744Z:~$ grep -E -o '\<(dos|tresdos|tres|cua
ro|cinco|siete|ocho|nueve|diez|once|doce|trece|catorce| quince|veinte)\>\sa\s\w
\s\w*' result.txt
dos a cuatro semanas
dos a siete dias
dos a cuatro cápsulas
dos a cuatro horas
dos a cuatro semanas
dos a cuatro semanas
dos a cuatro veces
tres a cinco días
```

Рисунок 2.5 – Шаблонный поиск ИГ с предлогом «a»

После операции поиска по расширенному шаблону: **«\<(dos|tresdos|tres|cuatro|cinco|siete|ocho|nueve|diez|once|doce|trece|catorce | quince|veinte)\>\sa\s\w*\s\w*\sde\s\w*\s\w***». Выяснилось, что среди 8

словосочетаний присутствует 2 расширенных с предлогом «de» (см. рисунок 2.6).

```
whiterabbit@whiterabbit-TravelMate-5744Z:~$ grep -E -o '\<(dos|tresdos|tres|cu
ro|cinco|siete|ocho|nueve|diez|once|doce|trece|catorce| quince|veinte)\>\sa\s'
\s\w*\sde\s\w*\s\w*' result.txt
dos a cuatro horas de la administración
tres a cinco dias de tratamiento no
```

Рисунок 2.6 – Результаты поиска расширенных ИГ с предлогом «a»

Раннее мы не брали во внимание количественные числительные от 200 до 900 для которых характерно согласование по роду (setecientos – setecientas gatos; ochocientas gatas – ochocientos gatos), они попадают под исключения в испанском языке. Также необходимо учитывать и то, что многие числительные являются составными. Грамматической особенностью таких числительных является наличие союза «y» между составляющими компонентами (напр.: doscientos treinta y seis gatos).

Для поиска составных нумеративных словосочетаний (3 компонента с союзом y): «**\w*(cientos|cientas)\sy\s\w*\s\w***». В результате обработки была выявлена 1 ИГ (см. рисунок 2.7).

```
whiterabbit@whiterabbit-TravelMate-5744Z:~$ grep -E -o '\w*(cientos|cientas)\sy
s\w*\s\w*' result.txt
cuatrocientas y tres pastillas
```

Рисунок 2.7 – Результат поиска составной нумеративной ИГ

Для поиска составных нумеративных словосочетаний (4 вершинных компонента с союзом y): **'\w*(cientos|cientas)\s\w*\sy\s\w*\s\w*'**. В нашем корпусе не содержатся ИГ данного типа.

Особенностью числа 100 в испанском языке является потеря окончания –to перед существительным. Мы не учитывали этот фактор в предыдущем шаблоне.

Шаблон для поиска: **'\Wcien\s\w*'**. Результаты поиска (см. рисунок 2.6).

```
whiterabbit@whiterabbit-TravelMate-5744Z:~$ grep -E -o '\Wcien\s\w*' result.txt
 cien pacientes
```

Рисунок 2.8 – Результат поиска ИС по схеме: количественное числительное (100) и существительное

Выявление порядковых числительных тысяча (mil) и миллион (millón), которые при образовании словосочетания могут иметь в составе более одного вершинного слова, затруднено без анализа контекста. Однако, мы можем с точностью говорить о наличии как менее одного порядкового числительного до слова mil или millón (к примеру: nueve mil gatos, doscientos millónes).

Шаблон для поиска ИГ с вершинным элементом mil или millón(es): **$ grep -E '\Wmil|millónes\W\s\w*'**. Всего найдено 7 образований, из которых лишь 1 представляет полную ИГ (diez mil pacientes), и 1 входит в состав расширенной ИГ с предлогом «de» (см. рисунок 2.9).

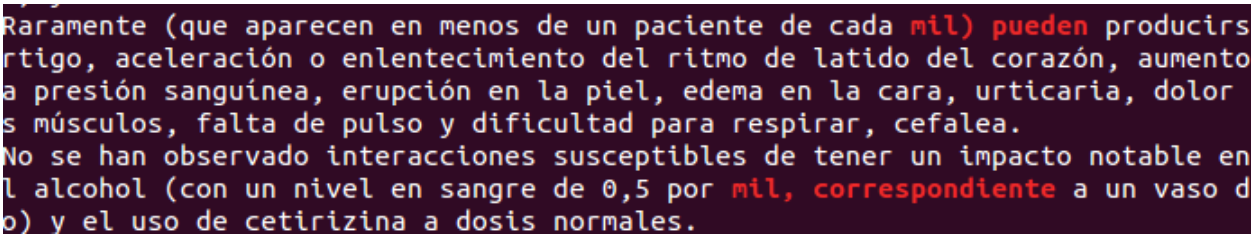


Рисунок 2.9 – Результат поиска ИГ с mil

При построении шаблона не была учтена многозначность слова mil и наличие графических знаков в предложении. В результате, при анализе выходных данных, было обнаружено наличие лишь одной полной нумеративной ИГ: mil pacientes (тысяча пациентов). Данный поисковый шаблон показал крайне малую вероятность, всего 1 к 7 или 0,14.

2.3Автоматизированный поиск субстантивных словосочетаний

Согласно классической грамматике большинство субстантивных ИГ содержат детерминат перед стержневым компонентом. Можно составить шаблон с опорой на артикли, так как они неизменно находятся в предпозиции к вершине ИГ.

Шаблон: «**\\<(el|la|los|las|un|una|unos|unas)\>\s\w*\s\w*».**

В корпусе всего было найдено 30686 образований данного типа. Среди них были найдены как трехкомпонентные субстантивные ИГ, так и лишь начальные фрагменты ИГ с предлогами и союзами: con, de, y, del, en, para, sin, o, que, a, no – указывающими на расширенную ИГ. Также по шаблону

нашлись полные грамматические основы (el antibiótico deberá, las quinolonas pueden, el médico realizará, el paciente sufría) (см. рисунок 2.10).

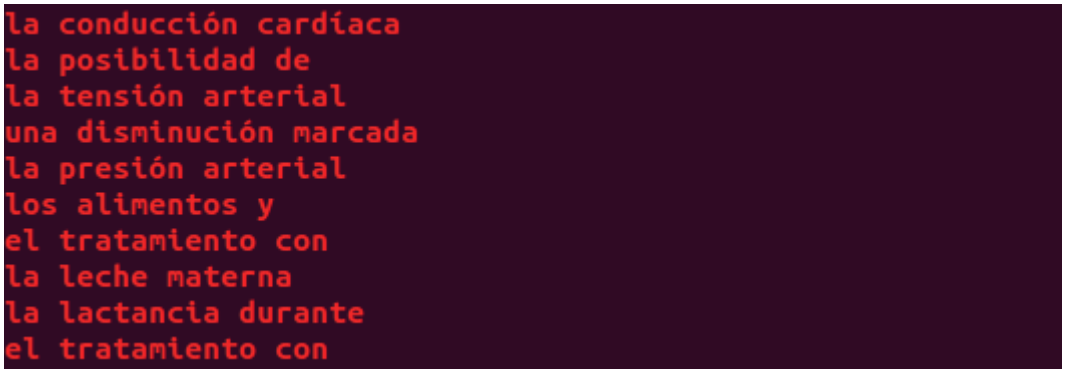


Рисунок 2.10 – Фрагмент поиска с использованием артиклей

Для более точного выявления субстантивных ИГ необходимо учитывать связь компонентов с предлогами и артиклями и последующие элементы конструкции.

Словосочетания с союзами и сочинительными предлогами «y», «o» и «a» требуют после себя дополнительный компонент.

Шаблон для поиска субстантивного именного словосочетания с y, o, a: **«\<(el|la|los|las|un|una|unos|unas)\>\s\w*\s(a|o|y)\s\w*\s\w*».**

В результате мы получили: 4670 образований данного типа (см. рисунок 2.11). После анализа исходного материала, оказалось что на 100 образований, приходится лишь 48 полных самостоятельных словосочетаний, в остальных случаях за предлогами следовали и такие части речи как глаголы (la cabeza o ha tenido, el periodo o tenerlo de, los impulsos a consumir cocaína и т.д.) или ИГ являлась лишь частью более крупной группы (la conducción o al manejo de herramientas o máquinas, la sensibilidad a la luz y al ruido, la piel o urticaria en cualquier parte del cuerpo). Вероятность точного нахождения ИГ по данному шаблону составляет всего лишь 0,48 к 1.

Для поиска субстантивной ИГ с «de» и «con» необходимо учитывать также что перед вторым компонентом может возникать артикль.

Шаблон для поиска субстантивной ИГ в случае наличия артикля после «de» или «con»: **«\<(el|la|los|las|un|una|unos|unas)\>\s\w*\s(de|con)**

\s\<(el|la|los|las|un|una|unos|unas)\>\s\w*». По поисковому шаблону было обнаружено 3592 словосочетания (см. рисунок 2.12).

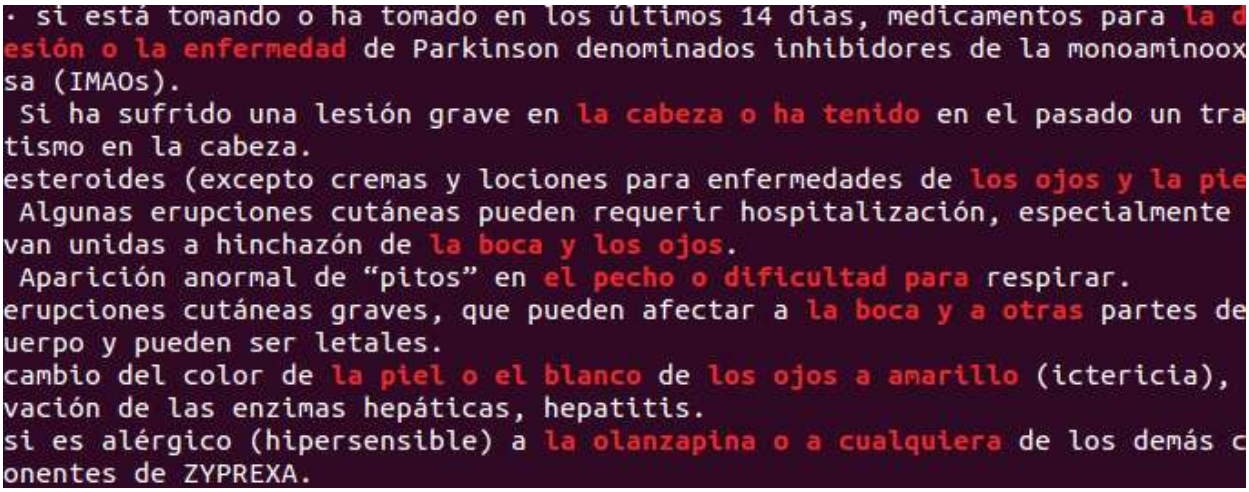


Рисунок 2.11 – Фрагмент выходного текста после работы шаблона для ИС с предлогами a, o, y

Некоторые из них справа имели предлог «de», что указывало на продолжение ИГ и вложенность конструкции (пример: el aumento de los niveles de hormona tiroide A, el tratamiento de la obesidad con pastillas como MySimba, el blanco de los ojos a amarillo, el mantenimiento de la abstinencia del alcohol).

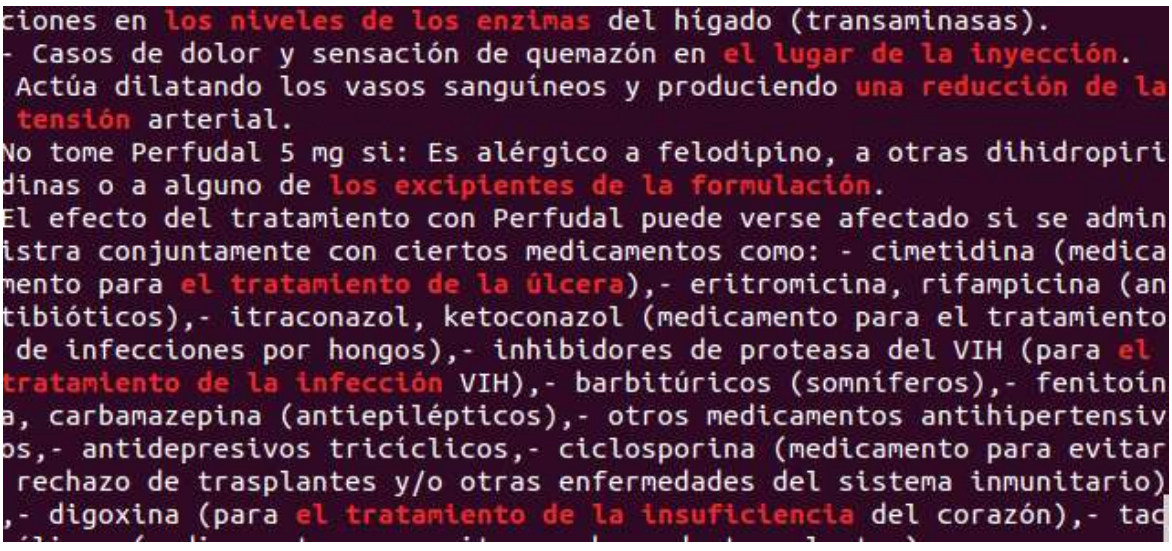


Рисунок 2.12 – Фрагмент извлечения ИГ с артиклем и de или con

Также шаблон не учитывает наличие прилагательного после последнего слова справа, что существенно влияет на последующие этапы обработки текста и его перевод на другие языки (пример: una evaluación de la función

visual, un incremento de los enzimas hepáticos, la valoración de los efectos adversos).

Для поиска ИГ, где вершиной является имя существительное, а зависимым компонентом выступает порядковое числительное необходима предварительна составленная база знаний порядковых числительных. Так как порядковые числительные согласуются в роде и числе, необходимо учитывать все типы окончаний при составлении поискового шаблона, а также их предпозицию по отношению к существительному (segundo chico, segunda chica, segundos chicos, segundas chucas).

Для эксперимента мы взяли первые 5 порядковых числительных: **«\<(primer|segund|tercer|cuart|quint)\s\w*»**. По шаблону было найдено 364 совпадений (см. рисунок 2.13). Большинство из них, а именно 293 являются полностью самостоятельными двухкомпонентными ИС. Остальные же исключения – 71 образование – представляют сочетание порядкового числительного и предлога или союза (primer y, tercer o).

Для полноты поиска необходимо в дальнейшей учитывать данные образования и включить в регулярное выражение под знаком исключения.

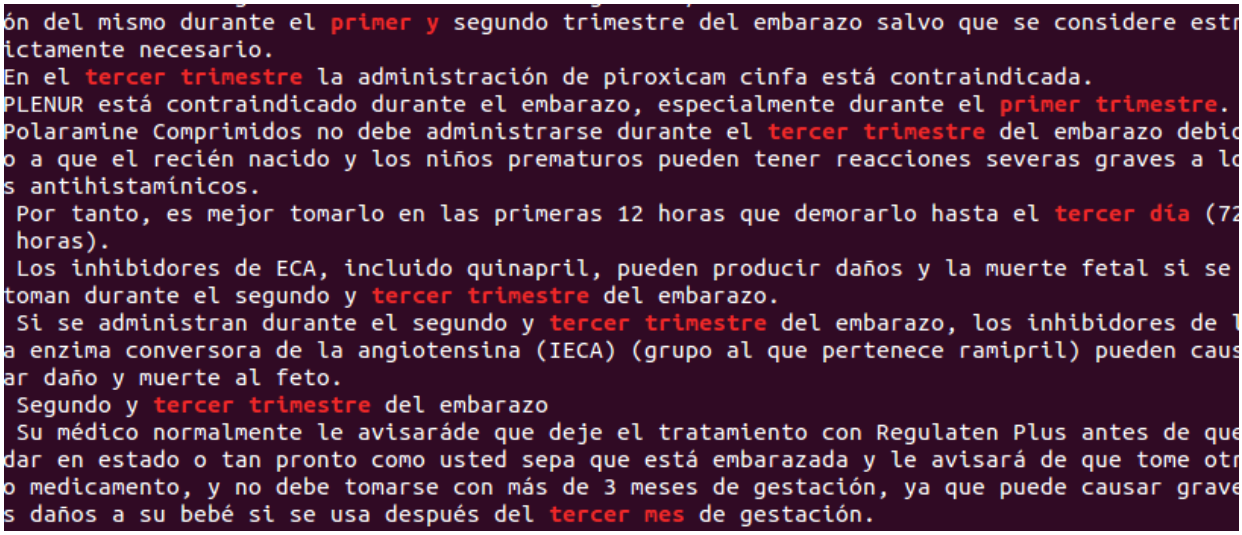


Рисунок 2.13 – Результаты поиска по шаблону числительных

В некоторых случаях местоимение является зависимым элементом в субстантивной ИГ. Они находятся в предпозиции к существительному (стержневому слову).

Для поиска ИГ с притяжательными местоимениями (mi, mis, tu, tus, su, sus, nuestro(s), nuestra(s), vuestro(a,as), мы составили шаблон: «**\<(mis?|tus?|su?|nuestros?|nuestras?|vuestros?|vuestras?)\>\s\w***». Всего было найдено 10758 образований данного типа. Среди них, наиболее часто употребительная форма с местоимением su (Ваш/Вашего) (su sangre – вашей крови, su médico – вашего медика, su reacción - вашу реакцию, su bebé - ваш ребенок, su estómago - ваш желудок), кроме них так же встречаются словосочетания с местоимениями tu (tu psiquiatra – твой психиатр, tu enfermedad – твоё заболевание), nuestros (nuestros pacientes – наши пациенты), nuestro (nuestro pais – нашей страны) (см. рисунок 2.14).

Во всех случаях была полностью выдержена структура ИГ (притяжательное местоимение + имя существительное). В некоторых случаях найденная ИГ входила в состав более расширенной ИГ с предлогом «de» (a su médico de que está tomando Zytram, su centro de salud).

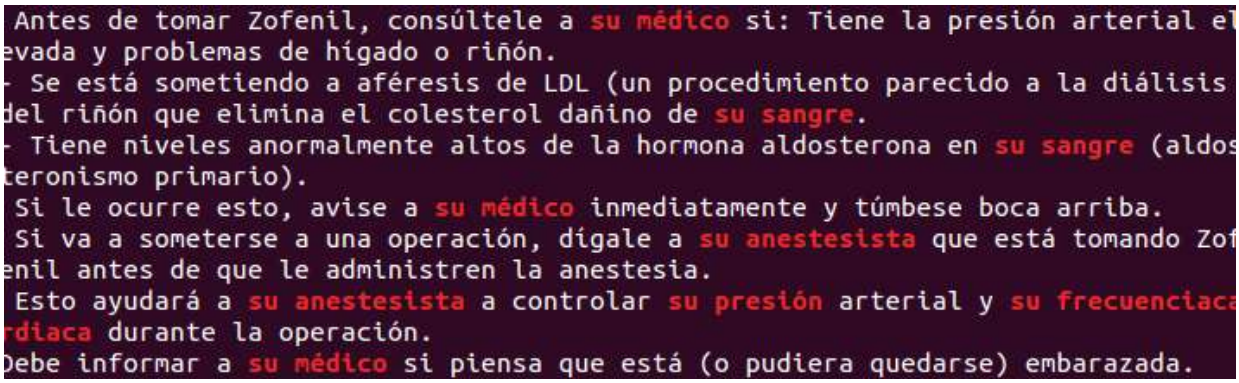


Рисунок 2.14 – Фрагмента поиска ИГ по сх по схеме притяжательное местоимение + имя существительное

Поисковая точность данного шаблона описывается тем, что притяжательные местоимения всегда находятся в предпозиции по отношению к существительному и согласуются с ним в роде и числе, также при наличии притяжательного местоимения артикль перед именем существительным опускается.

В испанском языке насчитывается всего три основных указательных местоимения, которые согласуются с существительному в роде и числе: (este – esta – estos – estas; ese – esa – esos – esas; aquel – aquella – aquellos –

aquellas). Для поиска словосочетаний с указательными местоимениями, имеющими близкое значение к артиклю и также находящиеся в предпозиции к имени существительному, был создан поисковый шаблон следующего типа: «\<**(ese|esas?|esos|aquel|aquella|aquellos|aquellas)**\>\s\w*».

Всего было найдено 310 образований в составе которых входит притяжательное местоимение (см. рисунок 2.15). 215 из них являются полноценными ИГ (esas horas, ese problema, aquellos medicamentos , aquellas mujeres , ese día, esos casos,esos vasos). Поисковый шаблон работает в вероятностью 0,69. Среди, конструкций не попадающих под ИГ есть указательные местоимения с предлогами, союзами (aquellos con, aquellos en, aquellos que, aquellas para, aquellos sin).

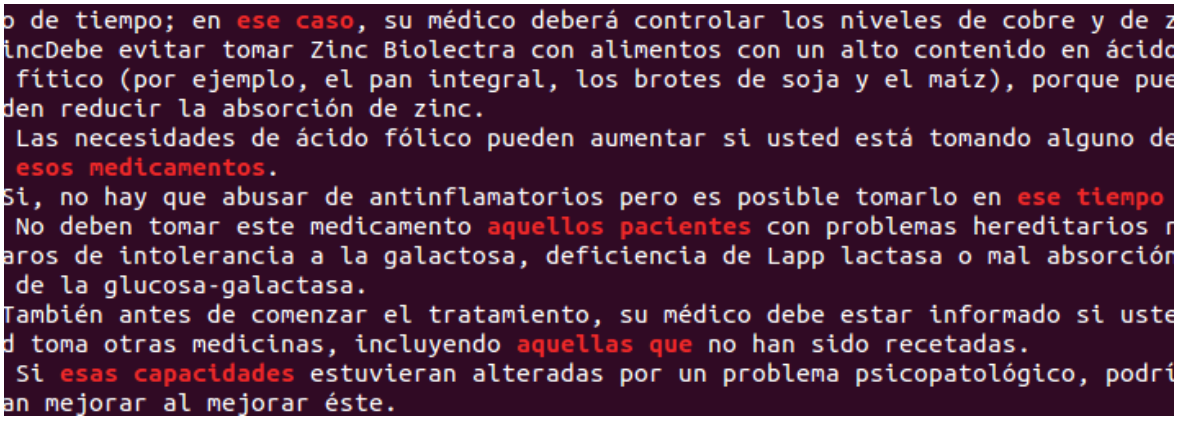


Рисунок 2.15 – Фрагмент работы поискового шаблона с указательными местоимениями

При сочетании с существительным относительные местоимения (Cuyo – cuya – cuyos – cuyas – чей, чья, чьи; который (–ая, –ые)) могут быть показателями придаточных предложений. Таким образом местоимение cuyo выполняет две функции: относительного местоимения и притяжательного-прилагательного. Оно связывает два имени, одно из которых всегда называет лицо (или предмет), обладающее чем-либо, а другое – предмет обладания:

Шаблон для поиска: «**\w*\<(cuyo|cuya|cuyos|cuyas)**\>**\s\w***». Всего было выявлено: 135 конструкций (см. рисунок 2.16).

Большинство из них, а именно 93 имеют перед указательным местоимением существительное, согласующееся с придаточным предложением (gástricas cuyos síntomas, medicamentos cuya acción, adversos cuya frecuencia, niños cuyas madres).

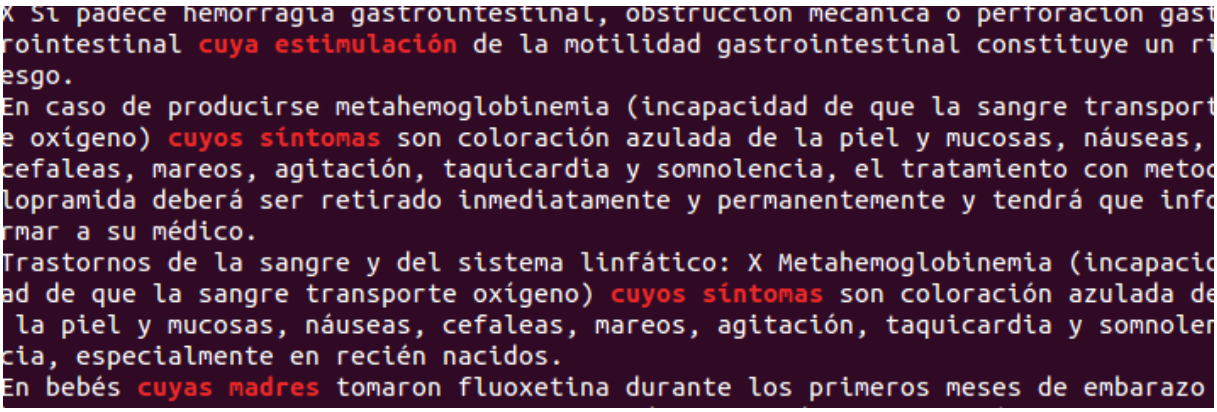


Рисунок 2.16 – Шаблон поиска относительных местоимений

Неопределенные местоимения-прилагательные также являются одним из показателей ИГ. Они всегда занимают предпозицию относительно существительного. Для поиска мы отобрали наиболее распространенные из них (alguno,todo, otro, mismo, varies, cualquier).

Поисковый шаблон: **«\w*\<(algunos|algunas?|algún|todo|todas?|todos|otros?|otras|mismos?|mismas?|varies|varias|cualquiera?)\>\s\ w*»**.

Всего найдено: 10285 конструкций (см. рисунок 2.16). На 100 единиц выявленных сущностей, приходится лишь 57 полноценных ИС (пример: alguna enfermedad, otras causas, mismo tiempo, otros antisépticos, cualquier origen, algunas personas). Среди остальных обработанных конструкции встречались неопределенные местоимения с артиклями (todos los, todas las); с предлогом «de», что указывало на незавершенность ИГ (alguna de, cualquiera de); с различными союзами и предлогами (mismo o, todo en, cualquier otra, mismo que, algún otro). Вероятность выявления ИС с неопределенным местоимение, даже без предварительно включенных стоп-слов в шаблон, достаточна высока 0,57 к 1.

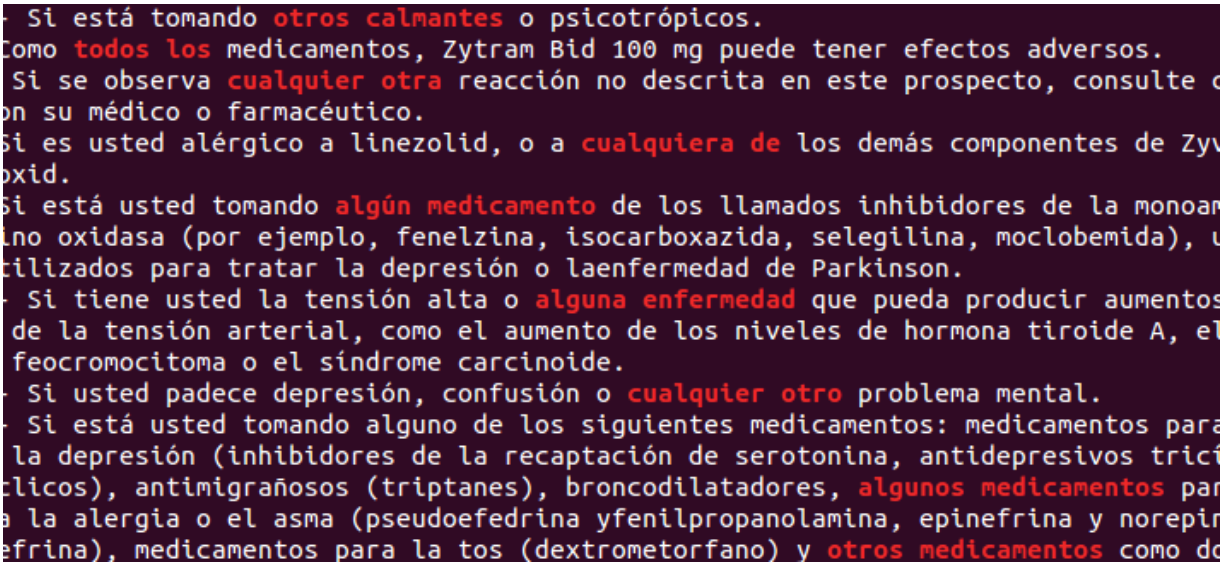


Рисунок 2.17 – Фрагмент текста по поисковому шаблону с неопределенным местоимением

Предыдущие наши шаблоны не включали в себя отрицательные местоимения ningún, ninguna(o).

Шаблон для поиска ИГ с отрицательными местоимениями: **«\w*\<(ningún|ningun(a|o)?)\>\s\w*»**. Обнаружено: 215 образований (см. рисунок 2.17). При этом все конструкции с ningún являются полноценными ИГ (пример: ningún componente , ningún problema , ningún efecto , ningún país). А среди конструкций с ninguna и ninguno был выявлен признак (предлог «de») расширенной именной групп. Из 215 наименований лишь 168 являются полноценными ИГ (ninguna reacción, ninguna temperatura, ninguna herramienta). Поисковый шаблон работает с точностью 0,78 .

Наречие также может быть зависимым словом в субстантивной ИГ. Для эксперимента были отобраны 5 самых распространенных наречий места в испанском языке (dentro – внутри, arriba – вверху, detrás – позади, fuera – снаружи, lejos – далеко).

Поисковый шаблон: **«\w*\s(\<(dentro|arriba|detrás| fuera|lejos)\>)».**

По поисковому шаблону нашлось: 174 конструкции (см. рисунок 2.18).

Полными самостоятельными именными словосочетаниями среди них являются 84 конструкции (producto dentro, debilidad fuera, inyección dentro, lesión dentro, medicamentos arriba).

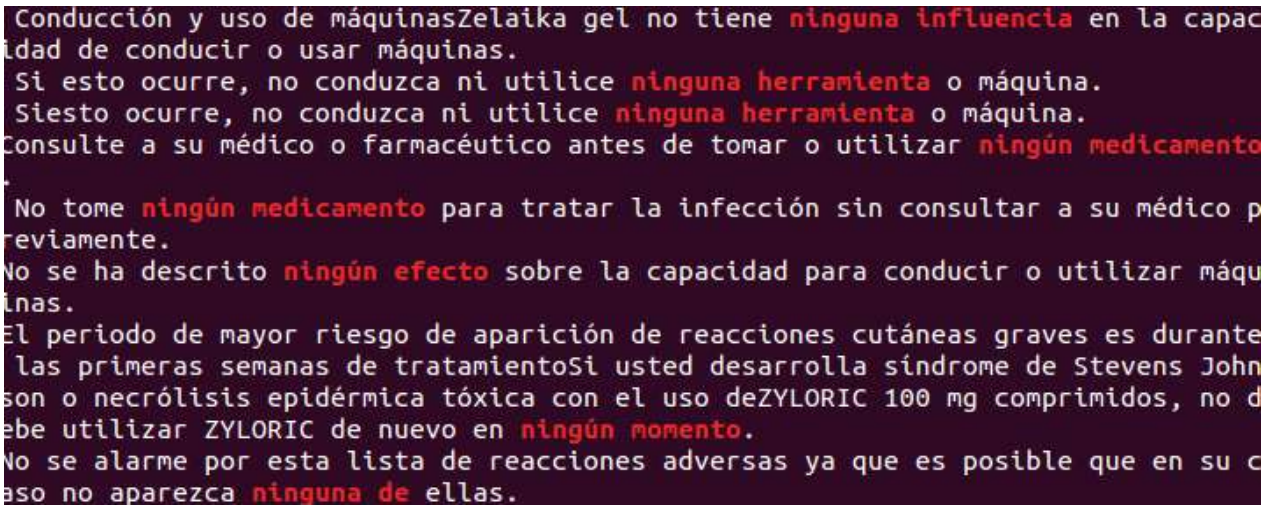


Рисунок 2.18 – Фрагмент работы поискового шаблона с отрицательными местоимениями

Также по шаблону были найдены наречия с союзом si (si fuera), с предлогом por (por dentro, por fuera), с наречием más (más arriba) и другие фрагменты конструкций (que fuera ,así fuera, o dentro).

Особенностью некоторых испанских наречий, является наличие суффикса -mente. Построим шаблон для поиса ИС с наречием на -mente: **«\w*(mente)\>\s\w*»**.

В корпусе было обнаружено 6179 конструкции с наречиями на –mente (см. рисунок 2.19). На 100 конструкций приходится всего 35 полноценных ИГ (наречие + существительное) (пример: esporádicamente niveles, recientemente alcohol, estrechamente relacionados, predominantemente hematomas, previamente pensamientos, generalmente diarrea, habitualmente alcohol, conjuntamente medicamentos). Среди остальных были найдены сочетания наречия с артиклем (mensualmente el), с местоимениями (normalmente su), с предлогами и союзами (especialmente en, especialmente de , habitualmente y, únicamente para) и с другими служебными частями речи (recientemente otros, periódicamente ya, adicionalmente si).

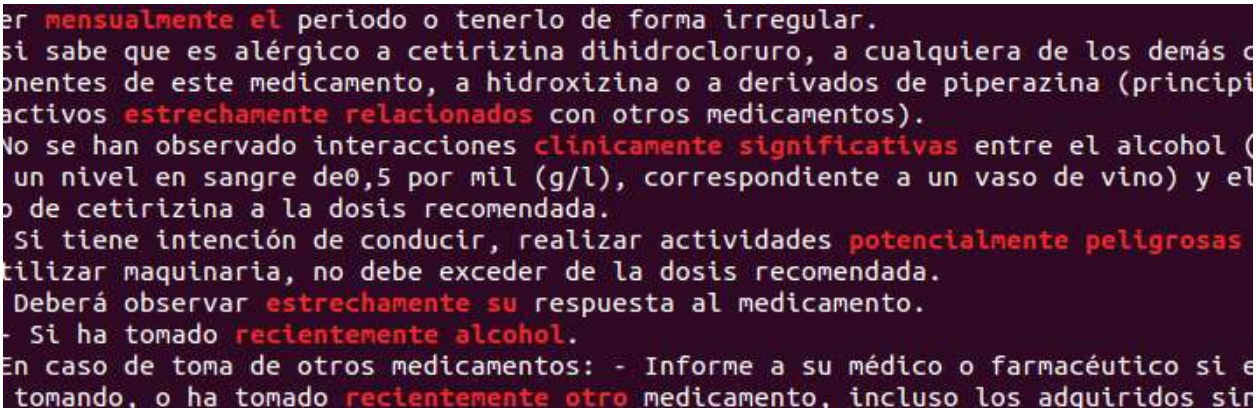


Рисунок 2.19 – Поиск ИГ с наречиями на –mente в корпусе

2.7 Автоматизированный поиск прономинальных словосочетаний

Особенностью фармацевтических рецептов является употребление вежливых местоимений. Рецепты не изобилируют ИГ содержащие прилагательные.

Проверим наличие личных местоимений в обрабатываемом корпусе. Шаблон для поиска: «**(\<(yo|tú|él|ella|usted|nosotros|nosotras|vosotros|vosotras|ellos|ellas|ustedes)\>)\s\w***». В результате поиска были выявлены конструкции содержащие лишь местоимение usted (см. рисунок 2.20). Остальные местоимения можно исключить из последующих поисковых шаблонов.

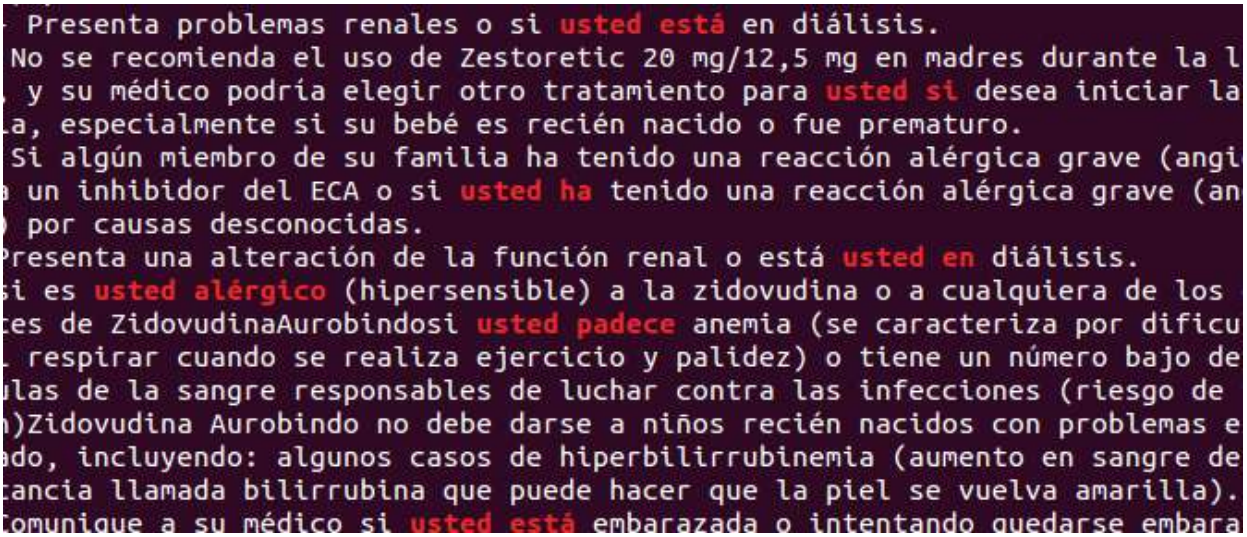


Рисунок 2.20 – Фрагмент поиска местоимений в корпусе

Анализ найденных конструкций, показал что для нашего корпуса характерно наличие прилагательных, описывающих состояние больного. Данные прилагательные оканчиваются одинаково, что можно использовать

для дальнейшего их нахождения в корпусе (usted embaraz**ada**, usted alérg**ico,** usted anc**iano**) (см. рисунок 2.21). Данная особенность поможет отделить прилагательные от других частей речи, которые следуют за usted и обнаружить полную ИГ. Кроме того, в корпусе также содержатся ИГ содержащие в себе прилагательное mayor (usted mayor – вы старший (старше).

Рисунок 2.21 – Фрагмент поиска сочетания с личным местоимением usted

Шаблон для поиска прономинальных словосочетаний с прилагательными: **«(\<(usted|ustedes)\>)\s(\w*((ico|ada)\>)|mayor)»**. В корпусе нашлась 131 конструкция (см. рисунок 2.22). Все из них являются полноценными прономинальными словосочетаниями (usted alérgico, usted diabético, usted epiléptico, usted asmático, usted embarazada , usted mayor).

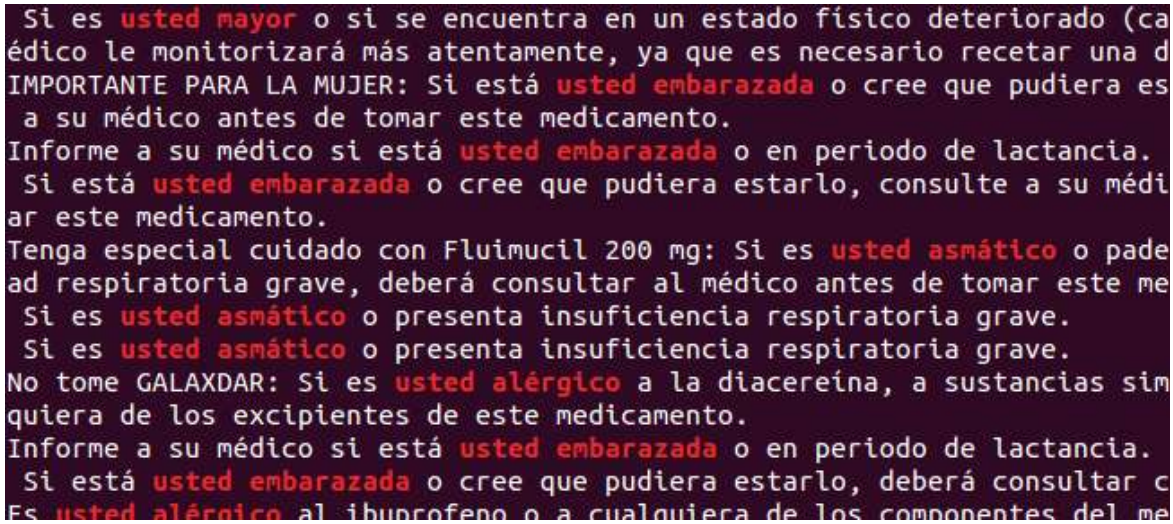


Рисунок 2.22 – Фрагмент поиска словосочетания с прилагательными

2.8 Автоматизированный поиск адъективных словосочетаний

Для поиска прилагательного в сравнительной степени, с наречием más (более) мы использовали шаблон: **'(\<(mas)\>)\s\w*'**.

Как оказалось, в корпусе содержится 37 конструкций, удовлетворяющих поисковому шаблону. Большинство из них построено по принципу наречие más и существительное (más datos, más informacíon, más picor), глагол (más ser), с предлогом «de» (más de 2 semanas, más de lo normal) (см. рисунок 2.23). Настоящими двухсоставными ИГ среди выявленных конструкций является 24 элемента, среди них: más eficaz, más definitivos, más sensible, más complejas, más despierto, más difícil, más baja, más adecuado, más fáciles, más altas. Точность поискового шаблона можно оценить в 0,64.

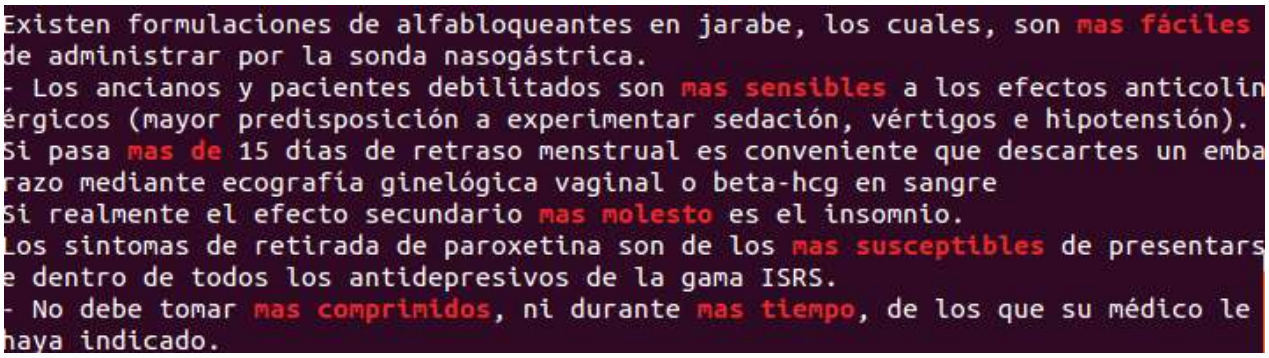


Рисунок 2.23 – Фрагмент поиска адъективного словосочетания с наречием más

Особенностью собранного фармацевтического корпуса рецептов

является не распространенность имен прилагательных и отсутствие адъективных именных групп.

2.9 Алгоритм постобработки для повышения точности

Большинство поисковых шаблонов имеют низкую точность определения ИГ. Это связанно с тем, что предварительно не были составлены списки стоп-слов, включающие закрытые части речи, союзы, артикли и предлоги. Составление поисковых шаблонов проводилось без учета глагольных форм (отсутствовали инициальный и финальный список), и признаков вхождения искомого элемента в состав более крупной именной группы (напр. Наличие предлога «de» или союза «a»).

Имея в наличии извлеченные фрагменты текстов из обрабатываемого корпуса с искомыми конструкциями, можно провести постанализ и избавиться от нерелевантных частей речи, неудовлетворяющих структуре ИГ, без предварительно составленных списков стоп-слов. Для проведения такой обработки необходимо наличие интернета на ЭВМ и API любой электронный словарной базы на оригинальном языке корпуса с тегами (с пометами частей речи) (см. рисунок 2.24).

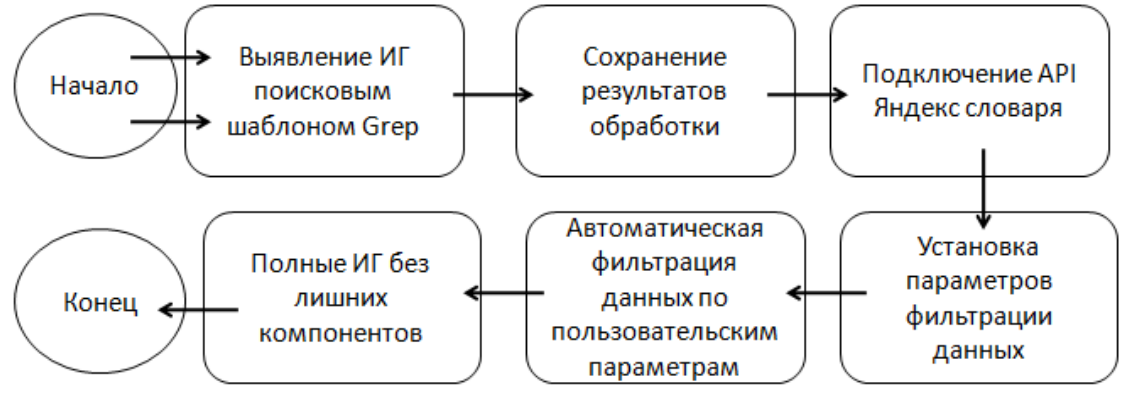


Рисунок 2.24 – Блок-схема алгоритма постобработки текста

В приложении со внутренним кодом API посылается запрос на поиск в базе зависимого компонента ИГ и установление его части речи. Предварительно необходимо задать определенные параметры для анализируемого зависимого компонента (указать определенную часть речи). После анализа, скрипт выдает готовый обработанный зависимый

компонент будущей ИГ, полностью удовлетворяющий первоначально заявленному поисковому параметру. Например, после поиска по шаблону: $ grep -E '\<(el|la|los|las|un|una|unos|unas)\>\s\w*\s\w*', было обнаружено, что существительное следующее за артиклем может иметь за собой глагол (el antibiótico deberá, las quinolonas pueden, el médico realizará, el paciente sufría), что непосредственно указывает на выявление грамматической основы, а не именную группу. После обращения к скрипту с API и установкой запрета на глагол, данная выявленная конструкция будет удалена из текста и на выходе получены полноценные именные словосочетания без нарушения структуры. Данная процедура будет проведена и со всеми закрытыми частями речи. Метод также может быть применен для распознавания ИГ определенной структуры при задании определенных параметров.

Выводы по главе 2

В главе был представлен детерминированный конечный автомат, который позволяет рассматривать различные варианты алгоритмов для построения правил по извлечению именных словосочетаний.

В ходе работы нами были созданы 14 правил и 24 шаблона в терминах регулярных выражений для извлечения именных словосочетаний. На экспертном корпусе мы апробировали составленную базу знаний шаблонов и лексических единиц. В ходе апробации точность реализованной модели составила 67%, а полнота 44%, результаты представлены в форме таблицы в приложении 7.

Для повышения точности мы предложили алгоритм с возможностью фильтрации и последующей обработки выходного материала.

ЗАКЛЮЧЕНИЕ

В данной дипломной работе были рассмотрены различные лингвистические подходы в определении сущности «словосочетания», которые показывают неоднозначность понимания синтаксической группы как отдельной синтаксической единицы.

Словосочетание является достаточно сложным объектом для машинного анализа в испанском языке.

В ходе исследования, мы выяснили, что рационалистический подход (подход основанный на правилах с использованием регулярных выражений) не позволяет создать точную модель поиска и выделения именных словосочетаний в корпусах испанского языка. В-основном, это обусловлено флективностью языка и особенностями выбранного метода.

Основываясь на конечном детерминированном автомате появляется возможность создания различных алгоритмов и поисковых вариантов для выявления именных словосочетаний в текстовых корпусах.

В ходе практической работы нами были выявлены 14 правил, составленных с опорой на классификацию П.А. Леканта, для извлечения именных словосочетаний.

В ходе работы мы выяснили, что большинство поисковых шаблонов имеют низкую точность определения именных словосочетаний. Это связанно с тем, что предварительно нами не были составлены списки стоп-слов (закрытые части речи, союзы, артикли и предлоги), списки глагольных форм (отсутствовали инициальный и финальный список), не учтены признаки вхождения искомого элемента в состав более крупной именной группы (напр. Наличие предлога «de» или союза «a»).

В ходе работы была разработана модель для извлечения именных словосочетаний, включающая в себя 11 правил и 24 шаблона в терминах регулярных выражений.

Модель реализована в форме базы знаний, состоящей из правил, шаблонов и списка лексических единиц. В ходе апробации мы получили достаточно высокие показатели точности, которые указывают на то, что при

составлении шаблонов была максимальна учтена структура именного словосочетания, а также грамматические показатели стержневого компонента. Мы имеем средний уровень полноты, который указывает на то, что выбранные варианты шаблонов не покрывают всевозможные вариации именных словосочетаний и требуют дальнейшей разработки и пополнения базы знаний.

При обработке достаточно весомого корпуса была достигнута высокая скорость извлечения, что указывает на возможность обработки массивных корпусов данных на современных компьютерных устройствах даже вне лаборатории.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Маслов, Ю.С. Введение в языкознание [Текст] / Ю.С. Маслов. – Москва: Изд-во Высшая школа, 1987. – 272 с.

2. Виноградов, В. В. Русский язык [Текст] / В. В. Виноградов. – Москва: Изд-во Высшая школа, 1972. – 478 с.

3. Noreen, A. O. О словах и классах слов [Текст] / A. O. Noreen. – Nordisk Tidskrift, 1879. – 136 с.

4. Сепир, Э. Язык. Введение в изучение речи [Текст] / Э. Сепир. – Москва: Изд-во Прогресс, 1993. – 656 с.

5. Ганеев, Б. Т. Язык [Текст] / Б. Т. Ганеев. – Москва: Изд-во Академия, 2004. – 368 с.

6. Ярцева, В. Н. Лингвистический энциклопедический словарь [Текст] / В. Н. Ярцева. – Москва: Изд-во Советская энциклопедия, 1990. – 342 с.

7. Ефремова, Т. Ф. Современный словарь русского языка три в одном: орфографический, словообразовательный, морфемный: около 20 000 слов, около 1200 словообразовательных единиц [Текст] / Т. Ф. Ефремова. – Москва: Изд-во АСТ, 2010. – 699 с.

8. Ушаков, Д. Н. Электронное издание. Толковый словарь русского языка Ушакова [Текст] / Д. Н. Ушаков. – Москва: Изд-во ЭТС, 1999. – 734 с.

9. Смирницкий, А. И. Лексикология английского языка [Текст] / А. И. Смирницкий. – Москва: Изд-во Московский Государственный Университет, 1956. – 260 с.

10. Слово [Электронный ресурс] – URL: http://tapemark.narod.ru/les/464c.html (дата обращения: 05.11.2017). – Загл. с экрана.

11. Виноградов, В. В. О формах слова [Текст] / В. В. Виноградов. – СССР: Изд-во Язык, 1944. – 248 с.

12. Шмелёв, Д. Н. Проблемы семантического анализа лексики [Текст] / Д. Н. Шмелёв. – Москва: Изд-во Аскмо, 1973. – 62 с.

13. Вандриес, Ж. Язык. Лингвистическое введение в историю [Текст] / Ж. Вандриес. – Москва: Изд-во Высшая школа, 1937. – 178 с.

14. Ярцева, В. Н. Лингвистический энциклопедический словарь [Текст] / В. Н. Ярцева. – Москва: Изд-во Советская энциклопедия, 1990. – 342 с.

15. Зализняк, А. А. Русское именное словоизменение с приложением избранных работ по современному русскому языку и общему языкознанию [Текст] / А. А Зализняк. – Москва: Изд-во Языки славянской культуры, 2002. – 372 с.

16. Шмелёв, Д. Н. Избранные труды по русскому языку [Текст] / Д. Н. Шмелёв. – Москва: Изд-во Академия, 2008. – 154 с.

17. Стернин, И. А. Лексическое значение слова и его компоненты [Текст] / И. А. Стернин. – Воронеж: Изд-во Воронежский университет, 1985. – 137 с.

18. Шмелёв, Д. Н. Избранные труды по русскому языку [Текст] / Д. Н. Шмелёв. – Москва: Изд-во Аскмо, 1973. –184 с.

19. Фортунатов, Ф.Ф. О преподавании грамматики русского языка в средней школе [Текст] / Ф.Ф. Фортунатов // Избранные труды. – Т.2. – Москва: Изд-во Высшая школа, 1957. – 247 с.

20. Жеребило, Т. В. Словарь лингвистических терминов [Текст] / Т. В. Жеребило. – Назрань: Изд-во Пилигрим, 2010. – 293 с.

21. Шахматов, А. А. Синтаксис русского языка [Текст] / А. А. Шахматов. – Л., 1941. – 214 с.

22. Виноградов, В. В. Русский язык [Текст] / В. В. Виноградов. – 3-е изд. – М., 1986. – 720 с.

23. Белошапкова, В. А. Современный русский язык [Текст] / В. А. Белошапкова. – М.: Высш.шк., 1989. – 800 с.

24. Тестелец, Я. Г. Введение в общий синтаксис [Текст] / Я. Г. Тестелец. – М., 2001. – 798 с.

25. Розенталь, Д. Э Словарь-справочник лингвистических терминов [Текст] / Д. Э. Розенталь, М. А. Теленкова. – Москва: Изд-во Просвещение, 1976. – 457 с.

26. Ярцева, В. Н. Лингвистический энциклопедический словарь [Текст] / В. Н. Ярцева. – Москва: Изд-во Советская энциклопедия, 1990. – 342 с.

27. Тестелец, Я. Г. Введение в общий синтаксис [Текст] / Я. Г. Тестелец. – М., 2001. – 798 с.

28. Золотова, Г.А. Очерк функционального синтаксиса русского язык [Текст] / Г. А. Золотова. – Москва: Изд-во Наука, 1973. – 32 с.

29. Мельчук, И. А. Курс общей морфологии [Текст] / И. А. Мельчук. – Т.2. – Вена: Изд-во Прогресс 1997. – 339 с.

30. Белошапкова, В. А. Современный русский язык [Текст] / В. А. Белошапкова. – Москва: Изд-во Высш.шк., 1989. – 800 с.

31. Левитан, К. М. Юридический перевод: основы теории и практики. [Текст]: учебное пособие / К. М. Левитан. – Москва: Изд-во Проспект, 2005. – 103 с.

32. Синтаксическая связь в словосочетании [Электронный ресурс] – URL: http://koi.tspu.ru/koi_books/kurysheva3/ccvc.htm (дата обращения: 15.04.2018). – Загл. с экрана.

33. Шуба, П. П. Современный русский язык. Синтаксис. Пунктуация. Стилистика. [Текст]: учебное пособие / П. П. Шуба. – Минск: Изд-во Поппури, 1998. – 68 с.

34. Шведова, Н. Ю. Грамматика современного русского литературного языка [Текст] / Н. Ю. Шведова. – Москва: Изд-во Наука, 1970. – 478 с.

35. Валгина, Н. С. Современный русский язык [Текст]: учебное пособие / Н. С. Валгина, Д. Э. Розенталь, М. И. Фомина. – Москва: Изд-во Логос, 2002. – 432 с.

36. Казанцева, Я. Н. Теоретическая грамматика английского языка [Текст]: учебное пособие / Я. Н. Казанцева, Н. В. Немчинова, Е. В. Семенова. – Красноярск: Изд-во Сибирский федеральный университет, 2015. – 135 с.

37. Розенталь, Д. Э. Словарь-справочник лингвистических терминов [Текст] / Д. Э. Розенталь, М. А. Теленкова. – Москва: Изд-во Просвещение, 1976. – 457 с.

38. Семантическая сочетаемость [Электронный ресурс] – URL: http://studbooks.net/777372/literatura/semanticheskaya_sochetaemost (дата обращения: 20.01.2018). – Загл. с экрана.

39. Морковин, В. В. Основы теории учебной лексикографии [Текст] / В.В. Морковин. –Смоленск: Изд-во Концепт, 1990. – 169 с.

40. Апресян, Ю. Д. Лексическая семантика [Текст] / Ю. Д. Апресян. – Москва: Изд-во Восточная литература, 1974. – 287 с.

41. Жеребило, Т. В. Словарь лингвистических терминов [Текст] / Т.В.Жеребило. – Назрань: Изд-во Пилигрим, 2010. –593 с.

42. Гуренко, В. В. Введение в теорию автоматов [Текст] / В. В. Гуренко. – Москва: Изд-во МГТУ им. Н. Э. Баумана, 2013. –154 с.

43. Форта, Б. Регулярные выражения за 10 минут [Текст] / Б. Форта. – Москва: Изд-во Вильямс, 2017. –184 с.

44. Кубрякова, Е. С. Основы морфологического анализа [Текст] / Е. С. Кубрякова – Москва: Изд-во Наука, 1974. –136 с.

45. Клышинский, Э. С. Начальные этапы анализа текста [Текст]: учебное пособие / Э. С. Клышинский. – Киев: Изд-во Вища шк., 1983. – 112 с.

46. Захаров, В. П. Корпусная лингвистика [Текст]: учебное пособие / В. П. Захаров. – Иркутск: Изд-во ИГЛУ, 2005. – 160 с.

47. Sahlgren, M. The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces [Текст] – M. Sahlegen. – Stockholm: Изд-во Stockholm University Department of Linguistics, 2006. – 156 с.

48. Abney, S. Parsing by chunks. Principle-based parsing [Текст] / S. Abney. – Kluwer: Изд-во Academic Publishers, 1991. – 278 с.

49. Schütze, H. Automatic word sense discrimination. Computational Linguistics [Текст] / H. Schütze – USA: Изд-во MIT Press Cambridge, 1998. – 123 с.

50. Salton, G. Introduction to modern information retrieval [Текст] / G. Salton, M. McGill – New York: Изд-во McGraw-Hill, 1986. – 448 с.

ПРИЛОЖЕНИЕ 1

Регулярные выражения

Якоря

^	Начало строки +
\A	Начало текста +
$	Конец строки +
\Z	Конец текста +
\b	Граница слова +
\B	Не граница слова +
\<	Начало слова
\>	Конец слова

Символьные классы

\c	Управляющий символ
\s	Пробел
\S	Не пробел
\d	Цифра
\D	Не цифра
\w	Слово
\W	Не слово
\xhh	Шестнадцатиричный символ hh
\Oxxx	Восьмиричный символ xxx

Символьные классы POSIX

[:upper:]	Буквы в верхнем регистре
[:lower:]	Буквы в нижнем регистре
[:alpha:]	Все буквы
[:alnum:]	Буквы и цифры
[:digit:]	Цифры
[:xdigit:]	Шестнадцатиричные цифры
[:punct:]	Пунктуация
[:blank:]	Пробел и табуляция
[:space:]	Пустые символы
[:cntrl:]	Управляющие символы
[:graph:]	Печатные символы
[:print:]	Печатные символы и пробелы
[:word:]	Буквы, цифры и подчеркивание

Утверждения

?=	Вперед смотрящее +
?!	Отрицательное вперед смотрящее +
?<=	Назад смотрящее +
?!= или ?	Отрицательное назад смотрящее +
?>	Однократное подвыражение
?()	Условие [если, то]
?()|	Условие [если, то, а иначе]
?#	Комментарий

Примечание *Отмеченное + работает в большинстве языков программирования.*

Образцы шаблонов

([A-Za-z0-9-]+)	Буквы, числа и знаки переноса
(\d{1,2}\/\d{1,2}\/\d{4})	Дата (напр., 21/3/2006)
([^\s]+(?=\.(jpg|gif|png))\.\2)	Имя файла jpg, gif или png
(^[1-9]{1}$\|^[1-4]{1}[0-9]{1}$|^50$)	Любое число от 1 до 50 включительно
(#?([A-Fa-f0-9]){3}(([A-Fa-f0-9]){3})?)	Шестнадцатиричный код цвета
((?=.*\d)(?=.*[a-z])(?=.*[A-Z]).{8,15})	От 8 до 15 символов с минимум одной цифрой, одной заглавной и одной строчной буквой (полезно для паролей).
(\w+@[a-zA-Z_]+?\.[a-zA-Z]{2,6})	Адрес email
(\<(/?[^\>]+)\>)	HTML теги

Примечание *Эти шаблоны предназначены для ознакомительных целей и основательно не проверялись. Используйте их с осторожностью и предварительно тестируйте.*

Кванторы

*	0 или больше +
*?	0 или больше, нежадный +
+	1 или больше +
+?	1 или больше, нежадный +
?	0 или 1 +
??	0 или 1, нежадный +
{3}	Ровно 3 +
{3,}	3 или больше +
{3,5}	3, 4 или 5 +
{3,5}?	3, 4 или 5, нежадный +

Специальные символы

\	Экранирующий символ +
\n	Новая строка +
\r	Возврат каретки +
\t	Табуляция +
\v	Вертикальная табуляция +
\f	Новая страница +
\a	Звуковой сигнал
[\b]	Возврат на один символ
\e	Escape-символ
\N{name}	Именованый символ

Подстановка строк

$n	n-ая непассивная группа
$2	«xyz» в /^(abc(xyz))$/
$1	«xyz» в /^(?:abc)(xyz)$/
$`	Перед найденной строкой
$'	После найденной строки
$+	Последняя найденная строка
$&	Найденная строка целиком
$_	Исходный текст целиком
$$	Символ «$»

Диапазоны

.	Любой символ, кроме переноса строки (\n) +
(a|b)	a или b +
(...)	Группа +
(?:...)	Пассивная группа +
[abc]	Диапазон (a или b или c) +
[^abc]	Не a, не b и не c +
[a-q]	Буква между a и q +
[A-Q]	Буква в верхнем регистре между A и Q +
[0-7]	Цифра между 0 и 7 +
\n	n-ая группа/подшаблон +

Примечание *Диапазоны включают граничные значения.*

Модификаторы шаблонов

g	Глобальный поиск
i	Регистронезависимый шаблон
m	Многострочный текст
s	Считать текст одной строкой
x	Разрешить комментарии и пробелы в шаблоне
e	Выполнение подстановки
U	Нежадный шаблон

Мета-символы (экранируются)

^	[	.
$	{	*
(	\	+
)	|	?
<	>	

Эта таблица доступна на www.exlab.net
Англоязычный оригинал на AddedBytes.com

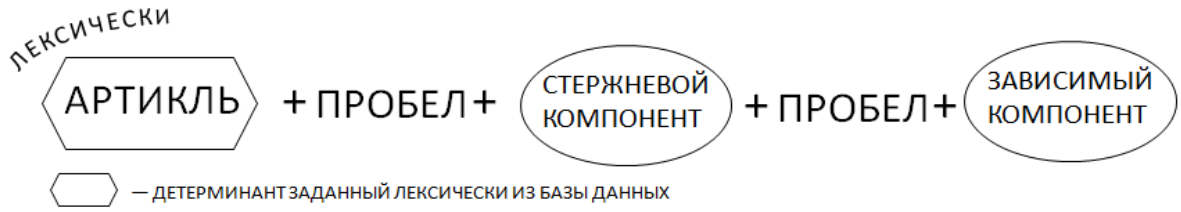
ПРИЛОЖЕНИЕ 2

Правила для субстантивных словосочетаний

1) Для двухкомпонентного словосочетания с зависимым числительным

2) Для двухкомпонентного словосочетания с зависимым местоимением

3) Двухкомпонентное словосочетание с детерминантом, определяющим стержневой компонент выраженный именем существительным

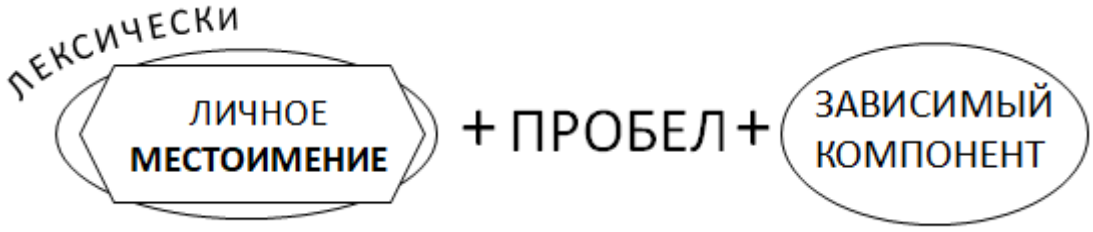


4) Двухкомпонентное словосочетание с детерминантом и предложной группой

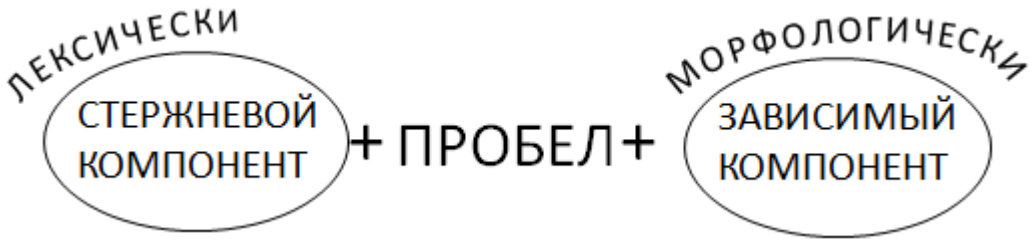
5) Двухкомпонентное словосочетание, заданное на основе морфологического признака

Правила для прономинальных словосочетаний

6) Двухкомпонентное словосочетание с личным местоимением

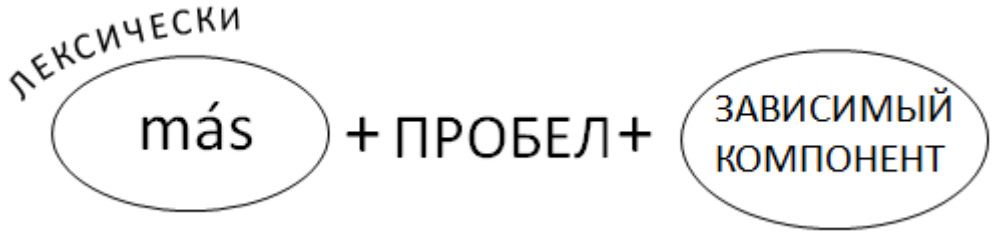


7) Двухкомпонентное словосочетание с выраженными морфологическими признаками у прилагательного (зависимого компонента)



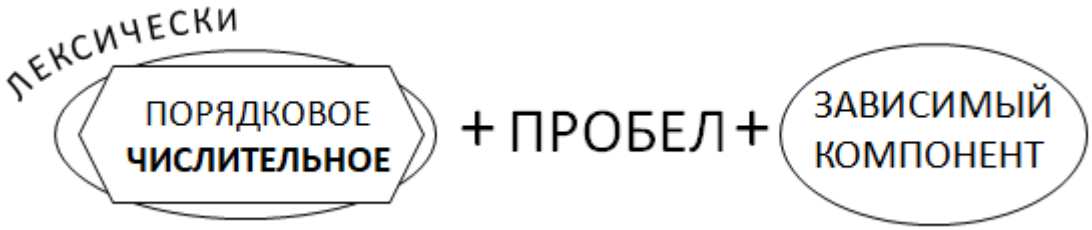
Правила для адъективных словосочетаний

8) Двухкомпонентное словосочетание с прилагательным в сравнительной степени

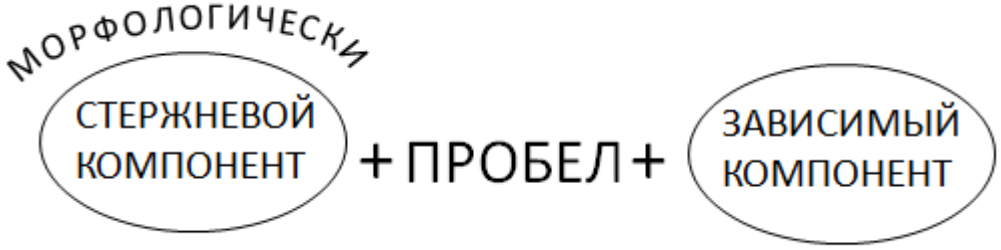


Правила для нумеративных словосочетаний

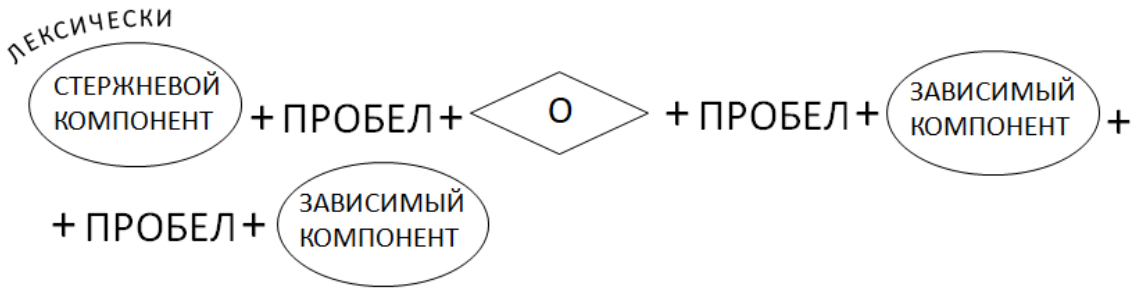
9) Для ИС, где стрежневой компонент может задаваться лексически



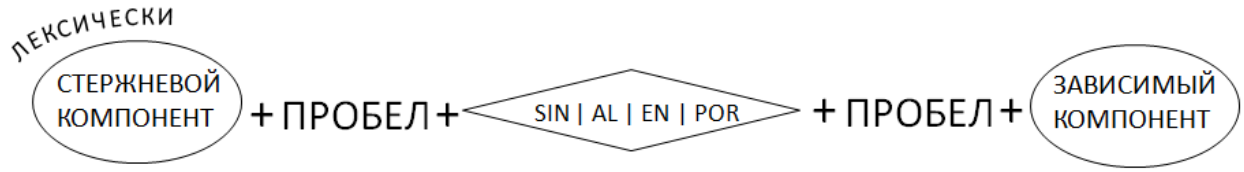
10) Для сложных порядковых числительных с одинаковыми уникальными приставками (числительные от 20 до 99)



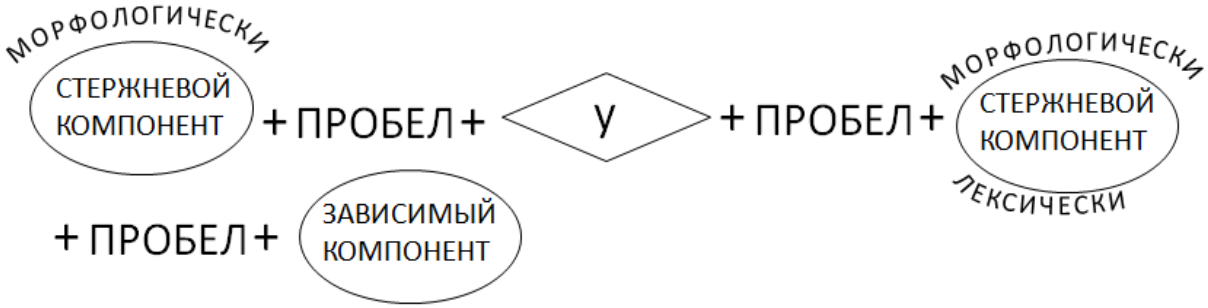
11) Для расширенной 3-х компонентной ИГ с союзом «о»



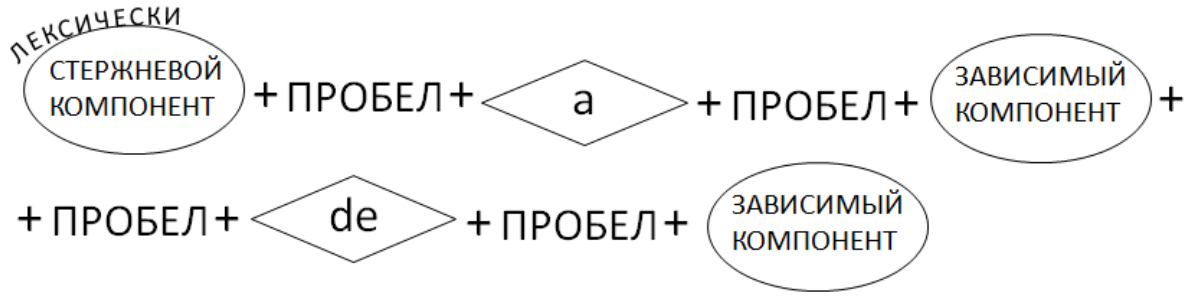
12) Для 2-х компонентного предложного словосочетания



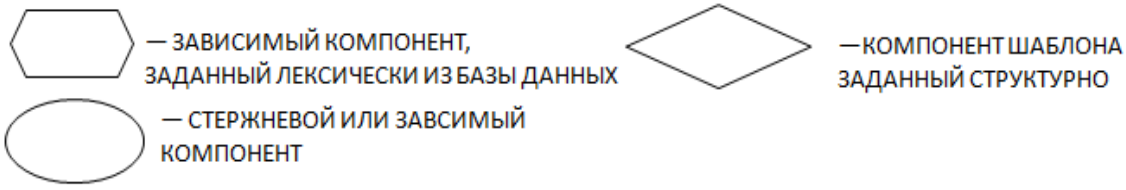
13) Для 3-х компонентных нумеративных сочетаний с предлогом «у» при согласовании по роду



14) Для расширенного ИС с предлогом «a» и «de» перед зависимым компонентом



Условные обозначения



ПРИЛОЖЕНИЕ 3

Таблица 1 – Перечень артиклей

	Неопределенный артикль		Определенный артикль	
	Единственное число	Множественное число	Единственное число	Множественное число
Мужской род	un	unos	el	los
Женский род	una	unas	la	las

Таблица 2 – Формы личных местоимений и соответствующих им притяжательных местоимений

Единственное число			Множественное число		
yo	tu	èl ella usted	nosotros(as)	vosotros(as)	ellos ellas ustedes
Мужской род					
(el) mìo	(el) tuyo	(el) suyo	(el) nuestro	(el) vuestro	(el) suyo
Женский род					
(la) mìa	(la) tuya	(la) suya	(la) nuestra	(la) vuestra	(la) suya
Мужской род					
(los) mìos	(los) tuyos	(los) suyos	(los) nuestros	(los) vuestros	(los) suyos
Женский род					
(las) mìas	(las) tuyas	(las) suyas	(las) nuestras	(las) vuestras	(las) suyas

Таблица 3 – Формы указательных местоимений в испанском языке

Число	Род	este (этот)	ese (этот)	aquel (тот)
Единственное	мужской	este	ese	aquel
	женский	esta	esa	aquella
	средний	esto	eso	aquello
Множественное	мужской	estos	esos	aquellos
	женский	estas	esas	aquellas

Таблица 4 – Относительные местоимения в испанском языке

единственное число	множественное число	Перевод (русский)
que	–	который(-ая, -ое)
quien	quienes	кто, который (-ая)
cual	cuales	тот/та или который/которая
cuyo/cuya	cuyos/cuyas	чей, который
cuanto/cuanta	cuantos/cuantas	столько, сколько

Таблица 5 – Отрицательные местоимения в испанском языке

единственное число	множественное число
ningún/ninguna	ningunos/ningunas
nada	–
nadie	–

ПРИЛОЖЕНИЕ 4

Испанские порядковые числительные

primer(o), -a, -os, -as (primo, -a, -os, -as)	1-й, -я, -е, (-е)
segundo, -a, -os, -as	2-й, -я, -е, (-е)
tercer(o), -a, -os, -as (tercio, -a, -os, -as)	3-й, -я, -и, (-е)
cuarto, -a, -os, -as	4-й, -я, -е, (-е)
quinto, -a, -os, -as	5-й, -я, -е, (-е)
sexto, a, os,-as	6-й, -я, -е, (-е)
séptimo, -a, -os, -as	7-й, -я, -е, (-е)
octavo, -a, -os, -as	8-й, -я, -е, (-е)
noveno, -a, -os, -as (nono, -a, -os, -as)	9-й, -я, -е, (-е)
décimo, -a, -os, -as	10-й, -я, -е, (-е)

undécimo, -a, -os, -as	11-й, -я, -е, (-е)
duodécimo, -a, -os, -as	12-й, -я, -е, (-е)
decimotercero, -a, -os, -as (decimotercio, -a, -os, -as)	13-й, -я, -е, (-е)
decimocuarto, -a, -os, -as	14-й, -я, -е, (-е)
decimoquinto, -a, -os, -as	15-й, -я, -е, (-е)
decimosexto, -a, -os, -as	16-й, -я, -е, (-е)
decimoséptimo, -a, -os, -as	17-й, -я, -е, (-е)
decimoctavo, -a, -os, -as	18-й, -я, -е, (-е)
decimonoveno, -a, -os, -as (decimonono, -a, -os, -as)	19-й, -я, -е, (-е)
vigésimo, -a, -os, -as	20-й, -я, -е, (-е)

ПРИЛОЖЕНИЕ 5

Двузначные числительные в испанском языке

vigésimo primero	21-й
vigésimo segundo	22-й
vigésimo tercero (tercio), etc.	23-й и т. д.
trigésimo, -a, -os, -as	30-й, -я, -е, (-е)
trigésimo primero, etc.	31-й и т. д.
cuadragésimo, -a, -os, -as	40-й, -я, -е, (-е)
quincuagésimo, -a, -os, -as	50-й, -я, -е, (-е)
sexagésimo, -a, -os, -as	60-й, -я, -е, (-е)
septuagésimo, -a, -os, -as	70-й, -я, -е, (-е)
octogésimo, -a, -os, -as	80-й, -я, -е, (-е)
nonagésimo, -a, -os, -as	90-й, -я, -е, (-е)

centésimo, -a, -os, -as	100-й, -я, -е, (-е)
centésimo primero, etc.	101-й и т. д.
ducentésimo, -a, -os, -as	200-й, -я, -е, (-е)
tricentésimo, -a, -os, -as	300-й, -я, -е, (-е)
cuadringentésimo, -a, -os, -as	400-й, -я, -е, (-е)
quingentésimo, -a, -os, -as	500-й, -я, -е, (-е)
sexcentésimo, -a, -os, -as	600-й, -я, -е, (-е)
septingentésimo, -a, -os, -as	700-й, -я, -е, (-е)
octingentésimo, -a, -os, -as	800-й, -я, -е, (-е)
noningentésimo, -a, -os, -as (nongentésimo, -a, -os, -as)	900-й, -я, -е, -(ое)
milésimo, -a, -os, -as	1000-й, -я, -е, (-е)
milésimo primero, etc.	1001-й и т. д.
millonésimo, -a, -os, -as	миллионный, -я, -е, (-е)

ПРИЛОЖЕНИЕ 6

Количественные числительные в испанском языке

0	cero
1	uno
2	dos
3	tres
4	cuarto
5	cinco
6	seis
7	siete
8	ocho
9	nueve
10	diez

11	once
12	doce
13	trece
14	catorce
15	quince
16	dieciséis, diez y seis
17	diecisiete, diez y siete
18	dieciocho, diez y ocho
19	diecinueve, diez y nueve
20	veinte

21	veintiuno, veinte y uno
22	ventidós, veinte y dos
23	veintitrés, veinte y tres
24	veinticuatro, veinte y cuatro

10	diez
20	veinte
30	treinta
40	cuarenta
50	cincuenta
60	sesenta
70	setenta
80	ochenta
90	noventa

100	cien, ciento
101	ciento uno
200	doscientos
300	trescientos
400	cuatrocientos
500	quinientos
600	seiscientos
700	setecientos
800	ochentos
900	novecientos

1000	mil
2000	dos mil
1000000	un millón
2000000	dos millones
1000000000	mil millones
2000000000	dos mil millones
1116	mil ciento diez y seis
2217	dos mil doscientos diecisiete
23358	veititrés mil trescientos cincuenta y ocho

ПРИЛОЖЕНИЕ 7

Шаблоны по группам	Точность	Полнота
Нумеративные	81,5	33,6
\<(dos|tres|cuatro|cinco|siete|ocho|nueve|diez|once|doce|trece|catorce|quince|veinte)\>\s\w*|\<(dieci*|veinti*)\w*\s\w*	0,87	0,69
\<(dos|tres|cuatro|cinco|siete|ocho|nueve|diez|once|doce|trece|catorce|quince|veinte)\>\so\s\w*\s\w*|\<(dieci*|veinti*)\so\s\w*\s\w*	1	0,20
\<(dos|tres|cuatro|cinco|siete|ocho|nueve|diez|once|doce|trece|catorce|quince)\>\so\s\w*\s\w*\sde{0,1}\s\w*|\<(dieci*|veinti*)\so\s\w*\s\w*\sde{0,1}\s\w*	1	0,03
\<(dos|tres|cuatro|cinco|siete|ocho|nueve|diez|once|doce|trece|catorce|quince|veinte)\>\s(sin|al|en|por)\s\ w*	1	0,05
\<(dos|tres|cuatro|cinco|siete|ocho|nueve|diez|once|doce|trece|catorce|quince|veinte)\>\sa\s\w*\s\w*	1	0,66
\<(dos|tresdos|tres|cuatro|cinco|siete|ocho|nueve|diez|once|doce|trece|catorce|quince|veinte)\>\sa\s\w*\s\w*\sde\s\w*\s\w*	1	0,28
\w*(cientos|cientas)\sy\s\w*\s\w*	1	0,1
\w*(cientos|cientas)\s\w*\sy\s\w*\s\w*.	0	0

\Wcien\s\w*	1	1
\Wmil|millónes\W\s\w*	0,28	0,35
Субстантивные	0,66	0,42
\<(el|la|los|las|un|una|unos|unas)\>\s\w*\s\w*	0,62	0,12
\<(el|la|los|las|un|una|unos|unas)\>\s\w*\s(a|o|y)\s\w*\s\w*	0,71	0,34
\<(el|la|los|las|un|una|unos|unas)\>\s\w*\s(de|con)\s\<(el|la|los|las|un|una|unos|unas)\>\s\w*	0,77	0,52
\<(primer|segund|tercer|cuart|quint)\s\w*	0,80	0,29
\<(mis?|tus?|su?|nuestros?|nuestras?|vuestros?|vuestras?)\>\s\w*	0,9	0,58
\<(ese|esas?|esos|aquel|aquella|aquellos|aquellas)\>\s\w*	0,69	0,215
\w*\<(cuyo|cuya|cuyos|cuyas)\>\s\w*	0,68	0,93
\w*\<(algunos|algunas?|algún|todo|todas?|todos|otros?|otras|mismos?|mismas?|varies|varias|cualquiera?)\>\s\ w*	0,38	0,14
\w*\<(ningún|ningun(a|o)?)\>\s\w*	0,78	0,16
\w*\s(\<(dentro|arriba|detrás|fuera|lejos)\>)	0,48	0,84
\w*(mente)\>\s\w*	0,73	0,72
Прономинальные	1,01	1,07

(\<(yo|tú|él|ella|usted|nosotros|nosotras|vosotros|vosotras|ellos|ellas|ustedes)\>)\s\w	0,8	0,6
(\<(usted|ustedes)\>)\s(\w*((ico|ada)\>)|mayor)	0,21	0,47
Адъективные	0,37	0,24
(\<(mas)\>)\s\w*	0,37	0,24
Общее	67,49%	44,78%