Санкт-Петербургский государственный университет

Филологический факультет

Кафедра математической лингвистики

ИЗВЛЕЧЕНИЕ ТЕРМИНОВ АВТОМАТИЧЕСКИМИ МЕТОДАМИ
 (НА МАТЕРИАЛЕ ФИНСКИХ ТЕКСТОВ)

Направление: «Лингвистика»
Образовательная программа: «Прикладная и экспериментальная лингвистика»
Профиль: «Компьютерная лингвистика и интеллектуальные технологии»

Выпускная квалификационная работа

соискателя на степень магистра

филологии

Коноваловой Александры Владимировны

Научный руководитель:

канд. филол. наук, доц. Хохлова Мария Владимировна

Санкт-Петербург

2016

Содержание

[Введение 4](#_Toc451857234)

[Глава 1. Понятие термина в лингвистике 7](#_Toc451857235)

[1.1. Определение понятия «термин» 7](#_Toc451857236)

[1.2. Особенности термина. Требования к термину 8](#_Toc451857237)

[1.3. Происхождение терминов 12](#_Toc451857238)

[1.4. Классификации терминов 15](#_Toc451857239)

[1.5. Термин в финноязычной лингвистике 19](#_Toc451857240)

[1.6. Экономические термины 22](#_Toc451857241)

[1.7. Выводы 23](#_Toc451857242)

[Глава 2. Способы автоматического извлечения терминов 25](#_Toc451857243)

[2.1. Общая характеристика систем для автоматического извлечения терминов 25](#_Toc451857244)

[2.2. Классификация систем для автоматического извлечения терминов 27](#_Toc451857245)

[2.3.1. Онлайн-системы 29](#_Toc451857246)

[2.3.2. Программы, находящиеся в свободном доступе 32](#_Toc451857247)

[2.3.3. Программы с закрытым кодом 34](#_Toc451857248)

[2.4. Выводы 36](#_Toc451857249)

[Глава 3. Правила извлечения терминов 37](#_Toc451857250)

[3.1. Материал исследования 37](#_Toc451857251)

[3.2. Разметка текстов 38](#_Toc451857252)

[3.3. SketchEngine и грамматики SketchEngine 39](#_Toc451857253)

[3.3. Грамматика для извлечения терминов. 42](#_Toc451857254)

[3.4. Результаты 46](#_Toc451857255)

[3.4.1. Выделение терминологических словосочетаний 46](#_Toc451857256)

[3.4.2. Выделение терминов 47](#_Toc451857257)

[3.5. Ошибки при автоматическом выделении терминов 50](#_Toc451857258)

[3.6. Выводы 60](#_Toc451857259)

[Заключение 62](#_Toc451857260)

[Список использованной литературы 65](#_Toc451857261)

[Список источников 69](#_Toc451857262)

[Приложение 1. Грамматика SketchEngine для выделения терминов 71](#_Toc451857263)

Введение

Основными функциями языка являются коммуникативная (язык предназначен для взаимодействия людей), когнитивная (язык служит для передачи смысла) и аккумулятивная (язык обеспечивает сохранение и передачу знаний) [Жеребило, 2010, с. 442].

Терминология как часть специальной лексики языка обеспечивает осуществление аккумулятивной функции языка, позволяя передавать информацию как о конкретных объектах физического мира, имеющих особое назначение и определенные свойства и качества, так и об абстрактных идеях или понятиях. Изучением терминологии занимается терминоведение, в рамках которой специальные единицы языка анализируются «с точки зрения их типологии, происхождения, формы, содержания (значения) и функционирования, а также использования, упорядочения и создания» [Гринев-Гриневич, 2008, с. 9].

В последнее время в терминоведении все большую актуальность приобретает задача автоматического извлечения терминов в связи с огромными и, более того, постоянно растущими объемами информации [Браславский, Соколов]. Автоматическое выделение терминов может использоваться для создания терминологических словарей различных тематик, а также при переводе [Герд, 2014, с. 4]. Это определяет **практическую значимость** нашей работы.

Поскольку к терминам, в отличие от остальной лексики, предъявляются определенные требования (в частности, к форме и к значению термина, о которых мы будем говорить далее), автоматизация для их извлечения может быть применима с бо́льшим успехом, чем к остальной лексике языка. Согласно А. С. Герду, особенности терминов позволяют «заранее задать определенную модель параметров описания семантики терминов данного типа» [Герд, 2014, с. 4].

Тема извлечения терминов из текстов разрабатывалась еще с 1990х годов. Начало изучения ознаменовалось появлением в 1990 году TERMINO — первой широкой известной программы для извлечения терминов [Cabré, Estopà, Vivaldi, 2001, p. 2]. В дальнейшем исследователями применялись различные подходы, на их основе создавались алгоритмы, но и по сей день проблема автоматического извлечения терминов не потеряла своей **актуальности**.

**Объектом** нашего исследования являются термины финского языка (конкретнее — экономическая терминология, встречающаяся в газетных статьях), а **предметом** — автоматическое извлечение данных терминов из текстов. **Целью** нашего исследования является исследование терминов и терминологических словосочетаний в финском языке.

Для достижения данной цели нами были поставлены следующие **задачи**:

1. теоретическое описание терминов в финском языке (на материале созданного нами корпуса и корпуса *fiTenTen*);
2. изучение программного обеспечения для выделения терминов;
3. описание моделей терминов в рамках грамматики *SketchEngine;*
4. выявление выделенных терминов на основе разработанной грамматики из корпуса экономических текстов и последующий анализ результатов.

Мы использовали такие методы как корпусный анализ данных, статистические методы и подход с использованием правил (*rule-based approach*).

В качестве материала для нашего исследования были отобраны газетные статьи по тематике «Экономика». Объем корпуса составляет 50 тысяч токенов.

Для извлечения терминов нами была написана и протестирована грамматика *SketchEngine*.

В первой главе мы рассматриваем термин как лексическую единицу, его характеристики и особенности (в частности, морфологические особенности финского термина).

Во второй главе мы приводим краткий обзор существующих способов автоматического извлечения терминов. При рассмотрении программных средств мы ограничились только теми инструментами для выделения ключевых слов и терминов, которые позволяют выявлять именно термины как единицы специального языка.

В третьей главе мы описываем грамматику *SketchEngine*, с помощью которой нами проводилось извлечение терминов, и материал исследования, анализируем результаты исследования и ошибки, возникшие при автоматическом извлечении терминов.

Глава 1. Понятие термина в лингвистике

1.1. Определение понятия «термин»

Довольно часто в науке можно столкнуться с тем, что общеупотребительное понятие определяется разными учеными по-разному: в частности, понятия концепта в различных областях гуманитарных наук отличаются друг от друга. Даже в рамках одного направления нередки случаи, когда существует несколько определений одного и того же понятия.

 Так, в настоящее время нельзя говорить о едином определении понятия «термин» в лингвистике [Лейчик, 2007, с. 20]. Согласно С. Д. Шелову, «далеко не во всех работах, обсуждающих терминологические темы, предлагается дефиниция понятия “термин”…даже если они активно комментируют определения своих коллег», а некоторые авторы «предлагают свое собственное определение этого понятия без учета уже имеющихся дефиниций» [Шелов, 2010]. В данной главе мы приведем несколько существующих определений термина.

Согласно А. С. Герду термин представляет собой единицу какого-либо конкретного естественного или искусственного языка (слово или словосочетание), которая либо уже существовала ранее, либо была специально создана и которая обладает специальным терминологическим значением, выраженным в словесной форме или в каком-либо формализованном виде, и при этом достаточно полно отражает основные признаки существующего научного понятия [Герд, 1991, с. 1–4]. Специфичность термина как лексической единицы обусловлена не планом выражения, а планом содержания [Герд, 1986, с. 4–5]. Связь термина с мышлением осуществляется через понятие [Герд, 2005, с. 81].

С.  В. Гринев-Гриневич определяет термин как «номинативную специальную лексическую единицу (слово или словосочетание), принимаемую для точного наименования понятий». Это определение, по его мнению, содержит основные свойства и признаки термина**,** отличающие его от нетерминов: «обозначение понятия, принадлежность к специальной области знания, дефинированность, точность значения, контекстуальная независимость, конвенциональность и целенаправленный характер появления, устойчивость и воспроизводимость в речи, номинативность, стилистическая нейтральность» [Гринев-Гриневич, 2008, с. 30]. С. Д. Гринев-Гриневич также отрицает возможность того, чтобы термином являлась часть речи, отличная от существительного [Гринев-Гриневич, 2008, с. 29].

С. Д. Шелов формулирует следующее определение термина: термин – это «языковой знак (слово, словосочетание, сочетание слова или словосочетания с особыми символами), выражающий понятие какой-либо области знания и в силу этого имеющий дефиницию (толкование, объяснение), на которую сознательно ориентируются использующие этот языковой знак» [1Шелов, 2010].

В работе [Лантюхова, Загоровская, Литвинова, 2013] термин определяется как «слово или словосочетание, соотнесенное со специальным понятием, явлением или предметом в системе какой-либо области знания» [Лантюхова, Загоровская, Литвинова, 2013].

Во всех этих определениях подчеркивается то, что термин является либо словом, либо словосочетанием (распространенность того или другого типа термина в языке зависит от множества разных факторов, например, от типа стандартного языка или терминосистемы, в которую входит термин), его принадлежность к специальному языку и связь с каким-либо понятием. Таким образом, можно определить термин как единицу специального языка, обозначающую какое-либо понятие.

1.2. Особенности термина. Требования к термину

Лексика специальных языков преимущественно состоит из номинативных единиц — названий объектов и действий, относящихся к специальным областям, то есть денотатами таких языковых единиц являются специальные понятия — как абстрактные, так и конкретные [Лейчик, 2007, с. 76–77]. Таким образом, первой особенностью термина является его номинативность.

Второй особенностью термина является его принадлежность и к естественному языку, и к языку для специальных целей. Следовательно, термины могут служить связующим звеном между разными языками, обеспечивая переход от специального языка к естественному языку, поскольку их содержательная и формальная структура часто схожи в одинаковых терминосистемах разных языков, но значительно отличаются в различных терминосистемах одного и того же языка [Лейчик, 2007, с. 76–77].

К термину как знаковой единице предъявляются определенные требования, которые можно разделить на три группы в соответствии с тремя аспектами: требования к форме (синтаксический аспект), к значению (семантический аспект) и специфические требования, связанные с особенностями употребления термина (прагматический аспект) [Гринев-Гриневич, 2008, с. 30]. Тем не менее, необходимость некоторых из приведенных ниже критериев все еще является предметом дискуссий [Лантюхова, Загоровская, Литвинова, 2013].

К форме термина предъявляются следующие требования:

1. соответствие фонетическим, грамматическим и стилистическим нормам языка (данное требование скорее относится к терминам, заимствованным из другого языка);
2. краткость (избыточно длинный термин неудобен в использовании);
3. хорошая деривационная способность (чтобы образовывать новые термины от уже имеющихся, а не вводить другие слова для терминов, схожих по значению);
4. неизменность (инвариантность) [Гринев-Гриневич, 2008, с. 32–33].

Значение термина должно соответствовать следующим требованиям:

1) соответствие термина отражаемому им понятию [Лотте, 1961, с. 76] или непротиворечивость семантики термина;

2) однозначность термина в данной терминологии (невозможно требовать однозначности термина в языке, поскольку многие термины многозначны);

Данное требование представляет для нас особый интерес, поскольку в нашем исследовании при автоматическом выделении терминов мы не принимаем во внимание семантический аспект, возможны ошибки, связанные с неправильным выделением термина: алгоритм может выделить как термин слово, не являющееся термином в данном контексте, или наоборот.

3) полнозначность (значение термина должно отражать минимальное количество признаков, достаточных для идентификации обозначаемого им понятия);

4) отсутствие синонимов(тем не менее, синонимия — довольно частое явление в терминологиях, особенно между заимствованными и автохтонными терминами, неологизмами и устаревающими терминами) [Гринев-Гриневич, 2008, с. 31–32].

Среди прагматических требований выделяют следующие:

1) внедренность термина в язык, характеризующаяся общепринятостью или употребительностью термина специалистами;

2) интернациональность (в нескольких национальных языках термины совпадают или достаточно близки по форме и содержанию);

3) современность (терминосистема должна меняться вместе с языком и отвечать запросам тех, кто ей пользуется);

4) орфоэпическая и ассоциативная благозвучность термина (данное требование перекликается с требованием соответствия различным языковым нормам) [Гринев-Гриневич, 2008, с. 36].

Структуру термина можно разделить на содержательную и формальную. К содержательной структуре относятся:

1) собственно семантика;

2) языковая и терминологическая мотивированность;

3) сигнификативное значение (обозначение специального понятия или концепта, а также их признаков) [Лейчик, 2007, с. 32].

Формальная структура термина вызывает много вопросов, один из которых — проблема оптимальной длины термина [Лейчик, 2007, с. 50] — имеет особое значение для нашего исследования, проводимого на материале финских текстов. Как было указано выше, одним из требований к форме термина является краткость. Согласно В. М. Лейчику, целесообразнее говорить не о краткости термина, а о его оптимальной длине или оптимальном наборе терминоэлементов, из которых состоит термин. Таким образом, формальная структура термина будет связана с содержательной структурой, поскольку каждый терминоэлемент соотносится с понятием или же признаком понятия [Лейчик, 2007, с. 50].

Понятие терминоэлемента впервые было введено Д. С. Лотте в работе [Лотте, 1961]: «Под “терминоэлементами” понимаются слова, имеющие самостоятельное значение и входящие в состав простого или сложного термина» [Лотте, 1961, с. 15]. В дальнейшем под терминоэлементом Д. С. Лотте понимал «любой термин — словосочетание, или термин-слово, или, наконец, часть (частичку) термина-слова, не делимые в терминологически-смысловом отношении» [Лотте, 1961, с. 88].

По В. М. Лейчику, термин обладает идеальной длиной, если каждый его терминоэлемент обозначает одно понятие из системы понятий данной области знания. Длину можно обозначить как I + n (или n + I). Если термин является основным для данной терминосистемы, то n — количество этапов деления понятия — равняется нулю. Оптимальная длина термина, при которой возможно обозначить основное понятие терминоэлементами в количестве m, равняется m + n, причем в случае мотивированного термина m не может равняться нулю [Лейчик, 2007, с. 51]. В термине выделяется количество терминоэлементов, соответствующее количеству морфем (или слов), которые являются означающими понятий, относящихся к определенной системе понятий [Лейчик, 2007, с. 85].

Следует различать идеальную и оптимальную структуру термина. Термин обладает идеальной структурой, если связи между его терминоэлементами соответствуют логическим связям между соответствующими понятиями. Оптимальной структурой В. М. Лейчик считает такую идеальную структуру термина, при которой количество его терминоэлементов является минимальным для данной терминосистемы [Лейчик, 2007, с. 52].

1.3. Происхождение терминов

Как мы писали выше, ключевой характеристикой термина является его принадлежность к специальному языку. Тем не менее, термин может употребляться и в повседневной речи. Согласно С. А. Гриневу-Гриневичу, в любой терминологии есть консубстанциональные термины — такие лексические единицы, которые встречаются как в обыденной, так и в профессиональной речи и которые могут затруднить автоматическое выделение терминологической лексики из корпуса [Гринев-Гриневич, 2008, с. 25], например:

*vero* — ‘налог’

*lasku* — ‘счет’

*osake* — ‘акция’.

Тем не менее, термин в первую очередь является единицей языка, а его принадлежность к специальной лексике вторична [Лейчик, 2007, с. 24]. То есть термины, как и другие лексические единицы, могут использоваться на различных уровнях языка, при этом их значение будет изменяться [Лейчик, 2007, с. 88].

Кроме терминов говорят также о квазитерминах, под которыми понимают такие номинативные единицы, которые обозначают понятия с расплывчатым содержанием и объемом, а, следовательно, не могут относиться к терминосистеме. Например, ***raha****hana* —‘денежный кран’ (буквально ‘деньги-кран’).

С ними тесно связаны опорные, или «пустые» слова, входящие в состав терминов, но не являются терминоэлементами, поскольку не обозначают понятий и их признаков [Лейчик, 2007, с. 77].

Внутренние границы между различными группами терминов могут быть размыты: зачастую слово может относиться сразу к нескольким областям человеческого знания. Согласно В. М. Лейчику, «граница между терминологической и общеупотребительной лексикой нестабильна…постоянно происходит как процесс превращения терминов в общеупотребительные слова, так и использование бытовой лексики для формирования терминологий, когда на основе представлений формируются понятия» [Лейчик, 2007, с. 26].

Существует гипотеза, что в момент своего возникновения любое слово было термином, но не всякое слово сохранило свое положение в терминологии к настоящему моменту. Соответственно, в ходе развития языка термины становятся бытовыми словами по мере того, как новые изобретения, новые понятия входят в обыденную жизнь, в массовое употребление, то есть с течением времени терминология динамично развивается [Яковлев, 1948, с. 139]. В языке постоянно происходит процесс превращения терминов в общеупотребительные слова и наоборот [Гринев-Гриневич, 2008, с. 25]. Любая лексическая единица естественного языка может стать лексической единицей специального языка; при этом она становится термином, если начинает выполнять определенные функции [Лейчик, 2007, с. 79].

Одним из основных средств пополнения терминологии считается заимствование из общеупотребительной лексики в специальную лексику [Гринев-Гриневич, 2008, с. 25].

В. М. Лейчик предлагает следующую классификацию источников терминов, согласно которой они делятся на две группы: лексические единицы определенного специального языка и лексические единицы, не относящиеся к какому-либо специальному языку [Лейчик, 2007, с. 81].

Из не относящейся к специальным языкам лексики в термины переходят:

1) лексические единицы литературного языка;

2) лексические единицы диалектов [Лейчик, 2007, с. 81–82].

Из лексики одних специальных языков в состав других переходят:

1) квазитермины (общенаучные или общетехнические термины);

2) элементы профессиональной лексики;

3) элементы профессионального просторечия;

4) элементы профессиональных жаргонов;

5) заимствования из других терминосистем (этот процесс также называется межсистемным заимствованием терминов);

6) номенклатурные единицы;

7) имена собственные, а также имена собственные, перешедшие сначала в разряд номенклатурных единиц, а затем — в термины;

8) заимствования из другого языка, проходящие терминологизацию в принимающем языке;

9) интернационализмы (уже существуют в качестве терминов в нескольких языках);

10) гибридотермины (термины, в состав которых входит заимствованный или интернациональный элементы и элементы принимающего языка, которые затем в процессе терминологизации объединяются);

11) псевдозаимствования (термины, образованные уже в принимающем языке, но на основе заимствованных или интернациональных элементов) [Лейчик, 2007, с. 82–83].

Изучение терминов, заимствованных из другого языка, имеет большую важность для исторического языкознания, поскольку в рамках такого изучения можно рассматривать исследовать также взаимодействие языков и народов, говорящих на этих языках. Более того, исходя из характера заимствованной лексики, можно говорить о характере взаимоотношений между народами и их культурном уровне относительно друг друга (зачастую термины заимствуются из других языков вместе с понятиями, основные характеристики которых они выражают). Тем не менее, Лаури Хакулинен утверждал, что «при отсутствии в языке собственного термина для какого-либо понятия нельзя еще делать вывода о том, что это понятие было неизвестно носителям данного языка» [Хакулинен, 1955, с. 23].

Таким образом, существует три пути становления термина как лексической единицы специального языка:

1. Изменение уже имеющихся лексических единиц естественного языка (обычно не являющихся терминами) путем применения к ним различных способов словообразования (например, через словосложение).
2. Терминологизация нетерминов.
3. Заимствование терминов из других языков, либо заимствование с одновременной терминологизацией [Лейчик, 2007, с. 83–84].

1.4. Классификации терминов

Существует несколько классификаций (или типологий) терминов, но, в отличие от типологий, в которых термины делятся по одному существенному признаку, в классификациях термины группируются по нескольким различным признакам, независимым друг от друга: по содержанию, по языковой форме, по функции, по внутриязыковым и внеязыковым признакам. Тем не менее, классификации терминов могут быть взаимосвязаны [Лейчик, Шелов, 1990, с. 22, 24].

В качестве оснований классификаций используются различные характеристики терминов, связанные с содержанием терминов, их формальной структурой и функциями. Основание классификации может зависеть от области знаний, к которой относится терминология [Литовченко, 2006].

В зависимости от области знания или деятельности термины могут быть научными (которые, в свою очередь, делятся согласно терминологиям, соответствующим имеющимся наукам), техническими (которые могут также использоваться и в научной сфере), терминами языка описания и языка обслуживания экономики, терминами языка управления (например, дипломатические термины, термины делопроизводства или военных языков) и общественно-политическими [Литовченко, 2006].

По логической категории означаемого выделяют термины, обозначающие предметы, процессы, признаки, свойства, величины и их единицы. В зависимости от содержательной структуры термины могут быть однозначными или многозначными [Лейчик, Шелов, 1990, с. 27].

Существуют также лингвистические классификации терминов, основанные на лексических признаках терминов. Подобную классификацию можно найти в работе Б. Н. Головина [Литовченко, 2006].

Б. Н. Головин предлагает следующую классификацию терминов на основе их морфолого-синтаксической структуры. В этом случае термины делятся на два типа: термины-слова и термины-словосочетания [Головин, Кобрин, 1987, с. 70]. Лейчик и Шелов выделяют также термины, выраженные словами с символами [Лейчик, Шелов, 1990, с. 34].

Дальнейшая классификация терминов-слов имеет своим основанием морфемную структуру слова. Таким образом, термины-слова могут быть:

1. непроизводными (или корневыми), например, *lasku* — ‘счет’;
2. производными (***lasku****tus* — ‘выставление счета’);
3. сложными (*sähkö****lasku*** — ‘счет за электроэнергию’);
4. аббревиатурами (*ALV* — сокращение от *arvonlisä****vero***: налог на добавочную стоимость, сокращенно НДС) [Головин, Кобрин, 1987, с. 70].

Также среди терминов-слов можно выделить телескопические слова, ‘цепочечные образования’ или символо-слова [Лейчик, Шелов, 1990, с. 29].

С точки зрения морфологической структуры термины могут быть существительными (*lasku* — ‘счет’), глаголами (***lasku****ttaa* — ‘выставлять счет’), прилагательными (*kvalitatiivinen* — ‘качественный’) и наречиями [Головин, Кобрин, 1987, с. 70]. Согласно Б. Н. Головину и Р. Ю Кобрину, часть речи термина зависит от той области знания, в которой используется термин: так, для математических текстов характерны терминологические предикаты, выраженные глаголом, кратким прилагательным или кратким причастием; прилагательные преимущественно выступают в качестве элементов сложных терминов, но субстантивированные прилагательные являются терминами сами по себе; в определенных терминологиях (например, в музыкальной) терминами являются в том числе наречия или деепричастия, выражающие признак действия [Головин, Кобрин, 1987, с. 71].

Тем не менее, в качестве словарных единиц используются преимущественно термины, выраженные существительными или именными словосочетаниями, поскольку они являются означающими понятий и часто встречаются в текстах научно-технического содержания [Головин, Кобрин, 1987, с. 72]. В данной работе мы будем рассматривать исключительно термины-слова, выраженные существительными, и далее рассмотрим более подробно сложные термины.

В зависимости от типа структуры термины-словосочетания могут быть простыми словосочетаниями, в состав которых входит два знаменательных слова, при этом одно из них будет главным, а другое — зависимым, и сложными словосочетаниями, в которых несколько зависимых слов, отражающих различные аспекты значения главного слова [Головин, Кобрин, 1987, с. 72]. С точки зрения семантики термины-словосочетания могут быть свободными или устойчивыми [Лейчик, Шелов, 1990, с. 29–30].

По морфологическому типу главного слова Б. Н. Головин выделяет:

1. субстантивные словосочетания; в роли главного слова в таких словосочетаниях выступает существительное, в роли зависимых слов могут выступать имена существительные (с предлогом и без), имена прилагательные, порядковые числительные и причастия;
2. адъективные словосочетания; главным словом является прилагательное или причастие, зависимым — имя существительное или наречие;
3. глагольные словосочетания; главное слово — глагол, зависимое — имя существительное.

В данной работе мы рассматриваем исключительно субстантивные словосочетания, в роли зависимых слов в которых выступают прилагательные.

В зависимости от языка-источника термины бывают исконными (уже имевшимися в составе литературного языка, но не специального), заимствованными (из другого языка) и гибридными или гибридотерминами, как мы писали выше [Лейчик, Шелов, 1990, с. 34].

По сфере использования выделяются универсальные (используемые в нескольких областях знания), уникальные (для одной области) и авторские термины [Лейчик, Шелов, 1990, с. 33].

С историко-лексикологической точки зрения термины делятся на термины-архаизмы, входящие в устаревшие терминосистемы или относящиеся к устаревшим областям знаний, и термины-неологизмы, появляющиеся в связи с новыми явлениями и предметами [Литовченко, 2006].

Поскольку специальные языки выполняют более конкретные функции, чем литературный язык, а их лексический состав постоянно пополняется новыми терминами, они периодически нуждаются в стандартизации. По степени нормативности термины могут быть стандартизованными (прошедшими стандартизацию), стандартизированными (проходящими стандартизацию), недопустимыми (не прошедшими стандартизацию), рекомендуемыми, рекомендованными, параллельно допустимыми (в качестве дополнительного варианта) и отклоненными. Также в зависимости от частоты встречаемости в тексте можно выделить высокочастотные и низкочастотные термины [Литовченко, 2006]. Частота термина имеет ключевое значение для статистического подхода к автоматическому извлечению терминов.

Существует три основных этапа стандартизации:

1. полная систематизация всех названий;
2. оценка и унификация реально существующей терминологии;
3. подлинная стандартизация [Герд, 2005, с. 85–86].

По мотивированности термины могут быть полностью мотивированными, полностью немотивированными, частично мотивированными или же ложномотивированными [Лейчик, Шелов, 1990, с. 34].

1.5. Термин в финноязычной лингвистике

Финский язык относится к финно-угорской семье языков и, как и остальные финно-угорские языки (в частности, венгерский, эстонский), относится к агглютинативным языкам. Это означает, что наиболее эффективным способом словоизменения и словообразования в финском языке является агглютинация, или приклеивание к основе различных аффиксов, причем основа слова сама по себе зачастую является полноценной лексемой [Маслов, 1987, с. 234–236].

В финноязычной лингвистике с понятием «*termi*» — ‘термин’ тесно связаны такие понятия как «*erikoiskieli*» — ‘специальный язык’ или «*erikoisalasanasto*» — ‘лексика специальной области’. Специальный язык — это форма языка, которая используется в какой-либо конкретной специальной области и имеет такие характеристики как особая терминология и определенные стилистические и синтаксические черты [Terminologian sanasto, 2006, s. 30]. Специальная область, согласно Словарю терминологии (*Terminologian sanasto*) — область, для которой требуются специальные познания [ibid: s. 30].

По своей структуре термин может быть простым словом, производным словом (образованным из слова-основы путем прибавления суффикса), сложным словом (состоящим из двух или нескольких основ), или словосочетанием (состоящим из двух и более слов). Последнее не стоит путать с терминологическим выражением — устойчивым сочетанием слов, в которое входит один или несколько терминов [Terminologian sanasto, 2006: s. 22–24]. Мы рассмотрим подробнее сложные слова.

Сложное слово — слово, состоящее из двух или более слов, но являющееся одной лексической единицей [VISK, §398]. Тем не менее, при извлечении фактов из текста с практической точки зрения представляется возможным выделять отдельные значимые части сложного слова [Jackson, Moulinier, 2002, p. 10], что мы и сделали при написании грамматики для автоматического выделения терминов.

Существует два типа сложных слов (композитов):

1. композиты подчинительного типа или *määritysyhdyssanat,* которые состоят из атрибута и основной части [VISK: §403], при этом атрибут обычно стоит перед основной частью [Korpela].

Например, ***talous****tilanne* — ‘экономическая ситуация’; *tilanne* — ‘ситуация’, *talous* (как лексема) — ‘экономика’, *talous-* (как часть сложного слова) — ‘экономический, имеющий отношение к экономике’.

В качестве частей композита подчинительного типа могут выступать слова различных частей речи (например, *takaisin****maksu*** —‘погашение’*, maksu* —‘платеж’ (существительное), *takaisin* —‘обратно’ (наречие)) или стоящие в различных формах (например, ***laina****nhakija* — ‘проситель ссуды/займа’, *hakija* — ‘соискатель’, *lainan* — генитивная форма от *laina* — ‘ссуда/займ’).

1. композиты сочинительного типа или *summayhdyssanat,* в состав которых входит два или более слов, находящихся в семантически равнозначных отношениях [VISK: §432].В нашей работе мы рассматриваем преимущественно композиты подчинительного типа, поскольку композиты сочинительного типа встречаются реже и их семантика ограничена конкретными кластерами лексики (например, названиями профессий или цветов) [VISK: §432].

Согласно исследованию Лауры Тююстери, направленному на изучение составных терминов, можно говорить о следующих свойствах сложных слов в финском языке:

1. сложные слова встречаются чаще в терминологиях, чем в стандартном языке;

Хотя сложные слова являются наиболее распространенным типом слова как в стандартном финском языке в силу его агглютинативности, так и в специальных языках, в лексике специального языка встречается гораздо больше сложных слов по сравнению со стандартной лексикой, поскольку они в большей степени удовлетворяют требованиям, предъявляемым к специальной лексике. Кроме того, в последнее время в официальных документах проявляется тенденция к слитному написанию терминов и терминологических выражений, даже если они и писались раздельно, что приводит к образованию новых сложных слов [Korpela].

1. сложные слова-термины чаще, чем в стандартном языке, являются существительными;

Это напрямую связано с самой природой терминов, большая часть которых является существительными в силу определения: термин обозначает понятие, а не его свойство или действие, которое с ним можно совершить.

1. сложные слова-термины в среднем длиннее, чем сложные слова, употребляющиеся в стандартном языке;
2. сложные слова-термины чаще, чем в стандартном языке, содержат в себе другие слова и аббревиатуры;
3. сложные слова-термины реже содержат в себе префиксы, чем сложные слова, употребляющиеся в стандартном языке [Tyysteri, 2010, s. 350].

По своей употребительности термины делятся на предпочтительные (термины, которые признаны наиболее подходящими для обозначения данного понятия), общепризнанные (термины, которые оцениваются как подходящие для обозначения данного понятия и могут употребляться помимо предпочтительного термина) и непригодные (термины, признанные неподходящими для обозначения данного понятия). С точки зрения актуальности термин может быть устаревшим, а по происхождению — заимствованным [Terminologian sanasto, 2006, s. 22–26].

Следует отметить разницу в основаниях, используемых для классификации терминов между русским и финским терминоведением. В частности, в [Terminologian sanasto] не приводится отдельного названия для терминов-архаизмов и терминов-неологизмов. Также предпочтительность термина не выделяется как основание для классификации в русском терминоведении: вместо этого термины классифицируются в зависимости от степени их стандартизованности.

1.6. Экономические термины

Под экономическими терминами мы понимаем термины, употребляющиеся в экономической среде, т.е. в экономической сфере и в областях, каким-либо образом связанных с экономической наукой. Язык экономической сферы, в зависимости от отрасли, делится на несколько специальных подъязыков: язык гостиничного дела, язык торговли, язык горного дела, язык финансов, язык банковского дела, язык экономики сельского хозяйства, язык экономики строительства, язык экономики транспорта, язык менеджмента и язык экономики производства [Бабенко]. Корпус, собранный нами, преимущественно состоит из текстов, относящихся к банковскому делу и торговле.

Поскольку экономика является социальной наукой, она подвержена изменениям, соответствующим изменениям в обществе, что также влияет и на экономическую терминологию [Resche, 2000, p. 159]. Кроме того, по сравнению с другими специальными языками в экономическом дискурсе значительно больше метафор, находящих свое воплощение в различных идиомах и образных выражениях, которые, тем не менее, тоже могут считаться терминами [Corcodel, Corcodel, 2014, p. 123], как, например, ***pääoma****virrat* —‘потоки капитала’.

Можно выделить два направления экономического дискурса в зависимости от целевой аудитории текста: собственно научный и ориентированный на непрофессионалов [Museanu, p. 253]. Поскольку данное исследование базируется не на специальных экономических текстах, а на газетных статьях, посвященных различным экономическим темам, предметом нашего исследования будут термины, относящиеся к повседневному языку экономики.

От собственно научного языка язык экономики отличает близость к стандартному языку, что объясняется общедоступностью данной терминосистемы для неспециалистов (в частности, через газетные статьи) [Resche, 2000, 160]. Таким образом, большинство терминов из корпуса понятны неспециалисту, то есть являются консубстанциональными (например, *vero* — ‘налог’, *lasku* — ‘счет’, *osake* — ‘акция’).

1.7. Выводы

В данной главе мы рассмотрели несколько определений терминов, их свойства, характеристики и особенности.

Выделяя общее у всех рассмотренных определений, мы понимаем термин как единицу какого-либо специального языка, выражающую основные характеристики научного понятия.

Что касается образования термина, то переход лексической единицы из стандартного языка в специальный язык может осуществиться несколькими способами, например, через словообразование, терминологизацию или заимствование из других языков, а также через сочетание этих способов.

Мы рассмотрели несколько классификаций терминов, созданных в рамках русского языкознания и финноязычной лингвистики. Так, с точки зрения морфемной структуры термины-слова могут быть непроизводными, производными, сложными и аббревиатурами, а с точки зрения части речи — существительными, глаголами, прилагательными и наречиями. Стоит также отметить, что, по мнению некоторых исследователей, термином может быть исключительно существительное.

На основании того, что наше исследование проводится на корпусе финских экономических текстов, мы можем сделать вывод, что большая часть терминов, существующая в нашем корпусе и подлежащая выделению, является:

1. существительными;
2. сложными словами;
3. консубстанциональными терминами.

Глава 2. Способы автоматического извлечения терминов

2.1. Общая характеристика систем для автоматического извлечения терминов

Существует несколько различных подходов к автоматическому извлечению терминов, но к настоящему моменту ни один метод не зарекомендовал себя как однозначно эффективный. Тем не менее, можно выделить некоторые общие характеристики для систем автоматического извлечения терминов. Так, на вход системе подается электронный корпус специальных текстов (либо, в случае систем, функционирующих онлайн, просто тексты), на выходе система формирует списки терминов-кандидатов в различных форматах (поддерживаемые форматы зависят от системы), которые подлежат дальнейшей ручной проверке. В некоторых случаях также для каждого термина предоставляется контекст либо другая дополнительная информация, необходимая для исследователя, например, частота термина [Cabré, Estopà, Vivaldi, 2001, p. 2].

Иногда в литературе, посвященной обработке естественного языка, под терминами понимают не единицу специального языка, а семантически значимые токены (ключевые слова) [Fkih, Omri, 2013, p. 1]. В данной главе мы приводим информацию о системах, которые извлекают не ключевые слова, а именно термины как лексические единицы, связанные с каким-либо понятием.

Основными проблемами при автоматическом извлечении терминов являются:

1. определение границ составных терминов и терминов-словосочетаний, состоящих из двух и более слов;
2. распознавание составных терминов и терминов-словосочетаний, состоящих из двух и более слов; в частности, распознавание лексической единицы как части составного термина или как свободной лексической единицы;
3. определение лексической единицы как термина в зависимости от контекста и тематики текста, в котором данная лексическая единица употребляется;
4. слишком длинные списки терминов-кандидатов, которые необходимо проверять вручную, поскольку частота не является достаточным критерием для оценки того, является ли выделенное слово термином или нет [Cabré, Estopà, Vivaldi, 2001, p. 28].

Также существует проблема уместности термина в данном словаре, но попытки решить эту проблему почти не предпринимались, поэтому мы не будем ее рассматривать как релевантную для нашего исследования.

Определение границ термина является одной из основных задач при выделении терминов. Существуют различные стратегии для определения начала и конца терминов: списки терминов, шаблоны структур, синтаксическая разметка, типографские знаки (а также пробелы) и другие [Cabré, Estopà, Vivaldi, 2001, p. 25–26].

После извлечения терминов некоторые из систем производят классификацию полученных результатов для того, чтобы связанные между собой термины находились рядом. Для этих целей может быть построена семантическая сеть из извлеченных терминов, граф, частичная онтология, а также терминологическая сеть, разбивающая термины на основную часть и дополнительные расширения [Cabré, Estopà, Vivaldi, 2001, p. 27].

Большая часть систем автоматического извлечения терминов предназначена только для одного языка (английского или французского) и содержит лингвистическую информацию, относящуюся только к конкретному языку, что усложняет использование системы по отношению к другому языку [Cabré, Estopà, Vivaldi, 2001, p. 28]. Тем не менее, нередки случаи, когда алгоритм, написанный для одного языка, применялся в отношении других языков, в том числе и родственных ему. Так, гибридный метод, использованный в отношении английского, японского, словенского и сербского языков (два последних относятся к группе славянских языков) [Kupść, 2007, p. 1], и полученные при его использовании результаты были сопоставимы с соответствующими результатами, полученными в ходе экспериментов с польским языком, также относящимся к славянским языкам, что доказывает переносимость данного алгоритма в частности [Kupść, 2007, p. 6].

Поскольку термин, как мы писали в первой главе, в большинстве случаев является существительным или словосочетанием, главным словом которого выступает существительное, большая часть систем автоматического извлечения терминов работает исключительно с именными группами, а не с глагольными [Cabré, Estopà, Vivaldi, 2001, p. 28]. Мы, в свою очередь, тоже рассматривали только термины, представленные именными группами, а именно — сложными словами и словосочетаниями.

2.2. Классификация систем для автоматического извлечения терминов

Как и системы для машинного перевода, системы для автоматического выделения терминов делятся на три группы в зависимости от используемой технологии:

1. лингвистические системы;

Под лингвистическими системами понимают системы, в которых применяются только лингвистические данные. Они в свою очередь делятся на два вида: системы, которые используют информацию, касающуюся конкретно термина (в этом случае с помощью регулярных выражений и конечных автоматов задается выделение повторяющихся структур терминологических сочетаний), и системы, использующие общеязыковую информацию (выделяются базовые языковые структуры, например, именные группы) [Cabré, Estopà, Vivaldi, 2001, p. 3]. Как правило, в обоих типах систем используется морфологически размеченный корпус.

1. статистические системы;

Системы, использующие статистический подход, считаются традиционными для извлечения терминов в силу того, что они независимы от языка. Такие системы работают за счет выделения двух или более лексических единиц, частота совместной встречаемости которых больше некоторого заданного уровня. Основным недостатком статистических систем является неспособность извлекать низкочастотные термины [Cabré, Estopà, Vivaldi, 2001, p. 3].

При оценке результатов в подобных системах используются такие меры как *точность* и *полнота*. *Полнота* описывает способность системы извлекать все термины из корпуса и задается через отношение количества извлеченных кандидатов в термины к общему количества терминов в корпусе. *Точность* характеризует способность системы отличать термины от нетерминов и рассчитывается через отношение количества извлеченных терминов к количеству извлеченных кандидатов в термины [Cabré, Estopà, Vivaldi, 2001, p. 3]:

*точность =* $\frac{количество извлеченных терминов}{количество извлеченных кандидатов в термины}$;

*полнота =* $\frac{количество извлеченных кандидатов в термины}{общее количество терминов в корпусе}$

Большинство систем для автоматического извлечения терминов не использует чисто статистический подход, а добавляет к нему дополнительную лингвистическую информацию, например морфологическую (разметка) или семантическую. Также может быть использован список стоп-слов [Cabré, Estopà, Vivaldi, 2001, p. 25].

1. гибридные системы [Cabré, Estopà, Vivaldi, 2001, p. 2].

Гибридные системы сочетают в себе статистический метод с использованием некоторой лингвистической информации. Чаще всего используются специальные правила (в нашем случае они учитывают морфологическую разметку).

 Алгоритм, которым мы будем пользоваться для автоматического извлечения терминов, является гибридным и не зависит от контекста, но учитывает частоту термина, что позволяет нам выделять и низкочастотные термины. Мы опишем его подробней в следующей главе.

**2.3. Существующие системы автоматического выделения и извлечения терминов**

2.3.1. Онлайн-системы

Существуют системы для извлечения терминов, с которыми можно работать онлайн. К таким инструментам относятся:

1. *Term Extraction* компании *FiveFilters.org* — программное обеспечение, позволяющее извлекать термины с помощью веб-сервиса.

На вход подается текст или ссылка на сайт, из которого извлекаются термины. На выходе пользователь получает списки терминов, отсортированные по релевантности. Выходной формат списков может быть различным (.html, .xml, .json или .txt). В данный момент поддерживается исключительно английский язык [Term Extraction | fivefilters.org].



Рис. 1 — Входной интерфейс *Term Extraction*



Рис. 2 — Выходной интерфейс *Term Extraction*: список терминов

В первом столбце (*Term*) приводится список наиболее употребительных терминов, во втором столбце (*Occurence*) ­— количество употреблений данного термина в тексте, в третьем столбце (*Word Count*) — количество слов в терминологическом словосочетании.

1. *TerMine* компании *NaCTem* — онлайн-сервис, позволяющий извлекать термины из текста, файла (.txt или .pdf) или с интернет-страницы.

Поддерживается только кодировка ASCII. Подход является гибридным: используется морфологическая разметка и метод *C-value* [TerMine].



Рис. 3 — Входной интерфейс *TerMine*



Рис. 4 — Выходной интерфейс *TerMine*

Красным цветом выделены найденные термины. В левом верхнем углу окна указывается их количество.

1. *Terminology Extraction* компании *Translated Labs*.

Основная идея использованного метода — сравнение частоты слов, встречающихся в тексте, с частотой слов в языке, то есть слово, которое чаще встречается в тексте, но редко — в языке, предположительно является термином. Поддерживаются английский, французский и итальянский языки [Terminology Extraction].



Рис. 5 — Входной интерфейс *Terminology Extraction*



Рис. 6 — Выходной интерфейс *Terminology Extraction*

Приводится список из 20 наиболее встречаемых терминов. Кроме этого, термины выделяются разным цветом в тексте.

2.3.2. Программы, находящиеся в свободном доступе

Некоторые программы или библиотеки для извлечения терминов находятся в свободном доступе, но их необходимо загружать. К подобным программам относятся следующие:

1. *topia.termextract 1.1.0* — достаточно простой алгоритм, написанный на языке *Python* и сочетающий морфологическую разметку и статистический метод. Для того чтобы слово было выделено алгоритмом как термин, оно должно встретиться в тексте не менее трех раз [topia.termextract 1.1.0: Python Package Index].



Рис. 7 — Пример работы *topia.termextract 1.1.0*: извлечение *Term Extractor*



Рис. 8 — Пример работы *topia.termextract 1.1.0:* результат работы теггера



Рис. 9 — Примеры работы *topia.termextract 1.1.0:* применение *Term Extractor*

1. *Araya Bilingual Term Extraction Tool* —программа, извлекающая из файлов формата .tmx кандидаты в термины надвух языках с дополнительными характеристиками (например, частота). Пользователь может отмечать правильность/неправильность извлеченного термина. В тестовой версии доступно извлечение только 20 пар терминов [Terminology Extraction < Heartsome Europe GmbH].



Рис. 10 — Интерфейс *Araya Bilingual Term Extraction Tool*

2.3.3. Программы с закрытым кодом

Также существуют системы извлечения терминов, которые созданы для коммерческих целей и не находятся в свободном доступе. К таким системам относятся:

1) *Sandstone’s term extraсtion solution* — сервис, предоставляемый компанией *Sandstone*. Система ориентирована на интернациональные компании и извлекает из различных документов термины, характерные для данной организации. Также возможно параллельное извлечение терминов (то есть терминов и их переводов на другой язык). Термины на одном языке извлекаются из файлов формата .doc, .docx, .rtf и .txt, на двух — из файлов формата .tmx, .csv и .txt. Поддерживаются английский, финский и шведский языки. Качество работы сервиса улучшается по мере наращивания объемов текстов, предоставляемых на вход [Terms Extraction].

#### 2) Системы, созданные компанией *TermCoord*.

#### а) *SynchroTerm*. Термины извлекаются как на одном, так и на двух языках из файлов формата .doc, .xls, .rtf, .txt, .html, .pdf, .tmx (точнее всего извлечение для двух языков происходит на файлах формата .tmx, в остальных случаях возникает проблема с выравниванием). Для выделенных терминов доступен просмотр контекста. Поддерживается множество языков, в том числе английский, русский, финский, шведский, норвежский и греческий [Term extraction analysis done by TermCoord - Terminology Coordination Unit [DGTRAD] - European Parliament].

б) *SDL MultiTerm Extract.* Поддерживаются все языки, совместимые с кодировкой *Unicode*. Совместимыми форматами файлов являются .txt, .doc, .html, .htm, .tmx, .rtf, .xml, .sgm, sgml, .ppt, .xls, .tmx, .tmw, .ttx. Списки терминов выдаются в форматах .txt, .xml и в существующих терминологических базах. Алгоритм нечувствителен к регистру [Term extraction analysis done by TermCoord - Terminology Coordination Unit [DGTRAD] - European Parliament].

в) *TaaS (Terminology as a Service).* Алгоритм работает для 24 языков Европейского Союза, а также русского языка. Есть несколько вариантов извлечения терминов (сохранение словоформы, нормализация, визуализация, система *Kilgray*, *TWSC* (лингвистический алгоритм, дополненный статистическими чертами)). Поддерживаются форматы .pdf, .doc, .docx, .xls, .xlsx, .pptx, .rtf, .txt, .xliff, .xlf, .xml, .html, .htm, .mif. В системе имеется список стоп-слов. Есть возможность изменять термины, добавлять к ним определения и многое другое [Term extraction analysis done by TermCoord - Terminology Coordination Unit [DGTRAD] - European Parliament].

г) *TermoStat Web 3.0.* Поддерживаются только файлы форматов .txt и .rtf и только французский, английский, испанский, итальянский и португальский языки. Списки терминов (включая частоту употребления термина, варианты написания, формы множественного числа) выдаются только в формате .txt [Term extraction analysis done by TermCoord - Terminology Coordination Unit [DGTRAD] - European Parliament].

д) *Promt Terminology Manager* (*Promt TerM*). Используется статистический метод. Можно задавать дополнительные параметры для извлечения терминов. Программа встроена в профессиональную версию системы перевода *@promt Expert* [ПРОМТ объявляет о выпуске нового продукта ПРОМТ Terminology Manager]. По нашим сведениям, проект до сих пор развивается.

2.4. Выводы

В данной главе мы рассмотрели общие черты систем и алгоритмов для автоматического извлечения терминов.

Большинство систем использует статистический подход, хотя в некоторых случаях привлекается дополнительная лингвистическая информация.

Несмотря на то, что вышеуказанные системы показывают неплохие результаты, для нашего исследования мы выбрали инструмент *SketchEngine*, поскольку:

— система *SketchEngine* позволяет написать свою грамматику, загрузить ее и применить к любому корпусу;

— в систему *SketchEngine* можнозагрузить свой корпус любого объема;

— системой SketchEngine поддерживается множество языков, в том числе и финский язык.

Подробнее об инструменте *SketchEngine* и его применении пойдет речь в следующей главе.

Глава 3. Правила извлечения терминов

3.1. Материал исследования

Для нашего исследования мы собрали корпус (50 тыс. токенов), состоящий из газетных статей на экономическую тему. Размер корпуса был обусловлен желанием провести как можно более тщательное исследование для отдельных лексем и предоставить подробный анализ результатов. В качестве источников экономических текстов мы использовали ряд финноязычных новостных сайтов.

Табл. 1 — Список источников экономических текстов

|  |  |  |
| --- | --- | --- |
| **Название сайта** | **Ссылка** | **Количество статей** |
| *Yle Uutiset* (‘Новости Yle’) | yle.fi | 13 |
| *Taloussanomat | Talousuutiset* (‘Экономическая газета: Экономические новости’) | taloussanomat.fi | 63 |
| *Talouselämä* (‘Экономическая жизнь’) | talouselama.fi | 12 |
| *Savon Sanomat* (‘Газета Саво’) | savonsanomat.fi | 2 |
| *Yrittäjät.fi* (‘Предприниматели’) | yrittajat.fi | 16 |
| *Arvopaperi* (‘Ценная бумага’) | arvopaperi.fi | 3 |
| *Kauppalehti: Tärkeimmät talousuutiset*(‘Торговая газета: Самые важные новости экономики’) | kauppalehti.fi | 1 |

Мы дополнительно проверяли работоспособность запросов для выделения словосочетаний, содержащих заданный термин, и некоторых запросов для извлечения терминов на основе финского корпуса *fiTenTen*.

Корпус *fiTenTen*, как и другие корпуса группы *TenTen*, был создан с помощью программы-краулера *SpiderLing*, который извлекает с различных интернет-страниц тексты для создания корпуса. Данные тексты были очищены на следующем этапе от гиперссылок, заголовков и сносок с помощью инструмента *jusText*, разбиты на токены с помощью программы *unitok*, с помощью инструмента *onion* были удалены повторы. В конце корпус был размечен размечен с помощью теггера *TreeTagger* версии *TreeTagger for Finnish v2* [Language resources and tools | Sketch Engine]. Объем корпуса *fiTenTen* составляет 1,7 млн токенов.

Для последующей проверки выделенных терминов использовался словарь экономических терминов *Taloussanakirja* [Taloussanakirja | Talousuutiset], доступный онлайн на сайте *taloussanomat.fi*, а также другие словари, в том числе финско-русские (Большой финско-русский словарь [Вахрос, Щербаков, 2007]) и финско-английские (*Wiktionary* [Wiktionary, the free dictionary], *Sivistysanakirja* [Sivistysanakirja], *Sanakirja.org* [Sanakirja.org – Ilmainen Sanakirja. Suomi, englanti, ruotsi jne. nettisanakirja]). Выбор данного словаря экономической лексики был обусловлен его доступностью, а также тем, что он непосредственно связан с экономической прессой, поскольку размещен на сайте экономической газеты, и, соответственно, содержит лексику, ориентированную на читателя газетных статей. Мы также использовали данный словарь при написании грамматики.

Кроме этого, для проверки кандидатов в термины на терминологичность мы использовали *Yleinen suomalainen ontologia* (‘Общая финская онтология’), сокращенно *YSO* [Finto: YSO – Yleinen suomalainen ontologia].

3.2. Разметка текстов

Собранный нами корпус был вручную очищен от гиперссылок и размечен с помощью инструмента частеречной разметки *TreeTagger*, встроенного в платформу *SketchEngine*. Мы использовали для разметки версию *TreeTagger for Finnish v2*.

*TreeTagger* — инструмент для морфологической разметки текста, основанный на статистическом подходе. Он был разработан Г. Шмидом (*H. Schmid*) в Институте компьютерной лингвистики в Штутгартском университете. *TreeTagger* успешно применяется для разметки текстов на английском, немецком, французском, итальянском, испанском, русском, китайском, эстонском и других языках. Для обучения *TreeTagger* используется лексикон и вручную размеченный учебный корпус [TreeTagger].

Морфологические характеристики, используемые в финской версии *TreeTagger*, основаны на анализаторе *OMorfi* — инструменте для морфологического анализа с открытым исходным кодом, созданном на кафедре современных языков Хельсинкского университета. Краткие формы тегов *TreeTagger* получены из более длинных тегов, используемых в *Omorfi* [Voutilainen, Purtonen, Muhonen, 2012, s. 11].

Ниже приведен пример морфологической разметки программой *TreeTagger* в вертикальном формате.

Табл. 2 — Пример морфологической разметки *TreeTagger*

|  |  |  |
| --- | --- | --- |
| **Словоформа** | **Часть речи (тег)** | **Лемма** |
| Suomen | *N\_Prop\_Gen\_Sg* | Suomi |
| sisukas | *A\_Nom\_Sg* | sisukas |
| kansa | *N\_Nom\_Sg* | kansa |
| on | *V\_Prs\_Act\_Sg3* | olla |
| jäänyt | *PrfPrc\_Act\_Nom\_Sg* | jäädä |
| jyrän | *N\_Gen\_Sg* | jyrä |
| alle | *Adp\_Po* | alle |

Примечание: изначально словоформа *alle* была размечена как *Adv,* что в данном контексте неверно.

3.3. SketchEngine и грамматики SketchEngine

*SketchEngine* — онлайн-инструмент, широко используемый для различных лексикографических работ. Платформа *SketchEngine* позволяет работать с уже имеющимися корпусами (в частности, осуществлять поиск по лемме, строить конкордансы, делать запросы и так далее), а также создавать или загружать свои корпуса и обрабатывать их. В самом *SketchEngine* существует несколько корпусов на финском языке (*fiTenTen*, *EUROPARL7*, *OPUS2*, *DGT*, *Araneum Finnicum Maius*), но ни один из них не отвечал целям данной работы в силу своих размеров и тематики, поэтому мы загрузили в *SketchEngine* собранный нами корпус экономических текстов.

Табл. 3 — Список корпусов финского языка, доступных в *SketchEngine*

|  |  |  |
| --- | --- | --- |
| **Название корпуса** | **Объем в токенах** | **Объем в словах** |
| *fiTenTen* | 1 703 429 270 | 1,404,100,049 |
| *EUROPARL7* | 40, 979, 520 | 34,182,031 |
| *OPUS2* | 180,134, 681 | 131,985,872 |
| *DGT* | 47,397,459 | 35,129,923 |
| *Araneum Finnicum Maius* | 1,200,000,486 | 817,453,523 |

Одной из функций *SketchEngine* является извлечение ключевых слов и терминов (*keywords and* *terminology extraction*). Подход, используемый в платформе *SketchEngine*, является гибридным, поскольку извлечение терминов из размеченного корпуса происходит за счет выполнения следующих условий:

1. наличие частеречной разметки;
2. наличие размеченного эталонного корпуса (в нашем случае этим корпусом являлся *fiTenTen*);
3. наличие скетч-грамматики для извлечения терминов, с помощью которой в специальном корпусе выделяются элементы, соответствующие правилам грамматики, после чего частота их встречаемости в специальном корпусе сравнивается с частотой их встречаемости в эталонном корпусе[[1]](#footnote-1) [Kilgarriff et al., 2014, p. 53].

На 2013 год существовали подобные скетч-грамматики для китайского, английского, французского, немецкого, японского, корейского, русского, испанского и португальского языков.

Скетч-грамматика (*sketch grammar*) представляет собой набор правил, основанный на языке CQL (*Corpus Query Language* — язык запросов к корпусу) и используемый в платформе *SketchEngine* для создания коллокаций и списков слов и словосочетаний (*word sketches*).

Запрос CQL описывается в виде шаблона, согласно которому из корпуса извлекаются токены. Каждому токену при разметке присваивается набор атрибутов (лемма, частеречный тег и так далее). Обычный CQL запрос выглядит следующим образом:

[attribute="value"]([атрибут="значение"]),

где *attribute* — атрибут, присвоенный токену, а *value* представляет собой регулярное выражение. В значении также могут быть использованы логические операторы ‘и’ (&) и ‘или’ (|). [Jakubíček M. et al., 2010, p. 2]

Правило скетч-грамматики состоит из указания на тип правила, отмеченного астериском (\*), названия правила, перед которым стоит знак равенства (=) и списка выражений языка запросов CQL. Также в начале правила может стоять пояснение, обозначенное знаком # [Benko, 2014, p. 421]. Например:

*#economics*

*\*UNARY*

*=talous*

*1:[(lemma=".\*talou(s|d|t|ks).\*")&tag="N\_.\*"]*

 где *#economics* — дополнительный комментарий, *\*UNARY* означает, что отношение унарно, то есть правило должно применяться только один раз (в этом случае будет использована только одна метка ‘1:’).

Метка ‘1:’ обозначает ‘ключевое слово’ или лемму, для которого и создано правило грамматики.

*tag* задает часть речи и ее морфологические характеристики: в данном случае *N\_.\** — *noun* (существительное), собственное или нарицательное, стоящее в любом падеже, единственного или множественного числа [Voutilainen, Purtonen, Muhonen, 2012, s. 22]. Так, тег для нарицательного существительного, стоящего в номинативе единственного числа, будет выглядеть как:

*N\_Nom\_Sg* [теги].

Тег для имени собственного, стоящего в номинативе единственного числа, будет выглядеть как:

*N\_Prop\_Nom\_Sg* [теги].

Регулярное выражение *.\*talou(s|d|t|ks).\** задает любую лемму, в состав которой входят *-talous-,* *-taloud-*, *-talout-* или *-talouks-.*

Таким образом, согласно этому правилу должны выделяться существительные единственного или множественного числа, стоящие в любом падеже и имеющие в своем составе основы *-talous-,* *-taloud-*, *-talout-* или *-talouks-.*

3.3. Грамматика для извлечения терминов.

Созданная нами грамматика *SketchEngine* извлекает списки слов и словосочетаний (*word sketches*) согласно 50 правилам для слов и одному – для словосочетаний (см. приложение 1).

Для написания набора грамматических правил мы отобрали, ориентируясь на [Taloussanakirja | Talousuutiset], ключевые слова-основы в количестве 50 единиц, которые:

1. являются терминами сами по себе:

*talous* ‘экономика’, *osake* ‘акция’*, lasku* ‘счет’, *vero* ‘налог’, *luotto* ‘кредит’, *markkina* ‘рынок’, *pankki* ‘банк’;

1. являются основной частью сложного слова-термина:

*sähkö****lasku*** ‘счет за электроэнергию’, буквально ‘электричество-счет’, *pankki****laina*** (по запросу на основу *laina*) ‘банковский кредит’, буквально ‘банк-кредит’, *pörssi****osake*** ‘биржевая акция’, буквально ‘биржа-акция’;

1. являются атрибутом сложного слова-термина:

***osake****yhtiö* ‘акционерное общество’, буквально ‘акция-общество’, ***velka****kirja* ‘долговое обязательство’, буквально ‘долг-книга’,
***lasku****kausi* ‘период экономического спада’, буквально ‘спад-период’.

Таким образом, мы отобрали термины с хорошей деривационной способностью.

Дополнительным требованием являлась относительная однозначность лексической единицы: например, лексема *etu* (‘выгода, преимущество’) не вошла в наш список по причине длины и слишком большой вероятности того, что слово, содержащее данную лексему, не будет однокоренным с ним. По той же причине не вошла в наш список лексема *otto* (‘снятие <денежных средств>’); кроме того, по своему составу она пересекалась с лексемой *lu****otto*** (‘кредит’).

Для всех 50 ключевых слов мы написали правила, позволяющие извлекать из корпуса лексемы, в состав которых входят данные ключевые слова. Похожий метод используется в поисковой системе самого экономического словаря *Taloussanakirja*: так, по запросу на слово *liike* (‘торговля’) выдается список статей, ключевое слово которых содержит данную лексему в качестве части сложного слова [Taloussanakirja | Talousuutiset].

Например, для лексемы *hinta* (‘цена’) написано следующее правило:

*\*UNARY*

*=hinta*

*1:[(lemma=".\*hint(a|o).\*"|lemma=".\*hinn(a|o).\*")&tag="N\_.\*"]*

Данное правило позволяет выделять словоформы, в состав которых входят основы -*hinta- , -hinto- , -hinna- , -hinno-*. Например:

**hinta**taso[[2]](#footnote-2) — ‘уровень цен’, буквально ‘цена-уровень’;

varallisuus**hinto**ja (партитив множественного числа от varallisuus**hinta**) — ‘цены активов’;

neliö**hinnat** — ‘цены за квадратный метр’, буквально ‘квадратный метр-цены’;

**hinno**ittelu — ‘расценка’ (в отличие от предыдущих случаев, это производное слово, а не сложное).

Наличие в значении различных основ (*hinta- ,hinto- ,hinno- ,hinna-*) обусловлено чередованием основы (*-nt* в слабой основе переходит в *-nn*) [Klusiilien astevaihtelu], а также особенностями образования множественного числа у данного слова (в двусложных словах, оканчивающихся на *-a* и имеющих в первом слоге гласную *a, i* или *e*, при присоединении показателя множественного числа *-i* происходит чередование гласных *a: o*) [VISK: §80]; к тому же, это позволяет выделять однокоренные термины, в которых данная лексема выступает в качестве атрибута сложного слова, поскольку существительное в составе сложного слова может выступать в разных падежах (чаще всего в номинативе или генитиве, но так же могут употребляться партитив, иллатив, адессив или инструктив) [VISK: §403, §415].

Указание части речи (*tag="N\_.\*"*)ограничивает результат исключительно существительными, что помогает избежать однокоренных прилагательных и глаголов (которые, тем не менее, тоже могут быть терминами, но мы не ставим задачу по их выделению).

Существуют и более простые правила:

*\*UNARY*

*=lasku*

*1:[lemma=".\*lasku.\*"&tag="N\_.\*"]*

Согласно данному правилу будут выделяться словоформы, содержащие основу *-lasku-*. Наличие в правиле грамматики выражения только для одной основы обусловлено тем, что у лексемы *lasku* (‘счет’) не происходит ни чередования основ, ни чередования гласных, как в предыдущем случае с лексемой *hinta*.

Для извлечения словосочетаний, содержащих термин, написано одно правило на примере лексемы *talous* (‘экономика’):

*\*DUAL*

*=termi/adj*

*2:[tag="A\_Nom\_.\*"] [tag="A\_Nom\_.\*"]? 1:[lemma=".\*talou(s|d).\*"&tag="N\_Nom\_.\*"]*

*2:[tag="A\_Gen\_.\*"] [tag="A\_Gen\_.\*"]? 1:[lemma=".\*talou(d|ks).\*"&tag="N\_Gen\_.\*"]*

*2:[tag="A\_Par\_.\*"] [tag="A\_Par\_.\*"]? 1:[lemma=".\*talou(t|ks).\*"&tag="N\_Par\_.\*"]*

*2:[tag="A\_Acc\_.\*"] [tag="A\_Acc\_.\*"]? 1:[lemma=".\*talou(s|d).\*"&tag="N\_Acc\_.\*"]*

*2:[tag="A\_Ill\_.\*"] [tag="A\_Ill\_.\*"]? 1:[lemma=".\*talou(t|ks).\*"&tag="N\_Ill\_.\*"]*

*2:[tag="A\_Ine\_.\*"] [tag="A\_Ine\_.\*"]? 1:[lemma=".\*talou(d|ks).\*"&tag="N\_Ine\_.\*"]*

*2:[tag="A\_Ela\_.\*"] [tag="A\_Ela\_.\*"]? 1:[lemma=".\*talou(d|ks).\*"&tag="N\_Ela\_.\*"]*

*2:[tag="A\_All\_.\*"] [tag="A\_All\_.\*"]? 1:[lemma=".\*talou(d|ks).\*"&tag="N\_All\_.\*"]*

*2:[tag="A\_Ade\_.\*"] [tag="A\_Ade\_.\*"]? 1 :[lemma=".\*talou(d|ks).\*"&tag="N\_Ade\_.\*"]*

*2:[tag="A\_Abl\_.\*"] [tag="A\_Abl\_.\*"]? 1:[lemma=".\*talou(d|ks).\*"&tag="N\_Abl\_.\*"]*

*2:[tag="A\_Tra\_.\*"] [tag="A\_Tra\_.\*"]? 1:[lemma=".\*talou(d|ks).\*"&tag="N\_Tra\_.\*"]*

*2:[tag="A\_Ess\_.\*"] [tag="A\_Ess\_.\*"]? 1:[lemma=".\*talou(t|ks).\*"&tag="N\_Ess\_.\*"]*

*2:[tag="A\_Abe\_.\*"] [tag="A\_Abe\_.\*"]? 1:[lemma=".\*talou(d|ks).\*"&tag="N\_Abe\_.\*"]*

*2:[tag="A\_Com\_.\*"] [tag="A\_Com\_.\*"]? 1:[lemma=".\*talou(t|ks).\*"&tag="N\_Com\_.\*"]*

*2:[tag="A\_Ins\_.\*"] [tag="A\_Ins\_.\*"]? 1:[lemma=".\*talouks.\*"&tag="N\_Ins\_.\*"]*

*\*DUAL* означает, что выделяться будет больше одного слова, причем слово, помеченное маркером *1:* , будет главным. В нашем случае это все слова, содержащие в своей структуре лемму *talous* в качестве основной части сложного слова, или же само слово *talous.*

Выражениятипа *2:[tag="A\_Nom\_.\*"] [tag="A\_Nom\_.\*"]?* могут задавать в качестве зависимого слова одно или больше прилагательных, согласованных с главным словом в падеже и числе и стоящих перед главным словом в силу особенностей синтаксиса финского языка. В отличие от правил для извлечения отдельных слов, мы не могли задать падеж прилагательного с помощью регулярного выражения *tag="A\_.\*"*, поскольку в этом случае не выполнялось требование согласованности.

Таким образом, приведенное выше правило нашей скетч-грамматики задает поиск всех субстантивных словосочетаний, в состав главного слова которого входят основы *-talous-,* *-taloud-*, *-talout-* или *-talouks-* (в зависимости от падежа), зависимыми словами являются любые прилагательные (одно и больше), и данные словосочетания стоят в любом из пятнадцати падежей финского языка.

3.4. Результаты

3.4.1. Выделение терминологических словосочетаний

Проверка на нашем корпусе показала, что для лексемы *talous* (‘экономика’) в нем есть только 15 случаев субстантивных словосочетаний с прилагательным в качестве зависимого слова в номинативе и одно словосочетание, стоящее в аллативной форме. Мы дополнительно проверили работоспособность данного правила на корпусе *fiTenTen* и получили результаты для всех падежей, кроме абессива и комитатива, которые достаточно малоупотребительны и наряду с инструктивом считаются маргинальными (для инструктива в *fiTenTen* было выделено два словосочетания).

Из 15 словосочетаний, стоявших в номинативе, трижды встретилось устойчивое словосочетание *julkinen* ***talous*** (‘экономика публичного сектора’) и один раз — словосочетание *kotimainen* ***talous*** (‘национальная экономика’). В остальных случаях прилагательные либо выражали характеристику главного слова, как оценочную:

*hidas <****talous****kasvu>* — ‘медленный <экономический рост, буквально ‘экономика-рост’>’

*vaikea <****talous****tilanne>* — ‘сложная <экономическая ситуация, буквально ‘экономика-ситуация’>’

*vastaiset <****talous****pakotteet>* — ‘ответные <экономические санкции, буквально ‘экономика-санкции’>’,

так и количественную характеристику:

*useimmat <koti****taloud****et>* — ‘бо́льшая часть <домашних хозяйств>‘, буквально ‘б*о*льшая часть домашних экономик’,

либо представляли собой эпитет:

*aneeminen <****talous****>* — ‘анемичная <экономика>‘.

В случае аллатива в выделенном словосочетании прилагательное также выражало количественную характеристику:

*useimmille <palkansaajakoti****talouks****ille>* — бо́льшей части <дословно: домашних хозяйств получателей зарплаты>.

 Таким образом, было выделено всего два терминологических словосочетания: *julkinen* ***talous*** (‘экономика публичного сектора’) и *kotimainen* ***talous*** (‘национальная экономика’).

3.4.2. Выделение терминов

В полученный список терминов также вошли лексемы, извлечение которых как кандидатов в термины было безошибочным.

Под безошибочным извлечением мы понимаем такое извлечение терминов, при котором все кандидаты в термины, находящиеся в конечном списке:

а) являются существительными (как мы уже писали раньше, мы понимаем термин как единицу специального языка, связанную с каким-либо научным понятием, и рассматриваем его исключительно как существительное);

б) являются производными от исходного слова (в процессе выделения были случаи, когда искомая основа совпадала с частью сложного снова, не имеющего отношения к лексеме, для которой проводился поиск);

в) их значение совпадает со значением термина в языке экономики (в случае многозначных и консубстанциональных терминов).

Ниже в таблице (Табл. 4) мы приводим полученные результаты для терминов, которые были выделены без ошибок.

Табл. 4 — Результаты для терминов, выделение которых было безошибочным

|  |  |  |  |
| --- | --- | --- | --- |
| **Ключевое слово** | **Перевод** | **Число выделенных** **кандидатов в термины** | **Точность** |
| *indeksi* | индекс | 10 | 0,80 |
| *investointi* | капиталовложение, инвестирование | 86 | 0,77 |
| *kilpailu* | конкуренция | 91 | 0,58 |
| *kulut* | расходы | 90 | 0,67 |
| *kustannus* | покрытие расходов | 66 | 0,82 |
| *lasku* | счет | 74 | 0,65 |
| *luotto* | кредит | 37 | 0,78 |
| *markkina* | рынок | 241 | 0,71 |
| *osake* | акция | 79 | 0,71 |
| *rahasto* | денежный фонд | 39 | 0,54 |
| *siirto* | перенос, передача | 41 | 0,70 |
| *tappio* | убыток | 28 | 0,93 |
| *tuotanto* | производство, продукция | 44 | 0,68 |
| *vakuus* | залог, гарантия | 20 | 0,65 |
| *valuutta* | валюта | 77 | 0,76 |
| *vekseli* | вексель | 0[[3]](#footnote-3) | – |
| *vienti* | экспорт | 95 | 0,52 |

В некоторых случаях в результатах встречались слова, которые являются терминами как в английском, так и в русском языках, но не входили ни в один словарь финского языка из тех, которыми мы пользовались для оценки. Например:

*arvonlisä****vero****vapautus* —‘*освобождение от НДС*’*,*

***vero****nmaksaja* —‘*налогоплательщик*’*,*

***pääoma****virrat* —‘*потоки капитала*’*.*

В Таблице 5 (Табл. 5) мы приводим результаты для таких терминов-кандидатов, при выделении которых мы столкнулись с определенными ошибками, о которых будет сказано далее.

Табл. 5 — Результаты для терминов, которые были выделены с ошибками

|  |  |  |  |
| --- | --- | --- | --- |
| **Ключевое слово** | **Перевод** | **Число выделенных** **кандидатов в термины** | **Точность** |
| *arvo* | стоимость | 117 | 0,85 |
| *finanssi* | финансы | 79 | 0,68 |
| *hinta* | цена | 183 | 0,82 |
| *kauppa* | торговля | 140 | 0,76 |
| *korko* | процент | 353 | 0,57 |
| *kurssi* | курс | 39 | 0,49 |
| *laina* | ссуда, займ | 149 | 0,53 |
| *lama* | кризис | 44 | 0,66 |
| *liike* | бизнес | 149 | 0,72 |
| *maksu* | платеж | 58 | 0,41 |
| *menot* | расходы | 36 | 0,69 |
| *omistaa[[4]](#footnote-4)* | владеть | 78 | 0,65 |
| *osto* | покупка | 135 | 0,22 |
| *palkka* | заработная плата | 87 | 0,36 |
| *palvelu* | обслуживание | 167 | 0,61 |
| *pankki* | банк | 449 | 0,76 |
| *pääoma* | капитал | 47 | 0,51 |
| *pörssi* | биржа | 31 | 0,65 |
| *raha* | деньги | 531 | 0,55 |
| *seteli* | купюра | 56 | 0,64 |
| *säätiö* | фонд | 14 | 0,86 |
| *talous* | экономика | 472 | 0,72 |
| *tase* | баланс | 35 | 0,91 |
| *tili* | счет | 43 | 0,47 |
| *tuki* | финансовая поддержка | 96 | 0,31 |
| *tulo* | доход | 188 | 0,62 |
| *tuote* | продукт | 61 | 0,75 |
| *vara* | запас, средства | 124 | 0,48 |
| *velka* | долг | 226 | 0,55 |
| *vero* | налог | 158 | 0,66 |
| *voitto* | прибыль | 38 | 0,79 |
| *yhtiö* | компания  | 241 | 0,83 |
| *yritys* | предприятие, фирма | 316 | 0,77 |

3.5. Ошибки при автоматическом выделении терминов

При автоматическом выделении терминов мы столкнулись с несколькими типами ошибок.

1. Ошибки, связанные с морфологической разметкой *TreeTagger*, или ошибки в тегах.

Они в тегах были двух видов: в падеже или числе и в части речи.

1) Ошибки в падеже и/или числе.

В первом случае существенно ошибка не влияла на выделение: так, существительному ***talous****tieteilijöillekin*был присвоен тег *N\_All\_Sg*, что неверно — ***talous****tieteilijöillekin* является существительным множественного числа и, следовательно, тег должен быть *N\_All\_Pl.* Аналогичная ошибка была допущена при разметке словоформы ***talous****näkemyksistään:* вместо тега *N\_Ill\_Sg* должен быть *N\_Ela\_Pl,* то есть, и падеж лексемы, и ее число были распознаны неверно. Но, как мы уже писали выше, подобные ошибки не оказали влияния на конечный результат.

2) Ошибки в части речи.

Во втором случае ошибка привела к тому, что были выделены слова, не являющиеся существительными. Например, словоформа *maa****talous****valtaisissa*была определена как существительное единственного числа, стоящее в инессиве (тег *N\_Ine\_Sg*), что неверно: *maa****talous****valtainen* является прилагательным, в данном случае стоящим в инессиве множественного числа, следовательно, тег должен быть *A\_Ine\_Pl*.

Аналогично в случае со словоформами ***velka****isin,* ***velka****isimpien* и ***velka****isimmista*, определенными как существительное (*N\_Nom\_Sg, N\_Gen\_Pl* и *N\_Par\_Sg* соответственно)*,* хотя данная лексема является прилагательным (начальная форма — ***velka****inen*), следовательно, теги должны быть *A\_Nom\_Sg, A\_Gen\_Pl* и *A\_Ela\_Sg.* Кроме того, данное прилагательное в обоих случаях стоит в превосходной форме, но в конечной разметке это не отражено, хотя тег для суперлативной формы прилагательных есть в списке тегов, используемых в разметке *TreeTagger*.

Употребленное в причастном предложном обороте причастие ***laina****avansa* было определено как существительное (*N\_Gen\_Sg;* правильный тег — *PrsPrc\_Act\_Gen\_Sg*).

Также при выделении слов, содержащих словоформы лексемы *tuki* (‘<финансовая> поддержка’), как существительные были определены причастия *oio****tui****sta* (*N\_Ela\_Pl*; правильный тег — *PrfPrc\_Pass\_Ela\_Pl*) и *vaihde****tui****n* (*N\_Nom\_Sg*; правильный тег — *PrfPrc\_Pass\_Nom\_Sg*).

Аналогичный случай имел место при выделении слов, содержащих лексему *palkka* (‘заработная плата’): причастию ***palkka****avansa* был присвоен тег *N\_Gen\_Sg* (правильный тег — *PrsPrc\_Act\_Gen\_Sg*).

При выделении терминов-кандидатов, связанных с лексемой *vero* (‘налог’), были также выделены словоформы ***vero****tettavaakaan* (*N\_Par\_Sg*, правильный тег — *PrsPrc\_Pass\_Par\_Sg* или *Adj\_Par\_Sg*, поскольку данная словоформа может быть либо пассивным причастием настоящего времени, либо прилагательным) и ***vero****tettavissa* (*N\_Ine\_Sg*, правильный тег — *PrsPrc\_Pass\_Ine\_Pl*).

Словоформа ***kaupa****takseen* была выделена по правилу для лексемы *kauppa* (‘торговля’) с тегом *N\_Ill\_Sg*, хотя ***kaupa****takseen* является долгой формой первого инфинитива глагола *kaupata* (‘продавать’), соответственно, правильным тегом будет *V\_Inf1\_Tra\_PxSg3.*

При выделении терминов-кандидатов, содержащих лексему *raha*, из-за неверной разметки были выделены прилагательные ***raha****kas* (*N\_Gen\_Sg*, правильный тег — *A\_Nom\_Sg*) и *vapaa****raho****itteisesta* (*N\_Ela\_Sg*, правильный тег — *A\_Ela\_Sg*)*.*

Также возникла проблема с разметкой имен собственных. Так, например, по грамматическому правилу для лексемы *pankki* (‘банк’) выделилось несколько названий банков:

*Sberbank-****pankki****, Lloyds-****pankki****, HSBC-****pankki****, Hansa****pankki****, Apple-****pankki****, Saksalais****pankki***.

 Аналогично для лексемы *yhtiö* (‘компания’) были выделены такие названия компаний как *Berner-****yhtiö***, *<Migrit>* *Solarna-****yhtiö***, *EQT-pääomasijoitus****yhtiö***.

Для лексемы *palvelu* (‘обслуживание, сервис’) были извлечены названия сервисов: *Uber-taksi****palvelu*** и *<Underhood>. co-****palvelu*.**

В случае лексемы *säätiö* (‘фонд’) были извлечены названия *Nuoriso****säätiö***и *Y-****säätiö***;в случае лексемы *tuote* (‘продукт’) — название *InnovFin-****tuote****perhe*.

Также было выделено название банковской карты *Mastercard* (*Mastercard-****maksu****kortti*)при извлечении терминов, в состав которых входит лексема *maksu* (‘платеж’).

При выделении терминов-кандидатов, содержащих основу *kauppa* (‘торговля’), было выделено название компании *Finsilva* в сложном слове *Finsilva-****kauppa,*** а также названия торговых центров: *Itäkeskus****-kauppa****keskus*и*Forum****-kauppa****keskus*.

Во всех вышеописанных случаях названия не были размечены как имена собственные (*N\_Prop\_.\**), что и привело к ошибкам.

1. Ошибки, связанные с выполнением правил грамматики.
2. Ошибки, связанные с семантикой слова, использованного в грамматике.

Зачастую ключевые слова, использованные нами в правилах грамматики, не являлись однозначными, причем не всегда их значения были связаны между собой.

Так, *kauppa* в стандартном языке означает обычный магазин, но в экономическом специальном языке данная лексема означает также торговлю, коммерцию, сделку. Это привело к тому, что среди выделенных терминов оказалось слово *ruokakauppa* (‘продуктовый магазин’).

В отличие от лексемы *kauppa*, значения которой все же относятся к одной и той же области, два значения лексемы *yritys* (‘попытка’ в стандартном языке и ‘предприятие, фирма’ в специальном языке экономики) относятся к разным семантическим полям. В нашем корпусе *yritys* в значении ‘попытка’ встретилось 8 раз (на 316 извлеченных терминов).

Извлекались также авторские слова, не являвшиеся авторскими терминами в силу того, что они зачастую несли дополнительный эмоциональный оттенок. Например, лексема *tiedemies****laina*** (дословно: ‘займ у ученого’, буквально: ‘ученый-займ’) в авторском контексте означало ‘цитата’:

Kun kerran komeasti Einsteinilla aloitin, pistänpä vielä toisenkin *tiedemies****laina****n*: Hulluutta on se, että tekee samat asiat uudelleen ja uudelleen ja odottaa eri tuloksia.

 Раз я так красиво начал с Эйнштейна, вставлю-ка еще одну цитату ученого: ‘Безумие: делать то же самое снова и снова, ожидая различные результаты’.

Также в текстах один раз встретилось такое сложное слово как ***tase****miina* (*tase* — ‘баланс’, *miina* — ‘мина’, буквально: ‘балансовая мина’). Данную лексему сложно классифицировать как термин, но она несомненно является авторской метафорой.

В сложном слове***finanssi****kämmi* главная часть — *kämmi* (‘ошибка’) дается в словаре с пометкой *colloquial* (‘разговорный’), что не позволяет говорить о данной лексеме как о термине. Относящейся к разговорному языку также можно считать лексему *taala****seteli***, где *taala* —разговорное название доллара, *seteli* — купюра.

Много экспрессивной лексики встретилось в результате выделения лексем, содержащих в себе лексему *raha* (‘деньги’), как то:

***raha****hana* (‘денежный кран’, буквально ‘деньги-кран’), ***raha****koe* (‘денежное испытание’, буквально ‘деньги-испытание’), ***raha****kopteri* (буквально: ‘деньгокоптеры’ или ‘деньголеты’), ***raha****ruiske*(‘денежное впрыскивание’, буквально ‘деньги-впрыскивание’), ***raha****temppu* (‘денежный трюк’, буквально ‘деньги-трюк’), *suosikki****raha*** (дословно: ‘любимые деньги’).

 Также к экспрессивной лексике можно отнести такие слова как***seteli****soppa* (‘суп из купюр’, ‘денежный суп’, буквально ‘купюра-суп’) и ***seteli****silppuri* (‘денежный шредер’, буквально ‘купюра-шредер’).

При выделении лексем, содержащих лексему *arvo* (‘ценность, стоимость’) была неоднократно выделена лексема *eri****arvo****isuus* (‘неравенство’). Хотя она действительно является в какой-то степени производным словом от *arvo* (*arvo > arvoinen > arvoisuus > eriarvoisuus*), ее значение не вполне соответствует значению изначального слова.

Лексема *tulo* (‘доход’) является многозначной; в частности, в значении, обозначающем прибытие куда-либо, она используется для образования слов, которые и были выделены, например: *mukaan****tulo***(‘вхождение <в компанию, концерн и т.д.>‘), *voimaan****tulo***(‘вступление в силу’), ***tulo****kas* (‘новичок, пришелец’), ***tulo****kasauma* (‘куча новичков’).

Кроме того, существуют достаточно общеупотребительные выражения *olla* ***tulo****ssa и tehdä* ***tulo****a* (‘приходить, быть на подходе’; буквально ‘находиться в прибытии’, ‘совершать прибытие’), которые также были выделены. Подобные конструкции были выделены и в случае лексемы *meno* (в множественном числе — ‘расходы’): *olla* ***meno****ssa* (‘идти’).

Похожая ситуация сложилась с лексемой *liike* (‘движение’, ‘бизнес’). Были выделены слова, образованные от *liike* со значением ‘движение’ (как в прямом, так и в переносном смысле):

 *ay-****liike*** или *ammattiyhdistys****liike*** — ‘профсоюзное движение’,

*ryhti****liike*** — ‘гимнастические движения, направленные на исправление осанки’,

 *muutto****liike***— ‘переселение’, ‘миграция’,

*kapina****liike*** — ‘повстанческое движение’.

Кроме этого, выделялись слова, содержащие лексему ***liike****nne* (‘движение, транспорт’) или производные от нее, например, *meri****liike****nne* (‘морской транспорт’), ***liike****nnöinti* (‘курсирование, движение’).

Также была выделена сама лексема *liike* в значении ‘движение’ – <*lähteä>* ***liike****elle* (‘<отправиться> в путь’).

Интересен случай с лексемой *lotto****voitto***(‘выигрыш в лотерею’), выделенной согласно правилу, написанному для лексемы *voitto* (‘прибыль’). В данном контексте лексема *lotto****voitto*** используется в переносном значении, а *voitto* в его составе выступает в своем другом значении — ‘выигрыш’.

При выделении слов, содержащих лексему *kustannus* (‘покрытие расходов’), была выделена лексема ***kustannus****toiminta* (‘издательское дело’), что было связано со вторым значением лексемы *kustannus* (‘издание’).

Кроме этого, помимо явных терминов, извлекались также метафоры – как устоявшиеся, так и нет, например:

***hinta****hissi* — ‘ценовой лифт’, означает постоянное изменение цен (обычно повышение).

1. Ошибки, связанные с составом слова, использованного в грамматике.

Как мы уже писали выше, мы старались подбирать ключевые слова с таким расчетом, чтобы избежать двусмысленности при выделении терминов. К сожалению, совсем исключить неоднозначность не получилось. Далее мы подробно рассмотрим все случаи подобных ошибок.

а)Первой лексемой, при извлечении которой мы столкнулись с такой проблемой, была лексема *korko* (‘процент’). В данном слове происходит чередование основы: *korko-* (сильная основа)/*koro-* (слабая основа), поэтому изначально правило для выделения терминов выглядело так:

*\*UNARY*

*=korko*

*1:[(lemma=".\*korko.\*"|lemma=".\*koro.\*")&tag="N\_.\*"]*

Согласно этому правилу выделялось достаточно много слов, содержащих лексему *korotus* (‘повышение’). Всего таких слов было 34 на 387 извлеченных терминов.

Поэтому мы решили модифицировать правило с учетом полученных результатов:

*\*UNARY*

*=korko*

*1:[(lemma=".\*korko.\*"|lemma=".\*koro.\*")&lemma!=".\*korotu.\*"&tag="N\_.\*"]*

Исключение из запроса основы *korotu-* никак не повлияло на производные от *korko* (‘процент’)в силу особенностей словообразования и чередования основ. *Koro-* является слабой основой слова, используемой в начале закрытого слога [Klusiilinen astevaihtelu], а следовательно, во всех местных падежах (кроме иллатива), в транслативе, абессиве, инструктиве, генитиве и для образования номинатива и аккузатива множественного числа. Но *-tu-* не является показателем какого-либо из этих падежей, следовательно, ни одна словоформа лексемы *korko* не исключается из поиска.

Что касается сложных слов, то в них используется либо номинативная, либо генитивная форма лексемы *korko*, например: *ohjaus****korko*** (‘процентная ставка’, буквально ‘управление-процент’) или ***korko****kate* (‘чистый процентный доход’, буквально ‘процент-покрытие’), где *korko* — номинативная форма, или ***koron****nosto* (‘повышение процентов’), где *koron* — генитивная форма.

Производные слова от *korko* могут образовываться от слабой основы, но в качестве производных будут выступать либо глаголы, либо прилагательные, которые будут отсеяны уже на этапе определения части речи. Кроме этого, *-tu-* не является словообразовательным суффиксом. Следовательно, ни одно производное сложное слово не будет пропущено при извлечении терминов.

б) При выделении лексем, содержащих лексему *hinta* (‘цена’), была выделена словоформа *louhinnan* (начальная форма *louhinta ‘*дробление, добыча <полезных ископаемых>‘). Лексема *louhinta* не имеет никакого отношения к лексеме *hinta* и образована от глагола *louhia* (‘дробить’, ‘добывать’) путем прибавления суффикса *-nta*.

в) В случае с уже упомянутой выше лексемой *tulo* (‘доход’)выделились слова, содержащие лексему *tulos* (‘результат’), например, *loppu****tulos*** (‘конечный результат’). Хотя доля подобных слов в нашем корпусе была невелика, проверка на корпусе *fiTenTen* показала, что количество слов, содержащих лексему *tulo*, и количество слов, содержащих лексему *tulos*, примерно одинаково.

В отличие от случая с лексемами *korko* и *korotus*, невозможно задать выделение лексем, содержащих исключительно лексему *tulo* и не содержащих лексему *tulos*, потому что если добавить в правило выражение *lemma!=".\*tulos.\*"* , то как минимум не будут выделяться инессивные и элативные формы лексемы *tulo* (показатели падежей *-ssa* и *-sta* соответственно), не говоря уже о сложных словах, в которых *tulos* будет образовываться на стыке морфем.

г) По правилу, написанному для лексемы *tili* (‘счет’) дважды была выделена лексема *ko****tili****esiasiakas*. В данном случае имеет место другое морфемное членение слова:

*kotiliesiasiakas* — *koti* (‘дом’) *+ liesi* (‘очаг’) *+ asiakas* (‘клиент’)

Искомое *tili* образуется на стыке двух основ сложного слова: *koti* и *liesi*.

Аналогично в случае с лексемами *brit****tile****hti* и *ryh****tili****ikkeeltä*, выделенными по тому же запросу:

*brittilehti* — *britti* (‘британский’) *+ lehti* (‘газета’)

*ryhtiliike* — ryhti (‘осанка’)+ liike (‘движение’)

Лексема *brittilehti* была выделена, поскольку содержит сочетание *tile-* — основу, указанную в правиле для форм множественного числа лексемы *tili*.

д) В случае уже упоминавшейся лексемы *tuki* (‘<финансовая> поддержка’), было выделено несколько сложных слов, содержавших в себе лексему *is****tui****n* (‘местонахождение, центр, сиденье’), образованную от глагола *istua* (‘сидеть’) путем прибавления суффикса *-in.* Также была выделена лексема *kier****tue***(‘обращение’, ‘оборот’, ‘цикл’), словоформы *johdannaisvas****tui****ta* и *johdannaisvas****tui****den* —партитив и генитив соответственно множественного числа леммы *vastuu* (‘ответственность, обязательство, гарантия’) и лексема ***tui****jottaminen* (‘пристальное разглядывание’).

Слова, в которые входили сочетания *tui*- и *tue-* выделялись за счет того, что при написании правила мы учли чередование основы данной лексемы (*tue-* — слабая основа единственного числа, *tui-* — слабая основа множественного числа).

Все вышеперечисленные лексемы по смыслу и значению никак не связаны с лексемой *tuki.*

е) При извлечении слов, содержащих лексему *lama* (‘кризис’, ‘спад’, ‘<экономическая> депрессия’) была выделена словоформа *Uudel****lama****alla* (начальная форма *Uusimaa* — Уусимаа, область Финляндии). Совпадение произошло из-за того, что в данной лексеме склоняются обе части сложного слова, что и привело в адессивной форме (показатель *-lla*) к появлению *lama* на стыке морфем. Кроме того, лексема *Uusimaa* вообще не должна была быть выделена, поскольку она является именем собственным (названием административной единицы).

ж) Много проблем возникло при выделении слов, содержащих лексему *osto* (‘покупка’): примерно половина результатов (50 из 135) не имеют отношения к данной лексеме. Чаще всего среди неверных результатов встречались лексемы, содержащие в своем составе *n****osto*** (‘подъем’) и *neuv****osto*** (‘совет’), но были и другие, например, *verk****osto***(‘сеть’), *tied****osto***(‘файл’). Это связано с тем, что *-sto* — очень продуктивный словообразующий суффикс в финском языке, с помощью которого можно образовывать коллективные существительные от существительных, обозначающих отдельные предметы.

з) При выделении слов, содержащих лексему *kurssi* (‘курс’), также выделялись слова, содержащие лексему *kon****kurssi*** (‘банкротство’). Хотя *konkurssi* тоже можно считать экономическим термином, данная лексема никак не связана с лексемой *kurssi*, следовательно, ее извлечение согласно данному правилу нежелательно.

и) При выделении слов, содержащих лексему *tase* (‘баланс’), из корпуса была извлечена лексема *taus****tase****lvitys* (‘наведение справок’)*.* Как и в случаях с лексемами *tili* и *lama*, причина заключается в другом морфологическом составе слова: данное сложное слово состоит из основ *tausta* (‘фон’) и *selvitys* (‘выяснение’); на стыке морфем получается *tase*. Также по непонятной причине по данному запросу была выделена лексема *markkina-asema*.

к) В случае с выделением производных от лексемы *vara* (‘запас’; во множественном числе — ‘средства’, ‘ресурсы’) были выделены слова, содержащие лексему *ta****vara*** (‘товар’), слова, содержащие лексему ***varo****itus* (‘предупреждение’), а также слова, содержащие префикс ***vara****-* (‘запасной’, ‘вице-’, ‘заместитель’).

л) При выделении лексем, содержащих лексему *raha* (‘деньги’), была выделена словоформа *vuok****raho****ivatilaratkaisuja*, что неверно, поскольку данная лексема имеет другое морфемное членение:

*vuok****raho****ivatilaratkaisu* — *vuokra* (‘аренда’) *+ hoiva* (‘опека’) *+ tila* (‘пространство’) *+ ratkaisu* (‘решение’)

м) При извлечении кандидатов в термины, содержащих основу *omist-* (от *omistaa* — ‘владеть’), были выделены термины *ekon****omist****i* (‘экономист’) и *pääekon****omist****i* (‘главный экономист’), которые не являются производными от данной основы.

3.6. Выводы

В данной главе мы описали грамматику *SketchEngine*, с помощью которой нами проводилось извлечение терминов, описали материал исследования, результаты исследования и ошибки, возникшие при автоматическом извлечении терминов.

Типичная грамматика *SketchEngine* состоит из правил, написанных на *CQL* (*Corpus Query Language* — языке запросов к корпусу, который используется в *SketchEngine* для создания коллокаций и списков слов и словосочетаний).

Мы проводили наше исследование на корпусе экономических текстов, собранном на различных сайтах экономических новостей и размеченном с помощью морфологического теггера *TreeTagger,* существующего в системе *SketchEngine*.

Результаты мы оцениваем как положительные: в среднем точность извлеченных кандидатов в термины превышает 0,60. Надо отметить, что в нашей оценке могут быть погрешности, поскольку при проверке результатов и вычислении точности мы не пользовались полноценным словарем экономической лексики. В частности, при проверке по онтологии мы неоднократно сталкивались со случаями, когда термин-кандидат входил в состав сложного слова-термина, но сам по себе как термин обозначен не был.

К сожалению, нам не удалось избежать определенных ошибок — как на уровне разметки, так и на уровне функционирования грамматики.

На уровне разметки ошибки касались падежа или числа лексемы, что никак не влияло на конечный результат, и части речи, что приводило к лишним словам в списках слов: в частности, как существительные были распознаны прилагательные и причастия настоящего времени, входящие в причастный предложный оборот. В случае причастий причиной неправильной разметки может быть притяжательный показатель, характерный для данного оборота, в то время как в обычных предложениях притяжательный показатель присоединяется исключительно к существительным и никогда не присоединяется ни к прилагательным, ни к причастиям.

На уровне функционирования грамматики ошибки были связаны либо с семантическим значением выделяемого термина (в случае его многозначности), либо с его написанием и морфологической структурой (в случае другого морфемного членения кандидата в термины). Такого рода ошибок было больше, чем ошибок, связанных с разметкой, что обусловлено наличием в правиле для одной лексемы нескольких основ.

Заключение

Количество специальной лексики в языке с каждым годом увеличивается в силу того, что все сферы человеческой жизни быстро развиваются: возникают новые объекты и явления, для которых необходимы новые наименования, изменяется содержание уже существовавших понятий, что приводит к необходимости изменений в названиях.

В большинстве случаев все эти новоприобретенные языком слова могут считаться терминами — единицами специального языка, которые выражают основные характеристики научного понятия.

Термин может образоваться через словообразование, терминологизацию или заимствование из других языков, а также через сочетание этих способов.

Существует несколько классификаций терминов как в рамках русского языкознания, так и в рамках финской лингвистики. С точки зрения морфемной структуры термины-слова могут быть непроизводными, производными, сложными и аббревиатурами, а с точки зрения части речи — существительными, глаголами, прилагательными и наречиями. Стоит также отметить, что, по мнению некоторых исследователей, термином может быть исключительно существительное. В нашем исследовании мы выделяли только термины-существительные, большая часть которых являлась сложными словами.

Корпус, которым мы пользовались для исследования, состоит из газетных статей на экономическую тему (в качестве источников нами использовались финноязычные новостные сайты). Объем корпуса — 50 тыс. токенов. Корпус был размечен с помощью морфологического теггера *TreeTagger,* существующего в системе *SketchEngine*.

Мы также рассмотрели существующие системы для автоматического выделения терминов. В большинстве из них используется статистический подход, хотя в некоторых привлекается дополнительная лингвистическая информация (например, морфологическая разметка).

Наш подход, использующий инструмент *SketchEngine*, также является гибридным, поскольку в нем используются правила, опирающиеся на морфологические характеристики.

Для извлечения терминов мы загрузили свой корпус в систему *SketchEngine*, разметили его с помощью *TreeTagger* и написали скетч-грамматику из 51 правила: 50 — для выделения слов, 1 — для словосочетания (см. приложение 1). В некоторых случаях мы дополнительно проверяли работоспособность правил грамматики на корпусе *fiTenTen.*

Результаты мы оцениваем в целом как положительные: хотя точность извлеченных кандидатов в термины зависит от конкретного случая, (максимальная точность – 0,93, минимальная – 0,22) в целом точность колебалась в районе 0,50 – 0,70, а поскольку наш подход подразумевал выделение всех терминов, содержащих заданную основу, мы можем говорить о 100% полноте. Тем не менее, наша оценка весьма приблизительна, поскольку мы не использовали полноценный словарь экономической лексики для проверки результатов. Мы также не пользовались помощью специалиста в данной области.

Кроме терминов по правилам нашей грамматики выделялись также лексемы, несшие дополнительный экспрессивный оттенок или имевшие метафорическое значение. Мы считаем, что дальнейшее изучение подобных случаев представляет особый интерес с точки зрения семантики, стилистики и словообразования финского языка.

К сожалению, нам не удалось избежать определенных ошибок — как на уровне разметки, так и на уровне функционирования грамматики. Мы сделали небольшую классификацию ошибок и подробно рассмотрели каждый случай.

Мы бы хотели отметить значение нашего исследования для морфологической разметки *TreeTagger for Finnish v2*. В ходе исследования нами были выявлены определенные ошибки, связанные с неправильной разметкой. Мы надеемся, что наши замечания будут полезны при дальнейшей ручной проверке данной разметки.

Список использованной литературы

1. Бабенко О. В. Linguistic aspects of the economic term system. URL: <http://elibrary.nubip.edu.ua/10430/1/11bov.pdf> (дата обращения: 19.05.2016).
2. Браславский П. И., Соколов Е. А. Сравнение пяти методов извлечения терминов произвольной длины. URL: <http://www.dialog-21.ru/dialog2008/materials/html/11.htm> (дата обращения: 19.05.2016).
3. Герд А. С*.* Ещё раз об автоматизации лексикографических работ // Советская лексикография, Научный совет по лексикологии и лексикографии Академии Наук СССР. М.: Русский язык. С. 225-232 // La Filológica por la Causa, 2014.
4. Герд А. С. Значение термина и научное знание / А. С. Герд // Научно-техническая информация. Сер. 2. – 1991. – № 10. – С. 1–4.
5. Герд А. С. Основы научно-технической лексикографии. Л., 1986.
6. Герд А. С. Прикладная лингвистика. СПб., 2005. Разделы I и II.
7. Головин Б. Н., Кобрин Р. Ю. Лингвистические основы учения о терминах. М.: Высшая школа, 1987. — 105 с.
8. Гринев-Гриневич С. В. Терминоведение. М.: Издательский центр ‘Академия’, 2008. — 304 с. — (Серия: Высшее профессиональное образование)
9. Жеребило Т. В. Словарь лингвистических терминов, издание 5-е, исправленное и дополненное. Назрань: ООО ‘Пилигрим’, 2010. – 486 с.
10. Лантюхова Н. Н., Загоровская О. В., Литвинова Т. А. Термин: определение понятия и его сущностные признаки // Вестник Воронежского института ГПС МЧС России, 2013, выпуск № 1 (6). URL: <http://cyberleninka.ru/article/n/termin-opredelenie-ponyatiya-i-ego-suschnostnye-priznaki> (дата обращения: 19.05.2016).
11. Лейчик В. М. Терминоведение: предмет, методы, структура. М., 2006.
12. Лейчик В. М., Шелов С. Д. Лингвистические проблемы терминологии и научно-технический перевод // Серия ‘Теория и практика научно-технического перевода’: Обзор информации. Вып. 18. Часть II. М.: Всесоюзный центр переводов научно-технической информации и документации, 1990. 80 с.
13. Литовченко В. И. Классификация и систематизация терминов // Вестник Сибирского государственного аэрокосмического университета им. академика М.Ф. Решетнева, 2006, выпуск № 3, с. 156-159. URL: <http://cyberleninka.ru/article/n/klassifikatsiya-i-sistematizatsiya-terminov> (дата обращения: 19.05.2016).
14. Лотте Д. С. Основы построения научно-технической терминологии. М.: АН СССР, 1961.
15. Маслов Ю. С. Введение в языкознание: Учеб. для филол. спец. вузов. —2-изд., перераб. и доп. — М.:Высш. шк., 1987. 272 с.
16. ПРОМТ объявляет о выпуске нового продукта ПРОМТ Terminology Manager. URL: <http://www.promt.ru/press/news/10111/?phrase_id=20603485> (дата обращения: 21.05.2016).
17. Хакулинен Л. Развитие и структура финского языка. Часть II // Лексикология и синтаксис. М., 1955.
18. Шелов С. Д. Еще раз об определении понятия ‘термин’ // Вестник Нижегородского университета им. Н.И. Лобачевского, 2010, выпуск № 4-2, с. 795-799. URL: <http://cyberleninka.ru/article/n/esche-raz-ob-opredelenii-ponyatiya-termin> (дата обращения: 19.05.2016).
19. Яковлев Н. Ф. Грамматика литературного кабардино-черкесского языка. М., 1948.
20. Benko V. Compatible Sketch Grammars for Comparable Corpora // Proceedings of the XVI EURALEX International Congress: The User in Focus, Bolzano/Bozen, 2014. URL: <http://www.academia.edu/7781794/Compatible_Sketch_Grammars_for_Comparable_Corpora> (дата обращения: 22.05.2016).
21. Cabré, M. T., Estopà, R., Vivaldi, J. Automatic term detection: a review of current systems // Bourigault, D.; Jacquemin, C.; L’Homme, M-C. (2001) Recent Advances in Computational Terminology, p. 53-88.
22. Corcodel D., Corcodel A. On the specialized language of economic field // STUDIA UNIVERSITATIS MOLDAVIAE, 2014, nr.10(80), p.122-124. URL: <http://oaji.net/articles/2015/2055-1432569893.pdf> (дата обращения: 19.05.2016).
23. Fkih F., Omri M. N., Complex Terminology Extraction Model from Unstructured Web Text based Linguistic and Statistical Knowledge // IJIRR: International Journal of Information Retrieval Research. 2(3), 1-18, 2013.
24. Ison suomen kieliopin verkkoversio (VISK). Kotimaisten kielten tutkimuskeskus, 2008 [Электронный ресурс]. URL: <http://scripta.kotus.fi/visk/etusivu.php> (дата обращения: 19.05.2016).
25. Jackson, P., Moulinier, I. Natural Language Processing for online applications: text retrieval, extraction and categorization. John Benjamins Publishing Company, 2002. – 226 p.
26. Jakubíček M., Kilgarriff A., McCarthy D., Rychlý P. Fast syntactic searching in very large corpora for many languages. Japan, PACLIC 2010.
27. Kilgarriff A., Jakubíček M., Kovář V., Rychlý P., Suchomel V. Finding Terms in Corpora for Many Languages with the Sketch Engine // Proceedings of the Demonstrations at the 14th Conference the European Chapter of the Association for Computational Linguistics*,* Sweden, April 2014, p. 53–56. URL: <https://www.sketchengine.co.uk/wp-content/uploads/Finding_Terms_2014.pdf> (дата обращения: 21.05.2016).
28. Klusiilien astevaihtelu // Verkkokielioppi. Suomen kielen äänne-, muoto- ja lauseoppia. URL: <https://www.finnlectura.fi/verkkokielioppi/Morfologia/sivu223.htm> (дата обращения: 22.05.2016).
29. [Korpela](https://www.cs.tut.fi/~jkorpela/henkkoht.html), J. Suomen kielen yhdyssanamuodot (kompositiivit). URL: <https://www.cs.tut.fi/~jkorpela/suomi/kompos.html> (дата обращения: 19.05.2016).
30. Kupść, A. Extraction automatique de termes à partir de textes polonaise // TALN 2007. Toulouse: 2007. URL: http://llf.linguist.jussieu.fr/llf/Gens/Kupsc/kupsc-taln07.pdf (дата обращения: 20.04.2016).
31. Language resources and tools | Sketch Engine. URL: <https://www.sketchengine.co.uk/language-resources-and-tools/> (дата обращения: 22.05.2016).
32. Museanu, E.  Economic terminology — new perspectives. URL: <http://eujournal.org/index.php/esj/article/viewFile/1254/1263> (дата обращения: 19.05.2016).
33. Resche, C.  Equivocal Economic Terms or Terminology Revisited // Journal des traducteurs /Translators' Journal, vol. 45, n° 1, 2000, p. 158-173.  URL: <http://id.erudit.org/iderudit/003940ar> (дата обращения: 19.05.2016).
34. Term Extraction | fivefilters.org. URL: http://fivefilters.org/term-extraction/ (дата обращения: 21.05.2016).
35. Term extraction analysis done by TermCoord - Terminology Coordination Unit [DGTRAD] - European Parliament. URL: http://termcoord.eu/discover/free-term-extractors/term-extraction-analysis-done-termcoord/
36. Termine. URL: <http://www.nactem.ac.uk/software/termine> (дата обращения: 15.05.2016).
37. Terminologian sanasto. Helsinki: Sanastokeskus TSK ry. 2006 – 40 s.
38. Terminology Extraction. URL: <http://labs.translated.net/terminology-extraction/> (дата обращения: 22.05.2016).
39. Terminology Extraction < Heartsome Europe GmbH. URL: <http://www.heartsome.de/en/termextraction.php> (дата обращения: 22.05.2016).
40. Terms Extraction. URL: <http://www.sanakirja.net/terms-extraction?___store=gb&___from_store=fi> (дата обращения: 22.05.2016).
41. TreeTagger. URL: http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
42. topia.termextract 1.1.0: Python Package Index. URL: <https://pypi.python.org/pypi/topia.termextract/1.1.0> (дата обращения: 21.05.2016).
43. Tyysteri, L. Yhdyssanatermien rakenne erikoiskielen erityisluonteen kuvaajana. // Kieli ja tunteet. Käännösteoria, ammattikielet ja monikielisyys. VAKKI-juhlasymposiumi XXX. Vaasa 12.–13.2.2010. Vaasan yliopiston käännöstieteiden, ammattikielten ja monikielisyyden tutkijaryhmän julkaisut 37, s. 350–360.
44. Voutilainen A., Purtonen T., Muhonen K. FinnTreeBank2 Manual. University of Helsinki, Department of Modern Languages, 2012 – 159 s.

Список источников

1. Вахрос И., Щербаков А. Большой финско-русский словарь. М., 2007.
2. Arvopaperi (‘Ценная бумага’). URL: <http://www.arvopaperi.fi/>(дата обращения: 22.05.2016)**.**
3. Finto: YSO — Yleinen suomalainen ontologia. URL: <https://finto.fi/yso/fi/> (дата обращения: 22.05.2016)**.**
4. Kauppalehti: Tärkeimmät talousuutiset(‘Торговая газета: Самые важные новости экономики’). URL: <http://www.kauppalehti.fi/>.(дата обращения: 22.05.2016)**.**
5. Sanakirja.org — Ilmainen Sanakirja. Suomi, englanti, ruotsi jne. nettisanakirja. URL: <http://www.sanakirja.org/> (дата обращения: 22.05.2016)**.**
6. Savon Sanomat (‘Газета Саво’). URL: http://www.savonsanomat.fi/**.**
7. Sivistysanakirja. URL: http://www.suomisanakirja.fi/ (дата обращения: 19.05.2016)
8. Talouselämä (‘Экономическая жизнь’). URL: http://www.talouselama.fi/.(дата обращения: 20.05.2016)
9. Taloussanakirja | Talousuutiset(‘Экономический словарь | Экономические новости’). URL: <http://www.taloussanomat.fi/porssi/sanakirja> (дата обращения: 19.05.2016)
10. Taloussanomat | Talousuutiset(‘Экономическая газета: Экономические новости’). URL: taloussanomat.fi (дата обращения: 19.05.2016)**.**
11. Wiktionary, the free dictionary. URL: [https://en.wiktionary.org/wiki/Wiktionary:Main\_Page](https://en.wiktionary.org/wiki/Wiktionary%3AMain_Page) (дата обращения: 22.05.2016)**.**
12. Yle Uutiset (‘Новости Yle’). URL: http://yle.fi/ (дата обращения: 19.05.2016**).**
13. Yrittäjät.fi (‘Предприниматели’). URL: http://www.yrittajat.fi/ (дата обращения: 23.04.2016)**.**

Приложение 1. Грамматика SketchEngine для выделения терминов

Word Sketch Engine#Word Sketch Engine Sketch Grammar for Finnish

\*STRUCTLIMIT s

\*DEFAULTATTR tag

\*UNARY

=talous

1:[(lemma=".\*talou(s|d|t|ks).\*")&tag="N\_.\*"]

\*UNARY

=vero

1:[lemma=".\*vero.\*"&tag="N\_.\*"]

\*UNARY

=korko

1:[(lemma=".\*korko.\*"|lemma=".\*koro.\*")&lemma!=".\*korotu.\*"&tag="N\_.\*"]

\*UNARY

=velka

1:[(lemma=".\*velk(a|o).\*"|lemma=".\*vel(a|oi).\*")&tag="N\_.\*"]

\*UNARY

=kauppa

1:[(lemma=".\*kaupp(a|o).\*"|lemma=".\*kaup(a|o).\*")&tag="N\_.\*"]

\*UNARY

=hinta

1:[(lemma=".\*hint(a|o).\*"|lemma=".\*hinn(a|o).\*")&tag="N\_.\*"]

\*UNARY

=vienti

1:[(lemma=".\*vient(i|e).\*"|lemma=".\*vienni.\*")&tag="N\_.\*"]

\*UNARY

=laina

1:[(lemma=".\*lain(a|o).\*")&tag="N\_.\*"]

\*UNARY

=yritys

1:[lemma=".\*yrity(s|ks).\*"&tag="N\_.\*"]

\*UNARY

=tuotanto

1:[(lemma=".\*tuotanto.\*"|lemma=".\*tuotanno.\*")&tag="N\_.\*"]

\*UNARY

=arvo

1:[lemma=".\*arvo.\*"&tag="N\_.\*"]

\*UNARY

=markkina

1:[lemma=".\*markkin(a|oi).\*"&tag="N\_.\*"]

\*UNARY

=pankki

1:[(lemma=".\*pankk(i|e).\*"|lemma=".\*pank(i|e).\*")&tag="N\_.\*"]

\*UNARY

=rahasto

1:[lemma=".\*rahasto.\*"&tag="N\_.\*"]

\*UNARY

=tulo

1:[lemma=".\*tulo.\*"&tag="N\_.\*"]

\*UNARY

=osake

1:[(lemma=".\*osake.\*"|lemma=".\*osakke.\*")&tag="N\_.\*"]

\*UNARY

=tili

1:[lemma=".\*til(i|e).\*"&tag="N\_.\*"]

\*UNARY

=finanssi

1:[lemma=".\*finanss(i|e).\*"&tag="N\_.\*"]

\*UNARY

=pörssi

1:[lemma=".\*pörss(i|e).\*"&tag="N\_.\*"]

\*UNARY

=raha

1:[lemma=".\*rah(a|o).\*"&tag="N\_.\*"]

\*UNARY

=tuki

1:[(lemma=".\*tuk(i|e).\*"|lemma=".\*tu(e|i).\*")&tag="N\_.\*"]

\*UNARY

=investointi

1:[(lemma=".\*investoint(i|e).\*"|lemma=".\*investoinn.\*")&tag="N\_.\*"]

\*UNARY

=yhtiö

1:[lemma=".\*yhtiö.\*"&tag="N\_.\*"]

\*UNARY

=liike

1:[(lemma=".\*liike.\*"|lemma=".\*liikke.\*")&tag="N\_.\*"]

\*UNARY

=kilpailu

1:[lemma=".\*kilpailu.\*"&tag="N\_.\*"]

\*UNARY

=kulu

1:[lemma=".\*kulu.\*"&tag="N\_.\*"]

\*UNARY

=kustannus

1:[lemma=".\*kustannu(s|ks).\*"&tag="N\_.\*"]

\*UNARY

=pääoma

1:[lemma=".\*pääom(a|i).\*"&tag="N\_.\*"]

\*UNARY

=lama

1:[lemma=".\*lam(a|o).\*"&tag="N\_.\*"]

\*UNARY

=luotto

1:[(lemma=".\*luotto.\*"|lemma=".\*luoto.\*")&tag="N\_.\*"]

\*UNARY

=tappio

1:[lemma=".\*tappio.\*"&tag="N\_.\*"]

\*UNARY

=voitto

1:[(lemma=".\*voitto.\*"|lemma=".\*voito.\*")&tag="N\_.\*"]

\*UNARY

=meno

1:[lemma=".\*meno.\*"&tag="N\_.\*"]

\*UNARY

=lasku

1:[lemma=".\*lasku.\*"&tag="N\_.\*"]

\*UNARY

=osto

1:[lemma=".\*osto.\*"&tag="N\_.\*"]

\*UNARY

=omistaminen

1:[lemma=".\*omist.\*"&tag="N\_.\*"]

\*UNARY

=palkka

1:[(lemma=".\*palkk(a|o).\*"|lemma=".\*palka.\*")&tag="N\_.\*"]

\*UNARY

=palvelu

1:[lemma=".\*palvelu.\*"&tag="N\_.\*"]

\*UNARY

=seteli

1:[lemma=".\*setel(i|ei).\*"&tag="N\_.\*"]

\*UNARY

=kurssi

1:[lemma=".\*kurss(i|e).\*"&tag="N\_.\*"]

\*UNARY

=siirto

1:[(lemma=".\*siirto.\*"|lemma=".\*siirro.\*")&tag="N\_.\*"]

\*UNARY

=säätiö

1:[lemma=".\*säätiö.\*"&tag="N\_.\*"]

\*UNARY

=tase

1:[lemma=".\*tase.\*"&tag="N\_.\*"]

\*UNARY

=tuote

1:[(lemma=".\*tuote.\*"|lemma=".\*tuotte.\*")&tag="N\_.\*"]

\*UNARY

=vakuus

1:[lemma=".\*vakuu(s|t|d|ks).\*"&tag="N\_.\*"]

\*UNARY

=valuutta

1:[(lemma=".\*valuutt(a|o).\*"|lemma=".\*valuut(a|o).\*")&tag="N\_.\*"]

\*UNARY

=vara

1:[lemma=".\*var(a|o).\*"&tag="N\_.\*"]

\*UNARY

=vekseli

1:[lemma=".\*veksel(i|ei).\*"&tag="N\_.\*"]

\*UNARY

=indeksi

1:[lemma=".\*indeks(i|e).\*"&tag="N\_.\*"]

\*UNARY

=maksu

1:[lemma=".\*maksu.\*"&tag="N\_.\*"]

\*DUAL

=termi\adj

2:[tag="A\_Nom\_.\*"] [tag="A\_Nom\_.\*"]? 1:[lemma=".\*talou(s|d).\*"&tag="N\_Nom\_.\*"]

2:[tag="A\_Gen\_.\*"] [tag="A\_Gen\_.\*"]? 1:[lemma=".\*talou(d|ks).\*"&tag="N\_Gen\_.\*"]

2:[tag="A\_Par\_.\*"] [tag="A\_Par\_.\*"]? 1:[lemma=".\*talou(t|ks).\*"&tag="N\_Par\_.\*"]

2:[tag="A\_Acc\_.\*"] [tag="A\_Acc\_.\*"]? 1:[lemma=".\*talou(s|d).\*"&tag="N\_Acc\_.\*"]

2:[tag="A\_Ill\_.\*"] [tag="A\_Ill\_.\*"]? 1:[lemma=".\*talou(t|ks).\*"&tag="N\_Ill\_.\*"]

2:[tag="A\_Ine\_.\*"] [tag="A\_Ine\_.\*"]? 1:[lemma=".\*talou(d|ks).\*"&tag="N\_Ine\_.\*"]

2:[tag="A\_Ela\_.\*"] [tag="A\_Ela\_.\*"]? 1:[lemma=".\*talou(d|ks).\*"&tag="N\_Ela\_.\*"]

2:[tag="A\_All\_.\*"] [tag="A\_All\_.\*"]? 1:[lemma=".\*talou(d|ks).\*"&tag="N\_All\_.\*"]

2:[tag="A\_Ade\_.\*"] [tag="A\_Ade\_.\*"]? 1:[lemma=".\*talou(d|ks).\*"&tag="N\_Ade\_.\*"]

2:[tag="A\_Abl\_.\*"] [tag="A\_Abl\_.\*"]? 1:[lemma=".\*talou(d|ks).\*"&tag="N\_Abl\_.\*"]

2:[tag="A\_Tra\_.\*"] [tag="A\_Tra\_.\*"]? 1:[lemma=".\*talou(d|ks).\*"&tag="N\_Tra\_.\*"]

2:[tag="A\_Ess\_.\*"] [tag="A\_Ess\_.\*"]? 1:[lemma=".\*talou(t|ks).\*"&tag="N\_Ess\_.\*"]

2:[tag="A\_Abe\_.\*"] [tag="A\_Abe\_.\*"]? 1:[lemma=".\*talou(d|ks).\*"&tag="N\_Abe\_.\*"]

2:[tag="A\_Com\_.\*"] [tag="A\_Com\_.\*"]? 1:[lemma=".\*talou(t|ks).\*"&tag="N\_Com\_.\*"]

2:[tag="A\_Ins\_.\*"] [tag="A\_Ins\_.\*"]? 1:[lemma=".\*talouks.\*"&tag="N\_Ins\_.\*"]

1. Похожий подход к выделению терминов используется в *Terminology Extraction* компании *Translated Labs* с одним отличием: вместо специального корпуса используется входной текст. [↑](#footnote-ref-1)
2. При написании терминов мы выделяем жирным шрифтом те части сложных слов, которые использовались в нашей грамматике. В угловых скобках (<>) мы пишем контекст, который не был выделен в процессе извлечения терминов, но который имеет значение. [↑](#footnote-ref-2)
3. Лексема *vekseli* («вексель») и ее производные в нашем корпусе отсутствуют, поэтому мы проверили работоспособность правила на корпусе *fiTenTen* и получили 1411 результат (0.83 на миллион). [↑](#footnote-ref-3)
4. Хотя мы указываем глагол как ключевое слово, по правилу, написанному для него, будут выделяться существительные, при этом основа, которая будет выделяться в кандидатах в термины — *omist-*. [↑](#footnote-ref-4)