Моделирование и анализ информационных систем. Т. 22, №6 (2015), с. 834–851
Modeling and Analysis of Information Systems. Vol. 22, No 6 (2015), pp. 834–851

DOI: 10.18255/1818-1015-2015-6-834-851

УДК 004.853

Лингвостатистический анализ терминологии для построения тезауруса предметной области

Каряева М. С.[1]

получена 15 мая 2015

Работа посвящена анализу корпуса терминов и терминологических источников с целью дальнейшей автоматизации построения тезауруса данной предметной области, в качестве которой рассматривается поэтология. Предварительная систематизация терминологии с использованием лингвостатистического подхода формирует корпус семантически связанных понятий для автоматизации извлечения семантических отношений между терминами, определяющих структуру тезауруса указанной предметной области.

Ключевые слова: тезаурус, метрики семантической близости, data mining, компьютерная лингвистика

Для цитирования: Каряева М. С., "Лингвостатистический анализ терминологии для построения тезауруса предметной области", *Моделирование и анализ информационных систем*, **22**:6 (2015), 834–851.

Об авторах:
Каряева Мария Сергеевна, orcid.org/0000-0003-4466-1735, аспирант,
Ярославский государственный университет им. П.Г. Демидова,
ул. Советская, 14, г. Ярославль, 150000 Россия,
e-mail: mari.karyaeva@gmail.com

Благодарности:
[1]Работа поддержана грантом РФФИ № 16-07-01180

Введение

В последнее время возрос интерес к такому виду представления знаний, как тезаурус. Под тезаурусом обычно понимают словарь концептов с определенной структурой хранения данных и набором семантических отношений, указывающих на общность (например, синонимический ряд) или противопоставление значений лексических единиц. Главное отличие тезауруса от словаря – это система представления данных, которая позволяет использовать тезаурус не только как средство для отображения информации в удобочитаемом виде, но и для дальнейшей работы с ним, как с источником знаний для задач, связанных с компьютерной лингвистикой и информационным поиском. Например, российский тезаурус РуТез [1] используется в качестве ресурса для автоматической обработки текстов в проектах для государственных и коммерческих организаций. РуТез относится к классу тезаурусов типа WordNet [2], то есть базы знаний знаменательных частей речи (существительных, глаголов, наречий, прилагательных), где лексической единицей является «синсет», или синонимический ряд, то есть набор слов со схожим значением. Тем самым, за счет семантической сети между концептами возможен автоматический анализ текстов и ряд других задач, которые возможно решить с помощью такого мощного инструмента.

Наличие предметно-ориентированного тезауруса позволяет значительно упростить процесс сбора, формализации, хранения, оценки и использования знаний, что способствует повышению эффективности работы специалиста или рабочей группы выбранной предметной области.

Создание тезаурусов предметных областей – достаточно трудоемкий и дорогостоящий процесс, поскольку при этом необходимо объединение усилий целых групп соответствующих специалистов и экспертов для обработки большого числа объемных источников информации: словарей, справочников, научных публикаций и других текстов. Естественным подходом к созданию тезауруса в связи с развитием информационных технологий является комбинирование как ручных, так и автоматических методов.

В данном случае в качестве предметной области выбрана поэтология, под которой понимается группа дисциплин, ориентированных на всестороннее теоретическое и историческое изучение поэзии. Для тезаврирования поэтологии были заложены достаточные основы [3, 4], в которых экспертами предметной области был разработан базовый терминологический словник тезауруса, насчитывающий полторы тысячи специальных терминов, представленных в 10 предметных подобластях (кластерах).

Представлялось логичным для ускорения заполнения основных информационных полей терминологических статей тезауруса (ТСТ) создавать тезаурус как открытый сетевой ресурс с полноценным доступом пользователей к составлению и редактированию корпуса ТСТ [5]. Однако после размещения в сети Интернет прототипа тезауруса [6] с использованием Wiki-технологий стало ясно, что на этом этапе такой краудсорсинговый подход к созданию тезауруса не эффективен в связи с недостатком мотивации у пользователей самостоятельно развивать лингвистические ресурсы. Поэтому оказался методологически актуальным подход с автоматическим составлением тезауруса на основе терминологии предметной области, а имен-

но имеющегося базового терминологического словника тезауруса и ряда источников первичной информации, содержащих информацию для заполнения основных полей ТСТ, как явную – для определения термина, так и неявную – для семантических связей данного термина с другими. Автоматизация лингвостатистического анализа этой терминологии с необходимостью предваряет задачу автоматического составления тезауруса.

Изучение основных закономерностей при построении тезауруса поможет формализовать основные шаги и методики для создания баз знаний и способствовать развитию тезаурусов в других предметных областях. В качестве метода для автоматического извлечения семантических отношений был выбран метод, основанный на извлечении информации из определений рассматриваемых терминов. Семантическая близость терминов оценивается с помощью двух метрик близости и ручной оценки эксперта предметной области.

В качестве отношений между терминами для автоматического извлечения могут быть рассмотрены:

1. Родо-видовые отношения;
2. Целое–части;
3. Отношение синонимии.

Автоматическое распознавание семантических отношений возможно с помощью лексико-синтаксических шаблонов [7]. В данной статье были представлены достаточно простые конструкции для извлечения родо-видовых отношений при построении тезауруса WordNet. Извлечение семантических отношений без идентификации по типам представлено в работе [8]. Метод заключается в использовании статей Википедии для реализации методов машинного обучения, основанных на алгоритмах ближайших и взаимных ближайших соседей и двух метриках семантической близости слов. Результаты извлечения используются в системе Serelex [9] для поиска семантически связанных слов.

Для извлечения родо-видовых отношений был достаточно успешно использован подход [10] на основе использования определений толковых словарей с наложением ряда правил для распознавания отношений между определяемым словом в качестве рода и одним из слов дефиниции в качестве вида.

1. Формальная постановка задачи

Пусть $C = \{c_1, c_2, \ldots, c_N\}$ — множество терминов, $D = \{d_1, d_2, \ldots, d_N\}$ — множество определений, R — пары семантически связанных терминов.

Тогда задачей алгоритма служит задача распознавания множества семантических отношений, представленных в виде $R = \{(c_i, c_j), (c_{i+1}, c_{j+1})\}$ из всевозможных пар терминов. Т.е. необходимо построить функцию $F : R \subseteq C \times C \to \{0, 1\}$ и выбрать пары терминов, для которых $F = 1$.

2. Исходные данные

2.1. Базовый терминологический словник

Экспертами предметной области была разработана база тезауруса, состоящая из списка разделов и терминологического словника, который насчитывает 1545 уникальных терминов.

2.2. Источники терминологических данных

В качестве источников для извлечения семантических отношений и автоматического заполнения полей терминологических статей тезауруса оцифрованы и преобразованы в единую машиночитаемую форму представления следующие источники:

1. Краткая литературная энциклопедия: В 9 т. (КЛЭ) [11];
2. Литературная энциклопедия: В 11 т. (ЛЭ) [12];
3. Словарь литературных терминов: В 2-х т. (СЛТ) [13];
4. Квятковский А.П. Поэтический словарь. (ПСК) [14];
5. Большая советская энциклопедия: В 30 т. (БСЭ) [15].

3. Выбор инструментария для анализа данных

Известен целый ряд различных процедур анализа текста на естественном языке:

- графематический анализ с разделением входного текста на слова и разделители и с выделением предложений, абзацев, заголовков и примечаний;
- лексический анализ (токенизация) входного текста с выделением из него лексем (токенов);
- частеречная разметка, задачей которой является определение части речи и грамматических характеристик слов в тексте;
- морфологический анализ с распознаванием грамматической формы слов и словосочетаний и снятием омонимии;
- лемматизация слова (словоформы), т.е. приведение словоизменительной формы слова к лемме — к её нормальной (словарной) форме;
- стемминг — отбрасывание изменяемых частей слов, преимущественно окончаний;
- синтаксический анализ (парсинг) с распознаванием синтаксического строения предложения;
- семантический анализ, с установлением семантических связей (отношений) между элементами текста, которые могут включать более одного слова.

Автоматизация названных процедур для больших объемов данных реализуется множеством способов с помощью разнообразных инструментов и всевозможных анализаторов.

Инструмент автоматической обработки текстов под названием АОТ содержит графематический, морфологический, синтаксический и семантический анализаторы. Морфологический анализатор построен с использованием грамматического словаря А.А. Зализняка [16].

Программа от компании «Яндекс» Mystem [17] позволяет производить морфологический анализ текста на русском языке. Данный анализатор [18] разработан с использованием словаря в виде леса инвертированных префиксных деревьев суффиксов и инвертированного префиксного дерева основ. Mystem позволяет определять начальную форму слова, т.е. производить лемматизацию, определять грамматические характеристики. Дополнительным плюсом данного продукта является возможность построения гипотетических разборов для слов, не входящих в словарь. Для данного инструмента существует wrapper (обертка) PyMystem [19], используемая в качестве библиотеки для языка Python.

Другой анализатор, который возможно интегрировать в Python-проект в качестве библиотеки, называется Pymorphy2 [20], с помощью которого возможно выполнять лемматизацию и анализ слов, осуществлять склонение по заданным грамматическим характеристикам слов. В процессе работы с русским языком используется словарь OpenCorpora [21]. Диапазон скорости разбора более 100 тыс. слов/сек. Кроме того, у Pymorphy2, как и у PyMystem, предусмотрена возможность интеграции с фреймворком Django [22].

Ниже представлен разбор термина «Стопа» средствами Mystem и PyMorphy2.

Морфологический анализ термина: **Стопа**

['Стопа']

Mystem: [{'analysis': [{'lex': 'стопа', 'gr': 'S,жен,неод=им,ед'}], 'text': 'Стопа'}, {'text': '\n'}]

PyMorphy2: [Parse(word='стопа', tag=OpencorporaTag('NOUN,inan,femn sing,nomn'), normal_form='стопа', score=1.0, methods_stack=((<DictionaryAnalyzer>, 'стопа', 55, 0),))]

В качестве инструмента для проведения исследований достаточно использовать высокоуровневый язык программирования Python. Реализация каждого логического шага представлена отдельными скриптами. Полученные результаты в виде извлеченной информации достаточно представить в файле с форматом .csv для удобной загрузки в базу данных.

4. Подготовка источников

Для работы с большим объемом данных в качестве корпуса для извлечения семантических отношений необходимо провести анализ исследуемой литературы.

4.1. Задачи унификации данных источников

После оцифровки источников каждого документа необходима отдельная обработка для унификации данных:

1. Конвертация источника в формат с расширением .txt.
2. Объединение нескольких текстовых файлов источника в один файл.
3. Удаление пустых строк в файле.
4. Удаление диакритик, знаков и символов с неправильной кодировкой.
5. Удаление номеров страниц, слогов, показывающих начало терминов, опечаток и др.
6. Скрипты написаны на языке Python, который подходит для обработки русского языка за счет большого выбора библиотек. В данном случае, была использована библиотека «re (Regular expression operations)».

4.2. Соотношение терминов и определений в источниках

Источник КЛЭ+ЛЭ представляет собой файл с расширением .txt размером 85 МБ и количеством строк 295 928. После очистки было удалено 64 889 строк, которые содержали символы, обозначающие номера страниц, ссылки на иллюстрации и библиографию. Библиография была удалена из определений, поскольку данная информация не пригодится для извлечения отношений между терминами и может помешать процедуре автоматического заполнения полей.

После анализа СЛТ было установлено, что из 286 терминов, совпадающих со словником, 38 терминов СЛТ не имеют определения, а имеют отсылку на другой термин, например, «Ускорение — см. Пиррихий». Определения терминов в данном источнике записаны в виде развернутого ответа, средняя длина определения составляет 5 996 символов, что занимает порядка 70 строк.

В таблице 1 представлены характеристики исследуемых источников, а именно: название, год издания, количество уникальных терминов в источнике, количество полноценных определений без учета ссылок на другие и краткое описание самого источника.

5. Анализ терминов

В словнике, составленном экспертами предметной области, представлено 1 544 уникальных термина, составленных вручную.

В таблице 2 приведена статистика встречаемости терминов из словника в источниках, объявленных ранее. Таким образом, в каждом из источников встретилось 128 терминов, представленных в словнике, данные термины являются общеупотребительными. 587 терминов не встретилось ни в одном источнике. Из КЛЭ был извлечен 571 термин, что составляет порядка 3,7% определений, использованных из всего источника. 401 термин встретился в ЛЭ, что составляет порядка 8,3% от всего

Таблица 1. Описание и характеристики исследуемых источников

Источник	Год издания	Кол-во терминов	Кол-во определений	Краткое описание
КЛЭ	1962–1978	15 228	14 755	Наличие большого количества терминов в источнике подразумевает отдаленность от предметной области. В данной энциклопедии собран материал по персоналиям и более общим терминам, частично связанным с поэтологией
ЛЭ	1929–1939	4 782	4 276	Источник содержит полные определения, порядка 30–60 строк в среднем
СЛТ	1925	739	679	Словарь представляет общеобразовательное пособие из области теории литературы, лингвистической поэтики, эстетики и социологии художественного творчества
ПСК	1966	673	615	Поэтический словарь содержит максимально приближенные термины к предметной области. Определения характеризуются небольшим размером и очевидным сходством в построении определений
БСЭ	1969–1978	30 томов	–	Источник использовался в качестве вспомогательного

Таблица 2. Статистика частоты появления термина в словнике и в источниках

Название источника	Количество терминов из словника	Количество терминов в источнике	Соотношение
КЛЭ	571	15 228	3,7%
ЛЭ	401	4 782	8,3%
СЛТ	286	739	38%
ПСК	601	673	89%
Ни в одном источнике	587	–	–
Во всех источниках	128	–	–

Таблица 3. Статистика слов в терминах из словника

Количество слов в термине	Количество терминов	Соотношение со словником
1	996	64,5%
2	478	31%
3	60	3,9%
4	6	0,39%
5	3	0,2%
6	1	0,01%

объема ЛЭ. 286 терминов было извлечено из СЛТ, что составляет порядка 38% от всего объема СЛТ. Наибольшbv по объему извлеченных терминов и определений оказался источник ПСК – извлечен 601 термин из 673, что составляет 89%.

В таблице 3 исследованы термины из словника на количество слов в термине. Данная количественная мера важна для дальнейшего исследования, поскольку автоматические извлечение семантических отношений обычно проверяется на терминах длины не более чем 1. В данном случае, количество однословных терминов составляет 64,5% от всего количества терминов в словнике, второе место занимают двухсловные термины – 31%, третье место – 3,9% – трехсловные термины. Кроме того, в словнике присутствует шестисловный термин в единичном экземпляре и 3 пятисловных термина.

6. Методика поиска термина и определения

Следующим шагом служит выделение термина и его определения из источника. Задача является нетривиальной, поскольку компьютер различает поступающий текст только в виде строки, необходимо наложить ряд правил и условий для корректного извлечения терминов и определений.

Для этого необходимо разбить задачу нахождения термина и определения на подзадачи:

1. Поиск начала термина;

2. Поиск конца термина;

3. Поиск начала определения;

4. Поиск конца определения.

Разбиение задачи поиска термина на две подзадачи (поиск начала и поиск конца термина) связано с тем, длина термина не фиксирована: термин может быть однословным, но может состоять и из нескольких слов. Начало определения не всегда совпадает с концом термина или опознавательным знаком, используемых в словарях, а именно тире. Между термином и определением можно встретить конструкции в круглых и/или квадратных скобках, в которых добавлена дополнительная информация:

АББРЕВИАТУРЫ (итал. abbreviatura — сокращение, от лат. abbrevio — сокращаю) — слова, образованные из первых букв или сокращенных частей слов...

Определение может состоять из одной или нескольких строк. Точка или символ конца строки не являются показателями окончания определения. Маркером окончания определения может служить начало следующего термина. Важно понять, где начинается следующий термин, так как термин может встретиться и в самом определении в роли поясняющего понятия. В определении часто можно встретить в качестве примеров стихотворения, цитаты, отсылки на другие термины. Определения могут состоять как из одного предложения, так и из нескольких десятков строк. Ниже продемонстрировано два термина с определениями. Первый термин состоит из трех слов и в качестве определения имеет отсылку на другой термин. Второй термин состоит из двух слов и имеет в определении виды Алкеевой строфы и пример – отрывок из произведения «Подражание Горацию».

Прежде чем разработать правила для алгоритма автоматического извлечения терминов и определений, необходимо изучить исследуемые источники. Каждый словарь имеет свою уникальную специфику интерпретации термина. Важно разработать совокупность правил, которые позволили бы с большой точностью и полнотой определять начало и конец терминов и определений. Для данной задачи нет смысла подключать методы машинного обучения, так как при использовании шаблонов с регулярными выражениями и ряда правил получился качественный результат, который стал достижим за несколько итераций.

КАЗАНСКО-ТАТАРСКАЯ ЛИТЕРАТУРА — см. «Татарские лит-ры».

АЛКЕЕВА СТРОФА — античная четырехстишная строфа, изобретенная Алкеем; состоит из стихов трех видов:
1) девятисложный ямбический стих ||||^^
2) десятисложный стих |||
3) одиннадцатисложный стих |||.
В следующем отрывке В. Брюсов имитировал на русском языке ритм А. с., причем первые две строки соответствуют третьему виду А. с., третья строка — первому виду и четвертая строка — второму виду:
Не тем горжусь я, | Фебом отмеченный,
Что стих мой звонкий | римские юноши
На шумном пире повторяют,
Ритм выбивая узорной чашей.
(«Подражание Горацию»)

ААНРУД Ганс [Hans Aanrud, 1863–] — норвежский поэт и драматург, лит-ый руководитель национального театра в Христиании. Написал несколько «плутовских» и фривольных комедий и повестей о «детях и подростках».

ВЕЛЬХАВЕН (Welhaven), Юхан Себастьян (22.XII.1807, Берген, — 21.X.1873, Христиания) — норв.
894
поэт и критик.

Таким образом, можно выделить несколько правил, выведенных эмпирическим путем, которые помогут извлечь из источника термин и его определение:

1. Новая строка может начинаться с термина.

2. Термин может быть написан заглавными буквами.

3. Термин может быть разделен символами и знаками, не относящимися к самому термину.

4. Слова (или слово), входящее в термин, имеет начальную форму (стоит в им.п., ед.ч.)

5. Термин может быть отделен от прилагательного знаком тире, равно или пробелом.

6. Между термином и определением может стоять в круглых скобках перевод термина с другого языка.

7. Между термином и определением может стоять в квадратных скобках поясняющая информация.

8. Термин может обозначать личность, т.е. в качестве термина будет фамилия, отделенная запятой от имени или другой информации, которая может быть расположена в скобках.

Таблица 4. Статистика слов в терминах из словника

Количество определений	Количество терминов
1	446
2	249
3	134
4	118
5	1

9. В используемых источниках термин, смысл которого раскрывается, в определении встречается в виде сокращения первой буквы с точкой. Если термин состоит из двух или более слов, тогда сокращение имеет вид перечисления первых букв разделенных точкой и пробелом (например, Алкеева строфа = А. с.).

В таблице 4 приведены количественные характеристики после извлечения всех определений из исследуемых источников, таким образом, 446 терминов имеют только 1 определение, 249 терминов имеют по 2 определения, 134 термина имеют по 3 определения и так далее.

7. Распознавание семантических отношений

Люди имеют уникальную способность определять взаимосвязь между двумя лексическими единицами, однако оценить близость двух терминов с помощью программного комплекса возможно с помощью метрик семантической близости. Метрики семантической близости являются неким индикатором для определения взаимосвязи между терминами на основе задаваемого правила. В работе [19] рассмотрены метрики близости для установки отношений между словами из статей Википедии.

7.1. Метрика «Количество общих слов в определении»

Метрика (1) использует меру семантической близости на основе общих слов определений двух терминов.

$$similarity(t_i, t_j) = \frac{2|(d_i \cap d_j)/stopwords|}{|d_i| + |d_j|} \tag{1}$$

Числитель дроби равен количеству общих слов, приведенных в начальную форму, в выбранных определениях с учетом списка слов, определяемых как стоп-слова. Знаменатель дроби соответствует сумме всех слов в каждом из двух выбранных определений.

В качестве источника стоп-слов был выбран «Частотный словарь современного русского языка» [23]. Удаление стоп-слов помогает снижать уровень шума, другими словами, повышается качество выбранной метрики, поскольку, например, союз «и»

не несет никакой смысловой нагрузки в определении для нахождения взаимосвязи двух терминов, однако может встретиться в обоих определениях.

Однако данная метрика не учитывает длину определений. Это является главным недостатком метрики, поскольку длина некоторых определений может достигать до ста строк и иметь большое количество общеупотребительных терминов, которые не относятся к списку стоп-слов, например: система, совокупность, ряд, количество и т.д.

7.2. Метрика «Косинусная мера сходства», или «Косинус угла между векторами определений»

Для того чтобы компенсировать влияние длины определений на связность между терминами, применяется метрика под названием косинусная мера сходства. Определение представляется как N-мерный вектор, и далее происходит оценка терминов с использованием полученных векторов.

Размерность вектора определяется в зависимости от характера исследуемой задачи. Для данного случая размерность вектора определения совпадает с количеством слов в словнике, а именно 1554. Таким образом, вектор определения будет включать только термины из словника. Если термин из словника не встретился в определении, то обозначенный элемент становится равным нулю, в других случаях элемент становится равным количеству появлений термина в определении.

В формуле косинусной меры сходства (2) числитель представляет собой скалярное произведение векторов двух рассматриваемых определений, а знаменатель равен произведению евклидовых норм этих векторов. Знаменатель в формуле нормирует по длине векторы, таким образом, результат можно интерпретировать как скалярное произведение нормированных векторов, соответствующих двум определениям.

Для реализации метрики необходимо провести предварительные процедуры:

1. Нормализация определений — приведение каждого слова определения в начальную форму;
2. Нормализация словника — приведение каждого слова термина из словника в начальную форму;
3. Поиск терминов в определении (поиск производится с учетом всех слов в терминах из словника);
4. Формирование векторов определений размерности, равной количеству терминов в словнике.

Минимальная близость терминов соответствует количественному показателю 0.0, а максимальная – 0.99.

$$similarity(t_i, t_j) = \frac{f_i \cdot f_j}{||f_i|| \cdot ||f_j||} = \frac{\sum_{k=1,N} f_{ik} f_{jk}}{\sqrt{\sum_{k=1,N} f_{ik}^2} \sqrt{\sum_{k=1,N} f_{jk}^2}} \quad (2)$$

где f_{ik} — частота леммы c_k в определении d_i [8].

8. Описание алгоритма

Входными данными служит множество терминов T, между элементами которого необходимо установить семантические отношения по их определениям из множества D. Таким образом, задача сводится к задаче распознавания множества семантических отношений R из всех возможных пар терминов. Например:

$$T = \{\text{стопа,новелла,единоначатие,ямб,анафора}\}$$

$$R = \{<\text{стопа,ямб}>,<\text{единоначатие,анафора}>\}$$

9. Сравнение метрик

Для интерпретации полученных результатов, необходимо использовать ручную оценку. Были выбраны случайным образом 26 пар терминов. В качестве асессора выступает эксперт предметной области. Ниже приведены нормированные результаты метрики «количество общих слов в определении» — «Метрика 1» и метрики «косинусная мера угла» — «Метрика 2», диапазон результатов по двум метрикам и оценке эксперта составляет от 0 до 0.99. В таблице 5 представлено сравнение двух метрик близости терминов с ручной оценкой эксперта.

10. Заключение

В работе представлены методы по автоматическому распознаванию взаимосвязи терминов путем создания алгоритма на основе семантических метрик близости. Метрика «Количество общих слов в определении» является универсальным и наиболее очевидным способом установления взаимосвязи терминов по определениям, однако данный метод зависит от длины определений, что является важным фактором ввиду специфики предметной области. Напротив, метрика «Косинус угла между векторами определений» не учитывает длину определений, так как в основу заложен набор слов, задаваемых в качестве векторного представления. В данном случае длина вектора совпадала с количеством терминов словника, что позволило детектировать общие термины из словника в рассматриваемых определениях. Для формирования данных была проведена обработка словарей предметной области.

Таблица 5. Сравнение двух метрик близости терминов с ручной оценкой эксперта

№	Термины	Метрика 1	Метрика 2	Оценка эксперта
1	АНТИБАККХИЙ, ЖЕЛЬДИРМЕ 1. Античный размер. 2. Казахский размер.	0.00	0.00	0.00
2	ПЯТИДОЛЬНИК, ЭПИГРАФ 1. Пятисложник, силлаб.-тоническ. размер. 2. Жанр литературы.	0.01	0.00	0.00
3	ХОРЕЙ, КАНТИЛЕНА 1. Стихотворный размер. 2. Музыкально-поэтический жанр.	0.01	0.22	0.05
4	ДОХМИЙ, КРАТА 1. Античный размер. 2.Обобщение равносложн. стоп в силлабо-тонических размерах.	0.01	0.32	0.00
5	ГЕКЗАМЕТР, КОБЗАРЬ 1.Античный размер, в русской метрике имитируется 6-стопным дактилем. 2. Певец народных украинских песен.	0.02	0.00	0.00
6	ШЕСТИДОЛЬНИК, ТОНИЧЕСКОЕ СТИХОСЛОЖЕНИЕ 1. 6-сложная 2-акцентная стопа в силлабо-тонике. 2. Изначальное название силлаботоники.	0.03	0.22	0.20
7	ДОЛГИЙ СЛОГ, АРУЗ 1. В античной и некоторых современных метриках. 2. Метрика на основе чередования долгих и кратких слогов.	0.04	0.72	0.50
8	БЫЛИНЫ, МАКАРОНИЧЕСКИЕ СТИХИ 1. Жанр русского народного эпоса. 2. Жанр сатирических стихов.	0.05	0.81	0.10
9	ЧАСТУШКА, ПОСЛОВИЦА 1. Жанр лирической поэтической формы фольклора. 2. Жанр краткой поэтической формы фольклора.	0.06	0.60	0.50
10	ПАУЗНИК, ЭВРИТМИЯ 1. Стихи с неполносложной 3-сложной стопой. 2. Благозвучие стиха.	0.07	0.77	0.05

№	Термины	Метрика 1	Метрика 2	Оценка эксперта
11	АНАФОРА, ЦЕНТОН 1. Звуковой, лексический, синтаксический повтор в начале смежных стихов или строф. 2. Составление стихотворений из различных известных стихов.	0.08	0.99	0.00
12	СТОПА, ПЭОН 1. В силлабо-тонике повторяемая группа слогов с акцентом на заданном месте (род). 2. В силлабо-тонике 4-сложная стопа с акцентом на заданном месте (вид).	0.13	0.76	0.95
14	ПИРРИХИЙ, МОЛОСС 1. Пропуск ударения в 2-сложной стопе, нарушающий правильность ритмической схемы. 2. 3 ударения (2 сверхсхемных) в 3-сложной стопе (тримакр).	0.15	0.36	0.10
15	ИПОСТАСА, ДИАСТОЛА 1. Ритмическая модификация метрической стопы. 2. В античной метрике замена долгого слога кратким.	0.29	0.62	0.00
16	ТРОХЕЙ, ДОХМИЙ 1. Античная стопа. 2. Античный размер.	0.33	0.49	0.10
17	ЭПИТРИТ, ИОНИК 1. Античная 4-сложная стопа (3 долгих и 1 краткий). 2. Античная 4-сложная стопа (2 долгих и 2 кратких).	0.42	0.82	0.50
18	ДИХОРЕЙ, ТРИМЕТР 1. Диподия, четырехсложная стопа с ударением на 3-м или 4-м слоге (пэон 3-й или 4-й) (часть). 2.В античной метрике размер из 3 диподий (целое).	0.46	0.90	0.50
19	АНТИБАККХИЙ, БРАХИХОРЕЙ 1. Античная 3-сложная стопа (2 долгих и 1 краткий). 2. 3-сложная стопа (краткий, долгий и краткий), амфибрахий.	0.53	0.18	0.50

№	Термины	Метрика 1	Метрика 2	Оценка эксперта
20	ПЕНТАМЕТР, ДИСТИХ 1.В русской метрике античный 5-стопный дактилический стих имитируется 6-стопным с мужской цезурой (часть). 2. Элегический дистих, двустишие из гекзаметра и пентаметра (целое).	0.68	0.96	0.80
21	ДИЯМБ, ДИХОРЕЙ 1. Двойная ямбическая стопа (диподия). 2. Двойная хореическая стопа (диподия).	0.99	0.86	0.70

Список литературы / References

[1] Лингвистическая онтология «Тезаурус Рутез», http://www.labinform.ru/pub/ruthes/.

[2] Тезаурус WordNet, http://wordnet.princeton.edu/.

[3] Бойков В. Н. и др., “Тезаурус как инструмент поэтологии”, *Моделирование и анализ информационных систем*, **17**:1 (2010), 5–24; [Boikov V. N. et al., “Thesaurus as a poetological tool”, *Modeling and Analysis of Information Systems*, **17**:1 (2010), 5–24, (in Russian).]

[4] Бойков В. Н., “Семантическая модель «Тезауруса по поэтологии» в составе информационно-аналитической системы”, *Материалы научной конференции «Интернет и современное общество»*, 2013, 273–279; [Boikov V. N., “Semanticheskaya model «Tezaurusa po poehtologii» v sostave informacionno-analiticheskoj sistemy”, *Materialy nauchnoj konferencii «Internet i sovremennoe obshchestvo»*, 2013, 273–279, (in Russian).]

[5] Бойков В. Н., “Предметно-ориентированный тезаурус в открытой информационно-аналитической системе”, *RCDL’2013 «Электронные библиотеки: перспективы, методы и технологии, электронные коллекции»*, 2013, 70–76; [Boikov V. N., “Predmetno-orientirovannyj tezaurus v otkrytoj informacionno-analiticheskoj sisteme”, *RCDL’2013 «Ehlektronnye biblioteki: perspektivy, metody i tekhnologii, ehlektronnye kollekcii»*, 2013, 70–76, (in Russian).]

[6] Тезаурус по поэтологии, http://wikipoetics.ru/.

[7] Hearst M. A., “Automated discovery of WordNet relations”, *WordNet: an electronic lexical database*, 1998, 131–153.

[8] Панченко А. И., “Извлечение семантических отношений из статей Википедии с помощью алгоритмов ближайших соседей”, *Открытые системы*, **16** (2012), 18–27; [Panchenko A. I., “Izvlechenie semanticheskih otnoshenij iz statej Vikipedii s pomoshchyu algoritmov blizhajshih sosedej”, *Otkrytye sistemy*, **16** (2012), 18–27, (in Russian).]

[9] Serelex: Поиск семантически связных слов, http://serelex.org/ru.

[10] Киселев Ю. А., “Метод извлечения родовидовых отношений между существительными из определений толковых словарей”, *Программная инженерия*, **10** (2015), 38–48; [Kiselev YU. A., “Metod izvlecheniya rodovidovyh otnoshenij mezhdu sushchestvitelnymi iz opredelenij tolkovyh slovarej”, *Programmnaya inzheneriya*, **10** (2015), 38–48, (in Russian).]

[11] *Краткая литературная энциклопедия*, Сов. Энцикл., 1962–1978; [*Kratkaya literaturnaya ehnciklopediya*, Sov. Ehncikl., 1962–1978, (in Russian).]

[12] *Литературная энциклопедия*, Ком. акад., 1929–1939; [*Literaturnaya ehnciklopediya*, Kom. akad., 1929–1939, (in Russian).]

[13] *Словарь литературных терминов*, Изд-во Л. Д. Френкель, 1925; [*Slovar literaturnyh terminov*, Izd-vo L. D. Frenkel, 1925, (in Russian).]

[14] Квятковский А. П., *Поэтический словарь*, Сов. Энцикл, 1966; [Kvyatkovskij A. P., *Poehticheskij slovar*, Sov. Ehncikl, 1966, (in Russian).]

[15] *Большая советская энциклопедия*, Сов. энцикл., 1969–1978; [*Bolshaya sovetskaya ehnciklopediya*, Sov. ehncikl, 1969–1978, (in Russian).]

[16] Зализняк А. А., *Грамматический словарь русского языка*, Словоизменение, 1980; [Zaliznyak A. A., *Grammaticheskij slovar russkogo yazyka*, Slovoizmenenie, 1980, (in Russian).]

[17] Mystem, https://tech.yandex.ru/mystem/.

[18] Segalovich I., "A Fast Morphological Algorithm with Unknown Word Guessing Induced by a Dictionary for a Web Search Engine", *MLMTA*, 2003, 273–280.

[19] PyMystem, https://pypi.python.org/pypi/pymystem3/0.1.1.

[20] PyMorphy2, https://pymorphy2.readthedocs.org/en/latest/.

[21] OpenCorpora, http://opencorpora.org/.

[22] FrameWork Django, https://www.djangoproject.com/.

[23] Ляшевская О. Н., *Частотный словарь современного русского языка (на материалах Национального корпуса русского языка)*, Азбуковник, 2009; [Lyashevskaya O. N., *Chastotnyj slovar sovremennogo russkogo yazyka (na materialah Nacionalnogo korpusa russkogo yazyka)*, Azbukovnik, 2009, (in Russian).]

DOI: 10.18255/1818-1015-2015-6-834-851

Linguistic and Statistical Analysis of the Terminology for Constructing the Thesaurus of a Specified Field

Karyaeva M. S.[1]

Received May 15, 2015

The paper is devoted to the analysis of the body of terms and terminological sources for further automation of constructing the thesaurus of a subject area, which is regarded as poetics in our work. Preliminary systematization of terminology with a linguistic and statistical approach forms the body of semantically related concepts to automate extraction of semantic relationships between terms that define the structure of the thesaurus of the specified field.

Keywords: thesaurus, semantic similarity metrics, data mining, computer linguistics

For citation: Karyaeva M. S, "Linguistic and Statistical Analysis of the Terminology for Constructing the Thesaurus of a Specified Field", *Modeling and Analysis of Information Systems*, **22**:6 (2015), 834–851.

On the authors:
Karyaeva Maria Sergeevna, orcid.org/0000-0003-4466-1735, graduate student,
P.G. Demidov Yaroslavl State University,
Sovetskaya str., 14, Yaroslavl, 150000, Russia, e-mail: mari.karyaeva@gmail.com

Acknowledgments:

[1]This work was supported by Russian Foundation for Basic Research RFBR Project 16-07-01180