# The Distributive and Statistical Analysis as a Tool to Automate the Formation of Semantic Fields (on the Example of the Linguocultural Concept of "Empire")

Victor Zakharov

Saint-Petersburg State University, Universitetskaya emb., 7-9-11, 199034 Saint-Petersburg, Russia[2]
v.zakharov@spbu.ru

**Abstract.** The paper presents ongoing results of automatic creation of a semantic field of «empire» in Russian based on distribution and statistical method using corpus data.

A semantic field is a collection of content units covering a certain area of human experience and forming relatively an autonomous microsystem with one or a few centers. The nature of relations within it is mostly named as an association. The idea is to extract from data on syntagmatic collocability a set of lexical units connected by semantic paradigmatic relations of various strength using distributional analyses techniques. Nowadays the presence of big corpora and sophisticated algorithms give the possibility and hope to reach a reasonable results.

The first goal of the study is to develop tools and methodology to fill semantic fields by lexical units on the basis of morphologically tagged corpora and special sketch grammar and then to measure the strength of relations between units and to evaluate the method. We were using a corpus system the Sketch Engine that implements the method of distributional statistical analysis. Text material was represented by own topical Russian corpora created from Russian texts of XVIII –XX centuries. In the course of work and to achieve the goal we have solved a number of tasks have received lists of items filling the semantic space around a concept of "empire" and we are evaluating the method as successive and promising one. At conclusion further steps were identified to clarify the perspective areas of work and to improve the results obtained.

**Keywords:** Distributive and statistical analysis, Semantic field, Concept of Empire in Russian.

# Дистрибутивно-статистический анализ как инструмент автоматизации формирования семантических полей (на примере поля «империя»)

Виктор П. Захаров

[1] Санкт-Петербургский государственный университет, Университетская наб., 7-9-11, 199034 Санкт-Петербург
v.zakharov@spbu.ru

## 1 Введение

Понятие «семантическое поле» применяется в лингвистике для обозначения совокупности языковых единиц, объединенных каким-то общим семантическим признаком; имеющих некоторый общий компонент значения. «Поле — совокупность содержательных единиц, покрывающая определенную область человеческого опыта и образующая более или менее автономную микросистему» [5]. В роли таких лексических единиц выступают слова и словосочетания, как нарицательные, так и имена собственные. Сам термин «семантическое поле» имеет различные модификации или синонимы, как-то: лексическое поле, лексико-семантическое поле, функционально-семантическое поле, кластер, тезаурус, онтология и т. п. Каждый из этих терминов по-своему задает тип языковых единиц, входящих в поле и/или тип связи между ними. В основе теории семантических полей лежит представление о существовании в языке некоторых семантических групп, словарный состав которых объединен различными отношениями, как лингвистическими, так и экстралингвистическими, которые представляют собой сложную систему оппозиций.

Семантический признак, лежащий в основе семантического поля, может рассматриваться как некоторая понятийная категория [6, 12]. В трактовке В.Г. Адмони поле характеризуется наличием инвентаря элементов, связанных системными отношениями [1]. В.Г. Адмони усматривает в поле центральную часть — ядро, элементы которого обладают полным набором признаков, определяющих данную группировку, и периферию, элементы которой обладают не всеми, характерными для поля признаками, но могут иметь и признаки, присущие соседним полям. Поле предполагает непрерывность связей объектов множества, причем на некоторых участках поля создаются области, в которых связи особенно интенсивны, а признаки особенно сильно выражены. Поле предполагает непрерывность связей объектов множества, причем на некоторых участках поля создаются области, в которых связи особенно интенсивны, а признаки особенно сильно выражены. Тогда говорят о лексико-семантических группах – элементарных микрополях, объединяющих слова, обычно

относящиеся к одной части речи и наиболее сильно связанные отношением семантической близости. В общем же случае для поля характерна нечеткость границ между частями речи. Теории семантического поля в лингвистике посвящено большое число работ ([2, 4, 8, 9, 20] и др.).

Для полей характерна возможность количественного выражения силы связи между элементами внутри поля, и поэтому эта задача давно является предметом компьютерной лингвистики. Причем в компьютерной лингвистике «семантическое поле» обычно заменяется понятиями «тезаурус» и «онтология».

Задачу моделирования понятийной или терминологической системы можно разбить на две части: выявление системы понятий (лексических идентификаторов понятий) и выявление отношений между ними. В данной работе нас интересует первая задача, а именно, автоматизированное наполнение лексико-семантических полей. Она может решаться «вручную» путем экспликации и формализации профессионального знания, накопленного в системе человеческой деятельности, на основе знаний специалистов и с использованием имеющихся словарей, учебников и других пособий. Этот путь долгий и трудоемкий. Однако поскольку наши знания о мире так или иначе находят отражение в текстах, то можно поставить задачу извлечения системы понятий из текстов. Минимальный набор требований при этом следующий: множество этих автоматически извлеченных понятий должно быть достаточно полным и сами понятия должны быть связаны между собой. Характер связей на этом первом этапе автоматически не устанавливается. В нашем случае можно говорить о принципе когнитивной однородности [14], когда на каждом этапе решается одна задача, в данной работе это выявление множества основных взаимосвязанных понятий вокруг выбранного ядерного элемента (ключевого слова).

В данной работе мы будем говорить о семантической поле “империя”, понимая под ним совокупность пересекающих лексико-семантических групп (слов или словосочетаний), непосредственно или опосредованно связанных по смыслу с концентом “империя”. Выбор данного концепта обусловлен, с одной стороны, его богатым содержанием, с другой стороны, это содержание по-разному наполняется в разных языках. Эта работа является частью более широкого исследования, посвященного сравнительному анализу наполнения этого поля в русском, английском, чешском и немецком языках.

## 2 Дистрибутивно-статистический анализ как основа выявления парадигматических отношений

Основная цель данного исследования – выработка методов и адаптация механизмов автоматического выявления набора базовых понятий, относящихся к заданной теме в корпусах русских текстов на основе дистрибутивно-статистических методов. Следующий шаг на этом пути - представление лексических элементов семантического поля “империя” в виде компьютерного

тезауруса с количественными характеристиками силы связи между элементами и примерами из корпусов.

Одним из старых и известных методов лингвистического исследования является дистрибутивно-статистический анализ, при котором используется информация о дистрибуции элементов текста и их числовых параметрах. Уже на заре компьютерной лингвистики предпринимались попытки на основе частотной информации о встречаемости лексических единиц в контекстах определенной величины получать по некоторой заданной формуле количественную характеристику их связанности, что впоследствии нашло выражение в методах выявления коллокаций и многословных единиц на основе мер ассоциации.

Одновременно выдвигались идеи распространения этого метода и на парадигматический аспект языка – идеи о том, что парадигматические связи могут выводиться из связей синтагматических [21, 3, 13, 11, 10]. Принцип перехода от изучения текстуальных связей (синтагматических) к системным (парадигматическим) лежит в основе различных дистрибутивно-статистических методик [19, 20. 32]. Считается, что два элемента связаны парадигматически, если оба они текстуально систематически связаны с какими-то третьими элементами. Соответственно, сила парадигматической связи должна возрастать с увеличением *числа* и силы *общих* синтагматических связей [20: 370].

Однако возможности вычислительной техники того времени не позволяли реализовать эти идеи в виде практически работающих алгоритмов и программ. Далее, чтобы можно было говорить о закономерностях любых статистических распределений, нужны очень большие массивы данных [35]. Таковые появились только с созданием больших корпусов текстов. Одновременно стали появляться и соответствующие программные средства [23, 26, 27, 34, 35].

## 3 Механизм формирования лексико-семантических групп и полей

Как уже было сказано, парадигматические связи можно вывести из синтагматических. Эта идея была высказана А.Я. Шайкевичем [21] и К.С. Джоунс (K.S. Jones) (PhD thesis) еще в 1960-х гг., но была реализована только сейчас в корпусной лингвистике, где на базе корпусов текстов появилась возможность создать большую базу сочетаемости лексических единиц и на ее основе "вычислять" множество «ближайших соседей» для каждого слова. Математический аппарат для вычисления такого сходства был разработан Д. Лином (D. Lin) [30].

Однако при «переводе» синтагматики в парадигматику также важно также учитывать наличие синтаксической связи между контекстно близкими элементами текста [24]; [31]. Наш подход предполагает описание сочетаемости с помощью лексико-синтаксических шаблонов (иногда их называют лексико-грамматическими или морфологическими шаблонами). В нашем понимании лексико-синтаксический шаблон — это модель языковой конструкции, в которой

указываются существенные грамматические характеристики множества лексем, которые входят в языковые выражения, принадлежащие данному классу, и синтаксические условия построения языкового выражения в соответствии с заданным шаблоном (например, учет морфологических признаков лексических единиц в зависимости от контекстных условий).

В системе Sketch Engine [27], которая использовалась нами для формирования корпусов и выявления синтагматических и парадигматических связей, идея лексико-синтаксических шаблонов реализована в форме так называемых «эскизов слов» (word sketch). По определению «эскиз слова» - это одностраничная, автоматически генерируемая на базе корпуса сводка лексическо-грамматической сочетаемости слова, по-другому, сочетаемости в пределах заданных синтаксических формул. Эти «портреты» слов базируются на наборах правил, описывающих грамматические отношения между словами в тексте, которые получили название Word Sketch grammar, или Грамматика шаблонов.

При создании корпуса на основе указанной грамматики и данных морфологической разметки корпуса формируется специальная база данных, представляющая собой триплеты лексико-грамматических отношений. Статистическая обработка этой базы и вычисляет данные для построения дистрибутивного тезауруса (thesaurus), который для нас является аналогом лексико-семантической группы для заданного термина. Алгоритм вычисления семантических расстояний между элементами группы (кандидатами в группу) и их внутренней кластеризацией описан в [36: sect. 3, 4].

Формализм для грамматики лексико-синтаксических шаблонов использует регулярные выражения над морфологическими тегами. Соответственно, в принципе любой пользователь-лингвист с некоторым опытом и знакомством с вычислительными формализмами может задать свой набор грамматических отношений. Очевидно, что он должен быть при этом знаком с набором тегов и грамматикой языка. Далее эта грамматике лексико-синтаксических шаблонов при создании корпуса подключается к нему, используя стандартный механизм, и тогда функциональные инструменты системы будут формировать результаты, исходя уже именно из этой пользовательской грамматики.

Формализм для грамматики лексико-синтаксических шаблонов основывается прежде всего на линейной последовательности единиц текста и, следовательно, более явно подходит для языков с жестким порядком слов, таких как английский, и менее - для языков со свободным порядком слов, например, для русского, для последнего требуется гораздо более гибкий подход для написания такой грамматики. Дистрибутивно-статистический анализ в нашем исследовании базируется на грамматике лексико-синтаксических шаблонов для русского языка, разработанной М.В. Хохловой [18]. Схожесть дистрибуции слов высчитывается статистически на основе меры ассоциации logDice [33] и с учетом грамматики лексико-синтаксических шаблонов [18, 26, 33, 36].

## 4 Материал и инструменты исследования

Основной материал исследования – это специально созданный нами совместно с М.В. Хохловой корпус по теме “империя” на основе текстов об империи в русской литературе и культуре конца XVIII – начала XX вв. (105 текстов, 9 млн. токенов). Таким образом подчеркнем, что мы выделяем концепты, существующие в русском языке на протяжении длительного времени и являющиеся отражением русской культуры.

Корпус делится на 4 подкорпуса по хронологическому принципу: 18-ый век (идентификатор подкорпуса XVIII), 1-ая половина 19-го века (XIX-1), 2-ая половина 19-го века (XIX-2) и 20-ый век (XX). Граничные даты подкорпусов выбраны как своего рода «вехи» в осознании понятия империи в развитии русской общественной мысли. Жанрово-тематическое наполнение – история, литература, публицистика, философия.

Для нашего исследования, как уже говорилось, мы использовали систему Sketch Engine. Главная ее особенность – это наличие специальных средств, реализующих методику дистрибутивно-статистического анализа — «Тезаурус» (построение тезауруса для заданного термина, другими словами, лексико-семантической группы) (см. Рис. 1), «Кластеризация» (группировка единиц тезауруса в кластеры) и «Дифференциация» (выявление сходства и разницы в сочетаемости для пар слов).

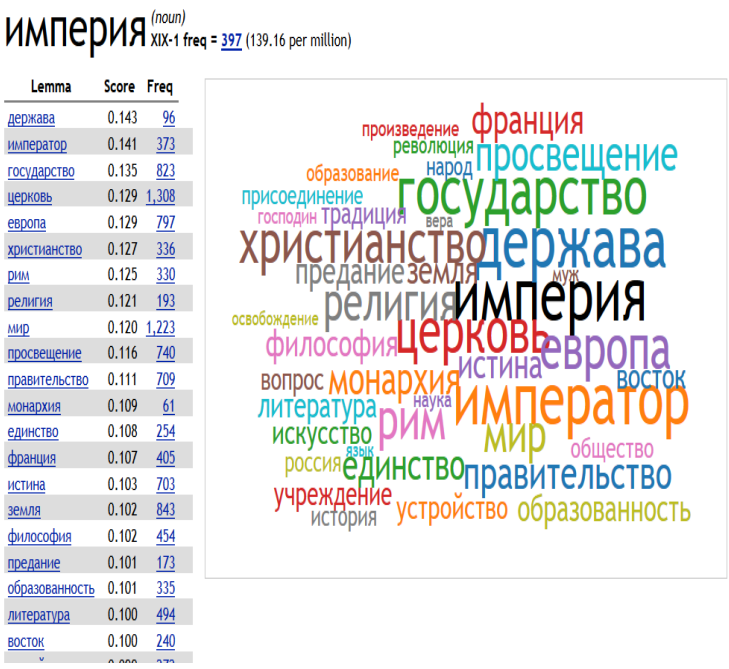



**Рис. 1.** Фрагмент дистрибутивного тезауруса для слова «империя» по подкорпусу 1-ой половины XIX века.

В состав микрополя (лексико-семантической группы) для термина *«империя»* вошли существительные, имеющие с данной лексемой похожую дистрибуцию (входят в одинаковые синтаксические отношения и часто встречаются в одинаковых контекстах): *«держава», «император», «государство», «церковь», «Европа» «христианство», «Рим»,* и др.

Тезаурус в системе Sketch Engine (или, как его называют, дистрибутивный тезаурус) показывает, какие слова имеют схожую дистрибуцию с заданным словом. В этом случае мы говорим о семантической близости или парадигматическом подобии слов. Единицы семантического поля обладают общими синтагматическими и парадигматическими свойствами, что отражает их семантическую близость.

В каждой предметной области значительная часть терминов, как правило, представлена словосочетаниями. Корпусные инструменты предоставляют нам возможность автоматического выявления коллокаций. Другой инструмент системы, выявляющий синтагматические связи между лексическими единицами – это «Коллокации», вычисляющий силу связанности единиц в линейной последовательности на основе 7 мер ассоциации. Но следует добавить, что этот инструмент выявляет не только синтагматические связи, но и парадигматические, выделяющий при достаточно большом «окне» анализа слова одного семантического поля с заданным.

Имеется также инструмент «Лексические портреты», выявляющий коллигации – коллокации в рамках заданных синтаксических моделей (лексико-синтаксических шаблонов). Если инструмент «Коллокации» вычисляет силу связи между словами по всему корпусу, то второй инструмент – в пределах заданной синтаксической формулы (шаблона). В рамках данного исследования грамматика лексико-синтаксических шаблонов использовалась в составе «Тезауруса».

И, наконец, Sketch Engine позволяет выдавать частотные списки лексических единиц, входящих в корпус, которые используются не сами по себе, а как входной материал для контрастивного анализа, когда данные нашего корпуса сравниваются с нейтральным фоновым. Т.е. лексические единицы, относительная частота которых в текстах исследуемого корпуса существенно превосходит частоту этих слов в фоновом неспециализированном корпусе, считаются ключевыми и включаются (могут быть включены) в формируемое семантическое поле.

# 5 Эксперименты и результаты

## 5.1 Методика исследования

Была проведена работа в соответствии со следующей методикой.

1) Получение ранжированного списка семантически связанных терминов (минитезаурус) для слова «империя» по каждому из подкорпусов с помощью инструмента «Тезаурус». Максимальное число единиц в гнезде тезауруса

задается равным 40. Каждому термину в каждом минитезаурусе присваивается ранг.

2) Объединение полученных минитезаурусов в один список, представляющий собой лексико-семантическое поле концепта «империя».

3) Выявление пересечения минитезаурусов, в результате чего каждому термину в объединенном списке присваивается «коэффициент стабильности» (k=1, 2, 3, 4, в зависимости от того, в скольких минитезаурусах тот или другой термин встретился). Термины с коэффициентом больше единицы образуют ядро семантического поля. Для этих терминов вычисляются средний и нормированный ранги силы семантической связи с заглавным словом «империя». Нормированный ранг получается умножением среднего ранга на «ранговый коэффициент нормализации»: 1 - для терминов, представленных во всех четырех минитезаурусах, 2 - для терминов из трех минитезаурусов и 3 - для терминов из двух минитезаурусов (Табл. 1). Таким образом, эти коэффициенты понижают ранги терминов, связанных со словом «империя» в большем числе подкорпусов (т. е. в большем числе временных периодов).

4) Ранжирование лексических единиц ядра семантического поля понятия «империя» по нормированному рангу.

5) Ранжирование лексических единиц поля по коэффициенту семантической близости (score).

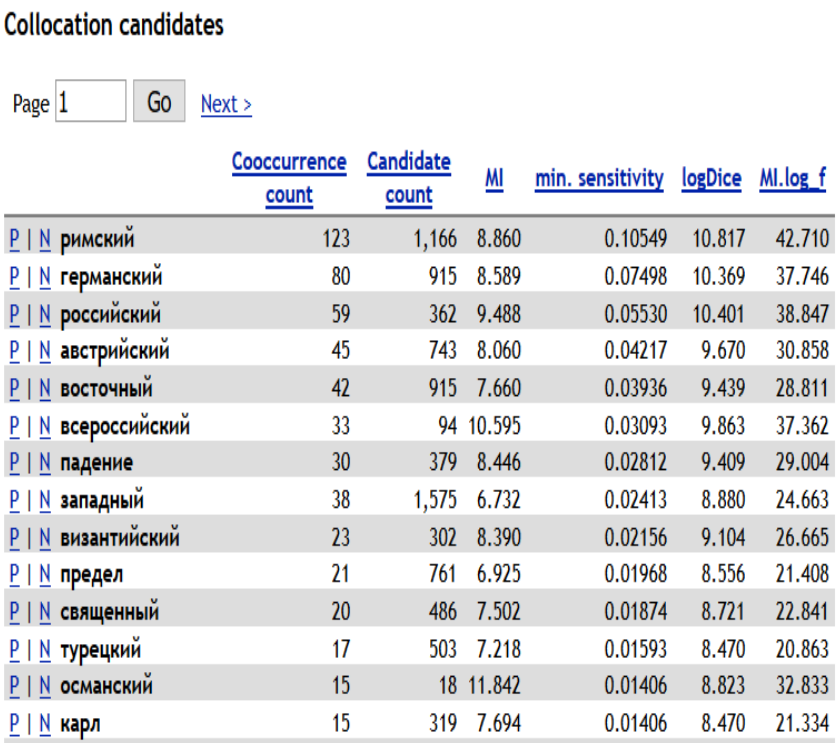
6) Подсчет относительной частоты (ipm) лексических единиц поля и ранжирование лексических единиц объединенного списка (поля) по относительной частоте.

**Table 1.** Сводный дистрибутивный тезаурус для слова «империя» (фрагмент).

| *Под-корпус* | *Ранг* | *Lemma* | *Score* | *Freq* | *Коэф-т стабильн.* | *Средн. ранг* | *Норм. ранг* |
|---|---|---|---|---|---|---|---|
| XIX-2 | 1. | австрия | 0,216 | 1014 | 1 | | |
| **XIX-2** | **36.** | **англия** | **0,131** | **1055** | **2** | **29** | **87** |
| XVIII | 22. | англия | 0,095 | 148 | 2 | | |
| XIX-2 | 19. | армия | 0,149 | 478 | 1 | | |
| …….. | …….. | | ……,, | …… | …….. | …….. | …….. |
| XIX-1 | 37. | господин | 0,085 | 363 | 1 | | |
| **XIX-2** | **24.** | **государственность** | **0,143** | **201** | **2** | **19** | **57** |
| XX | 14. | государственность | 0,141 | 143 | 2 | | |
| **XX** | **1.** | **государство** | **0,245** | **1016** | **4** | **2,25** | **2,25** |
| XIX-2 | 2. | государство | 0,200 | 4240 | 4 | | |
| XVIII | 3. | государство | 0,184 | 766 | 4 | | |
| XIX-1 | 3. | государство | 0,135 | 823 | 4 | | |
| XVIII | 19. | греция | 0,096 | 135 | 1 | | |
| XX | 2. | гуманизм | 0,188 | 195 | 1 | | |
| **XVIII** | **2.** | **держава** | **0,189** | **424** | **3** | **4,3** | **8,6** |

| | | | | | | |
|---|---|---|---|---|---|---|
| XIX-2 | 10.держава | 0,165 | 606 | 3 | | |
| XIX-1 | 1.держава | 0,143 | 96 | 3 | | |
| …….. | …….. | ……,, | …… | …….. | …….. | …….. |
| | ….. | | | | | |
| XIX-1 | 13.единство | 0,108 | 254 | 1 | | |
| **XIX-2** | **5.император** | **0,184** | **1381** | **3** | **4** | **8** |
| XX | 5.император | 0,177 | 295 | 3 | | |
| XIX-1 | 2.император | 0,141 | 373 | 3 | | |
| XX | 8.империализм | 0,166 | 297 | 1 | | |
| XX | 7.интеллигенция | 0,173 | 608 | 1 | | |
| …….. | …….. | …….. | …….. | …….. | …….. | …….. |

7) Формирование ранжированного списка коллокатов для слова «империя» для каждого из подкорпусов с помощью инструмента «Коллокации» (Рис. 2).



Collocation candidates

Page 1 Go Next >

| | Cooccurrence count | Candidate count | MI | min. sensitivity | logDice | MI.log_f |
|---|---|---|---|---|---|---|
| P \| N римский | 123 | 1,166 | 8.860 | 0.10549 | 10.817 | 42.710 |
| P \| N германский | 80 | 915 | 8.589 | 0.07498 | 10.369 | 37.746 |
| P \| N российский | 59 | 362 | 9.488 | 0.05530 | 10.401 | 38.847 |
| P \| N австрийский | 45 | 743 | 8.060 | 0.04217 | 9.670 | 30.858 |
| P \| N восточный | 42 | 915 | 7.660 | 0.03936 | 9.439 | 28.811 |
| P \| N всероссийский | 33 | 94 | 10.595 | 0.03093 | 9.863 | 37.362 |
| P \| N падение | 30 | 379 | 8.446 | 0.02812 | 9.409 | 29.004 |
| P \| N западный | 38 | 1,575 | 6.732 | 0.02413 | 8.880 | 24.663 |
| P \| N византийский | 23 | 302 | 8.390 | 0.02156 | 9.104 | 26.665 |
| P \| N предел | 21 | 761 | 6.925 | 0.01968 | 8.556 | 21.408 |
| P \| N священный | 20 | 486 | 7.502 | 0.01874 | 8.721 | 22.841 |
| P \| N турецкий | 17 | 503 | 7.218 | 0.01593 | 8.470 | 20.863 |
| P \| N османский | 15 | 18 | 11.842 | 0.01406 | 8.823 | 32.833 |
| P \| N карл | 15 | 319 | 7.694 | 0.01406 | 8.470 | 21.334 |

**Рис. 2.** Фрагмент списка коллокатов для ключевого слова «империя» (фрагмент).

Максимальное число коллокатов задается равным 50 (выбирается верхняя часть ранжированного списка). Для формирования списка коллокатов используется 4 наиболее эффективных меры: MI.l-og_f, logDice, min. sensitivity и MI, как это

было установлено нами в [37]. «Окно» для вычисления коллокаций задается равным от -3 до +3 (три слова влево и три слова вправо от заглавного).

8) Объединение полученных списков коллокатов в один.

9) Выявление пересечения в объединенном списке отдельных списков коллокатов (по подкорпусам) для каждого термина и приписывание им «коэффициент стабильности» (k=1, 2, 3, 4, в зависимости от того, в скольких списках тот или другой термин встретился). Коллокаты (коллокации) с коэффициентом больше единицы добавляются в ядро семантического поля (составные термины-батиграммы). Для этих терминов вычисляются средний и нормированный ранги силы синтагматической связи с заглавным словом. Нормированный ранг получается умножением среднего ранга на «коэффициент нормализации»: 1 – для коллокатов, представленных во всех четырех списках коллокатов, 2 - для коллокатов из трех списков и 3 - для коллокатов из двух списков (Таблица 2).

10) Ранжирование терминов-биграмм ядра семантического поля понятия «империя» по нормированному рангу.

**Table 2.** Сводный дистрибутивный тезаурус для слова «империя» (фрагмент).

| *Подкорпус* | *Ранг* | *Лемма* | *Коэф-т стабильн.* | *Средн.ранг* | *Норм. ранг* |
|---|---|---|---|---|---|
| **XIX-1** | **8.** | **австрийский** | **2** | **6,5** | **19,5** |
| **XIX-2** | 5. | австрийский | 2 | | |
| **XX** | 21. | англия | 1 | | |
| …… | …… | …… | …… | …… | …… |
| **XIX-2** | **8.** | **византийский** | **4** | **9,25** | **9,25** |
| **XVIII** | 6. | византийский | 4 | | |
| **XX** | 12. | византийский | 4 | | |
| **XIX-1** | 11. | византийский | 4 | | |
| …… | …… | …… | …… | …… | …… |
| **XIX-1** | 34. | Габсбурги | 1 | | |
| **XIX-1** | **5.** | **германский** | **3** | **9,3** | **18,6** |
| **XIX-2** | 3. | германский | 3 | | |
| **XX** | 10. | германский | 3 | | |
| **XVIII** | 33. | город | 1 | | |
| …… | …… | …… | …… | …… | …… |
| **XVIII** | **8.** | **могущество** | **2** | **13** | **39** |
| **XIX-2** | 18. | могущество | 2 | | |
| **XIX-2** | **37.** | **наполеон** | **2** | **30** | **90** |
| **XX** | 23. | наполеон | 2 | | |
| …… | …… | …… | …… | …… | …… |
| **XIX-2** | **27.** | **оттоманский** | **2** | **14** | **42** |
| **XVIII** | 1. | оттоманский | 2 | | |
| **XIX-2** | **7.** | **падение** | **3** | **7** | **14** |
| **XVIII** | 4. | падение | 3 | | |

| | | | | | |
|---|---|---|---|---|---|
| **XIX-1** | 10. | падение | 3 | | |
| …… | …… | …… | …… | …… | …… |

1) Формирование списка ключевых слов для каждого из подкорпусов с помощью инструмента «Word list (Output type: Keywords)». Сопоставимый корпус для этого – ruSkell 1.4 (см. https://www.sketchengine.eu/russian-skell-corpus/).

12) Объединение списков ключевых слов в один и сортировка объединенного списка по «коэффициенту уникальности» (score).

## 5.2 Результаты исследования

В результате выполнения пп. 1-2 (раздел 5.1) был получен список терминов, представляющий собой наполнение семантического поля «империя» по данным 4 подкорпусов. Этот список включает 112 разных слов (по алфавиту):

*Австрия, Англия, армия, варвар, Венгрия, вера, ветер, воинство, война, вопрос, восток, враг, Германия, герой, господин, государственность, государство, Греция, гуманизм, держава, Европа, единство, жар, земля, зло, злодей, император, империализм, интеллигенция, искусство, истина, история, Италия, Казань, католичество, княжение, королевство, культ, культура, Ливония, Литва, литература, луг, мир, мистика, монарх, монархия, мораль, муж, народ, народность, наука, национальность, нация, Новагород, обоз, образование, образованность, общественность, общество, община, орден, освобождение, отдохновение, отец, отечество, перевод, племя, подвиг, покупка, политика, польза, Польша, правительство, право, православие, предание, призвание, присоединение, продажа, произведение, просвещение, процесс, Пруссия, равнина, размышление, революция, религия, республика, Рим, Россия, Русь, Сибирь, социализм, союз, спокойствие, страна, султан, тип, тиран, традиция, Турция, устройство, учреждение, философия, Франция, христианство, царство, церковь, цивилизация, человечество, язык.*

В результате выполнения п. 3 было установлено, что из выше приведённого списка 79 слов (79 вхождений из 160) появляется единожды в одном из минитезаурусов, при этом распределение по подкорпусам следующее: XVIII: 32 слова, XIX-1: 16, XIX-2: 14, XX: 17.

33 слова (81 вхождение) появляются в 2, 3 или 4 минитезаурусах, при этом распределение по подкорпусам следующее: XVIII: 8 слов, XIX-1: 24, XIX-1I: 26, XX: 23. Эти 33 слова мы называем ядром семантического поля.

Вот этот список ядра семантического поля «империя» по данным 4 подкорпусов после ранжирования:

а) по алфавиту:

*Англия, государственность, государство, держава, Европа, император, искусство, история, культура, литература, мир, монархия, наука, нация, общество, община, политика, правительство, просвещение, революция, религия, Рим, Россия, союз, страна, традиция, учреждение, философия, Франция, христианство, царство, церковь.*

б) по нормированному рангу:

государство, император, держава, Европа, царство, церковь, Рим, Франция, христианство, монархия, правительство, страна, общество, философия, революция, культура, нация, Россия, литература, государственность, просвещение, религия, мир, искусство, община, политика, история, учреждение, Англия, союз, традиция, наука.

в) по коэффициенту семантической близости (score):

держава, государство, общество, союз, государственность, нация, император, политика, культура, страна, община, церковь, царство, христианство, религия, мир, просвещение, правительство, монархия, Европа, философия, Рим, литература, искусство, учреждение, традиция, англия, Франция, история, Россия, революция, наука.

г) по относительной частоте (ipm):

*Россия, общество, церковь, мир, история, государство, наука, просвещение, правительство, держава, политика, царство, литература, революция, философия, союз, страна, Европа, община, культура, император, искусство, христианство, нация, учреждение, Англия, религия, Рим, Франция, государственность, традиция, монархия.*

Выполнение пп. 7-10 дало следующие результаты.

Всего в сумме было выделено 115 биграмм, в подавляющем большинстве это биграммы типа *Adj+империя, империя+Ngen.*, *N+империи*. Биграммы контактные или разрывные. Еще одна группа слов – термины из парадигматического ряда, уже выявленные инструментом «Тезаурус». Количественные характеристики следующие: 78 биграмм характерны лишь для одного из подкорпусов, 13 – для двух, 10 – для трех и 4 – для четырех.

Ядро синтагматических коллокаций составляют 24 словосочетания:

*Российская империя, Византийская империя , империя германской нации, Восточная империя, Священная империя, падение империи, Австрийская империя, Великая империя, пределы империи, Турецкая империя, столица империи, Западная империя, могущество империи, Оттоманская империя, империя Карла, существование империи, восстановление империи, Латинская империя, область империи, империя Рима, империя Наполеона, разрушать империю, эпоха империи.*

В результате выполнения пп. 11-12 был сформирован объединенный список ключевых слов, полученный по частотному критерию: значительное превышение относительной частоты в наших подкорпусах по сравнению с нейтральным корпусом. Вот это список.

*князь, государь, Булгаков, царь, боярин, Иоанн, посол, отечество, россиянин, воевода, религиозный, Василий, король, войско, императрица, неприятель, Литва, Всеволод, славянин, Димитрий, русский, народ, церковь, бог, митрополит, Ярослав, Киев, крестьянин, хан, духовный, философия, Мстислав, польский, Святослав, Владимир, религия, воин, христианство, народ, государь, Христос, церковный, русский, дух, христианский, престол, царь, бытие, град, дружина, древний, двор, слава, грамота, откровение, литовский, свобода, император, мысль, учение, вельможа, мысль, святой, вера, народность,*

*свобода, царский, град, пленник, битва, граф, ум, князь, божественный, племя, грек, церковь, вера, Пётр, Франция, просвещение, поляк, душа, человечество, немец, граф, народ, сознание, небо, немецкий, французский, истина, император, Соловьев, Леонтьев, аполлон, Победоносцев, великий, наука, политический, дух, министр, цивилизация, царство, государь, царевич, мир, государство, Европа, смерть, Русь, Польша, православие, София, болгарин, Герцен, Вяземский, общество, воля, воля, римский, идеал, Австрия, мистический, сила, учение, мысль, разум, отечество, Киреевский, дух, истина, цензура, Тютчев, народ, церковь, сочинение, образованность.*

Мы можем назвать его периферией нашего семантического поля.

## 6 Заключение и выводы

Мы видим, что использование корпуса текстов и инструментов системы Sketch Engine позволяет выявлять в автоматизированном режиме синтагматические и парадигматические связи и создавать более адекватное наполнение терминосистемы. Были получены списки слов и словосочетаний, значительно расширяющие имеющиеся лексикографические пособия (Тезаурус РуТез, «Русский семантический словарь», [15: 13], [16: 475], «Константы: словарь русской культуры» [17]). Однако это «статистическое расширение» получилось чрезмерно широким (см. *посол, отечество, воевода, религиозный, религия, воин* и т.д.). Например, встает вопрос, правомерно ли включать в поле «империя» авторов, пишущих о ней (*Герцен, Киреевский, Тютчев* и др.). Очевидно, неправомерно включать в поле «империя» названия народов, населявших империи (*поляк, русский, россиянин, немец*) и соответствующие им прилагательные. Видимо, требуется продолжить эксперименты с другими более жесткими «техническими» параметрами. Очевидно, что эти списки должны быть соотнесены с экспертными знаниями.

Но уже сейчас на основе полученных результатов можно отметить, что по разным параметрам понятие “империя” в разные периоды времени в русской культуре имеет разные коннотации. Так, бросается в глаза существенное отличие текстов 18-го века. Это видно по составу лексики – см. раздел 4.2: из 79 слов тезауруса, «уникальных» только для одного периода, 32 относятся к 18-му веку. Это отличие проявляется и в именах собственных, вошедших в состав поля. И можно вообще сформулировать осторожный вывод, что несмотря на присутствие империи в 18-ом веке в реальности, сам концепт империи в русской культуре в 18-ом веке еще не сложился.

Далее, более глубокий анализ показывает изменение лексического наполнения нашего поля по данным подкорпуса 20-го века. И это при том, что тексты 20-го века в подавляющем большинстве ограничены 1917 годом.

Естественно, после работы автоматизированных механизмов необходимо привлекать экспертов, как для оценки результатов, так и для определения, если требуется, типов связей между элементами поля. Анализ лексики также показывает, что традиционные тезаурусные лексико-семантические отношения

для предметных областей в сфере культурно-литературного лексикона манифестируются недостаточно явно. Фактически, большую часть отношений между отобранными базовыми понятиями следует отнести к отношению "ассоциация". Предполагается разработка с привлечением экспертов специально ориентированного набора отношений для данного поля.

Направления дальнейшей работы следующие:

создать единый «ядерный» корпус, сбалансировав разные временные периоды;

создать подкорпус текстов после 1917 года и провести соответствующие эксперименты;

провести эксперименты с другими параметрами инструментов «Тезаурус» и «Коллокации» (в частности, уменьшить количество терминов, включаемых в дистрибутивный тезаурус, и увеличить размер окна выявления коллокаций);

выявить элементы семантического поля (дистрибутивного тезауруса) для терминов, вошедших в ядро поля «империя», т.е. создать тезаурусы (поля) второго уровня, и сформировать объединенный список, по возможности, в виде семантической сети;

провести лингвистическую и культурно-историческую интерпретацию полученных результатов;

разработать или адаптировать программное обеспечение для создания и ведения электронного тезауруса- компьютерного представления поля;

создать электронный тезаурус для семантического поля «империя» с указанием связей между его элементами, с частотными характеристиками и примерами употребления в корпусах.

## Благодарности

Исследование поддержано грантом РФФИ № 18-012-00474 «Семантическое поле «империя» в русском, английском и чешском языках» и грантом РФФИ № 17-04-00552-ОГН-А «Параметрическое моделирование лексической системы современного русского литературного языка».

## Литература

1. Адмони В. Г. Синтаксис современного немецкого языка: Система отношений и система построения. Л.: Наука, 1973.
2. Апресян Ю.Д. Образ человека по данным языка: Попытка системного описания // ВЯ. 1995, № 1.
3. Арапов М.В. Некоторые принципы построения словаря типа “тезаурус” // НТИ. Сер. 2. 1964. № 4. С. 40–46.
4. Аскольдов С.А. Концепт и слово / Русская словесность. От теории словесности к структуре текста. Антология. М., 1980.
5. Ахманова О.С. Словарь лингвистических терминов. М., 1966.
6. Бондарко А.В. Функциональная грамматика. Л., 1984

7. Васильев Л.М. Современная лингвистическая семантика. М., 1990
8. Вежбицкая А. Понимание культур через посредство ключевых слов М., Языки славянской культуры, 2001.
9. Вежбицкая А. Язык. Культура. Познание. М., Русские словари, 1997.
10. Войскунский В.Г., Захаров В.П., Мордовченко П.Г., Сороколетова Л.И. О некоторых лексико-семантических проблемах в "бестезаурусных" ИПС // Структурная и прикладная лингвистика: Межвузовский сборник. Вып. 2. Л., ЛГУ, 1983. С.170 - 177.
11. Караулов Ю.Н. Лингвистическое конструирование и тезаурус литературного языка. М., 1981.
12. Кобозева И.М. Лингвистическая семантика. М., 2000
13. Пиотровский Р. Г. Текст, машина, человек. Л., 1975.
14. Рубашкин В.Ш. Онтологическая семантика: Знания. Онтологии. Онтологически ориентированные методы информационного анализа текстов. М., 2012.
15. Русский семантический словарь. Т. 2. Москва, Азбуковник. 2002.
16. . Русский семантический словарь. Т. 3. Москва, Азбуковник. 2003.
17. Степанов Ю.С. Константы: Словарь русской культуры. М., 2001.
18. Хохлова М.В. Разработка грамматического модуля русского языка для специализированной системы обработки корпусных данных // Вестник Санкт-Петербургского государственного университета. Серия 9. Филология, востоковедение, журналистика. СПб., 2010. Выпуск 2. С. 162–169.
19. Шайкевич А. Я. Дистрибутивно-статистический анализ текстов. АДД. Л., 1982.
20. Шайкевич А.Я. Дистрибутивно-статистический анализ в семантике // Принципы и методы семантических исследований. М., 1976 . С. 353 378.
21. Шайкевич А.Я. Распределение слов в тексте и выделение семантических полей // Иностранные языки в высшей школе. М.,1963.
22. Щур Г.С. Теория поля в лингвистике. М.-Л., 1974.
23. Blancafort H. Daille B., Gornostay T., Heid U., Méchoulam C., Sharoff S. TTC: Terminology extraction, translation tools and comparable corpora // 14th EURALEX International Congress. 2010. P. 263 268.
24. Gamallo P., Gasperin C., Augustini A., Lopes G. P. Syntactic-Based Methods for Measuring Word Similarity // Text, Speech and Dialogue: Fourth International Conference TSD–2001. LNAI 2166. Springer-Verlag, 2001. P. 116–125.
25. Kilgarriff A., Miloš Husák, Katy McAdam, Michael Rundell and Pavel Rychlý (2008). GDEX: Automatically finding good dictionary examples in a corpus. In Proceedings of the 13th EURALEX International Congress. Spain, July 2008, pp. 425–432.
26. Kilgarriff A., Rychly P. An efficient algorithm for building a distributional thesaurus (and other Sketch Engine developments) // Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions. Czech Republic, June 2007, pp. 41–44.
27. Kilgarriff A., Rychlý P., Jakubíček M., Rundell M. et al., SketchEngine [Computer Software and Informatiom Resource], URL: http://www. sketchengine.co.uk/ Последнее обращение 3.12.2018.

28. Kilgarriff A., Rychly P., Smrz P., Tugwell D. The Sketch Engine // Proceedings of the XIth Euralex International Congress. — Lorient: Universite de Bretagne-Sud, 2004. — P. 105 116.
29. Kilgarriff Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, VítSuchomel. The Sketch Engine: ten years on. In Lexicography 2014? 1(1): 7–36. DOI: 10.1007/s40607-014-0009-9. ISSN 2197-4292.
30. Lin D. Automatic retrieval and clustering of similar words. Proc. COLING-ACL. Montreal: 1998. P. 768-774.
31. Pazienza M., Pennacchiotti M., and Zanzotto F. Terminology extraction: an analysis of linguistic and statistical approaches. Knowledge Mining Series: Studies in Fuzziness and Soft Computing, Springer Verlag, Berlin, 2005. P. 255–279.
32. Pekar V. Linguistic Preprocessing for Distributional Classification of Words // Proceedings of the COLING–04 Workshop on Enhancing and Using Electronic Dictionaries. Geneva: 2004. P. 15–21.
33. Rychlý P. A lexicographer-friendly association score // Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN. Brno. 2008. P. 6–9.
34. Sharoff S. Open-source corpora: Using the net to fish for linguistic data // International journal of corpus linguistics. John Benjamins Publishing Company. 2006. Vol. 11. No. 4. P. 435-462.
35. Smrž P., Rychlý P. Finding Semantically Related Words in Large Corpora // Text, Speech and Dialogue: Fourth International Conference (TSD–2001). LNAI 2166. Springer-Verlag, 2001. P. 108–115.
36. Statistics Used in Sketch Engine. URL: https://www.sketchengine.co.uk/documentation/statistics-used-in-sketch-engine/ (Accessed 03.12.2018).
37. Zakharov V. Comparative Evaluation and Integration of Collocation Extraction Metrics. In: Lecture Notes in Computer Science, vol. 10415 (Text, Speech, and Dialogue - 20th International Conference, TSD 2017, Prague, Czech Republic, August 27-31, 2017, Proceedings) / K. Ekstein, V. Matousek (Eds.). Springer International Publishing AG 2017. P. 255-262.

## Acknowledgement

This work was implemented with financial support of the Russian Foundation for Basic Research, the project No. 18-012-00474 «Semantic field "empire" in Russian, English and Czech» and the project No. 17-04-00552 «Parametric modeling of the lexical system of the modern Russian literary language».

## References


1. Admoni, V. G.: Syntax of modern German: The system of the relations and the system of construction, [Sintaksis sovremennogo nemeckogo jazyka: Sistema otnoshenij i sistema postroenija], Leningrad (1973).
2. Apresyan, Yu.D.: The image of a person according to the language: An attempt of the system description, [Obraz cheloveka po dannym jazyka: Popytka sistemnogo opisanija], Linguistics aspects, [Voprosy yazykoznaniya], 1 (1995).
3. Arapov, M.V.: Some principles of creation of the "thesaurus" dictionary NTI Serie 2(4), 40-46 (1964).
4. Askoldov, S.A.: Concept and word, [Koncept i slovo]. Moscow (1980).
5. Akhmanova, O.S.: Dictionary of Linguistic Terminology [Slovar' lingvistich-eskikh terminov]. . Moscow (1966).
6. Bondarko A.V.: Functional grammar, [Funkcional'naja grammatika]. Leningrad (1984).
7. Vasilyev, L.M.: Modern linguistic semantics, [Sovremennaja lingvisticheskaja semantika]/ Moscow (1990).
8. Vezhbitskaya, A.: Understanding of cultures through keywords, [Ponimanie kul'tur cherez posredstvo kljuchevyh slov]. Moscow (2001).
9. Vezhbitskaya, A.: Language. Culture. Knowledge, [Jazyk. Kul'tura. Poznanie]. Moscow (1997).
10. Voyskunsky, V.G., Zakharov, V. P., Mordovchenko, P.G., Sorokoletova, L.I.: About some lexico-semantic problems in "thesauruless" IRS, [O nekotoryh leksi-ko-semanticheskih problemah v "bestezaurusnyh" IPS]. In: Structural and applied linguistics. [Strukturnaja i prikladnaja lingvistika] vol. 2, pp. 170-177. LGU, Len-ingrad (1983).
11. Karaulov, Yu.N.: Linguistic designing and thesaurus of the literary language, [Lingvisticheskoe konstruirovanie i tezaurus literaturnogo jazyka]. Moscow (1983).
12. Kobozeva, I.M.: Linguistic semantics, [Lingvisticheskaja semantika]. Moscow (2000).
13. Piotrovsky, R. G.: Text, computer, human, [Tekst, mashina, chelovek]. Leningrad (1975).
14. Rubashkin, V.Sh.:Ontologic semantics [Ontologicheskaja semantika]. Moscow (2012).
15. Russian Sematic Dictionary [Russkiy semanticheskiy slovar'], vol.2, Moscow (2002).
16. Russian Sematic Dictionary [Russkiy semanticheskiy slovar'], vol.3, Moscow (2003).
17. Stepanov, Yu.S.: Constants: Dictionary of the Russian Culture [Konstanty: slovar' russkoy kultury]. Moscow (2002).
18. Khokhlova, M.V.: Development of the grammatical module of Russian for the specialized system of processing of corpus data [Razrabotka grammaticheskogo modulja russkogo jazyka dlja specializirovannoj sistemy obrabotki korpusnyh dannyh], Bulletin of St. Petersburg State University [Vestnik Sankt-Peterburgskogo gosudarstvennogo universiteta], Series 9, Philology, oriental stud-ies, journalism. 2(9), 162-169 (2010).

19. Shaykevich, A. Ya.: Distributive and statistical analysis of texts [Distributivno-statisticheskij analiz tekstov], PhD thesis. Leningrad (1982).
20. Shaykevich, A.Ya.: The distributive and statistical analysis in semantics [Distributivno-statisticheskij analiz v semantike]. In: Principles and methods of semantic researches [Principy i metody semanticheskih issledovanij], pp. 353-378. Moscow (1976).
21. Shaykevich, A.Ya.: Distribution of words in the text and allocation of semantic fields [Raspredelenie slov v tekste i vydelenie semanticheskih polej], In: Foreign languages in higher education. Moscow, 1963.
22. Shchur, G.S.: Field theory in linguistics, [Teorija polja v lingvistike], Moscow-Leningrad (1974).
23. Blancafor,t H. Daille, B., Gornostay, T., Heid, U., Méchoulam, C., Sharoff, S.: TTC: Terminology extraction, translation tools and comparable corpora. In: 14th EURALEX International Congress,. pp. 263-268. EURALEX (2010).
24. Gamallo, P., Gasperin, C., Augustini, A., Lopes, G. P.: Syntactic-Based Methods for Measuring Word Similarity, In: Text, Speech and Dialogue: Fourth International Conference TSD–2001. LNAI 2166, pp. 116–125. Springer-Verlag (2001).
25. Kilgarriff, A., Miloš Husák, Katy McAdam, Michael Rundell and Pavel Rychlý: GDEX: Automatically finding good dictionary examples in a corpus, In: Proceedings of the 13th EURALEX International Congress. Spain, July 2008. pp. 425–432. EURALEX (2008).
26. Kilgarriff, A., Rychly, P.: An efficient algorithm for building a distributional thesaurus (and other Sketch Engine developments), In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions. Czech Republic, June 2007, pp. 41–44. ACL (2007).
27. Kilgarriff, A., Rychlý, P., Jakubíček, M., Rundell, M. et al.: SketchEngine [Computer Software and Informatiom Resource], URL: http://www. sketchengine.co.uk, last accessed 2018/12/03.
28. Kilgarriff, A., Rychly, P., Smrz, P., Tugwel,l D. (2004), The Sketch Engine, In: Proceedings of the XIth Euralex International Congress, pp. 105-116. Lorient: Universite de Bretagne-Sud (2004).
29. Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, Vít Suchomel: The Sketch Engine: ten years on. In: Lexicography, 1(1), pp. 7–36. DOI: 10.1007/s40607-014-0009-9. ISSN 2197-4292 (2014).
30. Lin, D.: Automatic retrieval and clustering of similar words. In: Proc. COLING-ACL, pp. P. 768-774. Montreal (1998).
31. Pazienza, M., Pennacchiotti, M., and Zanzotto, F.: Terminology extraction: an analysis of linguistic and statistical approaches, In: Knowledge Mining Series: Studies in Fuzziness and Soft Computing, pp. 255–279. Springer Verlag, Berlin (2005).
32. Pekar, V.: Linguistic Preprocessing for Distributional Classification of Words. In: Proceedings of the COLING–04 Workshop on Enhancing and Using Electronic Dictionaries, pp. 15–21. Geneva (2004),
33. Rychlý, P.: A lexicographer-friendly association score, In: Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN, pp. 6–9. Brno (2008).

34. Sharoff S.: Open-source corpora: Using the net to fish for linguistic data, In: International journal of corpus linguistics, 4(11). pp. 435-462. John Benjamins Publishing Company (2006).
35. Smrž, P., Rychlý, P.: Finding Semantically Related Words in Large Corpora, In: Text, Speech and Dialogue: Fourth International Conference (TSD–2001), LNAI 2166, pp. 108-115. Springer-Verlag (2001).
36. Statistics Used in Sketch Engine. URL: https://www.sketchengine.co.uk/documentation/statistics-used-in-sketch-engine, last accessed 2018/12/3).
37. Zakharov V.: Comparative Evaluation and Integration of Collocation Extraction Metrics, In: K. Ekstein, V. Matousek (Eds.), Text, Speech, and Dialogue - 20th International Conference, TSD 2017, Prague, Czech Republic, August 27-31, 2017, Proceedings), LNCS, vol. 10415, pp. 255-262. Springer International Publishing AG, (2017).