Основное содержание отчета по гранту РФФИ 18-012-00474 А

«Семантическое поле «империя» в русском, английском и чешском языках»

рук. В.П.Захаров

Форма 501(итог). КРАТКИЙ НАУЧНЫЙ ОТЧЕТ

1.7. Аннотация, публикуемая на сайте РФФИ (кратко; описать содержание проведенных исследований и полученные результаты за весь срок реализации Проекта)

Целью проекта является формирование семантического поля «империя» на основе корпусов русского, английского и чешского языков и дистрибутивно-статистической методологии. Актуальность исследования заключается в создании лингвистического ресурса для указанных языков и в разработке методов анализа текста статистическими методами на основе корпусных данных. Понятие «семантическое поле» применяется в лингвистике для обозначения совокупности языковых единиц, объединенных каким-то общим семантическим признаком, то есть имеющих некоторый общий компонент значения. В роли таких лексических единиц выступают слова и словосочетания, как нарицательные, так и имена собственные. Понятие семантического поля интуитивно понятно, сами семантические поля разного типа представлены в разных областях знания в большом количестве, но формальных методов выявления наполнения полей не так много. Словарный состав полей представляет собой сложную систему отношений и оппозиций, как лингвистических, так и экстралингвистических. Обобщенно их можно назвать ассоциативными связями. Был проведен анализ лексикографических источников и намечены варианты их использования. Методом семантического развертывания было определено предъядро поля для каждого из языков. Далее было осуществлено исследование на базе корпусов, а также масштабная серия ассоциативных экспериментов. За период проведения исследования были созданы корпусы по теме проекта общим объемом 77 млн слов. Использовались также готовые корпусы и инструменты систем Sketch

Engine, Aranea Corpora, Corpus.Byu.Edu и Wortschatz. В результате работы сформировано лексическое наполнение семантического поля «империя» для трех языков. Одним из важнейших результатов исследования является технология формирования семантического поля на основе автоматизированных методов. Данная технология описана в отчете по проекту и в учебном пособии. Результаты исследований представлены в виде лексикографического продукта (базы данных и макета печатного издания) с количественными характеристиками связанности лексических единиц и примерами употреблений из корпусов. Произведена лингвистическая и культурно-историческая корреляция наполнения семантических полей "империя" для каждого из трех языков. Созданные за период проведения исследования корпусы выложены в общий доступ в текстовом формате и загружены в систему Sketch Engine.

Форма 502(итог). КРАТКИЙ НАУЧНЫЙ ОТЧЕТ на английском языке

2.3. Название Проекта (на англ. языке)

"Empire" semantic field in Russian, English and Czech languages

2.5. Аннотация, публикуемая на сайте РФФИ (на английском языке) (кратко; описать содержание проведенного исследования и полученные результаты за весь период реализации Проекта)

The aim of the project is the formation of the semantic field "empire" based on the corpus of Russian, English and Czech languages by means of distributional and statistical methods. The relevance of the study is explained by the creation of a unique linguistic resource for these languages and by the development of methods for analyzing texts using statistical methods based on corpus data. The concept of “semantic field” is used in linguistics to denote a set of linguistic units grouped by some common semantic attribute, that is, having some common component of meaning. Words and phrases are chosen as such lexical units, both common nouns and proper names. The concept of a semantic field is intuitively clear; semantic fields of various types themselves are represented in large amounts in different fields of knowledge, but there are not so many formal methods for detecting the

filling of fields. The vocabulary of the fields is a complex system of relations and oppositions, both linguistic and extralinguistic. In general, they can be called associative connections. An analysis of lexicographic sources was carried out and options for their use were outlined. Using semantic deployment, the core of the field was determined for each of the languages. At the next stage of our project, an associative experiment was carried out. In course of our study, corpora were created on the subject of the project with a total volume of 77 million words. Sketch Engine, Aranea Corpora, Corpus.Byu.Edu and Wortschatz were also used as off-the-shelf corpora and tools. As a result of the work, the lexical filling of the semantic field “empire” for three languages is formed. The result of the study includes a technology for the formation of a semantic field based on automated methods. It is described in the project report and in a manual for students. Research results are presented in the form of a lexicographic product (database and print layout) with quantitative characteristics of the connectedness of lexical units and examples of uses from the corpora. A linguistic and cultural-historical correlation of the filling of the semantic fields "empire" for each of the three languages was established. The resources created within the project are available in public access in text format and were uploaded to the Sketch Engine system.

Форма 503(итог). РАЗВЕРНУТЫЙ НАУЧНЫЙ ОТЧЕТ

3.4. Цель и задачи фундаментального исследования (указать как в заявке)

Основная цель проекта – описание семантического поля «империя» в русском, английском и чешском языках в виде взаимосвязанных тезаурусов с указанием связей между их элементами, с частотными характеристиками и примерами употребления в корпусах.

Выбор данных языков обусловлен тем, что в этих языках понятие «империя» сильно связано с исторической памятью и что оно «живо» в языковом сознании носителей языка. Также нам интересно исследовать разные языки, с одной стороны, принадлежащие к одной языковой семье, с другой – к разным.

Достижение поставленной цели предполагает решения ряда задач, в частности:

1) дать теоретическое освещение понятиям «семантическое поле» и «тезаурус» в современном лингвистическом знании, охарактеризовать смежные научные понятия и обосновать принятую в работе технологию анализа;

2) выявить смысловое наполнение понятия «империя» по данным различных словарей трех языков;

3) сформировать список исследуемых лингвистических и статистических параметров для синхронического и диахронического исследования, позволяющих оценить «поведение» лексических единиц и силу связей между ними количественно,

4) создать дополнительные корпуса для целей исследования;

5) провести корпусное исследование парадигматических и синтагматических связей слова «империя» в трех выбранных языках;

6) сформировать лексическое наполнение тезаурусов трех языков (собственно семантического поля «империя») на основе дистрибутивно-статистических методов;

7) исследовать парадигматику и синтагматику лексических единиц, вошедших в поле «империя»;

8) сопоставить смысловое наполнение поля «империя» для трех языков с данными лексикографического исследования;

9) провести квантитативное исследование поведения единиц сформированных полей в диахронии по корпусным данным;

10) разработать программно-лингвистическое обеспечение для сравнения наполнения поля «империя» и статистическую характеристику лексических единиц-элементов поля в трех языках;

11) провести сравнение наполнение поля «империя» и статистическую характеристику лексических единиц-элементов поля в трех языках;

12) провести лингвистическую и культурно-историческую интерпретацию полученных результатов:

13) разработать или адаптировать программное обеспечение для создания и ведения электронного тезауруса;

14) создать электронные тезаурусы для семантического поля «империя» для трех языков с указанием связей между его элементами, с частотными характеристиками и примерами употребления в корпусах;

15) предоставить созданные в ходе выполнения проекта продукты в общедоступное пользование;

16) подготовить к изданию макет тезаурусов для трех языков в печатной форме;

17) подготовить учебно-методические материалы для использования результатов проекта в учебном процессе.

3.5. Важнейшие результаты, полученные при реализации Проекта

По данным задачам были получены следующие результаты.

1. В результате работы над проектом мы дали теоретическое обоснование понятиям «семантическое поле» и «тезаурус», а также проанализировали такие смежные научные понятия, как «языковая картина мира», «языковое сознание», «концепт».

В современном лингвистическом знании понятие «семантическое поле» обозначает совокупность языковых единиц, которые объединены общим семантическим признаком, т. е. имеют некоторый общий компонент значения. В роли таких языковых единиц могут выступать слова и словосочетания, как имена нарицательные, так и имена собственные. С этим термином тесно связаны такие понятия, как поле, лексическое поле, лексико-семантическое поле, функционально-семантическое поле, тезаурус, онтология, кластер, терминосистема. Несмотря на то, что понятие семантического поля описано во многих работах (Щур 1974, Степанов 1975, Апресян 1957 и др., Караулов 1976, Николенко 2005, Уфимцева 1988, Воробьев 1991, Денисов 1993, Кобозева 2000, Шеина 2000), формальные методы наполнения семантического поля не были в достаточной степени разработаны.

Теория семантических полей опирается на представление о том, что в языке существуют некоторые семантические группы, элементы которых связаны между собой как лингвистическими, так и экстралингвистическими отношениями.

В нашем проекте мы прежде всего опирались на определение О. С. Ахмановой: «Поле — совокупность содержательных единиц, покрывающая определенную область человеческого опыта и образующая более или менее автономную микросистему» (Ахманова 1966). Важным в данном определении для нас является следующее: 1) большая или меньшая автономность поля; 2) поле как микросистема; 3) определенная область человеческого опыта.

Рассмотрим более подробно каждый из этих пунктов. Во-первых, для одной и той же области человеческого опыта можно сформировать поля, которая будут обладать большей или меньшей автономностью. Например, в Новом словаре иностранных слов понятие «империя» описывается следующим образом: «[от лат. imperium власть; государство] – 1) монархическое государство, главой которого является император; 2) государство, имевшее колониальные владения (Британская и., Французская колониальная и.); 3) * крупная монополия, осуществляющая контроль над целой отраслью промышленности, над какой-л. деятельностью, напр. медиаимперия, алмазная и.; 4) * о чем-л. как средоточии, господстве каких-л. качеств, свойств, напр. и. беззакония, и. зла (средоточие насилия, жестокости), и. грёз (о Голливуде)». Таким образом, выделяются 4 значения и ряд связанных с ними понятий: «император», «государство», «Британская империя», и т. д. Принадлежат ли эти понятия одному и тому же семантическому полю? В словаре (БСЭ, 1998) содержатся только 2 первых значения слова «империя», тогда как в энциклопедии «Российская империя : полная энциклопедия сословий, титулов, чинов» (2009) объём семантического поля значительно шире.

В связи с этим возникает вопрос о границах семантического поля. По определению В. Г. Адмони, в семантическом поле можно выделить ядро – центральную часть, элементы которой обладают полным набором признаков, определяющих данную группировку, – и периферию, элементы которой

могут быть связаны с другими полями. Сила этих связей может быть разной, в результате чего возникает необходимость вычислить эти значения. Стоит отметить, что те участки поля, на которых связи между элементами особенно сильны, можно считать элементарными микрополями внутри семантического поля (Адмони 1973).

Во-вторых, если речь идёт о семантическом поле как о микросистеме, то возникает вопрос о системных отношениях между его элементами. В лингвистике в целом и в лингвистической семантике в частности выделяют несколько типов отношений, однако далеко не все связи из человеческого опыта находят свое выражение в форме лингвистических категорий. Именно поэтому нам кажется особенно важным обратить внимание на третий аспект определения О. С. Ахмановой: «определенная область человеческого опыта».

Связь человеческого опыта и слова прослеживается Д. С. Лихачевым, который для характеристики потенций, заложенных в языковом тезаурусе, использует понятие концепта (Лихачев 1997).

Термин «концепт» широко используется в различных научных дисциплинах и является достаточно разработанным в лингвистике и культурологии. В частности, этот термин «покрывает» предметные области когнитивной психологии и когнитивной лингвистики – научных направлений, связанных с проблемами мышления и познания (Кубрякова 1996), – а также культурологии, в рамках которой концепт понимается как основная ячейка культуры в ментальном мире человека.

В рамках компьютерной и корпусной лингвистики термин «семантическое поле» расширяется до понятий «тезаурус» и «онтология». Построение тезаурусов и онтологий опирается на задачу автоматического выявления парадигматических и синтагматических связей между элементами семантического поля, т. е. его автоматического наполнения. Тогда как синтагматические связи между языковыми единицами отражены в большей или меньшей степени в традиционных словарях разного типа и представлены в тексте в явном виде, парадигматические отношения в тексте скрыты, и для их выделения требуются более сложные процедуры, чем при выделении синтагматических отношений. В этой связи тезаурусы занимают особое

место: они представляют собой словари понятий, фиксирующие человеческие знания, и могут считаться парадигматическими.

Важным типом представления знаний для когнитивных исследований является ассоциативный тезаурус – модель сознания, которая представляет собой набор правил оперирования знаниями определенной культуры (вербальными и невербальными значениями), отражающими образ мира данной культуры. Ассоциативный тезаурус можно назвать моделью сознания человека. На основании исследований, которые проводились в последнее время в московской психолингвистической школе на материале Русского ассоциативного словаря (Караулов и др., 1994–1998) и The Associative Thesaurus of English (Kiss G.& all., 1972).

В ходе нашего исследования мы опирались не только на анализ текстов, но и на ещё один способ построения полей с помощью метода ассоциативного эксперимента. Этот метод связан с понятиями «языковая картина мира», «языковое сознание» и «концепт». Концепты возникают в сознании человека не только на основе словарных значений слов, но и на основе культурно-исторического опыта (Лихачев 1993). По мнению Ю.С. Степанова, именно эта особенность позволяет определять концепт как «сгусток культуры в сознании человека; то, в виде чего культура входит в ментальный мир человека… то, посредством чего человек входит в культуру, а в некоторых случаях и влияет на нее» (Степанов 2001). Языковое сознание, таким образом, понимается как совокупность структур сознания, в формировании которых были использованы социальные знания, связанные с языковыми знаками, или как образы сознания, «овнешняемые» языковыми средствами: отдельными лексемами, словосочетаниями, фразеологизмами, текстами, ассоциативными полями и ассоциативными тезаурусами как совокупностью этих полей. На наш взгляд, понятие империи в сознании русских, британцев, чехов, живших, как известно, в Австро-Венгерской империи, является частью как раз этого языкового сознания. Именно поэтому часть нашего исследования отводится анализу элементов семантического поля «Империя» на основе ассоциативных словарей и эксперимента. В качестве альтернативного источника ассоциатов использовались также базы данных, в

которых сеть ассоциаций постоянно пополняется (wordassociations.net, sociation.org и др.).

2. С помощью анализа различных словарей русского, английского и чешского языков мы выявили смысловое наполнение понятия «империя». Были проанализированы следующие словарные источники.

(1) Русский язык:

Большая советская энциклопедия. 3-е изд. / Глав. ред. А. М. Прохоров. — М.: Сов. энциклопедия, 1969-1978.

Большой академический словарь русского языка: В 30 т. / Под ред. К. С. Горбачевича и др. СПб.: Изд-во «Наука», 2004—. Издание продолжается (в 2017 году вышел том 24).

Большой толковый словарь русского языка / РАН. Ин-т лингв. исслед.; Сост., гл. ред. канд. филол. наук С. А. Кузнецов. — СПб.: Норинт, 1998.

Большой энциклопедический словарь / Ред. А. М. Прохоров . – 2-е изд., перераб. и доп . – М.: 2000.

Энциклопедический словарь Ф.А. Брокгауза и И.А. Ефрона. — С.-Пб.: Брокгауз-Ефрон. 1890—1907.

Булыко А. Н. Большой словарь иностранных слов : 35 тысяч слов / А.Н. Булыко. - М., 2006.

Даль В. И. Толковый словарь живого великорусского языка: избр. ст. / В. И. Даль; совмещ. ред. изд. В. И. Даля и И. А. Бодуэна де Куртенэ; [науч. ред. Л. В. Беловинский]. - М. : ОЛМА 2009.

Захаренко Е. Н. Новый словарь иностранных слов : [свыше 25 000 слов и словосочетаний : толкование, этимология, примеры употребления] / Е. Н. Захаренко, Л. Н. Комарова, И. В. Нечаева. – Изд. 3-е, испр. и доп. – Москва : Азбуковник, 2008.

Караулов Ю. Н., Молчанов В. И., Афанасьев В. А., Михалев Н. В. Русский семантический словарь: Опыт автомат. построения тезауруса: от понятия к слову / Отв. ред. С. Г. Бархударов. М.: Наука, 1983.

Караулов Ю.Н., Сорокин Ю.А., Тарасов Е.Ф., Уфимцева Н.В., Черкасова Г.А. Русский ассоциативный словарь. Ассоциативный тезаурус современного русского языка. В 3-х частях, 6-ти книгах. М., 1994, 1996, 1998.

Комлев Н. Г. Словарь иностранных слов. М.: Эксмо-пресс, 2000.

Леонтьев А.А. Словарь ассоциативных норм русского языка. М : Изд-во Моск. ун-та, 1977. 192 с.

Музрукова Т. Г., Нечаева И. В. Популярный словарь иностранных слов. – М.: Азбуковник, 2002.

Абрамов Н. (Н.А.Переферкович) Словарь русских синонимов и сходных по смыслу выражений. – АСТ Астрель, 2006.

Ожегов С. И., Шведова Н. Ю. Толковый словарь русского языка: 80 000 слов и фразеологических выражений / Российская академия наук. Институт русского языка им. В. В. Виноградова. 4-е изд., дополненное. М.: Азбуковник, 1999.

Русский семантический словарь. Толковый словарь, систематизированный по классам слов и значений / Российская академия наук. Ин-т рус. яз. им. В. В. Виноградова; Под общей ред. Н. Ю. Шведовой. – М.: «Азбуковник», 1998.

Словарь иностранных слов / Ред. И.В. Лехин, Ф.Н. Петров. – М., 1979.

Словарь иностранных слов. Под ред. И.В.Лехина, С.М.Локшиной, Ф.Н.Петрова и Л.С.Шаумяна. Изд. 7-ое перераб. М.: Русский язык, 1979.

Словарь иностранных слов. Под ред.И.В.Лехина и проф. Ф.Н.Петрова. Изд.5-е стереотип. Изд. М.: Гос.изд-во иностр. и нац. словарей, 1955.

Словарь русского языка: В 4-х т. / АН СССР, Ин-т рус. яз.; Под ред. А. П. Евгеньевой. — 2-е изд., испр. и доп. М.: Русский язык, 1981—1984.

Словарь современного русского литературного языка: В 17 т. / Под ред. В. И. Чернышёва. М., Л.: Изд-во АН СССР, 1948—1965.

Словарь языка Пушкина: в 4 т / Отв. ред. акад. АН СССР В. В. Виноградов. — 2-е изд., доп. — М.: Азбуковник, 2000.

Советская историческая энциклопедия. — М.: Изд. Советская энциклопедия, 1973—1982.

Социологический энциклопедический словарь. На русском, английском, немецком, французском и чешском языках. Редактор-координатор — академик РАН Г. В. Осипов. — М.: Издательство НОРМА (Издательская группа НОРМА—ИНФРА) М., 2000.

Толковый словарь русского языка : в 4 т. / Под. ред. Д. Н. Ушакова. М., 1935—1940.

Толковый словарь русского языка / Д. В. Дмитриев. 2003.

Толковый словарь русского языка конца XX века. Языковые изменения / Под ред. Г. Н. Скляревской. — СПб., Фолио-Пресс, 1998.

Толковый словарь русского языка начала XXI века : актуальная лексика : около 8500 слов и устойчивых словосочетаний / [авт.-сост.: Г. Н. Скляревская и др.] ; под ред. Г. Н. Скляревской. Москва, 2008.

Толковый словарь современного русского языка Языковые изменения конца XX столетия. Под редакцией Г. Н. Скляревской, М., Астрель АСТ, 2001/

Фасмер М. Этимологический словарь русского языка: В 4-х т. / Перевод и дополнения О. Н. Трубачёва. — 4-е изд., стереотип. — М.: Астрель — АСТ, 2004.

Энциклопедический словарь Ф.А. Брокгауза и И.А. Ефрона. — С.-Пб.: Брокгауз-Ефрон. 1890—1907.

(2) Английский язык:

Cambridge Academic Content Dictionary. Cambridge University Press, 2017.

Cambridge Advanced Learner's Dictionary & Thesaurus. Cambridge University Press, 2008.

Cambridge Business English Dictionary. Cambridge University Press, 2011.

Collins English Dictionary. Collins Uk, 2006.

Compact Edition of the Oxford English Dictionary. Oxford University Press, 1979.

Concise Oxford American Dictionary. Oxford University Press, 2006.

Concise Oxford American Thesaurus. Oxford University Press, 2006.

Concise Oxford Thesaurus. Oxford University Press, 2002.

Encyclopedia Britannica. Encyclopaedia Britannica, 2003.

Longman Dictionary of Contemporary English. Pearson Longman, 2009.

Merriam-Webster Collegiate Dictionary. Merriam-Webster, 2004.

Merriam-Webster's Dictionary of English Usage. Merriam-Webster, 1989.

New Roget's Thesaurus in Dictionary Form. Berkley, 1992.

Online Etymology Dictionary. Available: https://www.etymonline.com.

Online OXFORD Collocation Dictionary. Available: http://www.freecollocation.com.

Oxford Dictionary Of Current English. Oxford University Press, 1993.

Oxford Dictionary of English. Oxford University Press, 2010.

Oxford Thesaurus of English. Oxford University Press, 2009.

Webster's New Dictionary of Synonyms. Merriam, 1978.

Webster’s Third New International Dictionary of the English Language. Merriam-Webster, 1993.

WordNet 2.0. Available: https://wordnet.princeton.edu.

WordNet 3.0. Available: https://wordnet.princeton.edu.

(3) Чешский язык:

Encyklopedie ceskych pravnich dejin. Dil 1-2-... Praha, 2013-...

Etymologicky slovnik jazyka ceskeho / Vaclav Machek. Praha, 2010.

Ilustrovany skolni slovnik ceskeho jazyka / [sestavili Jana Cerna ... et al.]. Plzen, 2008.

Mala ceskoslovenska encyklopedie. Praha, 1984–1987.

Muj prvni slovnik ceskeho jazyka / Jana Marie Schrimpfova. Plzen, 2008.

Prirucni slovnik naucny (PSN). Praha, 1962-1967.

Slovnik jazyka ceskeho / Frantisek Travnicek. 4. vyd. Praha, 1952.

Slovnik nareci ceskeho jazyka / Martina Ireinova, Hana Konecna. Praha, 2016.

Slovnik spisovneho jazyka ceskeho / za vedeni B. Havranka. Praha, 1989.

Slovnik spisovneho jazyka ceskeho. 1. vyd. Praha, 1971.

Strucny etymologicky slovnik jazyka ceskeho se zvlastnim zretelem k slovum kulturnim a cizim / Josef Holub, Stanislav Lyer. Praha, 1992.

Strucny slovnik ceskych synonym / J. Masin, J. V. Becka. Praha, 1947.

Tezaurus jazyka ceskeho : slovnik ceskych slov a frazi souznacnych, blizkych a pribuznych / Ales Klegr. Praha, 2007.

Universum. Praha, 1999–2001.

3. Для оценки «поведения» лексических единиц – элементов ядра семантического поля «империя» – и вычисления силы связей между ними были введены и проанализированы следующие лингвистические и статистические параметры:

1) по словарям:

- количество словарей, в которым представлена единица;

- тип словаря;

- объем словарной статьи;

- в каком месте статьи единица встретилась;

- плотность отдельного термина в словарной статье;

- плотность всех терминов поля в словарной статье;

- плотность терминов ядра поля в словарной статье.

На основе этих параметров была получена интеграционная характеристика качества единицы с точки зрения принадлежности к семантическому полю.

2) по ассоциативным словарям:

- количество стимульных слов и количество реакций, входящих в поле (ассоциативную статью);

- количество разных стимулов/реакций на слова;

- в конечном счете минисеть поля по данным ассоциативного словаря и эксперимента.

3) по корпусам (по текстам):

- с помощью мер tf*idf и/или ARF выявление базовых текстов (для переформирования корпусов);

- абсолютная частота единиц по всему массиву;

- относительная частота единиц (ipm) по всему массиву;

- частотность (ipm) элементов поля по подкорпусам (хронологическим и тематическим);

- оценка статистической значимости терминов и словосочетаний по мере logDice во всем массиве и по подкорпусам (keyness);

- рекомендуемый диапазон окна вычисления коллокаций:

переменный:

для существительных: от -3 до +1,

для прилагательных: от -1 до +3;

- максимальное число единиц в тезаурусе – 20.

4. Для целей исследования были созданы и использованы следующие корпусы текстов:

(1) Русский корпус по теме «империя» объёмом ок. 32,6 млн словоформ;

(2) Английский корпус по теме «империя» объёмом ок. 25,6 млн словоформ;

(3) Чешский корпус по теме «империя» объёмом ок. 19,6 млн словоформ.

Эти корпуса были созданы в системе Sketch Engine и могут быть предоставлены по запросу любому желающему в текстовом формате. Вместе с инструментами и готовыми корпусами систем Aranea Corpora (https://korpus.sk/semä, разработчик Радован Гарабик (Radovan Garabik)), Corpus.Byu.Edu (https://www.english-corpora.org/), Wortschatz (https://wortschatz.uni-leipzig.de/de) и параллельными корпусами InterCorp (https://treq.korpus.cz/), созданные нами корпуса использовались для

наполнения семантического поля «империя» для английского, русского и чешского языков.

5-7. Мы провели корпусное исследование парадигматических и синтагматических связей слова «империя» в русском, английском и чешском языках; на основе дистрибутивно-статистических методов сформировали лексическое наполнение собственно семантического поля «империя» для трех языков, а также провели исследование парадигматики и синтагматики лексических единиц, вошедших в поле «империя».

Для построения семантического поля «империя» мы отобрали толкования слов «империя» (empire, císařství, impérium, říše) из разных лингвистических и энциклопедических словарей, указанных выше. Методом семантического развертывания мы отобрали по 12 ключевых терминов, представляющих собой так называемое «предъядро».

Так, в предъядро семантического поля «империя» для русского языка вошли следующие лексические единицы: владение (владения), власть, государство, государь, держава, династия, император, империя, колония, монарх, монархия, правление. Предъядро семантического поля «империя» (empire) для английского языка: authority, colony, dynasty, emperor, empire, king, monarch, monarchy, power, rule, sovereign, state. Предъядро семантического поля «империя» (říše) для чешского языка: absolutismus, císař, císařství, dynastie, imperium, král, mocnářství, monarchie, panovník, říše, stát, vládce.

Затем на основе разработанных нами корпусов и инструмента «Thesaurus» в системе Sketch Engine были получены дистрибутивные минитезаурусы (по 20 слов) для каждого термина предъядра для русского, английского и чешского языков. Пересечение этих тезаурусов позволило выявить термины, наиболее часто встречающиеся во всех 12 минитезаурусах (статистический порог устанавливался эмпирически), которые вместе с исходными ключевыми терминами предъядра составили ядро семантического поля для каждого из языков. Подобная работа была также проведена с корпусами Aranea проекта Semantic similarity of words и корпусами проекта Wortschatz, в результате чего ядро семантического поля для каждого языка было еще раз расширено.

Наконец, с помощью инструмента «Keyword/Terms» по корпусу каждого языка был получен список из 100 ключевых терминов: в результате анализа этого списка к ядру семантического поля каждого из языков были добавлены недостающие элементы.

Таким образом, в состав ядра семантического поля «империя» для русского вошли 45 терминов: абсолютизм, боярин, вельможа, владение, власть, герцог, государственность, государство, государыня, государь, дворянин, дворянство, держава, династия, знать, император, императрица, империя, князь, колония, королевство, король, монарх, монархия, наместник, народ, нация, общество, папа, папство, подданный, правитель, правительство, правление, престол, престолонаследие, республика, самодержавие, самодержец, сословие, страна, трон, царь, церковь, элита.

Ядро семантического поля «империя» (empire) английского языка включает в себя 42 термина: administration, army, authority, church, colony, conquest, constitution, council, country, court, crown, dominion, dynasty, emperor, empire, family, fleet, garrison, government, king, kingdom, law, lord, majesty, minister, monarch, monarchy, nation, people, policy, power, prince, province, queen, reign, republic, rule, ruler, sovereign, state, territory, throne.

Ядро семантического поля «империя» (říše) для чешского языка насчитывает 46 терминов: absolutismus, arcikníže, arcivévoda, armáda, biskup, církev, císař, císařovna, císařství, država, dynastie, hrabě, impérium, kníže, knížectví, korunovace, král, království, kurfiřt, léno, magnát, markrabě, místodržící, místodržitel, moc, mocnář, mocnářství, monarcha , monarchie, národ, nevolnictví, panovník, papež, papežství, princ, říše, šlechta, stát, trůn, území, velkokníže, velkovévoda, velmož, vévoda, vláda, vládce.

Парадигматические связи между терминами ядра были выявлены с помощью дистрибутивно-статистического метода, реализованного в рамках инструмента «Thesaurus» в системе Sketch Engine: для каждого термина был получен минитезаурус из 20 связанных с ним слов, с указанием силы связи (коэффициент семантической близости logDice). Так, например, выглядит минитезаурус, отражающий парадигматические отношения для термина государство и ранжированной по значению logDice (приведено в скобках):

империя (0.410), страна (0.379), общество (0.375), Россия (0.364), власть (0.338), народ (0.337), церковь (0.305), мир (0.301), правительство (0.297), человек (0.296), система (0.291), город (0.285), союз (0.276), монархия (0.275), Германия (0.274), сила (0.272), политика (0.271), право (0.268), Европа (0.266), жизнь (0.266).

Синтагматические отношения для каждого термина были получены в форме коллокаций с помощью инструмента WordSketch в Sketch Engine. Сила связи между коллокатами высчитывалась с помощью меры logDice. Приводим пример списка коллокаций для чешского слова armáda (указывается коллокат, значение меры logDice и полная форма коллокации):

- sloužit (11.35, sloužit v armádě)
- velitel (10.78, velitel armády)
- rudý (10.38, Rudá armáda)
- císařský (9.94, císařské armády)
- velení (9.48, velení armády)
- ruský (9.42, ruské armády)
- důstojník (9.37, důstojník československé armády)
- čelo (9.35, v čele armády)
- pruský (9.09, pruské armády)
- rakouský (9.06, rakouské armády)
- francouzský (9.05, francouzské armády)
- voják (9, vojáků Rudé armády)
- spása (8.9, Armáda spásy)
- budovat (8.87, budovat armádu)
- švédský (8.82, švédská armáda)
- porážka (8.69, porážce osmanské armády)

- zbytek (8.68, zbytek armády)
- německý (8.65, německá armáda)
- generál (8.65, generál ruské armády)
- štáb (8.65, štábu Československé armády)
- rakousko-uherský (8.63, rakousko-uherské armády)
- jednotka (8.62, jednotky osmanské armády)
- silný (8.59, silnou armádu)
- reforma (8.58, reforma armády)
- liga (8.52, armádou Katolické ligy)
- síla (8.36, síly pruské armády)
- postup (8.34, postupu Rudé armády)
- typ (8.32, armáda nového typu)
- sovětský (8.3, sovětská armáda)
- československý (8.27, československá armáda)
- žoldnéř (8.18, armádu žoldnéřů)
- reorganizace (8.15, reorganizaci armády)
- turecký (8.14, turecká armáda)
- operace (8.13, operace Rudé armády)
- Napoleonův (8.05, Napoleonova armáda)
- arcivévoda (8.02, armádě arcivévody Leopolda Viléma)
- císařství (7.99, Armáda rakouského císařství)
- maršál (7.97, armádou maršála Tillyho)
- Rakouska-Uhersk (7.96, armády Rakouska-Uherska)
- britský (7.94, britské armády)

- sultán (7.69, armádou sultána Sulejmana I)
- kancléř (7.67, armádou švédského kancléře Axela Oxenstierna)
- vzor (7.67, vytvoření armády nového vzoru)
- klon (7.5, armádu klonů)
- car (7.48, ruským armádám cara Alexandra)
- velký (6.85, velká armáda)

Анализ лексики 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 в рамках существующей номенклатуры отношений следует отнести к отношению «ассоциация».

На основе полученных минитезаурусов для терминов ядра для каждого языка были сформированы так называемая «средняя зона» и периферия семантического поля. В среднюю зону вошли термины, которые встречают в двух и более минитезаурусах, тогда как термины, встречающиеся лишь в одном минитезаурусе, были отнесены к периферии.

Таким образом, с помощью корпусных методов было сформировано семантического поле «империя» для русского, английского и чешского языков, а также проанализированы парадигматические и синтагматические связи между элементами сформированных семантических полей. Полный список минитезаурусов и коллокаций представлен в созданных нами электронных тезаурусах, а также в Приложении [Файл 5, эскиз-макет словаря].

8. Сопоставление смыслового наполнения поля «империя» для русского, английского и чешского языков с данными лексикографического исследования позволяет сделать вывод о том, что выделенные нами элементы с помощью инструмента Sketch Engine в достаточно полной мере описывают семантическое поле «империя» для трех языков и позволяют дополнить данные, которые могут быть получены из других источников. В

частности, в то время как тезаурус WordNet выделяет многие лексемы из полученного нами списка, отражение в тезаурусе находит лишь часть выделенных нами элементов.

Уже выделенные три ядра ЛСП «империя» в русском, английском и чешском языках, визуализируют разницу в языковых картинах мира. Хотя обозначенные поля содержат в значительной мере эквивалентную лексику, их наполнение концептуально неравнозначно. В ходе анализа нами были обнаружены следующие закономерности:

1. Все три ядра включают в себя слова-синонимы и терминологические дублеты (císařství и impérium в чешском языке; царь, монарх и император в русском языке; king, emperor и monarch – в английском). При этом в чешском языке выстраивается синонимический ряд из трех лексических единиц (císařství – impérium – říše).
2. В русском языке лексические единицы, обозначающие персоналий, не обязательно семантически связаны с концептом «власти». Так, боярин, вельможа, герцог, дворянин, знать, элита – это существительные, обозначающие высокий статус либо лиц, которые обладают наследуемыми привилегиями. Для сравнения, в английском языке в СП сделан акцент на системности власти и на ее милитаристической природе (administration, army, fleet, garrison и др.). По данному параметру русское СП имеет больше сходства с чешским СП (см. п. 12 настоящего отчета).
3. В продолжение п. 2 в русском и чешском СП империя значительное количество лексических единиц связано с персональной властью (самодержец, самодержавие, абсолютизм и др.), тогда как в СП империя в английском языке персональное начало (king, emperor) уравновешивается имперсональным (council, law).
4. В ядре СП империя в русском языке фактически отсутствуют компоненты, связанные с национальной религией и ее представителями (существительные «папа» и «папизм» отсылают к реалиям католической церкви), в то время как в английском и чешском СП зафиксированы слова church, biskup, církev.

5. В ядре СП в английском и в чешском языках присутствует существительное со значением «территория» (territory, území), которого нет в русском сегменте ядра СП. Параллельно в английский и чешский тезаурусы интегрированы компоненты, отсылающие к травматическим историческим событиям, влияющим на формирование картины мира и на концептосферу носителя языка: nevolnictví (крепостное право), dominion, colony, conquest. В представленном русскоязычном СП империя отсутствует упоминание о крепостном праве (исключен целый пласт значений, начиная с понятия «территория»). В русскоязычном корпусе понятие колониализма, сопутствующее ему стремление к расширению границ, а также осмысление этого процесса нашли отражение в текстах XIX в. (в особенности, в текстах николаевской эпохи). Эти значения остаются на периферии современного СП.

Таким образом, разработанная нами технология наполнения семантического поля может быть использована для расширения состава существующих лексикографических источников в области семантики.

9. Квантитативное исследование поведения элементов семантического поля в диахронии было проведено на материале английского и русского корпусов текстов по тематике «империя», созданных нами ранее (русскоязычный корпус был создан совместно с М. В. Хохловой). Английский и русский корпуса текстов были разделены на 4 подкорпуса, соответствующие разным временным промежуткам, а именно XVIII в., первая половина XIX в., вторая половина XIX в. и XX в. Результаты этого исследования приводятся в статье «Corpus Methods and Semantic Fields: the Concept of Empire in English, Russian and Czech» (Zakharov et al. 2020). В частности, было сделано следующее наблюдение: поле «empire» для английского языка содержит большее количество лексем, обозначающих военные действия (army, conquest, enemy, troop, war, и т. д.), а также элементов, связанных с различными атрибутами государственности и власти (authority, court, crown, law, и т. д.). Что касается русского языка, в XVIII в. на протяжении всего столетия активно идет процесс формирования СП империя. Так, например, словосочетание «Российская Империя» вытесняет

использовавшиеся ранее синонимы «Российское царство» и «Московия» (взгляд на Россию извне). Тексты XIX в. характеризуются разветвлением системы номинации (одновременное функционирование слов царь, император, монарх, самодержец, кесарь и др.) – тенденция, которая нашла отражение в актуальном СП. Для чешского языка в силу недостаточного количества доступных исторических текстов для создания корпуса диахроническое исследование не было проведено. Однако сравнение семантических полей для русского и чешского языков позволило сделать интересное наблюдение о существовании в этих языках двух микрополей: одно из них связано с понятием «империя» (элементы «империя», «царство», «монархия» в русском языке и элементы «imperium», «císařství», «říše», «mocnářství» и т. д. в чешском языке); другое связано с понятием «император» (элементы «монарх», «правитель», «царь», «владыка» и т. д. в русском и элементы «panovník», «císař», «král», «vládce» и т. д. в чешском языке).

10–11. Одним из важнейших результатов нашего исследования стала разработка программно-лингвистического обеспечения для сравнения состава семантического поля «империя» и статистической характеристики элементов поля в трех языках. Данное программное обеспечение было реализовано в форме базы данных MySQL, с помощью которой осуществляется ввод и вывод лексико-семантической информации об элементах семантического поля «империя», включая статистические показатели (абсолютную и относительную частоту слова в корпусе, значение коэффициента семантического сходства logDice в дистрибутивных тезаурусах и коллокациях), примеры употребления, переводные эквиваленты, и т. д., что позволяет проводить сравнение состава семантических полей для трех языков. Более подробная информация о данном программном обеспечении указана ниже, в описании инструментов создания и ведения электронных тезаурусов для семантического поля «империя» в русском, английском и чешском языках.

12. В ходе исследования был проведен ассоциативный анализ поля «империя» в русском языке. Во-первых, были обследованы ассоциативные словари русского языка, в которых осуществлялся прямой и обратный поиск

на слова-стимулы, входящие в предъядро. Наибольшее количество реакций вызвали следующие слова-ассоциаты: власть, государство, император, правление. Так как существующие на сегодняшний день ассоциативные словари были опубликованы более 20 лет назад, имеется необходимость в более актуальных данных. С этой целью нами был проведен направленный ассоциативный эксперимент. В основу лег один из подкорпусов, сформированный на материале текстов статей, опубликованных в российском научном журнале «Ab Imperio». Издание посвящено вопросам «развития новой имперской истории, а также междисциплинарному и компаративистскому изучению истории и теории нации и национальных движений на постсоветском пространстве» [Ab Imperio. Основные принципы]. Журнал состоит из нескольких рубрик: вступительная заметка от редакции, разделы «История», «Архив», «Историография», «Методология и теория», «Новейшие мифологии», «Социология, этнология, политология», «ABC: Исследования империи и национализма», «Рецензии». Особенностью «Ab Imperio» является мультилингвизм – материалы принимаются на 5 языках (русском, английском, немецком, французском, украинском), однако публикуются только на русском и английском языках. При составлении корпуса было принято решение отказаться от влючения в него рецензий и архивных материалов и сосредоточиться на концептуальных текстах 2000-2019 гг. Корпус включает две части - тексты отечественных ученых (491) и переводы либо оригинальные тексы иностранных ученых (196): 2136710 и 727720 слов, соответственно. Сводный тезаурус этих двух корпусов выявил расхождение с основным предъядром: империя, Россия, государство, страна, общество, регион, культура, народ, власть, история, мир, нация. Эти слова стали стимулами в ассоциативном эксперименте. Кроме того, мы добавили однокоренные слова: император, императрица, империалист, империализм, имперскость, имперский (-ая,-ое,-ие), императорский (-ая,-ое,-ие) и империалистический (-ая,-ое,-ие). Опрос был расчитан на специалистов гуманитарного профиля. Мы обследовали ответы 44 респондентов (данные продолжают поступать) и сформировали промежуточное ассоциативное поле «империя», ядро которого представлено в приложении [Файл 6]. Что касается центрального концепта «империя», то он подтвердил определение, представленное в «Русском семантическом словаре» - «монархическое

государство во главе с императором». Как и в итоговом тезаурусе проекта, центральными реакциями являются «страна», «власть», «Россия», «держава». При этом концепты «народ», «церковь», «система», «политика» и др. оказались на периферии ассоциативного поля. Отсутствуют значения «распада» и «падения» империи, практические не встречается реакция «колониальный» / «колониалистический». В то же время актуализируются концепты «силы» и «власти». Для АП «империя» также характерно обилие прецедентных феноменов - имен правителей и названий государств (Петр I, Цезарь, Римская империя, Российская империя). Таким образом, эксперимент подтвердил данные тезауруса. С другой стороны, и в тезаурусе, и в эксперименте фактически зафиксировано отсутствие значения империи как крупной монополии (например, газетная империя) и как сферы, где господствует или царит что-либо (империя чувств, империя лжи).

Наиболее значимыми для ЛСП «империя» по результатам эксперимента стали концепты «власть» (появляется в качестве реакций в микрополях «империя», «государство», «император», «императрица», «имперскость») и «государство» (появляется в качестве реакций в микрополях «империя», «Россия», «страна», «власть», «имперское», «империалистическое»). Эксперимент также продемонстрировал 1) значительное количество реакций, которые имеют значение регалий и «предметов быта» (скипетр и держава, трон, корона, жезл; фарфор, платье, конюшни), 2) наличие стереотипизированных ассоциатов, которые вошли в язык в советский период (имперские замашки, буржуй и др.). Основные результаты ассоциативного эксперимента проиллюстрированы в приложении (см. Файл 6, данные о ядре ассоциативного поля «империя» по данным эксперимента).

13–15. На основе полученных корпусных данных была сформирована база данных для создания и ведения электронных тезаурусов для русского, английского и чешского языков. Один из важнейших результатов нашего проекта – создание электронных тезаурусов для семантического поля «империя» для русского, английского и чешского языков и предоставление этих материалов в общедоступное пользование.

Электронные тезаурусы находятся в общем доступе на сайте https://imperium.wordform.ru и содержат полученную в ходе нашего исследования информацию для семантического поля «империя» в русском, английском и чешском языках. Этот ресурс с технологической точки зрения представляет собой сайт с реляционной базой данных MySQL, в которой хранится лексико-семантическая информация. Сайт проекта построен на языке программирования PHP и CMS WordPress. Сайт обеспечивает возможность ввода лингвистических данных по проекту, а также отображения данных по запросу пользователя. Так как проект международный, в качестве языка интерфейса выбран английский язык.

Ввод данных: изначально лингвисты вводят данные в программе Google Sheets. Эти данные экспортируются в файлы формата TSV (Tab-separated values), которые в свою очередь загружаются на сайт через специальные скрипты в таблицы MySQL.

Вывод данных: сайт содержит 3 основных раздела для работы с данными – вывод индекса слов; отображение детальной информации по выбранному слову; отображение статистической информации по проекту.

На главной странице сайта (индекс слов) отображаются слова, входящие в ядро и среднюю зону семантического поля «империя». Детальная информация по слову содержит следующие разделы:

- указание зоны семантического поля, в которую входит данное слово (1 – ядро, 2 – средняя зона);
- относительную частоту в соответствующем корпусе (ipm);
- дефиниции, связанные с понятием «империя»;
- примеры употреблений данного слова в корпусе;
- список коллокаций с указанием коллоката, силы связи (значение меры logDice) и полной формы коллокации;
- минитезаурус с указанием коэффициента семантического сходства и частоты в корпусе;

- перевод между английский, русским и чешскими языками с долевым показателем переводных эквивалентов (информация о переводных эквивалентах получена на основе параллельных корпусов InterCorp).

Статистическая раздел дает возможность оценить актуальный объем введенных данных.

16. Подготовлен к изданию эскиз-макет созданных тезаурусов для русского, английского и чешского языков в печатной форме (см. Файл 5, эскиз-макет словаря). Данный словарь-справочник включает в себя всю информацию, полученную нами в ходе исследования: в нем представлены слова-элементы ядра семантического поля «империя» для трех языков с указанием статистических показателей (абсолютная и относительная частота слов в корпусе), дефиниций, примеров употребления в корпусе, наиболее характерных коллокаций, переводных эквивалентов, а также дистрибутивных минитезаурусов.

Таким образом, в результате нашей работы был получен уникальный лексикографический продукт, который может служить основой для дальнейших исследований в области построения семантических полей автоматизированными методами.

17. Для использования результатов проекта в учебном процессе были подготовлены учебно-методические материалы. Был создан эскиз-макет учебно-методического пособия (см. Файл 4, эскиз-макет пособия), в котором отражены отдельные этапы наполнения семантического поля (выделение слов ядра с помощью построения сводного тезауруса) и формирования тезауруса (создание списка коллокаций для ключевого термина). Данное пособие состоит из двух частей: в первой, теоретической, части затрагиваются вопросы теории семантического поля; вторая, практическая, часть включает в себя инструкции по выполнению конкретных задач, связанных с выделением ключевых терминов для семантического поля и формированием тезауруса. Материал пособия может быть использован при разработке теоретических и специальных курсов по корпусной лингвистике, сравнительной и типологической лингвистике, теоретическим проблемам семантики, а также в работах студентов и аспирантов.

Список литературы см. в приложении (Файл8_Захаров ВП_Literature.pdf).

3.6. Сопоставление результатов, полученных при реализации с мировым уровнем

Корпусные методы и статистика в области семантики используются довольно давно. Частые упоминания подходов, использующих корпуса для исследования структуры языка (в частности, подходы, использующие информацию о частотности и коллокациях), можно найти в работах Gries (2006), Evert (2009), Gries & Divjak (2012), Glynn (2014). Существуют также системы, позволяющие получить информацию о текстуальных и системных связях внутри текста, вычислять семантические отношения и изучать семантические поля при помощи их визуализации (Sketch Engine, RusVectōrēs, Semantic similarity of words (Slovenský národný korpus)).

Использованная в ходе работы над проектом технология построения семантических полей объединяет в себе традиционные лексикографические и современные дистрибутивно-статистические методы для получения наиболее полного и объективного списка терминов поля. Принципиально новым результатом работы явилось одновременное построение семантического поля для трех языков и проведение лингвокультурологического анализа полученных результатов для концепта «империя».

Полученные в ходе выполнения проекта результаты в целом соответствуют мировому уровню. Результаты работы опубликованы в 2018, 2019 и 2020 гг. в восьми статьях, не менее 2 из них с индексацией в базах данных Web of Science и Scopus.

Список литературы см. в приложении (Файл8_Захаров ВП_Literature.pdf).

3.7. Методы и подходы, использованные при реализации Проекта (описать, уделив особое внимание степени оригинальности и новизны)

Методологической основой нашего проекта являлся корпусно-ориентированный анализ парадигматики и синтагматики лексических единиц с использованием дистрибутивно-статистических методов, учитывающий

семантические связи разного типа. Материал и инструмент исследования составили существующие и специально созданные корпусы с лингвистической разметкой, корпусные лингвистические процессоры, лексикографические ресурсы.

Основной материал исследования – это корпусы русского, английского и чешского языков общим объемом ок. 77 млн словоформ, созданные с помощью корпусного менеджера Sketch Engine. Кроме того, были использованы инструменты и готовые корпуса Aranea Corpora, (https://korpus.sk/semä), Corpus.Byu.Edu (https://www.english-corpora.org/), Wortschatz (https://wortschatz.uni-leipzig.de/de), а также параллельные корпуса InterCorp (https://treq.korpus.cz/). Уникальность нашего исследования заключается в том, что методы корпусного анализа данных были впервые реализованы на такой обширной эмпирической базе.
В основе компьютерных методов исследования лежит дистрибутивно-статистический анализ, при котором используется информация о дистрибуции элементов текста и их числовых параметрах. Стандартными процедурами компьютерного анализа текстов на сегодняшний день являются разнообразные количественные оценки связанности лексических единиц в контекстах на основе частотных характеристик их совместной встречаемости, что имеет прикладное значение для реализации алгоритмов автоматического выделения ключевых слов, коллокаций (многословных единиц) и парадигматических рядов на основе мер ассоциации. Принцип перехода от изучения текстуальных связей (синтагматических) к системным (парадигматическим) лежит в основе различных дистрибутивно-статистических методик, как классических (Шайкевич 1982, Фрумкина 1992), так и современных (Sahlgren 2006, Baroni et al. 2014, Mikolov et al. 2013). Считается, что два элемента связаны парадигматически, если оба они текстуально систематически связаны с какими-то третьими элементами. Значит, представляется разумным предположить, что сила парадигматической связи должна возрастать с увеличением числа и силы общих синтагматических связей (Шайкевич 1976). В ходе реализации проекта нам удалось доказать данную гипотезу, ранее это не представлялось возможным сделать в силу недостаточности корпусных данных и

недостаточной разработанности инструментов компьютерного анализа текстовой информации.

Чтобы можно было говорить о закономерности любых статистических распределений, нужны очень большие массивы данных. Таковые появились только с созданием больших корпусов текстов. Одновременно стали появляться и соответствующие программные средства. В данном проекте впервые для решения задачи моделирования семантических полей была использована система Sketch Engine (Kilgarriff et al. 2004), которая представляет собой корпусный менеджер, работающий с морфологически размеченным корпусом.

Sketch Engine в числе прочих функций выдает частотные списки лексических единиц, входящих в корпус. Эти частоты, безусловно, характеризуют лексический состав анализируемых корпусов. Был проведен контрастивный анализ, когда данные нашего корпуса сравниваются с нейтральным фоновым. При этом относительные частоты в текстах исследуемого корпуса должны существенно превосходить частоту этих слов в некотором фоновом неспециализированном корпусе. Подобная возможность в Sketch Engine имеется (comparable frequency list).

Главная ценность использования этой системы для целей проекта – это наличие в ней специальных средств, реализующих методику дистрибутивно-статистического анализа — «Тезаурус» (построение тезауруса, другими словами, лексико-семантического поля), «Кластеризация» (группировка единиц тезауруса в кластеры – лексико-семантические группы) и «Дифференциация» (выявление сходства и разницы в сочетаемости для пар слов) и «Лексические шаблоны» (выявление коллигаций – коллокаций в рамках синтаксических моделей). Все они, разными способами, выявляют парадигматические (т.е. семантические) связи между терминами с количественным указанием силы этой связи.

Тезаурус в системе Sketch Engine (или, как его можно охарактеризовать, дистрибутивный тезаурус) позволяет увидеть, какие слова имеют схожую дистрибуцию с заданным словом. Для вычисления парадигматического подобия слов были рассмотрены наборы сочетаемости для пар слов с учетом синтаксического отношения (word sketches, или «лексические портреты»).

Единицы семантического поля обладают общими синтагматическими и парадигматическими свойствами, что отражает их семантическую близость. Схожесть дистрибуции слов высчитывается статистически на основе меры ассоциации logDice (Rychlý 2008) и с учетом лексико-синтаксических шаблонов (Kilgarriff, Rychly 2007; Kilgarriff et al. 2014). Процедура кластеризации позволила построить тезаурус с кластерами лексических единиц, которые соответствуют лексико-семантическим группам.

Однако мы знаем, что в каждой предметной области большая часть терминов, как правило, представлена словосочетаниями. Корпусные инструменты предоставляют нам возможность автоматического выявления коллокаций. В Sketch Engine имеется два механизма выявления устойчивых сочетаний: «Коллокации» и «Лексические шаблоны». Первый вычисляет силу связи между словами по всему корпусу, второй – в пределах заданной синтаксической формулы (шаблона). Подсистема «Лексические шаблоны» выдает наиболее устойчивые сочетания, вычисленные в соответствии со статистической мерой logDice, и общее количество сочетаний в корпусе, соответствующих каждому отдельному шаблону. Поскольку мы строим поля (тезаурусы) для разных языков, была проведена разработка программного обеспечения для их сравнения. Решение данной задачи нами рассматривалось не с технической, а с лингвистической точки зрения, учитывая, что лексические единицы в тезаурусах представлены на разных языках.

В нашем проекте использовалась уникальная комбинация разных подходов к выявлению наполнения семантических полей, позволившая привлечь статистические данные, полученные на представительном корпусном материале, что обеспечивает полноту и достоверность полученных результатов.

Список литературы см. в приложении (Файл8_Захаров ВП_Literature.pdf).

3.9. Апробация результатов реализации Проекта на научных мероприятиях (участие в научных мероприятиях по тематике Проекта за период, на который был предоставлен грант) (каждое мероприятие с

новой строки, указать название мероприятия, ФИО члена коллектива и тип доклада)

1. XV Международная конференция по компьютерной и когнитивной лингвистике TEL Казань, 31 октября – 3 ноября 2018 г. Захаров В.П. Доклад на дистанционной сессии.

2. 5th International Multidisciplinary Scientific Conference on Social Sciences and Arts SGEM 2018. Sofia, 24 August – 2 September 2018. Захаров В.П. Устный доклад.

3. XLVIII Международная филологическая научная конференция. Санкт-Петербург, 18 – 27 марта 2019. Захаров В.П., Гвоздёва Е.С. Устный секционный доклад.

4. XXII Международная объединенная конференция «Интернет и современное общество (IMS-2019). Санкт-Петербург, 19 – 22 июня 2019 г. Захаров В.П. 2 устных секционных доклада.

5. 10th International Conference NLP, Corpus Linguistics, Language Dynamics and Change. Bratislava, 23 – 25 October 2019. Захаров В.П. Устный секционный доклад.

6. Пиотровские чтения 2019. Санкт-Петербург, 27 ноября 2019 г. Захаров В.П., Семёнова Н.В., Гвоздёва Е.С. Устный пленарный доклад.

7. The Thirteenth Workshop on Recent Advances in Slavonic Natural Language Processing. Brno, 6 – 8 December 2019. Захаров В.П. Устный секционный доклад.

8. Летняя школа «Введение в историю понятий» / «Introduction to Conceptual History», University of Helsinki, Finland, 12-22 августа 2019, приглашенная лекция «The Concept of Empire in Russian Language: Model of Semantic Field». Семенова Н.В. Устное выступление.

3.12. Адреса (полностью) ресурсов в Интернете, подготовленных Проекту (например, http://www.somewhere.ru/mypub.html)

https://www.imperium.wordform.ru/