Компьютерная визуализация русской языковой картины мира

П. Чжан, В П. Захаров

***Аннотация* — В статье рассматривается визуализация русской языковой картины мира. Языковая картина мира является представлением о мире, сложившимся в обыденном сознании данного языкового коллектива. Концепт представляет собой основную единицу языковой картины мира, которая может реализоваться в слове словосочетании высказывании дискурсе тексте и т.д. Семантические поля с помощью которых можно описывать лексическую систему языка являются отражениями концептов. Существует много способов представления семантических полей: сервис RusVectōrēs Облако тегов и т.д. Анализ словарных толкований слов, в которых отражается заданный концепт, является одной из основных методик представления концептов.**

В статье мы анализируем русскую языковую картину мира на примере трех концептов: «империя», «государство», «власть». Идентификаторы концепта мы отбирали из словарных статей слов "империя" "государство" и "власть" из лингвистических и энциклопедических словарей. В работе описываются разные способов визуализации семантических полей. Рассматриваются облако тегов как один из важных способов визуализации семантических полей слов, в которых отражаются концептов. Также в работе описывается инструмент «Тезаурус» в системе Sketch Engine.

***Ключевые слова* — визуализация знаний, концепт, облако тегов, языковая картина мира, Sketch Engine**

I. Введение

Один и тот же язык один и тот же общественно-исторический опыт формирует у членов определенного общества сходные представления о мире. Картина мира лежит в основе как индивидуального, так и общественного сознания. Исторически представление о картине мира восходит к идеям немецкого лингвиста и философа В. фон Гумбольдта и опирается на идеи И. Гердера о природе и происхождении языка о взаимосвязи языка мышления и «духа народа» а также на идеи Ф. Шлегеля о сущности нации. Тесная связь мышления и языка позволяет выделить ее особую разновидность картины мира – языковую картину мира (ЯКМ). Гумбольдт писал, что представления человека о мире зависят от того языка, которым он пользуется. У каждого языка есть своя внутренняя форма; языки отличаются друг от друга не только звуковыми материальными оболочками смыслов, но и самим способом восприятия мира что позволяет говорить о некой обобщенной национальной языковой картине мира. Задачей описания ЯКМ занимаются когнитивная лингвистика психолингвистика компьютерная лингвистика каждая с использованием своих методов.

II. Концепты и семантические поля

Основной единицей языковой картины мира является концепт. Концепт как ментальная сущность имеет национально-специфические черты, соотносимые с мировидением культурой обычаями верованиями и историей народа. В концептах аккумулируется как культурный уровень каждой языковой личности, так и всех носителей языка в целом.

Концепты как основные элементы языковой картины мира могут реализоваться в слове словосочетании высказывании дискурсе тексте и т.д. Они образуют внутреннюю основу семантических полей. «Власть» «государство» и «империя» являются важными концептами у многих народов, а в русской языковой картине мира они занимают особое место.

Семантические поля с помощью которых можно описывать лексическую систему языка фактически являются «овеществленными» отражениями концептов. Впервые термин "семантическое поле" был введен Г. Ипсеном [1]. Поле обычно рассматривается как совокупность языковых единиц, объединенных каким-то общим семантическим признаком имеющих некоторый общий компонент значения. Семантический признак, лежащий в основе семантического поля, может также рассматриваться как некоторая понятийная категория (А.В. Бондарко Л.М. Васильев И.М. Кобозева). В трактовке В.Г. Адмони [2] поле характеризуется наличием инвентаря элементов, связанных системными отношениями. По мнению В. Г. Адмони в поле можно выделить 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.

III. Методы выявления семантических полей

Элементы семантического поля связаны между собой системными (парадигматическими) отношениями.

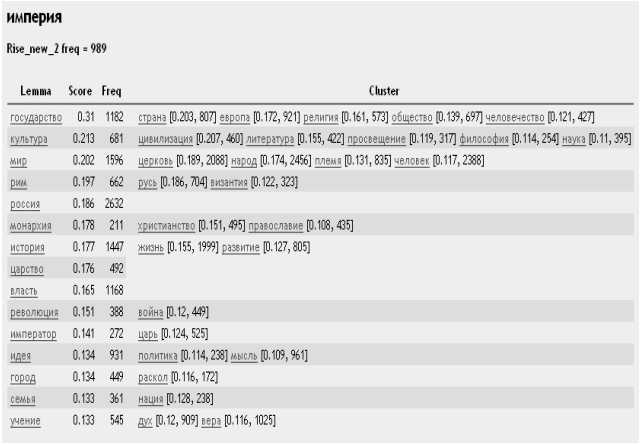
Семантический уровень языка представляет собой упорядоченную систему элементы которой находятся в отношениях взаимосвязи и взаимообусловленности. Имея в виду связи между значениями слов говорят о лексико-семантической системе языка или подъязыка. Элементами ее являются отобранные по определенным правилам лексические единицы естественного языка а структура изоморфна структуре логических связей между понятиями специальной области знаний и деятельности. Задача моделирования понятийной системы частично выходит за пределы лингвистики и является задачей описания и представления знаний. Ее можно разбить на две части: выявление системы понятий и выявление отношений между ними. Первая задача 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 материалов. Однако поскольку наши знания о мире так или иначе находят отражение в текстах то можно поставить задачу извлечения системы понятий из текстов. Характер связей на этом первом этапе автоматически не устанавливается. В данной работе это выявление множества основных взаимосвязанных понятий вокруг выбранного ядерного элемента (ключевого слова).

Одним из старых и известных методов лингвистического исследования является дистрибутивно-статистический анализ, при котором используется информация о дистрибуции элементов текста и их числовых параметрах.

Принцип перехода от изучения текстуальных связей (синтагматических) к системным (парадигматическим) лежит в основе различных дистрибутивно-статистических методик. Но чтобы можно было говорить о закономерности любых статистических распределений, нужны мощные вычислительные ресурсы и большие массивы текстовых данных. Таковые появились только с развитием вычислительной техники и созданием больших корпусов текстов.

Инструменты наполнения семантических полей встроены в некоторые корпусные менеджеры напр. в Sketch Engine, где исходным материалом является текстовый массив — корпус текстов. В системе Sketch Engine имеется сервис «Тезаурус» выявляющий 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 рассматриваются наборы сочетаемости для пар слов с учетом синтаксического отношения между ними. Единицы семантического поля обладают общими синтагматическими и парадигматическими свойствами что отражает их семантическую близость. Схожесть дистрибуции слов высчитывается статистически на основе меры ассоциации logDice и с учетом лексико-синтаксических шаблонов.

Ниже приведен пример автоматического построения дистрибутивного тезауруса по корпусу текстов «Понятие империи в русской культуре» для лексемы «империя» с автоматическим разбиением его на кластеры (рис. 1) [3]



империя

Rise_new_2 freq = 989

Lemma	Score	Freq	Cluster
государство	0.31	1182	страна [0.203, 807] европа [0.172, 921] религия [0.161, 573] общество [0.139, 697] человечество [0.121, 427]
культура	0.213	681	цивилизация [0.207, 460] литература [0.155, 422] просвещение [0.119, 317] философия [0.114, 254] наука [0.11, 395]
мир	0.202	1596	церковь [0.189, 2088] народ [0.174, 2456] племя [0.131, 835] человек [0.117, 2388]
рим	0.197	662	русь [0.186, 704] византия [0.122, 323]
россия	0.186	2632	
монархия	0.178	211	христианство [0.151, 495] православие [0.108, 435]
история	0.177	1447	жизнь [0.155, 1999] развитие [0.127, 805]
царство	0.176	492	
власть	0.165	1168	
революция	0.151	388	война [0.12, 449]
император	0.141	272	царь [0.124, 525]
идея	0.134	931	политика [0.114, 238] мысль [0.109, 961]
город	0.134	449	раскол [0.116, 172]
семья	0.133	361	нация [0.128, 238]
учение	0.133	545	дух [0.12, 909] вера [0.116, 1025]

Рис. 1. Гнездо тезауруса с выделенными кластерами для ключевого слова «империя»

В первом столбце приведены лексемы во втором — значение силы семантической связи данного слова с заглавным вычисляемое на основе статистической меры logDice в третьем — абсолютная частота лексемы в корпусе в четвертом — лексемы образующие с лексемой из первого столбца единый кластер (в квадратных скобках указаны значение меры и частота им соответствующие).

IV. Выявление семантических полей лексикографическими методами

Другой подход, как уже говорилось, базируется на использовании имеющихся словарей стандартов нормативно-технической документации и т.п.

З.Д. Попова и И.А. Стернин в работе [4] отметили одну из основных методик описания концептов — анализ словарных толкований ключевого слова (имя заданного концепта) по возможно большему числу словарей, когда из толкований слов делается выборка всех возможных характеристик концепта.

В данном исследовании мы попытались смоделировать семантические поля «власть», «государство» и «империя», используя словарные толкования этих слов из разных словарей (разд. 7 8).

V. Способы визуализации представления семантических полей

После выявления основных лексических единиц составляющих наполнение того или иного концепта (на самом деле это всегда не слова а отдельные значения) встает задача представить их в виде системы т. е. показать отношения между ними силу связи и возможно

использовать в речи (в текстах).

Одним из известных проектов для русского языка является сервис RusVectōrēs который «вычисляет семантические отношения между словами русского языка и позволяет скачать предобученные дистрибутивно-семантические модели (word embeddings)» [5]. Сервис строит лексические векторы представляющие значение слова автоматически извлеченное из статистики совместной встречаемости через связи с другими словами (рис. 2).

Рис. 2. Список из 10 ближайших семантических ассоциатов (квази-синонимов) слова «империя» сделанный при помощи сервиса RusVectōrēs

Одним из способов формального представления семантических полей является визуализация. Существует много способов компьютерной визуализации: облако тегов графы видеографы и т.д. Они основываются на списке выявленных лексических единиц и на числовых параметрах характеризующих функционирование этих единиц в языке.

В проекте Semantic similarity of words Института языкознания им. Л. Штура Словацкой академии наук (https://korpus.sk/semä/) (разработчик Радован Гарабик (Radovan Garabik)) сформированное автоматически семантическое поле может быть представлено даже в трехмерном пространстве (рис. 3)

.

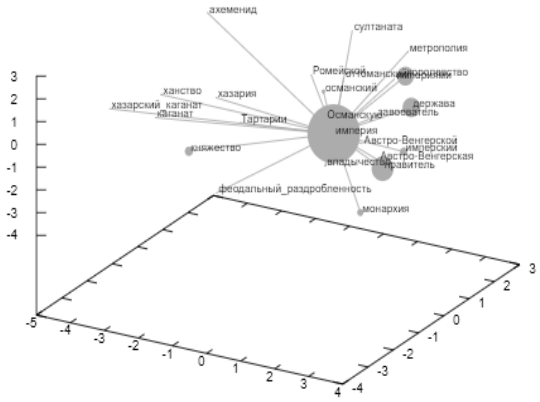


Рис. 3. Облако тегов для ключевого слова «империя» в проекте Semantic similarity of words Roman numerals.

VI. «Тезаурус» в системе Sketch Engine как инструмент построения облака тегов

Сервис «Тезаурус» в системе Sketch Engine как мы уже отмечали может показывать какие лексические единицы в корпусе имеют схожую дистрибуцию с заданным нами словом.

В результате применения функции «Тезаурус» мы сможем получить, с одной стороны, сводную таблицу с данными дистрибутивного тезауруса (рис. 1), с другой стороны, «облако тегов», включающее единицы дистрибутивного тезауруса (рис. 4). Чем крупнее шрифт включенного в облако тегов слова, тем выше значение его статистической меры, отражающей степень семантической близости данного слова к ключевому слову.

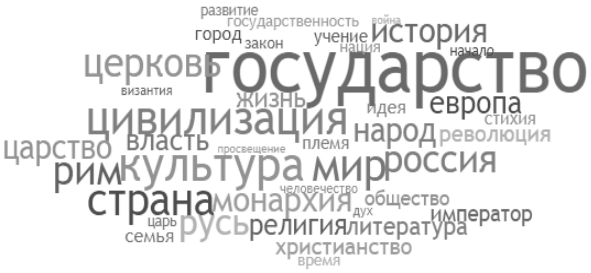


Рис. 4. Облако тегов для ключевого слова «империя» в системе Sketch Engine

VII. Компьютерная визуализация концептов «империя», «государство», «власть» на базе словарных толкований.

Для изучения концептов «империя» «государство» «власть» мы отобрали толкования слов «империя», «государство» и «власть» из разных лингвистических и энциклопедических словарей (см. Приложение).

Далее на базе отобранных словарных определений

для слов «империя», «государство» и «власть» в системе Sketch Engine были построены три корпуса соответственно для слов «империя», «государство» и «власть». На основе этих корпусов с помощью инструмента «Тезаурус» нами были представлены облака тега для указанных ключевых слов (рис. 5, 6,7).

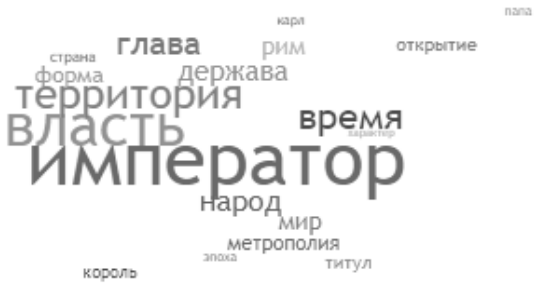


Рис. 5. Облако тегов для слова «империя» на базе корпуса его словарных толкований

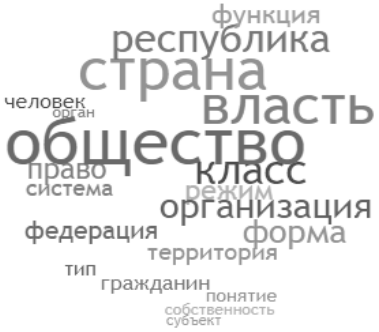


Рис. 6. Облако тегов для слова «государство» на базе корпуса его словарных толкований

Рис. 7. Облако тегов для слова «власть» на базе корпуса его словарных толкований

Облако тегов из всего набора лексических единиц, входящие в семантическое поле заданного ключевого слова, показывает слова, наиболее близкие к ключевому слову. Например, в семантическое поле слова «империя» (дистрибутивный тезаурус) на базе корпуса его словарных толкований в системе Sketch Engine входят следующие семантически близкие слова: *император, власть, территория, время, держава, глава, народ, мир, форма, страна, метрополия, Рим, титул, король* и др. Как мы видим, на облаке тегов «империя» (рис.4), здесь визуалирована лишь часть из них, при этом размер шрифта слов показывает, что семантически наиболее близкие слова по отношению к ключевому слову «империя» слова – это *император, власть, территория, время, держава, глава, народ*. Среди них на первом месте стоит *император* как самое близкое по значению.

VIII. Заключение

Мы проанализировали фрагмент русской языковой картины мира на примере трех концептов: «империя», «государство», «власть». Существуют разные способы выявления лексики семантических полей, как ручным, так и автоматизированным способом. Важный аспект представления семантических полей - это вычисление силы связи между его элементами. На основе этой характеристики строятся различные визуальные представления семантических полей. На основе корпуса словарных статей из лингвистических и энциклопедических словарей нами были построены облака тегов для анализируемых концептов.

Облако тегов является одним из популярных и удобных способов компьютерной визуализации. В рамках данной работы нами было рассмотрено применение этого способа для визуализации указанных концептов в русской языковой картине мира посредством сервиса «Тезаурус» в системе Sketch Engine. Построенные диаграммы наглядно отражают парадигматические связи разных единиц семантического полей, отражающих указанные концепты.

Исследование поддержано Российским фондом фундаментальных исследований, проект № 18-012-00474.

Библиография

[1] Ipsen G. The Ancient Orient and Indogermans Feast Scipts for W. Streitburg. Heidelberg 1924. p. 30-45.
[2] Адмони В. Г. Синтаксис современного немецкого языка: Система отношений и система построения. Л.: Наука 1973.
[3] Захаров В.П. Функциональность инструментов корпусной лингвистики // Труды международной конференции по компьютерной и когнитивной лингвистике TEL-2018. – В 2-х томах. Т 2. Казань: Изд-во Академии наук РТ 2018. С. 164 – 180.
[4] Попова З.Д. Стернин И.А. Очерки по когнитивной лингвистике. Воронеж: Изд-во Воронежского ун-та 2001.
[5] Kutuzov A. Kuzmenko E. WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models. In: Ignatov D. et al. (eds) Analysis of Images Social Networks and Texts. AIST 2016. Communications in Computer and Information Science vol 661. Springer Cham. 2017.

Приложение

Список словарных источников для дефиниционного анализа концептов «империя», «государство», «власть»

[1] Большая российская энциклопедия: [в 35 т.] / гл. ред. Ю. С. Осипов. М.: Большая российская энциклопедия, 2004—2017.
[2] Большая советская энциклопедия: [в 30 т.] / гл. ред. А. М. Прохоров. 3-е изд. М.: Советская энциклопедия, 1969—1978.
[3] Большой академический словарь русского языка / Рос. акад. наук, Ин-т лингвист. исслед.; гл. ред. А. С. Герд. М.; СПб.: Наука, 2004
[4] Большой толковый словарь русского языка / Под ред. С. А. Кузнецова. СПб.: Норинт, 1998.
[5] Большой энциклопедический словарь / Гл. ред. А. М. Прохоров. 1-е изд. М.: Большая российская энциклопедия, 1991.
[6] Борисов А. Б. Большой экономический словарь. М.: Книжный мир, 2006.
[7] Вербицкий А. А. Энциклопедический словарь по психологии и педагогике [Online]. Available: https://psychology_pedagogy.academic.ru/
[8] Википедия [Online]. Available: https://ru.wikipedia.org
[9] Даль В. И. Толковый словарь живого великорусского языка: В 4 ч. / Под ред. И. А. Бодуэна де Куртенэ. СПб.: Товарищество М. О. Вольфа, 1903—1909.
[10] Джери Д., Джери Дж. Большой толковый социологический словарь. В 2-х томах: Пер. с англ. Н.Н. Марчук. М.: Вече, АСТ, 1999
[11] Ефремова Т. Ф. Современный толковый словарь русского языка: В 3 т. М.: АСТ, Астрель, Харвест, 2006.
[12] Ильин И. П. Постмодернизм. Словарь терминов. М.: ИНИОН РАН— INTRADA, 2001.
[13] Исторический словарь [Online]. Available: https://dic.academic.ru/contents.nsf/hist_dic/
[14] Культурология. XX век. Энциклопедия: В 2 т. /гл. ред. С. Я. Левит. СПб.: Университет. кн., 1998.
[15] Мокиенко В. М., Никитина Т.Г. Толковый словарь языка Совдепии. СПб.: Фолио-Пресс, 1998.
[16] Музрукова Т. Г., Нечаева И. В. Популярный словарь иностранных слов. М.: Азбуковник,2002.
[17] Новая философская энциклопедия в 4-х томах/ Научно.-ред. совет: В. С. Стёпин, А. А. Гусейнов, Г. Ю. Семигин, А. П. Огурцов. М.: Мысль, 2000.
[18] Ожегов С. И., Шведова Н. Ю. Толковый словарь русского языка, М.: Издательство: ИТИ Технологии, 2008.
[19] Петрова Ф.Н. Словарь иностранных слов. М.: Сирин 1996.
[20] Православная энциклопедия [Online]. Available: http://www.pravenc.ru/
[21] Российская социологическая энциклопедия / под общ. ред. Г. В. Осипова. М.: Издат. группа НОРМА — ИНФРА-М, 1998.
[22] Российская юридическая энциклопедия / Гл. ред. А. Я. Сухарев. М.: ИНФРА-М, 1999.
[23] Сводная энциклопедия афоризмов [Online]. Available: http://endic.ru/aphorism/
[24] Словарь русского языка: В 4-х т. / РАН, Ин-т лингвистич. исследований; Под ред. А. П. Евгеньевой. — 4-е изд., стер. М.: Рус. яз.; Полиграфресурсы, 1999.
[25] Толковый словарь русского языка / Под ред. Д. В. Дмитриева. М.: Астрель: АСТ, 2003.
[26] Толковый словарь русского языка: В 4 т. / Под ред. Д. Н. Ушакова. М.: Сов. энцикл.: ОГИЗ, 1935—1940.
[27] Толковый словарь русского языка начала XXI века. Актуальная лексика / Под ред. Г.Н.Скляревской. / Ваулина, Е.Ю.; Скляревская, Г.Н.; Ткачева, И.О.; Фивейская, Е.А. ЭКСМО, 2006.
[28] Философия: Энциклопедический словарь/ под ред. А.А. Ивина. М.: Гардарики. 2004.
[29] Фритц Р., Герхард М. Библейская энциклопедия Брокгауза, Издатель О. С. Костюков, 2012
[30] Халтагарова О. Д. Краткий психологический словарь: Учебное пособие. Улан-Удэ: Изд-во ВСГТУ, 2006.
[31] Червонюк В. И., Иванец Г. И., Калинский И. В., Конституционное право России: энциклопедический словарь. М.: Юридическая литература, 2002.
[32] Энциклопедия «Русская цивилизация» /Отв. ред. О. А. Платонов. М.: Институт русской цивилизации, 2011.

Статья получена 12 ноября 2019.
Чжан Пэйлинь, Санкт-Петербургский Государственный Университет, аспирант (e-mail: zhangpl@yandex.ru)
Захаров Виктор Павлович, Санкт-Петербургский осударственный Университет, канд. филог. наук, доцент (e-mail: v.zakhrov@spbu.ru)

Computerized visualization of the Russian language picture of the world

P. Zhang, V. P. Zakharov

Abstract — **The article deals with the visualization of the Russian language picture of the world. The linguistic picture of the world is an idea of the world that has developed in the ordinary consciousness of a given linguistic group. The concept is the basic element of the linguistic picture of the world, which can be realized in a word, phrase, statement, discourse, text, etc. Semantic fields, with which you can describe the lexical system of the language, are reflections of concepts. There are many ways to represent semantic fields: the RusVectōrēs service, the word cloud, etc. The analysis of vocabulary interpretations of words in which a given concept is reflected is one of the main methods of concept presentation.**

In the article, we analyze the Russian language picture of the world on the example of three concepts: empire, state, power, using dictionary entries of the words empire, state and power from different linguistic and encyclopedic dictionaries. The paper describes various ways of visualizing semantic fields. Tag clould is considered as one of the important ways to visualize semantic word fields, which reflect concepts. In this study we also describe the thesaurus in Sketch Engine.

Keywords — **concept, knowledge visualization, language picture of the world, Sketch Engine, tag cloud**

References

[1] Ipsen G. The Ancient Orient and Indogermans. Feast Scipts for W. Streitburg. Heidelberg, 1924. pp. 30-45.

[2] Admoni V.G. The syntax of modern German: The system of relations and the system of construction [Sintaksis sovremennogo Leningrad: Nauka, 1973.

[3] Zakharov V.P. Functionality of corpus linguistics tools [Funktsional'nost' instrumentov korpusnoy lingvistiki]. In *Proceedings of the international conference on computer and cognitive linguistics TEL-2018* [Trudy mezhdunarodnoy konferentsii po komp'yuternoy i kognitivnoy lingvistike TEL-2018]. Vol. 2. Kazan: Izd-vo Akademii nauk RT, 2018, pp. 164-180.

[4] Popova Z.D. Sternin I.A. Essays on cognitive linguistics [Ocherki po kognitivnoy lingvistike]. Voronezh: Izd-vo Voronezhskogo un-taб 1980.

[5] Kutuzov A. Kuzmenko E. WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models. In: Ignatov D. et al. (eds) Analysis of Images Social Networks and Texts. AIST 2016. Communications in Computer and Information Science vol 661. Springer Cham. 2017.