УДК 004.8 **DOI** 10.17150/2500-2759.2021.31(3).382-390

ИССЛЕДОВАНИЕ ТОПОНИМОВ ИРКУТСКОЙ ОБЛАСТИ С ПРИМЕНЕНИЕМ МЕТОДОВ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

А.В. Боровский, Е.Е. Раковская
Байкальский государственный университет, г. Иркутск, Российская Федерация

Информация о статье

Дата поступления
29 июля 2021 г.

Дата принятия к печати
30 сентября 2021 г.

Дата онлайн-размещения
22 октября 2021 г.

Ключевые слова

Искусственный интеллект; обработка естественного языка; происхождение топонимов; эмбеддинговые модели; дистрибутивная семантика; векторное представление слов; метод дихотомии; метод трансформации

Аннотация

Актуальные проблемы топонимики подразумевают исследование отдельных слов с целью восстановления утраченного в современном языке понятийного значения географических названий, выяснения того, как в них отразились характерные особенности рельефа местности, род деятельности населяющих ее людей и т.п. Разрешение такого рода проблем возможно с помощью применения интеллектуальных методов анализа данных на основе информационных технологий, но в научных работах по топонимике таким методам практически не уделяется внимания. Статья посвящена исследованию происхождения и смысловых значений географических названий путем нахождения их семантических ассоциатов и вычисления семантического сходства слов с применением эмбеддинговой модели. По предлагаемому методу было определено происхождение некоторых топонимов Иркутской области, выявлены их семантические отношения. К топонимам, имеющим в своем составе два корня, был применен метод дихотомии, который улучшает работу модели за счет уточнения морфемного состава исходного слова. Для определения этимологии топонима «Москва» был применен метод трансформации слов, получены новые версии происхождения топонима. Показано, что использование методов на основе дистрибутивной семантики и векторного представления слов, полученного на основе больших массивов текстовых данных, значительно расширяет возможности исследований в области определения происхождения топонимов, уточнения их смысла.

RESEARCH OF TOPONYMS OF THE IRKUTSK REGION USING THE METHOD OF ARTIFICIAL INTELLIGENCE

Andrei V. Borovsky, Elena E. Rakovskaya
Baikal State University, Irkutsk, the Russian Federation

Article info

Received
July 29, 2021

Accepted
September 30, 2021

Available online
October 22, 2021

Keywords

Artificial intelligence; natural language processing; origin of toponyms; embedding models; distributive semantics; vector word representation; dichotomy method; transformation method

Abstract

Essential issues of toponymy presuppose studying separate words to reconstruct the denotative meaning of geographical names that were lost in the modern language and to find out how the peculiarities of the local topography, the inhabitants' activities, etc. are reflected in them. It is possible to solve this kind of problems using intellectual methods of data analysis on the basis of information technologies. However, in scientific literature on toponymy, such methods are practically ignored. The article is devoted to the study of the origin and semantic meanings of geographical names based on finding semantic associates and calculating the semantic similarity of words using the embedding model. According to the proposed method, the origin of some toponyms of the Irkutsk region was determined, their semantic relations were revealed. The dichotomy method was used for toponyms that have two roots in their structure. This made it possible to improve the

operation of the model by clarifying the morphemic composition of the original word. The method of word transformation was used to determine the etymology of the toponym «Moscow». We have received new versions of the origin of the toponym. It is shown that the application of the methods based on distributive semantics and vector representation of words, obtained on the basis of large arrays of text data, significantly expands the possibilities of research in the field of determining the origin of toponyms and clarifying their meaning.

Методы искусственного интеллекта (ИИ), в том числе нейросетевые технологии, успешно применяются для решения многих задач обработки естественного языка: машинный перевод, классификация и кластеризация текстов, построение семантических карт терминов, создание аннотаций, определение ключевых слов. Особое место здесь занимает работа с короткими текстами. Самым коротким текстом является отдельное слово. Исследование отдельных слов языка имеет выход на нерешенные проблемы топонимики. Последняя изучает географические названия (названия населенных пунктов, рек, озер, гор и т.д.): их происхождение, смысловое значение, изменение, современное состояние, написание и произношение. Она представляет интерес с точки зрения изучения исторической миграции народов и ареалов их проживания. К нерешенным проблемам топонимики можно отнести следующие. Во-первых, смысл многих названий (например, Москва, Тверь, Кимры, Коломна и др.) из-за изменения языка во времени забыт. В связи с этим возникает интерес к восстановлению смыслового значения многих топонимов. Во-вторых, в конкретном географическом ареале исторически пересекались различные племена и народы. Каждый из них оставил свои топонимы. Иногда непонятно, к какому языку принадлежит тот или иной топоним. Имеются также «спорные» топонимы, смысл которых можно интерпретировать на нескольких языках. Изучение таких топонимов представляет несомненный интерес (к примеру, Байкал, Ангара, Саяны — это русские топонимы или бурятские?). В-третьих, проблемы топонимики, особенно в исторической перспективе, не могут быть решены однозначно. Однако эти проблемы поддаются анализу с применением методов ИИ, теории вероятностей, математической статистики и других специальных методов. В рамках вероятностного подхода возможно разделение термина на части, замена и перестановка букв в топониме, что еще больше усложняет его изучение.

В научных работах по топонимике практически не уделяется внимания интеллектуальным методам анализа данных с применением информационных технологий. Вместе с тем для семантико-синтаксического анализа топонимов целесообразно применить современный метод, использующий распределенные (дистрибутивные) векторные представления слов [1; 2]. Цифровое векторное представление слова, предложения или текста называется эмбеддингом элемента языка. До последнего времени основная масса исследований по топонимике проводилась в ручном режиме с опорой на знания конкретных лингвистов в области различных языков. В настоящее время появилась возможность использовать векторные модели языка, математическое моделирование терминов с применением нейронных сетей и ИИ, что расширяет сферу исследования в топонимике.

Целью настоящей статьи является определение происхождения некоторых топонимов Иркутской области на основе нахождения при помощи ИИ их семантических ассоциатов с применением математической модели распределенной семантики, т.е. векторного представления слов.

Для достижения поставленной цели решались следующие задачи:

– выбор метода исследования топонимов, основанного на применении многовариантных семантических векторов;

– анализ некоторых топонимов Иркутской области, имеющих русскоязычное происхождение;

– разработка специальных методов подготовки топонимов к исследованию с помощью ИИ;

– интерпретация полученных результатов.

Инструментальная часть исследования

Для определения происхождения и смыслового значения топонимов Иркутской области было проведено их математическое исследование на основе эмбеддинговой модели fastText [3] для русского языка.

Модель получена на основе корпуса русскоязычных текстов GeoWAC (ресурс Common Crawl), сбалансированного авторами разработки по географии России (собран Джонатаном Дунном (Jonathan Dunn) и Бе-

ISSN 2500-2759

ном Адамсом (Ben Adams) в рамках проекта «RusVectōrēs: семантические модели для русского языка»[1]).

Параметры модели geowac_lemmas_none_fasttextskipgram_300_5_2020: корпус русского языка GeoWAC, размер корпуса — 2,1 млрд слов, объем словаря — 154 923 слова, средний частотный порог повторяемости слов — не менее 150, алгоритм — fastTextSkipgram (3–5-граммы), размерность вектора, содержащего ассоциаты, — 300, размер окна — 5 (количество слов в исходном термине), дата создания математического инструментария — октябрь 2020 г.

Тестирование метода

В ходе работы прежде всего было проведено тестирование математического инструментария применительно к задачам топонимики. Для анализа были взяты названия населенных пунктов Иркутского района, имеющие русскоязычное происхождение (табл. 1).

[1] RusVectōrēs: семантические модели для русского языка. URL: https://rusvectores.org.

Трактовка результатов

Голоустное. Название села происходит от названия мыса Голоустного на оз. Байкал, безлесой местности в устье реки, которая также называется Голоустной[2]. Дельта реки заболочена. До образования Прибайкальского национального парка здесь было место утиной охоты. Интересно проанализировать результаты поиска ассоциатов, осуществленного ИИ. Нейронная сеть обнаружила, что Голоустное находится на озере (Байкал) в пойме реки в окружении гор (*озеро 0,42*; *гор 0,42*; *пойменный 0,41*). В Голоустном имеется лесничество и отделение заповедника (*лесник 0,42*; *эколог 0,41*). В населенном пункте есть частные хозяйства (*частное 0,51*) и местное самоуправление (*домоуправление 0,43*; *коллегиальный 0,42*; *домовладение 0,42*; *самоорганизация 0,41*).

Грановщина — село, названное по имени Сеньки и Матюшки Граниных, которые в 1680-х гг. по указу Иркутской воеводческой канцелярии получили разрешение поселить-

[2] ИРКИПЕДИЯRU: энциклопедия и новости Приангарья. URL: http://irkipedia.ru.

Таблица 1

Семантические ассоциаты топонимов Иркутской области, полученные с применением модели GeoWAC fastText

Топоним	Семантические ассоциаты
Голоустное*	частное 0,51; домоуправление 0,43; лесник 0,42; гор 0,42; коллегиальный 0,42; домовладение 0,42; озеро 0,42; пойменный 0,41; эколог 0,41; самоорганизация 0,41
Грановщина*	грановский 0,68; орловщина 0,66; петровщина 0,61; бардинский 0,61; курасовщина 0,60; львовщина 0,59; тарановский 0,59; даровщина 0,59; черниговщина 0,58; духовщина 0,58
Добролёт*	звездолёт 0,53; радиолюбительский 0,47; радиотехнический 0,47; добролюбов 0,46; схемотехнический 0,46; антеть 0,45; воздухофлотский 0,45; самолетный 0,45; аэродромный 0,45; теплотехнический 0,45
Еловка*	еловый 0,65; михайловка 0,62; елочка 0,61; березовка 0,57; ель 0,56; елка 0,56; александровка 0,55; лебединовка 0,54; полянка 0,53
Жердовка*	жердь 0,64; подпорка 0,54; стремянка 0,54; колосник 0,53; настил 0,53; дощатый 0,53; дощечка 0,52; лежанка 0,52
Лисиха*	лиззи 0,50; джуди 0,50; лиса 0,49; лисичка 0,49; салли 0,49; ребекка 0,48; лисица 0,48; люся 0,48; дженни 0,48; энни 0,48
Лиственка	лиственный 0,66; сосновый 0,65; дубрава 0,61; хвойный 0,60; пойменный 0,59; сосна 0,58; лесистый 0,58; березовый 0,58; лиственница 0,58
Московщина*	харьковщина 0,76; орловщина 0,75; шарковщина 0,75; масюковщина 0,72; шарковщинский 0,71; петровщина 0,70; черниговщина 0,69; львовщина 0,67; курасовщина 0,66; киевщина 0,63
Падь	луг 0,56; дубрава 0,56; степной 0,54; болото 0,53; гора 0,53; боровое 0,52; луговой 0,52; озерный 0,52
Пивовариха*	пивовар 0,76; пивоварня 0,73; пивоваренный 0,64; пивоварение 0,60; винодел 0,57; пивзавод 0,57; пекарня 0,54; пивная 0,54; кондитерская 0,52; пиво 0,52
Поливаниха*	полив 0,61; обливание 0,60; опрыскиваний 0,60; орошение 0,58; поливать 0,57; аэрация 0,56; прополка 0,55; опрыскивать 0,55; подкормка 0,54
Разводная*	производная 0,67; разводной 0,63; разводиться 0,55; развести 0,54; развод 0,50; разведенный 0,50; разведять 0,48; производный 0,48; разводить 0,47; уравнение 0,46
Черемшанка*	черемша 0,92; щавель 0,73; брусника 0,70; квашеный 0,68; петрушка 0,68; краснокочанный 0,68; стручковый 0,67; капуста 0,67; земляника 0,67; кинзый 0,67

* Здесь и в табл. 2: слова нет в словаре модели. Цифры представляют собой косинусы углов между многомерными векторами — исходным и ассоциатом — в используемой математической модели русского языка.

ISSN 2500-2759

ся в этом месте для развития хлебопашества. Их отец Агей Федорович Гранин основал деревню Карлук близ Иркутска. ИИ нашел два кластера ассоциатов. Первый связан с фамилией (*грановский 0,68*) и ложным ассоциатом, имеющим одинаковое сочетание букв «-рановский» (*тарановский 0,59*). Второй кластер связан со словами, имеющими общее сочетание букв «-овщина» (*орловщина 0,66*; *петровщина 0,61*; *курасовщина 0,60*; *львовщина 0,59*; *черниговщина 0,58*; *духовщина 0,58*). В этот ряд затесался ложный ассоциат (*даровщина 0,59*).

Добролёт — поселок, образованный в период с 1920 по 1930 г. В то время объединенное управление сибирских воздушных линий общества «Добролёт» строило в Иркутске аэродром, ангары, авиаремонтные мастерские и другие объекты. Российское общество добровольного воздушного флота — «Добролёт» — было создано в СССР 17 марта 1923 г. для содействия развитию воздушного флота страны. 25 февраля 1932 г. ВО ГВФ при Совете труда и обороны было преобразовано в Главное управление гражданского воздушного флота при Совете народных комиссаров СССР. Ровно через месяц оно получило имя «Аэрофлот». Этот исторический факт «зафиксировался» в названии населенного пункта Иркутской области.

ИИ правильно определил ассоциаты, связанные с самолетами, аэродромами и техникой (*звездолёт 0,53*; *радиолюбительский 0,47*; *радиотехнический 0,47*; *схемотехнический 0,46*; *воздухофлотский 0,45*; *самолетный 0,45*; *аэродромный 0,45*; *теплотехнический 0,45*). Фамилия Добролюбов попала сюда как ложный ассоциат, имеющий общее сочетание букв «доброл-» с исходным словом. Еще один ложный ассоциат — слово (*антеть 0,45*) медицинского происхождения.

Еловка. Название связано с местностью, в которой произрастает еловый лес (*еловый 0,65*; *елочка 0,61*; *ель 0,56*; *елка 0,56*). В еловом лесу может быть (*полянка 0,53*). Еловка — это название населенного пункта, такое же, как (*михайловка 0,62*; *березовка 0,57*; *александровка 0,55*; *лебединовка 0,54*). Все названия содержат общий набор букв «-овка».

Жердовка. По рассказам старожилов, свое название Жердовка получила за то, что в свое время здесь рос мачтовый сосновый бор и из этого места возили жерди. ИИ установил основной ассоциат (*жердь 0,64*) и то, что из жердей делали подпорки, стремянки для облегчения посадки всадника на лошадь, колосники для удержания стогов сена от ветра, лежанки (*подпорка 0,54*; *стремянка 0,54*; *колосник 0,53*; *лежанка 0,52*). Кроме того, стволы молодых деревьев распиливали или раскалывали на две половинки. Отсюда происходит кластер ассоциатов (*настил 0,53*; *дощатый 0,53*; *дощечка 0,52*).

Лисиха. Микрорайон построен на месте д. Лисиха. Предание гласит, что еще в XIX в. в здешних лесах водилось много лис. Однажды купец Яковлев отправился подальше от города пристрелять новое ружье. Внезапно из чащи выбежала черно-бурая лиса. Меткий купец первым же выстрелом добыл ценного зверька. Место Яковлеву приглянулось, лисиц здесь оказалось не для одной охоты. Недолго думая, купец окрестил впадину, заросшую кустарником и деревьями, Лисихой. ИИ правильно определил основные ассоциаты (*лиса 0,49*; *лисичка 0,49*; *лисица 0,48*) и вычислил также вторую группу ассоциатов, а именно имена девочек, которых в каких-то текстах называли лисичками (*лиззи 0,50*; *джуди 0,50*; *салли 0,49*; *ребекка 0,48*; *люся 0,48*; *дженни 0,48*; *энни 0,48*).

Лиственка. Поселок назван по обилию здесь лиственницы. ИИ правильно определил основной ассоциат (*лиственница 0,58*). Кроме того, ИИ нашел, что Лиственка исторически располагалась в пойме ручья или реки (*пойменный 0,59*). Найден также кластер ассоциатов, связанный с понятием «лес» (*лиственный 0,66*; *сосновый 0,65*; *дубрава 0,61*; *хвойный 0,60*; *сосна 0,58*; *лесистый 0,58*; *березовый 0.58*).

Московщина. Название населенного пункта произошло от названия прежнего места жительства переселенцев — Москвы [4]. ИИ связал данный топоним с другими названиями местности, в которых используется сочетание букв «-овщина» или «-ковщина» (см. табл. 1).

Падь — деревня недалеко от Иркутска, расположенная в лощине. ИИ определил, что географически в пади могут находиться луг, дубрава, бор, болото, озеро (*луг 0,56*; *луговой 0,52*; *дубрава 0,56*; *боровое 0,52*; *болото 0,53*; *озерный 0,52*). Падь может располагаться в степи (*степной 0,54*). Падь — низкая местность между гор (*гора 0,53*).

Пивовариха — населенный пункт поблизости от Иркутска, в котором в старые времена варили пиво либо который был основан человеком по фамилии Пивоваров. ИИ нашел однокоренные слова (*пивовар 0,76*; *пивоварня 0,73*; *пивоваренный 0,64*; *пивоварение 0,60*; *пивзавод 0,57*; *пивная 0,54*;

ISSN 2500-2759

пиво 0,52); установил, что виноделие близко к пивоварению (*винодел 0,57*); установил также, что в старые времена в пекарне и кондитерской разливали пиво (*пекарня 0,54*; *кондитерская 0,52*).

Поливаниха — населенный пункт, в котором русские крестьяне занимались поливным земледелием, по-видимому поливали огороды. ИИ нашел целый ряд ассоциатов, связанных с поливом растений (*полив 0,61*; *обливание 0,60*; *опрыскиваний 0,60*; *орошение 0,58*; *поливать 0,57*; *аэрация 0,56*; *опрыскивать 0,55*). Слова «прополка» и «подкормка» попали в ассоциативный ряд, поскольку ими занимаются после полива земли.

Разводная — железнодорожная станция, на которой сортируют железнодорожные составы. ИИ нашел для указанного топонима два смысловых кластера. Первый кластер он связал с однокоренными словами (*разводной 0,63*; *разводиться 0,55*; *развести 0,54*; *развод 0,50*; *разведенный 0,50*; *разведять 0,48*; *разводить 0,47*). Слово «разведять», по-видимому, старое, уже утерянное современным русским языком. Второй кластер ИИ связал с математическими терминами (*производная 0,67*; *производный 0,48*; *уравнение 0,46*). ИИ нашел общее сочетание букв «-зводная» в словах «Разводная» и «производная» и определил, что производная может находиться в уравнении. Смысл второго кластера является ложным для изучаемого топонима.

Черемшанка. Название связано с произрастанием в данной местности дикого чеснока — черемши. Черемша используется населением для квашения капусты. Отсюда возникли ассоциаты, связанные с процессом квашения (*черемша 0,92*; *щавель 0,73*; *брусника 0,70*; *квашеный 0,68*; *петрушка 0,68*; *краснокочанный 0,68*; *стручковый 0,67*; *капуста 0,67*; *кинзый 0,67*). Слово «кинзый», по-видимому, старая форма слова «кинза». ИИ установил также, что наряду с черемшой в лесу может расти земляника (*земляника 0,67*).

Метод дихотомии слов

Некоторые топонимы являются составными словами. Яркий пример — топоним «Голоустное». Для исследования таких слов авторы предлагают использовать метод дихотомии, т.е. разделения слова на две части. Этот метод уточняет содержание исходного слова и улучшает работу ИИ. Разделим топоним на два слова: голое устье. В табл. 2 приведены результаты поиска ассоциатов при помощи программы fastText.

Отметим, что интерпретация исходного термина, полученная ИИ в рамках метода дихотомии, существенно изменилась. Программа сообщает, что голое устье (реки) имеется в причерноморье, семиречье, междуречье, а также на севере и северо-востоке (России). Этой трактовке соответствуют ассоциаты (*причерноморье 0,43*; *семиречье 0,41*; *междуречье 0,41*; *северо-восточный 0,41*; *севернее 0,41*; *север 0,40*; *северо-восток 0,40*). В самом деле, в степных и тундровых географических зонах устья рек голые. Программа сообщает также, что голое устье можно (*населить 0,39*).

Весьма интересной является связь ассоциата «улус» с понятием «голое устье». В последнем имеется три буквы — л, у, с, встречающиеся в слове «улус». Таким образом, ИИ прямо указывает на то, что слово «улус» имеет русское происхождение. Так русские называли населенные пункты в устьях рек в степной зоне.

Укажем также на ложный ассоциат (*богослов 0,40*), который, как мы думаем, связан со словосочетанием «голые уста».

Метод трансформации слов

Мощным методом исследования слов с неясным этимологическим происхождением является метод трансформации слов, предложенный в книгах А.Т. Фоменко и Г.В. Носовского [5]. В нашей статье впервые применим этот метод к исследованию с помощью ИИ важнейшего для русского человека топонима «Москва». Смысловое значение топонима «Москва» в настоящее время утеряно. Поставим задачу восстановить его в рамках корпуса русского языка GeoWAC с использованием программы ИИ fastText.

Рассмотрим следующие виды трансформации слов: отбрасывание изменяющегося

Таблица 2

Результаты применения метода дихотомии к топониму «Голоустное»

Топоним	Семантические ассоциаты
Голоустное*	частное 0,51; домоуправление 0,43; лесник 0,42; гор 0,42; коллегиальный 0,42; домовладение 0,42; озеро 0,42; пойменный 0,41; эколог 0,41; самоорганизация 0,41
Голое устье*	причерноморье 0,43; северо-восточный 0,41; семиречье 0,41; междуречье 0,41; севернее 0,41; улус 0,40; север 0,40; богослов 0,40; северо-восток 0,40; населить 0,39

ISSN 2500-2759

окончания; замена глухих и звонких согласных (в — б — п, г — к — х, с — з — ж, с — ш, ч — щ и т.д.); изменение гласных в корне слова (о — а, о — у, а — я и т.д.). Для изучения топонима «Москва» этого будет достаточно.

Слово «Москва» оканчивается на «-ва». Сегодня его считают сочетанием суффикса -в- и окончания -а. В качестве примеров можно привести следующие старые русские слова: трава, ботва, тыква, крапива, плотва, бритва, канава и т.д. Топоним подвергнем следующим видам трансформации: *Москва — моск — мошк — мушк.*

В табл. 3 собраны результаты моделирования трансформантов с применением нейросетевых технологий.

Взятые для анализа варианты *Москва* и *моск* с точки зрения изучения смысла топонима интереса не представляют. Получены ассоциаты, определяющие современное значение слова — крупный город и столица России.

При анализе вариантов *мошк* и *мушк* были получены семантические ряды (см. табл. 3), связанные с кровососущими насекомыми (*мошка*, *мошкара*, *комар*, *москит* (вероятность 0,19)), добычей меда и изготовлением медовухи (*пчеловек*, *пчитывать*, *пчела*, *пчелиный*, *насекомое*, *муравo*, *пчеловод*, *улей* (вероятность 0,28)), а также с огнестрельным оружием (*мушка*, *ружье*, *гладкоствольный*, *охотничий*, *мушкетер*, *карабин*, *ружейный*, *рогатка* (вероятность 0,53)) (табл. 4). Слово «мушка» означает «ружейный прицел». Слово «рогатка» возникло, поскольку мушкет при стрельбе ставили на рогатку.

Семантические ассоциаты указывают на сходство значений слов в сочетаниях «пчеловек — пчела» и «пчитывать — пчела». Видимо, раньше в русском языке существовали слова «пчеловек» — добытчик меда, «пчитывать» — добывать мед. В современном языке эти формы слов уже не используются.

В этимологическом словаре М. Фасмера [6]:

«**Му́ха** укр., блр. му́ха, др.-русск., ст.-слав. Моухаμоῖα… Связано чередованием гласного с *mъšьka (см. мо́шка)».

Таблица 3

Семантические ассоциаты топонима «Москва», полученные с применением модели GeoWAC fastText в рамках метода трансформации слова

Исходное слово	Семантические ассоциаты
Москва	Санкт-Петербург 0,82; Петербург 0,80; Казань 0,79; Калининград 0,76; Тверь 0,74; Екатеринбург 0,74; Питер 0,74; москво 0,74; Краснодар 0,73; санкт 0,73
моск	москва 0,61; москво 0,57; петербург 0,54; московский 0,53; московия 0,52; санкт 0,50; питер 0,50; санкт-петербург 0,49; саратов 0,47; тверь 0,47
мошк	мошка 0,70; мошкара 0,68; комар 0,59; насекомое 0,55; мошковский 0,52; таракан 0,52; муравo 0,51; слепень 0,51; кровососущий 0,50; москит 0,49
мушк	мушка 0,58; ружье 0,42; гладкоствольный 0,41; охотничий 0,39; мушкетер 0,38; рогатка 0,38; карабин 0,38; арбалет 0,37; шпага 0,37; ружейный 0,37

Таблица 4

Вероятность ассоциатов в семантическом ряду *мошк, мушк*

Ассоциат	Косинусное сходство с исходным словом	Количество словоупотреблений в корпусе n_i	Вероятность $Pr(n) = n / N$
мошка	0,7	3 190	0,023 330 31
мошкара	0,68	682	0,004 987 86
комар	0,59	16 657	0,121 822 25
муравo	0,51	3 621	0,026 482 46
слепень	0,51	650	0,004 753 83
кровососущий	0,5	679	0,004 965 92
москит	0,49	934	0,006 830 88
медоносный	0,46	772	0,005 646 08
пчеловек	0,44	5 908	0,043 208 61
пчитывать	0,42	1 631	0,011 928 44
пчела	0,42	11 639	0,085 122 72
пчеловод	0,42	3 849	0,028 149 96
улей	0,42	6 216	0,045 461 19
пчелиный	0,42	7 655	0,055 985 43
мушка	0,58	5 372	0,039 288 54

ISSN 2500-2759

Окончание табл. 4

Ассоциат	Косинусное сходство с исходным словом	Количество словоупотреблений в корпусе n_i	Вероятность $Pr(n) = n / N$
мушкет	0,53	832	0,006 084 9
ружье	0,42	23 788	0,173 975 37
гладкоствольный	0,41	1 899	0,013 888 48
охотничий	0,39	24 902	0,182 122 69
мушкетер	0,38	2 790	0,020 404 88
рогатка	0,38	2 739	0,020 031 89
карабин	0,38	9 006	0,065 866 07
ружейный	0,37	1 321	0,009 661 24
Всего	–	136 732	1

«**Мóшка** мошкарá ж., собир. (ср. детворá), др.-русск., цслав. мъшица, укр. мóшка, чеш. mšice «тля», польск. mszyca, н.-луж. pšyca «комар, мошка». Связано чередованием гласных с мýха».

«**Комáр**, род. п. -á, укр., блр. комáр, русск.-цслав. комаръ, болг. комáр, сербохорв. кòмâр, словен. komár, род. п. -árja, чеш., слвц. komár, польск. komar, в.-луж. komor. ...Другая ступень чередования: слав. *čьmelь (см. шмель)».

«**Шмель** род. п. шмеля́, диал. чмель, севск. (Преобр.), щéмель, псковск. (Даль), укр. чмíль, род. п. чмеля́, джмíль, род. п. -я́, также чмолá «шмель» (по аналогии слова бджолá), блр. чмель, витебск., словен. čmêlj, šmêlj, чеш. čmel, štmel, стар. ščmel, слвц. čmeľ, польск. czmiel, strzmiel, в.-луж. čmjeła, н.-луж. Tśḿel Праслав. *čьmeľь связано чередованием гласных с komarъ (см. комáр). Родственно лит. kamãnė «вид шмеля», kamìnė «дикая пчела», лтш. kamine, др.-прусск. camus «шмель», др.-инд. camarás «Bosgrunniens», д.-в.-н. humbal «шмель», далее — лит. kimĩnti «делать голос хриплым, глухим», kĩmti, kĩmstu «хрипнуть».

«**Пчелá** укр. пчолá, бджолá, др.-русск., ст.-слав. бьчела (Остром., Ассем., Ps. Sin.), бъчела (Мар.; см. Мейе, ниже), болг. бчелá, сербохорв. пчѐла, чѐла, словен. bə̀čə̑la, čəbə̑la, čbə̑la, čə̑la, чеш., слвц. včela, др.-польск. pczoła, польск. pszczoła, в.-луж. pčoła, н.-луж. coła, полаб. cü'öla. Праслав., скорее всего, *bьčela, расширение *bьko- (ср. весёлый), связанное с ирл. bech (*biko-) «пчела», лат. fūcus «трутень» (*bhoiko-)».

Из словаря Фасмера вытекает, что в старославянском языке слово «пчела» образовалось следующим образом: бить чело — бичела — пчела. Слово «муха» образовалось так: мычать корова — му ко — мýка — мýха. Отсюда же произошли слова «мýка», «мучиться», «мучение». Слово «муха» возникло вместе с изобретением скотоводства, слово «пчела» — значительно позже.

Обсуждение результатов

Анализ результатов, полученных при изучении некоторых русскоязычных топонимов Иркутской области с применением эмбеддинговых моделей, позволяет сделать следующие выводы:

1. Для определения происхождения слов целесообразно применять методы на основе эмбеддинговых моделей и математического моделирования естественного языка.

2. Для топонимов наилучшие результаты получены с применением семантических векторов с учетом *N*-грамм слов, так как топонимы и микротопонимы — это чаще всего слова, которые редко встречаются в общеупотребительной лексике и, соответственно, могут отсутствовать в сбалансированных корпусах русского языка.

3. Получены положительные результаты при анализе 13 топонимов Иркутской области (100 %). Это названия населенных пунктов: Голоустное, Грановщина, Добролёт, Еловка, Жердовка, Лисиха, Лиственка, Московщина, Падь, Пивовариха, Поливаниха, Разводная, Черемшанка. Семантические ассоциаты для Еловки, Лиственки, Пади, Черемшанки указывают на то, что происхождение топонимов связано с лесом, лесным хозяйством и названиями соответствующих растений и пород деревьев: ель, лиственница, черемша. Название «Жердовка» произошло от рода занятий людей, населяющих территорию, — изготовление жердей. Населенные пункты Разводная и Поливаниха также получили свои наименования по роду деятельности людей (разводить вагоны и поливать огороды). В обоих случаях семантические ассоциаты связаны с этимологией слов — от «разводить, развести» и «поливать, полив». Пивовариха названа по роду деятельности или фамилии Пивоваров («пивоварня», «пивовар»), Лиси-

ISSN 2500-2759

ха — местность, где в большом количестве водились лисы. Название поселка Добролёт произошло от названия российского общества добровольного воздушного флота «Добролёт» (в марте 1932 г. переименовано в «Аэрофлот»). Семантические связи «Добролёт — аэрофлот» представлены ассоциатами «аэродромный», «самолет».

4. Грановщина и Московщина — топонимы, названия которых происходят от имен собственных (Гранин, Москва). Модель, применяемая для анализа топонимов, была обучена на текстах корпуса GeoWAC без привязки к именам собственным и рассчитана в основном на вычисление семантических векторов общеупотребительных слов. Тем не менее семантические ассоциаты для топонимов «Грановщина», «Московщина» имеют смысл.

5. Интерпретация результатов анализа для названия «Голоустное» усложняется за счет морфемного состава самого топонима (имеется два корня). Поэтому авторам статьи для расширения анализа пришлось применить к данному слову метод дихотомии. Тем не менее в списке ассоциатов топонима «Голоустное» и словосочетания «голое устье» есть слова, указывающие на правдоподобные версии происхождения топонима:

– озеро, пойменный, горы (семантические ассоциаты топонима «Голоустное»);

– междуречье, семиречье, причерноморье (семантические ассоциаты словосочетания «голое устье»).

6. Формирование семантического ряда для топонима «Москва» не дало положительного результата с точки зрения определения его смыслового значения, так как в ходе анализа были обнаружены ассоциаты, связанные с наиболее употребительным значением слова — большой город, столица России. Важные закономерности были выявлены при применении к данному топониму метода трансформации слова. Исследования трансформантов *мошк* и *мушк*, которые были получены при видоизменении исходного топонима, привели к следующим трем классам ассоциатов: кровососущие насекомые; пчеловодство и добыча меда; огнестрельное оружие.

По информации из словаря М. Фасмера, старославянские и праславянские формы слов (муха — *моуха*, мошка — **тъšьka*, мошкара — *мъшица*, комар — *комаръ*, шмель — **čьteľь*, пчела — *бьчела* (*бъчела*)) косвенно указывают на возможное происхождение топонима «Москва» от *моуха*, **тъšьka*, *мъшица*, *комаръ* в значении «мошка, мошкара», а также на связанную версию его происхождения от *моуха*, **тъšьka*, *мъшица*, *комаръ*, **čьteľь*, *бьчела* (*бъчела*) в обобщенном значении «насекомое» и «пчела».

Можно сформулировать первую гипотезу происхождения топонима «Москва» следующим образом: географическое название местности и реки сформировалось от славянских слов *моуха*, **тъšьka*, *мъшица*, *комаръ*, **čьteľь*, *бьчела* (*бъчела*) в значении «мошка», «мошкара» и связанного с ним значения «насекомое», «пчела». Пойма р. Москвы и болотистые леса в округе всегда изобиловали мошкой, комарами и мухами, что не представляет труда проверить и в настоящее время.

Вторая гипотеза связана с добычей и переработкой меда в данной местности. Население, которое проживало в междуречье Волги и Оки в средние века, не знало виноградных вин. Занимались добычей меда в окружающих лесах и изготовлением веселящего напитка — медовухи. Мед производили пчелы, т.е. «мухи».

Не менее интересной и неожиданной является третья гипотеза происхождения топонима «Москва». Хорошо известно, что слова «мушка» (прицел), «мушкет», «мушкетер» русского происхождения. В XIV–XV вв. в метрополии Татаро-Монголии (в городах, расположенных между Волгой и Окой) было изобретено и развернуто производство огнестрельного оружия. Полки мушкетеров (стрельцов) наводили ужас на любого врага. Тысячи мушкетов выпускали тучи пуль — «мух». На концах стволов также сидели «мушки» — прицелы. Поэтому столица поздней Татаро-Монголии, крепость Москва, построенная во второй половине XVI в., могла получить название от первого вида огнестрельного оружия.

Отметим, что мы рассматриваем в статье только славянское происхождение топонима «Москва», используя результаты математического моделирования корпуса современного русского языка, имеющего объем в 2,1 млрд слов.

Заключение

В работе проведен анализ ряда русскоязычных топонимов Иркутской области. Опробована модель дистрибутивной семантики для определения происхождения географических названий.

Топонимика Иркутской области очень разнообразна, она представлена названиями географических объектов русскоязычного, бурятского, тунгусо-эвенкийского, тюрко-

ISSN 2500-2759

язычного происхождения. Именование объектов имеет взаимосвязь с их природно-географическими характеристиками, также в их названиях отражается индивидуальность человека, особенности его хозяйственной деятельности, нравственная и духовная сферы жизни.

Для определения происхождения топонимов впервые использованы методы на основе эмбеддингов слов. Эмбеддинговые модели для вычисления семантических ассоциатов — Word2vec с архитектурой CBOW и Skip-gram; модель fastText, основанная на построении семантических векторов *N*-грамм слов. Преимуществом модели fastText является возможность работать с редкими и устаревшими словами. Анализ топонимов Иркутской области в данной работе проводился с применением модели GeoWAC fastText русскоязычного корпуса GeoWAC, сбалансированного по географии России авторами разработки.

Получены положительные результаты применения модели для всех изученных топонимов.

К двухкоренным топонимам авторы статьи предлагают применять метод дихотомии, демонстрируя его использование на примере топонима «Голоустное». В результате были получены неожиданные ассоциаты. Например, слово «улус» оказалось русского происхождения и непосредственно связано со словосочетанием «голое устье».

Новые результаты получены и при изучении топонима «Москва». Авторы статьи предлагают старинные по происхождению топонимы с невыделяемым сегодня смысловым значением анализировать с использованием метода трансформации слова. Применение этого метода в сочетании с программой ИИ к топониму «Москва» привело к формулированию трех возможных гипотез происхождения термина: от кровососущих насекомых; от пчеловодства и добычи меда; от наименования огнестрельного оружия. Авторы статьи определили, что вероятности этих гипотез соотносятся как 0,19 : 0,28 : 0,53.

Проведенное исследование показало целесообразность применения современного метода, использующего дистрибутивные векторные представления слов — эмбеддинг элемента языка, для расширения сферы исследований в топонимике.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Efficient Estimation of Word Representations in Vector Space / T. Mikolov, K. Chen, G. Corrado, J. Dean // International Conference on Learning Representations. — Scottsdale, 2013. — URL: https://arxiv.org/abs/1301.3781.

2. Goldberg Y. Word2vec Explained: Deriving Mikolov et al.'s Negative-sampling Word-Embedding Method / Y. Goldberg, O. Levy // ArXiv. — 2014. — URL: https://arxiv.org/abs/1402.3722.

3. Enriching word vectors with subword information / P. Bojanowski, E. Grave, A. Joulin, T. Mikolov. — DOI 10.1162/tacl_a_00051 // Transactions of the Association for Computational Linguistics. — 2017. — Vol. 5, № 1. — P. 135–146.

4. Мильхеев М.Н. Топонимика Бурятии. История, система и происхождение географических названий / М.Н. Мильхеев. — Улан-Удэ : Бурят. кн. изд-во, 1969. — 150 с.

5. Носовский Г.В. Библейская Русь. В 4 т. / Г.В. Носовский, А.Т. Фоменко. — Москва : Римис, 2004.

6. Фасмер М. Этимологический словарь русского языка. В 4 т. / М. Фасмер ; пер. с нем. и доп. О.Н. Трубачева. — 2-е изд. — Москва : Прогресс, 1986–1987.

Информация об авторах

Боровский Андрей Викторович — доктор физико-математических наук, профессор, кафедра математических методов и цифровых технологий, Байкальский государственный университет, г. Иркутск, Российская Федерация, e-mail: andrei-borovskii@mail.ru.

Раковская Елена Евгеньевна — аспирант, кафедра математических методов и цифровых технологий, Байкальский государственный университет, г. Иркутск, Российская Федерация, e-mail: rakovskaya19@mail.ru.

Для цитирования

Боровский А.В. Исследование топонимов Иркутской области с применением методов искусственного интеллекта / А.В. Боровский, Е.Е. Раковская. — DOI 10.17150/2500-2759.2021.31(3).382-390 // Известия Байкальского государственного университета. — 2021. — Т. 31, № 3. — С. 382–390.

Authors

Andrei V. Borovsky — D.Sc. in Physical and Mathematical Sciences, Professor, Department of Mathematical Methods and Digital Technologies, Baikal State University, Irkutsk, the Russian Federation, e-mail: andrei-borovskii@mail.ru.

Elena E. Rakovskaya — Ph.D. Student, Department of Mathematical Methods and Digital Technologies, Baikal State University, Irkutsk, the Russian Federation, e-mail: rakovskaya19@mail.ru.

For Citation

Borovsky A.V., Rakovskaya E.E. Research of Toponyms of the Irkutsk Region Using the Method of Artificial Intelligence. *Izvestiya Baikal'skogo gosudarstvennogo universiteta = Bulletin of Baikal State University*, 2021, vol. 31, no. 3, pp. 382–390. DOI: 10.17150/2500-2759.2021.31(3).382-390. (In Russian).

ISSN 2500-2759