УДК 81'1. 1. DOI 10.51762/1FK-2021-26-02-07. ББК Ш10. ГРНТИ 16.21.07. КОД ВАК 10.02.19

О НЕКОТОРЫХ СЛЕДСТВИЯХ КОРПУСНОЙ ЛИНГВИСТИКИ ДЛЯ ОБЩЕЙ ТЕОРИИ ЯЗЫКА

Копотев М. В.
Высшая школа экономики (Санкт-Петербург, Россия)
Хельсинкский университет (Хельсинки, Финляндия)
ORCID ID: https://orcid.org/0000-0002-4998-2952

Аннотация. Статья посвящена обсуждению основных подходов в рамках корпусной лингвистики и их влиянию на общее развитие лингвистических знаний. На примере исследований, выполненных при участии автора, обсуждаются три подхода к изучению языка в корпусной лингвистике. Первый – анализ, использующий корпус, – предполагает, что данные, собранные в корпусе, используются как источник примеров на естественном языке. Второй – анализ, основанный на корпусе, – заключается в том, что корпусные данные исследуются не только качественно, но и количественно. Третий – анализ, направляемый корпусом, – предполагает, что задача исследователя состоит в создании алгоритма обработки языковых данных, результат которого требует теоретического осмысления или практического применения. Статья завершается обсуждением тех результатов, которые корпусная лингвистика привносит в общее представление о языке и лингвистике. Важнейшими из них являются: снижение роли интроспекции, увели- чение внимания к периферийным языковым феноменам, опора на количественные данные. Подводить итоги влияния корпусной лингвистики на общую теорию языка еще рано, но уже сейчас ясно, что синтагматические связи, в частности идиоматизация в широком смысле, продвинулись в центр лингвисти- ческого внимания и признаются одним из основных феноменов языка и его эволюции. Более того, адекватным описанием языка оказывается не моделирование правил взаимодействия единиц, разделенных на уровни, а описание всех – и индивидуальных, и самых общих – вероятностных параметров употребления, представляющих собой единый континуум, в котором разделение на язык и речь является условным.

Ключевые слова: корпусная лингвистика; теория языкознания; лингвистические исследования.

Для цитирования: Копотев, М. В. О некоторых следствиях корпусной лингвистики для общей теории языка / М. В. Копотев. – Текст : непосредственный // Филологический класс. – 2021. – Т. 26, № 2. – С. 90–102. – DOI: 10.51762/1FK-2021-26-02-07.

SOME THOUGHTS ON CORPUS AND GENERAL LINGUISTICS

Mikhail V. Kopotev
Higher School of Economics (Saint Petersburg, Russia) University of Helsinki (Helsinki, Finland)
ORCID ID: https://orcid.org/0000-0002-4998-2952

Abstract. The article is devoted to a discussion of dominant approaches developed within the framework of Corpus Linguistics (CL) and their influence on the general theory of language. Based on the research co-authored with his colleagues, the author describes three approaches to linguistic research in CL. First, corpus-*informed* analysis assumes that the data collected in the corpus are used as a source of examples in a natural language. Second, corpus-*based* analysis presupposes that the data are examined not only qualitatively but also quantitatively. Third, corpus-*driven* analysis assumes that the research task is to create an algorithm for data processing, the results of which require theoretical interpretation or practical application. The article concludes with a discussion of those implications that CL brings into the general understanding of language. The most important of them are: reduction of the role of introspection, increase of attention to peripheral linguistic phenomena, and reliance on quantitative data. It is still too early to sum up the impact of corpus linguistics on the general theory of language, but it is already clear that syntagmatic connections, in particular idiomatization in a broad sense, have moved into the focus of linguistic attention and are recognized as one of the main phenomena of language and its evolution. Moreover, an adequate description of a language is not limited to the rules of interaction of units divided into levels, but the description of all – both individual and the most general – probabilistic parameters of use, representing a single continuum in which the division into language and speech is conventional.

K e y w o r d s : corpus linguistics; theory of language; linguistic studies.

F o r c i t a t i o n : Kopotev, M. V. (2021). Some houghts on Corpus and General Linguistics. In *Philological Class*. Vol. 26. No. 2, pp. 90–102. DOI: 10.51762/1FK-2021-26- 02-07.

1. Введение

С некоторой долей субъективизма можно сказать, что методологическое развитие современной лингвистики во многом формируют два направления: экспериментальная и корпусная лингвистика. Последняя, являясь одним из самых молодых разделов языкознания, успела, однако, расширить объект изучения лингвистики, изменить представление о методах лингвистического анализа и внести свой вклад в общую теорию языка.

У самого термина *корпусная лингвистика* можно выделить два значения:

> *Корпусная лингвистика*$_1$ – раздел компьютерной лингвистики, посвященный созданию корпусов и инструментов для их обработки.

> *Корпусная лингвистика*$_2$ – раздел лингвистики (как частной, так и общей), использующий корпуса и методы корпусного анализа для проведения исследований.

Естественно, конкретный лингвист может представлять как первое, так и второе направление, хотя тех, кто занимается «корпусной лингвистикой$_2$», очевидно, больше. В этой статье я сосредоточусь на обсуждении тех результатов, которые корпусной анализ привносит в наше понимание языка и, соответственно, буду использовать термин во втором значении.

Основу корпусной лингвистики, ее главный объект, составляет *корпус*, или набор текстов, собранных и часто размеченных по определенным правилам. Формальное определение корпуса можно найти во множестве учебных пособиях по этой дисциплине. Приведу для примера такое:

> Corpus in modern linguistics, in contrast to being simply any body of text, might more accurately be described as a **finite-sized body** of **machine-readable** text, sampled in order to be **maximally representative** of the language variety under consideration [McEnery, Wilson 1996: 24].

> Корпус в современной лингвистике в отличие от любого набора текстов может быть более точно определен как **ограниченный по объему набор электронных текстов**, собранных с целью **максимально точно представлять** исследуемый вариант языка. (*Переводы здесь и далее выполнены автором статьи.*)

Выделенные мной фрагменты в этом определении задают ключевые особенности корпуса: электронно-читаемая форма, известный объем, сбалансированность и репрезентативность (см., подробнее [Копотев 2014]). На практике создание сбалансированного и репрезентативного корпуса является серьезной проблемой, которая широко обсуждается в научной литературе [см. например, Ädel 2020], и далеко не каждый корпус достигает нужной степени репрезентативности материала. Приведу в связи с этим слова Ч. Филлмора, актуальные, как представляется, до сих пор:

> I have two main observations to make. The first is that I don't think there can be any corpora, however large, that contain information about all of the areas of English lexicon and grammar that I want to explore; all that I have seen are inadequate. The second observation is that every corpus that I've had a chance to examine, however small, has taught me facts that I couldn't imagine finding out about in any other way [Fillmore 2011: 35].

> У меня есть два существенных наблюдения. Первое: я не думаю, что может существовать корпус, каким бы большим он ни был, который содержит информацию обо всех областях английского лексикона и грамматики, которые я хотел бы исследовать – все, что я видел, недостаточно. Второе: любой корпус, который

> я имел возможность исследовать, каким бы маленьким он ни был, предоставлял мне факты, про которые я не могу вообразить, что обнаружил бы их каким-нибудь другим образом.

На сегодняшний день существует огромное количество корпусов для разных языков. Не останавливаясь подробно на их описании, отмечу, что кроме Национального корпуса русского языка, доступного на сайте ruscorpora.ru, существуют и другие ресурсы, позволяющие решать множество лингвистических задач [см. наши обзоры в Kopotev et al. 2018, 2021]. Кроме этого, в распоряжении лингвистов существуют достаточно простые инструменты создания собственных корпусов. Наиболее популярными из них являются, пожалуй, SketchEngine [Kilgarriff и др. 2014], AntConc [Anthony 2004] и WordSmith [Scott 2008].

2. Виды корпусного анализа

Корпусная лингвистика естественным образом опирается на существующие подходы к исследованию, добавляя к ним новые. Ниже эти подходы и стоящие за ними методы анализа несколько искусственно разделены на три больших блока – от наиболее связанных с лингвистической традицией до новаторских, стоящих на границе лингвистики и информационных технологий. Каждый подход иллюстрируется конкретным исследованием.

2.1. Анализ, использующий корпус (англ. corpus-informed analysis)

Этот тип предполагает, что данные, собранные в корпусе, используются как источник примеров на естественном языке. В этом смысле такой подход мало чем отличается от сбора примеров как основы лингвистических исследований. Глубокого количественного анализа при таком подходе не предполагается. Этот подход в определенном смысле продолжает традицию лингвистики XX века, которая стремится к большей объективности лингвистического анализа: **в корпусном исследовании роль интроспекции (языковой интуиции исследователя) снижается**, она уступает место анализу языковых фактов во всем их разнообразии. В России этот поворот к языковой реальности совпал с критикой концепции литературного языка, развернувшейся в конце XX века, и позволил обратить внимание на те языковые явления, которые прежде не привлекали внимания исследователей. Для примера укажу на работу [Влахов 2010], в которой изучаются русские причастия будущего времени, долгое время существовавшие в русском языке незаконно – как «человек, не *предъявящий* никаких свидетельств и пашпортов» (Н. В. Гоголь; пример из [Влахов 2010]).

Одно из исследований [Janda, Kopotev, Nesset 2020], выполненное нами в рамках этого подхода, посвящено конструкциям, которые редко привлекали внимание исследователей в силу маргинальности и нечастотности их употребления [см. Nichols 1981; Величко 2016]. Речь идет о выражениях типа «Дурак дураком», которые исследователи обычно относят к фразеологии, считая их *лексическими* единицами, фразеологизмами. Однако собранный на основе НКРЯ материал позволил обнаружить 3088 примеров[1], в которых используются 1596 уникальных существительных. Таким образом, речь идет о *синтаксической фраземе*, включающей большой набор переменных, заполняющих позиции номинативной и инструментальной групп. Самые частотные из них приведены в таблице 1.

Таблица 1. Частотные существительные в конструкции X-nom X-ins

честь честью	338
дурак дураком	128
клин клином	90
дружба дружбой	80
чин чином	75
дура дурой	51
дело делом	39
туча тучей	37
служба службой	36
зверь зверем	28
любовь любовью	23
молодец молодцом	23
война войной	20

Анализ собранных примеров позволил выявить целое семейство конструкций, находящих-

[1] Полный набор данных доступен в репозитории The Tromsø Repository of Language and Linguistics archive по адресу: https://doi.org/10.18710/ZAM96S (дата обращения: 07.05.2021).

ся друг с другом в деривативных отношениях, имеющих разную семантику и, соответственно, разное лексическое наполнение в современном языке. Нам удалось установить, что указанные конструкции представляют собой радиальную категорию (в смысле Дж. Лакоффа, см. [Lakoff 2008]), состоящую из нескольких взаимосвязанных подтипов, из которых три самых частотных в соответствии с их значением можно назвать 'экстремальный' (пример 1), 'образцовый' (пример 2) и 'смена дискурса' (пример 3).

(1) А то ты теперь **дурак дураком ('полный дурак' – МК)**, хуже морковки на Луне, а я и вовсе трупом лежу, червей жду... (Юрий Буйда. У кошки девять смертей (2000) // Новый Мир, 2005).

(2) Красивей мужчины во всей губернии нет, умный, богатый, а никакого толку... **Дурак дураком ('выглядящий дураком' – МК)**... (А. П. Чехов. Леший, 1890).

(3) **Дурак дураком, а ('Х, возможно, дурак, но...' – МК)** дело говорит (Александр Вампилов. Прошлым летом в Чулимске, 1971).

Наше исследование не только вводит в оборот новые данные и описывает эти типы на большом диахроническом материале, но и демонстрирует две важные составляющие современного корпусного исследования. Во-первых, корпусные данные позволяют собирать большое количество примеров при относительно малых временных затратах, что дает возможность как переосмыслить существующие наблюдения, так и обратиться к тем явлениям, которые раньше считались периферийными. Во-вторых, корпусные исследования становятся проверяемыми в том смысле, что при желании любой может обратиться к корпусу, чтобы подтвердить результаты, или получить доступ к собранному материалу для возможного продолжения исследования, если, конечно, авторы это позволяют.

2.2. Анализ, основанный на корпусе (англ. corpus-based analysis)

Особенность этого вида корпусного анализа заключается в том, что корпусные данные исследуются не только качественно, но и количественно. Обычно под таким вариантом корпусной работы подразумевается проверка некоторой исходной гипотезы, теоретические положения которой заданы заранее и не меняются в ходе исследования. Этот тип корпусных исследований знаменует важное изменение исследовательской оптики: **в лингвистику вернулась статистика**. Конечно, количественные методы применялись в лингвистике задолго до возникновения корпусной лингвистики (см., например, пионерскую работу по определению авторского стиля [Морозов 1916]), однако в рамках этого подхода количественные данные и базовый статистический анализ становятся обязательным инструментом верификации и – главное – воспроизводимости результатов. Этот подход является наиболее распространенным в современной корпусной русистике. Не боясь ошибиться, можно сказать, что большинство современных исследований, содержащих в названии слова *корпус*, *корпусной* или *корпусный*, придерживаются именно этой исследовательской стратегии.

Пожалуй, наиболее последовательным и масштабным примером такого подхода является проект корпусного описания русской грамматики (http://rusgram.ru), участники которого декларируют количественное описание с опорой на НКРЯ:

> Авторами описания используются количественные методы анализа корпусных данных.
>
> Описание мыслится как строго эмпирически ориентированное и теоретически нейтральное: в нем по возможности учитываются данные, накопленные в рамках самых разных подходов и направлений современной лингвистики. <...>
>
> Настоящее описание во многом опирается на существующие грамматики русского языка: «академические» грамматики [Виноградов 1960] и [Шведова 1980]; монографические описания [Виноградов 1947; Исаченко 1954–1960; Timberlake 2004] и мн. др.
>
> (О проекте. (*Корпусная грамматика*, n.d.)).

В нашем исследовании [Kisselev, Klimov, Kopotev (в печати)] мы поставили задачу измерить, как синтаксическая сложность текста отражает уровень владения иностранным языком, в данном случае, русским. Для ана-

лиза были использованы данные ежегодного конкурса сочинений, проводимого в США (National Post-Secondary Russian Essay Contest (NPSREC)). Все сочинения были написаны в течение одного часа без использования компьютеров и словарей, проверены и оценены профессиональными педагогами по шкале WPT [см. ACTFL proficiency guidelines 2012]. Объем корпуса составил 46807 слов, или 2915 предложений, написанных 133 учениками. Таким образом, основой анализа стал уникальный по своей цельности, профессионально оцененный набор данных, который в процессе исследования был переведен в цифровую форму, обработан и проанализирован с помощью различных мер статистического анализа, целью которых стал поиска ответа на один вопрос: существует ли связь между синтаксическим явлением и уровнем владения языком.

Поскольку общая цель проекта состояла в том, чтобы создать систему автоматической проверки языковой компетенции, были выбраны такие параметры, которые можно проверить автоматически с помощью существующих в настоящее время систем синтаксического анализа[1]. В итоге, были выбраны следующие параметры:

– средняя длина предложения;
– максимальная глубина вложенных групп;
– минимальная глубина вложенных групп;
– сочинительные предложения;
– подчинительные предложения;
– причастные обороты;
– деепричастные обороты;
– относительные клаузы;
– инфинитивные клаузы;
– сентенциальные актанты (изъяснительные придаточные);
– сентенциальные сирконстанты (придаточные времени, места, уступки и т. п.);
– составные союзы (или скрепы, по терминологии М. И. Черемисиной и Т. А. Колосовой [Черемисина, Колосова 1987]).

Прежде всего, мы составили матрицу корреляций, чтобы убедиться в отсутствии существенных зависимостей разных мер друг от друга. Оказалось, что только три из них относительно хорошо коррелируют друг с другом (светлые квадратики на диаграмме 1). Это длина предложения, количество «плоских» предложений с небольшой глубиной вложения и количество предложений с большой глубиной вложения; последние два параметра, естественно, находятся в обратной зависимости. Корреляционный анализ показывает, что все параметры за исключением указанных трех фиксируют независимые друг от друга аспекты измерения.

Главный исследовательский вопрос решался с помощью теста ANOVA, цель которого – показать, какой вклад переменные величины (в нашем случае – синтаксические параметры) вносят в формирование постоянной величины (в нашем случае уровень – владение языком). Таблица ниже обобщает наши подсчеты в стандартной таблице результатов статистического текста ANOVA. Оказалось, что шесть параметров хорошо предсказывают уровень владения языком, еще три предсказывают уровень владения относительно слабо, а три параметра не могут считаться надежными показателями при определении уровня. Последняя колонка в таблице 2 показывает уровень надежности, где один астериск значит слабое, а три – максимальное надежное предсказание.

Совмещение двух представленных статистических расчетов позволяет нам сделать вывод, что оптимальный набор автоматизируемых синтаксических параметров, определяющих уровень владения русским языком, включает подсчет количества сирконстантных и относительных клауз, причастных оборотов и одного из трех пересекающихся параметров: средняя длина предложения, максимальная или минимальная глубина вложенных групп. Дополнительно можно учитывать количество составных союзов и инфинитивных и деепричастных оборотов.

Теоретическое обобщение этих результатов выходит за рамки настоящей статьи (см. нашу основную публикацию [Kisselev, Klimov, Koporev (в печати)]. Здесь же хочется обратить внимание на то, что в рамках корпусной лингвистики возможно создавать даже узкоспециализированные наборы данных и применять

[1] Часть выбранных параметров, очевидно, пересекается, например, сентенциальные актанты и сирконстанты являются частными случаями подчинительных клауз. Нас, однако, интересовали как обобщенные, так и более конкретные параметры и их влияние на уровень владения языком.

к ним разнообразные инструменты статистического анализа. С некоторым упрощением можно сказать, что чем лучше исследователь владеет статистическими инструментарием, тем более доказуемой оказывается гипотеза, которую он или она представляет, тем более тонкие закономерности можно выявить в анализируемых данных. Это, естественно, не отменяет исследовательской интуиции и теоретических предпочтений исследователя.

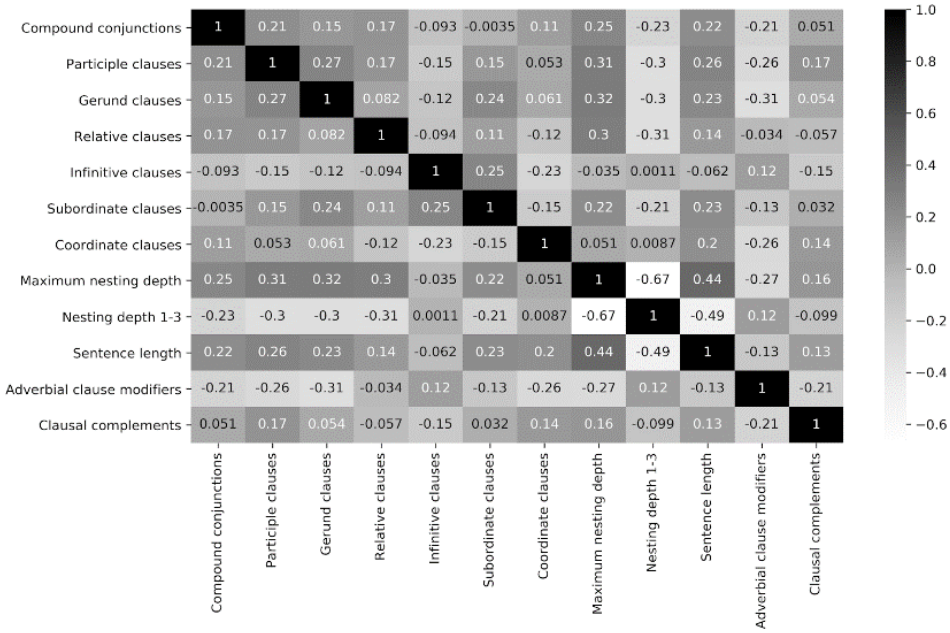

Диаграмма 1. Корреляция синтаксических параметров в учебных текстах

Таблица 2. Результаты теста ANOVA

	Степени свободы	Сумма квадратов	Среднее квадратов	Значение F-меры	Pr(>F)	Значимость результатов
Средняя длина предложения	5	262.7	52.53	**5.216**	0.000218	***
Максимальная глубина вложенных групп	5	138.0	27.61	**7.669**	<0.001	***
Минимальная глубина вложенных групп	5	2277	455.4	**6.745**	<0.001	***
Сочинительные предложения	5	523	104.52	**1.938**	0.0926	–
Подчинительные предложения	5	127	25.41	**0.386**	0.858	–
Деепричастные обороты	5	68.6	13.716	**2.863**	0.0175	*
Причастные обороты	5	191.9	38.38	**5.891**	<0.001	***
Инфинитивные клаузы	5	459	91.80	**2.751**	0.0215	*
Относительные клаузы	5	1410	282.05	**12.38**	<0.001	***
Сентенциальные актанты	5	76.5	15.30	**0.633**	0.675	–
Сентенциальные сирконстанты	5	0.185	0.03699	**8.287**	<0.001	***
Составные союзы	5	2.557	0.5113	**2.533**	0.032	*

2.3. Анализ, направляемый корпусом (англ. corpus-driven analysis)

Третий тип корпусных исследований предполагает исключение или минимальное включение заранее заданных теоретических положений. В этом случае предполагается, что задача исследователя состоит в создании алгоритма обработки языковых данных, результатом применения которого становится анализ, например классификация языковых явлений, проведенный без участия человека. Задачей исследователя в таком случае является интерпретация полученных результатов и корректировка теории. Этот подход радикально отличается от предыдущих тем, что позволяет по-новому взглянуть на языковой материал, заметить закономерности, которые прежде не привлекали внимания лингвистов. В этой части корпусная лингвистика является частью большого социогуманитарного поворота к алгоритмическому анализу больших данных.

В качестве яркого примера такого подхода может служить *дистрибутивная семантика* – направление, активно развивающееся в последнее десятилетие. Основы дистрибутивной семантики заложил американский лингвист Дж. Фёрс, еще в середине XX века сформулировавший: "You shall know a word by the company it keeps" («Слово следует опознавать по его окружению») [Firth 1957: 11]. С возникновением корпусов появилась возможность подтвердить это положение с помощью анализа очень большого количества контекстов, в которых употребляется то или иное слово. Приведу пример.

Интуитивно понятно, что лексемы *бегемот* и *гиппопотам* употребляются в сходных контекстах, тогда как *лингвистика* и *языкознание*, будучи очень близки, тем не менее встречаются в контрастных контекстах, ср.: *компьютерное языкознание* (2 примера в огромном корпусе ruTenTen[1]) и *сравнительно-историческая лингвистика* (9 примеров). Диаграмма 2 наглядно представляет этот контраст.

Диаграмма 2. Контексты употребления существительных «лингвистика» и «языкознание»

К настоящему моменту разработано несколько моделей дистрибутивной семантики: например, word2vec [Mikolov и др. 2013], ELMo [Peters и др. 2018], BERT [Devlin и др. 2018], в том числе и для русского языка [Kutuzov, Kuzmenko 2017]. Все они собирают контексты

[1] RuTenTen – аннотированный корпус русского языка объемом ок. 18 млрд слов [Jakubíček и др. 2013].

из огромного массива данных и помещают лексемы в многомерное семантическое пространство. Разные модели отличаются разными подходами к обработке контекста, к оптимизации семантических пространств и некоторыми другими параметрами, но суть их остается неизменной: созданный алгоритм без участия исследователя анализирует корпусные данные и классифицирует найденные единицы; чем чаще слова встречаются в одинаковых контекстах, тем семантически ближе они друг к другу. Диаграмма 3 ниже показывает взаимное расположение некоторых городов России, Беларуси и Финляндии в семантическом пространстве, как оно представлено в русскоязычном интернет-корпусе объемом в 2,1 миллиарда слов. Как кажется, эта диаграмма соответствует реальности – как минимум в контексте лингвистических связей между этими городами.

Диаграмма 3. Семантическое пространство существительных – названий городов (по данным https://rusvectores.org; Корпус GeoWAC (2.1 млрд слов), модель fastText)

Очевидный результат этого подхода, который лингвистам еще предстоит теоретически осмыслить, заключается в том, что степень семантической близости лексем можно численно измерить и синонимия оказывается многогранным и градуальным языковым феноменом. Таким образом, мысль Ю. Д. Апресяна о том, что язык избегает точных синонимов, не просто подтверждается, но становится более объемной: лексемы представляются разными наборами векторов в семантическом пространстве, вступающих в многомерные семантические отношения разной степени близости.

В нашем проекте CoCoCo (cococo.cosyco.ru) реализован интегрированный подход, который можно назвать надкорпусным: в нем сложный математический аппарат применен для автоматического решения теоретической задачи – проблемы сочетаемости слов, или идиоматизации в широком смысле. Классификация фразеологизмов, заимствованная В. В. Виноградовым у Шарля Балли и вошедшая в вузовский канон в версии Н. М. Шанского, оказывается, по данным корпусной лингвистики, слишком жесткой, не учитывающей многочисленные переходные случаи и исключения. В частности, выяснилось, что феномен *фразеологических сочетаний* [Шанский 2010], или *полуфразем* [Мельчук, Иорданская 2017], которому в традиционной фразеологии уделяется меньше внимания, оказывается намного сложнее и, как выясняется, фундаментальнее, чем был принято думать.

В корпусной лингвистике *такое «неслучайное* сочетание двух и более лексических единиц» [Ягунова, Пивоварова 2010: 30] принято называть *статистическими коллокациями*. Корпусная лингвистика предлагает несколько методов для их извлечения (для русского языка см. [Хохлова 2008; Pivovarova, Kormacheva, Kopotev 2017]). Эти единицы являются той первичной речевой стихией, из которой возникают и лексические фразеологизмы, и грамматические конструкции. Покажем это на одном примере: сочетания с предлогом без.

Статистические методы позволяют выявить целый ряд устойчивых выражений с предлогом *без* и управляемыми существительными: *без конца, без труда, без сомнения, без памяти, без дела, без исключения, без обиняков* и др., которые являются фразеологизмами и фиксируются в словарях фразеологизмов или эквивалентов слов. Однако список, извлеченный из НКРЯ с помощью статистических инструментов на сайте cococo.cosyco.ru, не ограничивается этими выражениями: он состоит из полутора тысяч существительных. Приведу только первые 50, отсортированные по статистической мере Dice[1].

[1] Подробнее о мерах можно узнать в работах [Evert 2008; Pivovarova, Kormacheva, Kopotev 2017].

Таблица 3. Устойчивые существительные после предлога *без* (по данным подкорпуса со снятой омонимией НКРЯ; сайт https://cococo.cosyco.ru)

словоформа	T-score	Dice
конца	12663	0,02590
труда	24654	0,01830
сомнения	30103	0,01500
памяти	16954	0,01300
исключения	30072	0,01110
дела	31168	0,01010
шапки	20210	0,01010
остатка	14001	0,00955
денег	18019	0,00926
вести	43221	0,00886
участия	44382	0,00840
учета	31868	0,00779
помощи	29677	0,00742
головы	27485	0,00716
слов	27485	0,00713
оглядки	21276	0,00694
устали	21276	0,00693
права	11780	0,00598
очереди	46844	0,00597
малого	45017	0,00594
следа	44652	0,00589
колебаний	44652	0,00589
улыбки	44652	0,00587
разрешения	44287	0,00582
сна	44231	0,00580
толку	35125	0,00518
проблем	33298	0,00502
свободы	33298	0,00501
внимания	25993	0,00453
окон	25263	0,00451
причины	24898	0,00447
ума	21610	0,00427
нужды	20880	0,00422
движения	20515	0,00419
сознания	19784	0,00415
надежды	19784	0,00415
преувеличения	16862	0,00396
погон	16862	0,00396
злобы	16132	0,00393
шума	15766	0,00391
работы	44289	0,00386
тени	13210	0,00376
хлеба	12479	0,00372
умолку	11383	0,00364
четверти	47178	0,00360
удовольствия	46447	0,00356
применения	46082	0,00354
ответа	44986	0,00350
ведома	42430	0,00331
галстука	42064	0,00330

Уже беглый взгляд на полный список позволяет заметить определенные закономерности. Первая – очевидная: предлог *без* управляет родительным и вторым родительным (в таблице выделен жирным) падежами. Более внимательный взгляд позволяет выделить лексико-семантические группы, которые образуют несколько групп, представленных десятком существительных в каждой группе:

БЕЗ + [одежда или деталь одежды]: *галстуков, манжет, капюшона, кителя, складок, пуговиц, халата, рубашки, чулок, перчатки, шали, юбки, рукавов, пальто, брюк, пиджака*;

БЕЗ + [промежуток времени, занятый неосновным действием]: *перерыва, выходных, отдыха, обеда, завтрака, вечера, антракта, тренировки*;

БЕЗ + [добавленный или удаленный фрагмент текста]: *исключения, изъятий, приписок, искажений, помарок, потерь, купюр*;

БЕЗ + [формат речи]: *толку, разговоров, расспросов, споров, объяснений, комментариев*.

Особого внимания заслуживает тот факт, что первые слова в каждой группе **одновременно** являются примерами конструкций и самостоятельными лексическими фразеологизмами: *без галстуков* ‘неформально’, *без перерыва* ‘усердно’, *без исключения* ‘эмфаза всеобщности’, *без разговоров* ‘не споря, без возражений’. В том же списке можно найти и фразеологизм *без толку*, уже потерявший связь с исходной конструкцией *без* + [формат речи], с которой он был, по-видимому, диахронически связан. Таким образом, фразеологические единицы не являются полностью изолированными, идиосинкратическими – они опираются на свободные конструкции, причем не только в диахронии, но и в синхронии.

Важно отметить, что вероятность сочетаемостных возможностей различных лексем варьируется от обязательности грамматической и/или лексической связи до полного запрета на сочетание с градуальным переходом между этими экстремумами. В этом смысле, грамматические правила и лексические ограничения являются частным случаем общей лексико-грамматической вероятностной модели,

в которой каждый установленный в процессе порождения речи параметр активирует набор вероятностных ограничений для последующих знаков – это в равной степени касается и грамматических, и лексических характеристик высказывания. Другими словами, произнеся слово, говорящий вызывает набор ограничений, который прямо коррелируют с частотностью употребления. Например, после предлога *без* вероятность появления существительного превышает вероятность появления глагола в 4000 раз; появление любого из трех родов существительного равновероятно; вероятность появления генитива примерно в 1000 раз выше, чем номинатива, и в 140 раз выше, чем второго родильного; наконец, вероятность появления словоформы *конца* в 1000 раз выше, чем *года*. Для контраста добавим, что вероятность появления, например, словоформы *в* или формы прошедшего времени в том же контексте стремится к нулю. Именно в этой сложной комбинации вероятного и невероятного рождается высказывание.

Важным результатом подобного рода исследований является не просто, так сказать, переход качества в количество и появление в лингвистических статьях таблиц и графиков. Принципиальным становится создание новой, многомерной модели языка, которая не укладывается в контрастное представление в формате грамматического и, отдельно, лексикографического описаний. Интегральное описание языка должно представлять собой базу данных, в которой одновременно описываются грамматические и лексические связи и степень их обязательности (вероятности) для каждого случая. В качестве первого, несовершенного, приближения к этому можно указать на базу данных cococo.cosyco.ru.

3. Вклад корпусной лингвистики в теорию языка

Может показаться, что корпусная лингвистика сводится к набору более или менее удобных инструментов для поиска примеров. В таком варианте корпусная лингвистика тоже представлена в современных исследованиях, но этим она далеко не ограничивается. За десятилетия своего существования корпусная лингвистика накопила достаточно наблюдений, которые подтверждают, уточняют и опровергают наши представления о языке. Ниже я обсужу те находки и идеи, вклад которых в общую теорию языка представляется мне наиболее весомым.

В 1991 году один из пионеров корпусной лингвистики Дж. Синклер сформулировал принцип идиоматичности (*idiom principle*):

> …a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analyzable into segments [Sinclair 1991: 110].

> …говорящий имеет в своем распоряжении большое число полуоформленных фраз, конституирующих однократный выбор, несмотря на то, что при анализе их можно разбить на сегменты.

К настоящему моменту появилось множество подтверждений того, что наша речь изобилует штампами, мы не порождаем (в генеративистском смысле этого слова) высказывание каждый раз с самого глубинного уровня. Безусловно, длинные синтаксические связи требуют грамматической поддержки на уровне правил, однако на уровне фразы сочетаемость единиц оказывается в гораздо большей степени идиоматизированной.

Линейная развертываемость речи играет огромную роль в производстве и восприятии языка: в реальности, сказав А, мы скорее скажем Й, чем Б. Это значит, что в потоке речи каждый произведенный языковой знак снижает энтропию, повышая тем самым предсказуемость высказывания. Синтагматические отношения, таким образом, играют более фундаментальную роль, чем это считалось ранее [Hunston, Francis 2000; Sinclair 2000].

Структуралистское представление об иерархической структуре языка постоянно сталкивалось с фактами, которые трудно расположить на ригидной шкале языковой системы. Они часто обозначались как исключения из правил, маргинализуясь в сфере фразеологии. Корпусная лингвистика и тесно связанные с ней модели языка, основанные на употреблении [Barlow, Kemmer 2000; Langacker 2010], и Грамматика конструкций [Рахилина 2010;

Goldberg 2006; Stefanowitsch, Gries 2003] продемонстрировали отсутствие границ между уровнями и их проницаемость.

Для обобщения новых явлений было предложено объяснение, согласно которому при порождении высказывания говорящий не выполняет последовательные операции от глубинных структур к фонетической реализации на поверхностном уровне, а постоянно выбирает между несколькими возможностями построения высказывания, не иерархизируя, а приоретизируя доступный языковой материал без разделения на уровни. Эта концепция получила название «конкурирующей мотивации» (*compiting motivations*; см. [Du Bois 1985; MacWhinney, Malchukov, Moravcsik 2014]).

Таким образом, в паре «язык-речь» роль синтагматических, речевых связей существенно возрастает. Речевая деятельность становится не просто первичной по отношению к языку – стирается граница между правилами и их реализацией, между языком и речью. Адекватным описанием оказывается не моделирование правил взаимодействия языковых единиц, разделенных на уровни, а описание всех – и индивидуальных, и самых общих – параметров употребления, представляющих собой единый континуум, в котором разделение на язык и речь является предельно условным.

Литература

Величко, А. В. Предложения фразеологизированной структуры в русском языке. Структурно-семантическое и функционально-коммуникативное исследование / А. В. Величко. – Москва : МАКС Пресс, 2016.

Влахов, А. В. Причастия будущего времени в русском языке : выпускная квалификационная работа бакалавра филологии / Влахов А. В. – СПб. : СПбГУ, 2010.

Копотев, М. В. Введение в корпусную лингвистику : учебное пособие для студентов филологических и лингвистических специальностей университетов / М. В. Копотев. – Praha : Animedia Company, 2014.

Материалы для проекта корпусного описания русской грамматики. – URL: http://rusgram.ru (дата обращения: 12.05.2021). – Текст : электронный.

Мельчук, И. А. Смысл и сочетаемость в словаре / И. А. Мельчук, Л. Н. Иорданская. – Москва : Языки славянских культур, 2017.

Морозов, Н. А. Лингвистические спектры: средство для отличения плагиатов от истинных произведений того или другого известного автора / Н. А. Морозов. – Петроград : Тип. Имп. Акад. наук, 1916. – 42 с.

Рахилина, Е. В. Лингвистика конструкций / Е. В. Рахилина. – Москва : Азбуковник, 2010.

Хохлова, М. В. Экспериментальная проверка методов выделения коллокаций / М. В. Хохлова // Slavica Helsingiensia. – Хельсинки : Unigrafia, 2008. – С. 343–357.

Черемисина, М. И. Очерки по теории сложного предложения / М. И. Черемисина, Т. А. Колосова. – Новосибирск : Наука, 1987.

Шанский, Н. М. Фразеология современного русского языка / Н. М. Шанский. – Москва : URSS, 2010.

Ягунова, Е. В. Природа коллокаций в русском языке. Опыт автоматического извлечения и классификации на материале новостных текстов / Е. В. Ягунова, Л. М. Пивоварова // Научно-техническая информация. Серия 2. – 2010. – Т. 2. – С. 30–40.

ACTFL proficiency guidelines. – Alexandria, VA, 2012.

Ädel, A. Corpus Compilation / A. Ädel // A Practical Handbook of Corpus Linguistics / ed. by M. Paquot, S. Gries. – New York : Springer, 2020. – P. 3–24.

Anthony, L. AntConc: A learner and classroom friendly, multi-platform corpus analysis toolkit / L. Anthony // Proceedings of IWLeL. – 2004. – P. 7–13.

Barlow, M. Usage-based models of language / M. Barlow, S. Kemmer. – Stanford, CA : Center for the Study of Language and Information, 2000.

Devlin, J. BERT: pre-training of deep bidirectional transformers for language understanding / J. Devlin et al. – Text : electronic // arXiv preprint arXiv:1810.04805. – 2018. – URL: https://arxiv.org/abs/1810.04805 (mode of access: 28.05.2021).

Du Bois, J. W. Competing motivations / J. W. Du Bois // Iconicity in syntax. – 1985. – Vol. 6. – P. 343–365.

Evert, S. Corpora and collocations / S. Evert // Corpus linguistics. An international handbook. – 2008. – Vol. 2. – P. 1212–1248.

Fillmore, C. J. Corpus linguistics or Computer-aided armchair linguistics / C. J. Fillmore // Directions in corpus linguistics / ed. by J. Svartvik. – Berlin ; New York : de Gruyter Mouton, 2011. – P. 35–60.

Firth, J. Papers in Linguistics / J. Firth. – London : Oxford University Press, 1957.

Goldberg, A. E. Constructions at work: The nature of generalization in language / A. E. Goldberg. – London : Oxford University Press, 2006.

Hunston, S. Pattern grammar: A corpus-driven approach to the lexical grammar of English / S. Hunston, G. Francis. – Amsterdam : John Benjamins Publishing, 2000.

Jakubíček, M. The TenTen corpus family / M. Jakubíček et al. // 7th International Corpus Linguistics Conference CL. – 2013. – P. 125–127.

Janda, L. A. Constructions, their families and their neighborhoods: the case of *durak durakom* 'a fool times two' / L. A. Janda, M. V. Kopotev, T. Nesset // Russian Linguistics. – 2020. – P. 1–19.

Kilgarriff, A. The Sketch Engine: ten years on / A. Kilgarriff et al. // Lexicography. – 2014. – Vol. 1, № 1. – P. 7–36.

Kisselev, O. Syntactic complexity measures as indices of language proficiency in writing: focus on heritage learners of Russian / O. Kisselev, A. Klimov, M. Kopotev // Heritage Language Journal. A Special Issue on Heritage Language Complexity. 2021 (в печати).

Kopotev, M. Corpora in Text-Based Russian Studies / M. Kopotev, A. Mustajoki, A. Bonch-Osmolovskaya // The Palgrave Handbook of Digital Russia Studies. – Cham : Palgrave Macmillan, 2021. – P. 299–317.

Kopotev, M. Russian challenges for quantitative research / M. Kopotev, O. Lyashevskaya, A. Mustajoki // Quantitative approaches to the Russian language. – Routledge, 2018. – P. 3–29.

Kutuzov, A. WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models / A. Kutuzov, E. Kuzmenko // International Conference on Analysis of Images, Social Networks and Texts. – 2017. – Vol. 661. – P. 155–161.

Lakoff, G. Women, fire, and dangerous things: What categories reveal about the mind / G. Lakoff. – University of Chicago press, 2008.

Langacker, R. W. A dynamic usage-based model / R. W. Langacker // Grammar and Conceptualization. – Amsterdam : De Gruyter Mouton, 2010. – P. 91–146.

MacWhinney, B. E. Competing motivations in grammar and usage / B. E. MacWhinney, A. E. Malchukov, E. E. Moravcsik. – London : Oxford University Press, 2014.

McEnery, T. Corpus Linguistics: An Introduction / T. McEnery, A. Wilson. – Edinburgh : Edinburgh University Press, 1996.

Mikolov, T. Efficient estimation of word representations in vector space / T. Mikolov et al. – Text : electronic // arXiv preprint arXiv:1301.3781. – 2013. – URL: https://arxiv.org/abs/1301.3781 (mode of access: 28.05.2021).

Nichols, J. Predicate nominals: A partial surface syntax of Russian / J. Nichols. – Los Angeles : Univ. of California Press, 1981.

Peters, M. E. Deep contextualized word representations / M. E. Peters et al. – Text : electronic // arXiv preprint arXiv:1802.05365. – 2018. – URL: https://arxiv.org/abs/1802.05365 (mode of access: 28.05.2021).

Pivovarova, L. Evaluation of collocation extraction methods for the Russian language / L. Pivovarova, D. Kormacheva, M. Kopotev // Quantitative Approaches to the Russian Language. – Routledge, 2017. – P. 137–157.

Scott, M. Developing Wordsmith / M. Scott // International Journal of English Studies. – 2008. – Vol. 8, № 1. – P. 95–106.

Sinclair J. Lexical grammar / J. Sinclair // Naujoji Metodologija. – 2000. – Vol. 24. – P. 191–203.

Sinclair, J. Corpus, concordance, collocation / J. Sinclair. – Oxford University Press, 1991.

Stefanowitsch, A. Collostructions: Investigating the interaction of words and constructions / A. Stefanowitsch, S. T. Gries // International journal of corpus linguistics. – 2003. – Vol. 8, № 2. – P. 209–243.

References

ACTFL proficiency guidelines. (2012). Alexandria, VA.

Ädel, A. (2020). Corpus Compilation. In Paquot, M., Gries, S. (Eds.). *A Practical Handbook of Corpus Linguistics*. New York, Springer, pp. 3–24.

Anthony, L. (2004). AntConc: A Learner and Classroom Friendly, Multi-Platform Corpus Analysis Toolkit. In *Proceedings of IWLeL*, pp. 7–13.

Barlow, M., Kemmer, S. (2000). *Usage-Based Models of Language*. Stanford, CA, Center for the Study of Language and Information.

Cheremisina, M. I., Kolosova, T. A. (1987). *Ocherki po teorii slozhnogo predlozheniya* [Essays on the Theory of Complex Sentences]. Novosibirsk, Nauka.

Devlin, J. et al. (2018). BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In *arXiv preprint arXiv:1810.04805*. URL: https://arxiv.org/abs/1810.04805 (mode of access: 28.05.2021).

Du Bois, J. W. (1985). Competing Motivations. In *Iconicity in syntax*. Vol. 6, pp. 343–365.

Evert, S. (2008). Corpora and Collocations. In *Corpus linguistics. An international handbook*. Vol. 2, pp. 1212–1248.

Fillmore, C. J. (2011). Corpus Linguistics or Computer-Aided Armchair Linguistics. In Svartvik, J. (Ed.). *Directions in corpus linguistics*. Berlin, New York, de Gruyter Mouton, pp. 35–60.

Firth, J. (1957). *Papers in Linguistics*. London, Oxford University Press.

Goldberg, A. E. (2006). *Constructions at Work: The Nature of Generalization in Language*. London, Oxford University Press.

Hunston, S. (2000). *Pattern grammar: A Corpus-Driven Approach to the Lexical Grammar of English*. Amsterdam, John Benjamins Publishing.

Jakubíček, M. et al. (2013). The TenTen Corpus Family. In 7th *International Corpus Linguistics Conference CL 2013*, pp. 125–127.

Janda, L. A., Kopotev, M. V., Nesset, T. (2020). Constructions, Their Families, and Their Neighborhoods: the Case of *durak durakom'a* fool times two'. In *Russian Linguistics*. Vol. 44, pp. 109-27.

Khokhlova, M. V. (2008). Eksperimental'naya proverka metodov vydeleniya kollokatsii [Experimental Evaluation of Collocation Extraction Methods]. In *Slavica Helsingiensia*. Helsinki, Unigrafia, pp. 343–357.

Kilgarriff, A. et al. (2014). The Sketch Engine: Ten Years on. In *Lexicography*. Vol. 1. No. 1, pp. 7–36.

Kisselev, O., Klimov, A., Kopotev, M. (2021). Syntactic Complexity Measures as Indices of Language Proficiency in Writing: Focus on Heritage Learners of Russian. Heritage Language Journal. A Special Issue on Heritage Language Complexity. (In print).

Kopotev, M. V. (2014). *Vvedenie v korpusnuyu lingvistiku* [Introduction to Corpus Linguistics]. Praha, Animedia Company.

Kopotev, M., Lyashevskaya, O., Mustajoki, A. (2018). Russian Challenges for Quantitative Research. In *Quantitative approaches to the Russian language*. Routledge, pp. 3–29.

Kopotev, M., Mustajoki, A., Bonch-Osmolovskaya, A. (2021). Corpora in Text-Based Russian Studies. In *The Palgrave Handbook of Digital Russia Studies*. Cham, Palgrave Macmillan, pp. 299–317.

Kutuzov, A. (2017). WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models. In *International Conference on Analysis of Images, Social Networks and Texts*. Vol. 661, pp. 155–161.

Lakoff, G. (2008). *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. University of Chicago press.

Langacker, R. W. (2010). A Dynamic Usage-Based Model. In *Grammar and Conceptualization*. Amsterdam, de Gruyter Mouton, pp. 91–146.

MacWhinney, B. E., Malchukov, A. E., Moravcsik, E. E. (2014). *Competing Motivations in Grammar and Usage*. London, Oxford University Press.

Materialy dlya proekta korpusnogo opisaniya russkoi grammatiki [Materials to the Description of Corpus-Based Russian Grammar]. URL: http: //rusgram.ru (mode of access: 12.05.2021).

McEnery, T., Wilson, A. (1996). *Corpus Linguistics: An Introduction*. Edinburgh, Edinburgh University Press.

Melchuk, I. A., Iordanskaya, L. N. (2017). *Smysl i sochetaemost' v slovare* [Meaning and Compatibility in the Dictionary]. Moscow, Yazyki slavyanskikh kul'tur.

Mikolov, T. et al. (2013). Efficient Estimation of Word Representations in Vector Space. In *arXiv preprint arXiv:1301.3781*. URL: https://arxiv.org/abs/1301.3781 (mode of access: 28.05.2021).

Morozov, N. A. (1916). *Lingvisticheskie spektry: Sredstvo dlya otlicheniya plagiatov ot istinnykh proizvedenii togo ili drugogo izvestnogo avtora* [Linguistic Spectra: A Tool for Distinguishing Plagiarism from the True Works of One or Another Famous Author]. Petrograd, Tipographiya Imperatorskoi Akademii nauk. 42 p.

Nichols, J. (1981). *Predicate Nominals: A Partial Surface Syntax of Russian*. Los Angeles, Univ. of California Press.

Peters, M. E. et al. (2018). Deep Contextualized Word Representations. In *arXiv preprint arXiv:1802.05365*. URL: https://arxiv.org/abs/1802.05365 (mode of access: 28.05.2021).

Pivovarova, L. (2017). Evaluation of Collocation Extraction Methods for the Russian Language. In *Quantitative Approaches to the Russian Language*. Routledge, pp. 137–157.

Rakhilina, E. V. (2010). *Lingvistika konstruktsii* [Construction Grammar]. Moscow, Azbukovnik, 2010.

Scott, M. (2008). Developing Wordsmith. In *International Journal of English Studies*. Vol. 8. No. 1, pp. 95–106.

Shansky, N. M. (2010). *Frazeologiya sovremennogo russkogo yazyka* [Phraseology of the Modern Russian Language]. Moscow, URSS.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford University Press.

Sinclair, J. (2000). Lexical Grammar. In *Naujoji metodologija*. Vol. 24, pp. 191–203.

Stefanowitsch, A., Gries, S. T. (2003). Collostructions: Investigating the Interaction of Words and Constructions. In *International Journal of Corpus Linguistics*. Vol. 8. No. 2, pp. 209–243.

Velichko, A. V. (2016). *Predlozheniya frazeologizirovannoi struktury v russkom yazyke. Strukturno-semanticheskoe i funktsional'no-kommunikativnoe issledovanie* [Sentences of the Phraseological Structure in the Russian Language. Structural-Semantic and Functional-Communicative Research]. Moscow, MAKS Press.

Vlakhov, A. V. (2010). *Prichastiya budushchego vremeni v russkom yazyke* [Future Tense Participles in Russian]. Vypusknaya kvalifikatsionnaya rabota bakalavra filologii. Saint Petersburg, SPbSU.

Yagunova, E. V., Pivovarova, L. M. (2010). Priroda kollokatsii v russkom yazyke. Opyt avtomaticheskogo izvlecheniya i klassifikatsii na materiale novostnykh tekstov [The Nature of Collocations in the Russian Language. Experience of Automatic Extraction and Classification Based on News Texts]. In *Nauchno-tekhnicheskaya informatsiya. Seriya* 2. Vol. 2, pp. 30-40.

Данные об авторе

Копотев Михаил Вячеславович – PhD, профессор, школа гуманитарных наук и искусств НИУ ВШЭ (Санкт-Петербург, Россия); адъюнкт-профессор, Отделение языков, Хельсинкский университет (Хельсинки, Финляндия).

Адрес: P.O. Box 24 (Unioninkatu 40), 00014 University of Helsinki, Finland.

E-mail: mihail.kopotev@helsinki.fi.

Author's information

Kopotev Mikhail Vyacheslavovich – PhD, Professor, School of Arts and Humanities at the HSE University (Saint Petersburg, Russia); Associate Professor, Department of Languages, University of Helsinki (Helsinki, Finland).

Дата поступления: 31.05.2021; дата публикации: 30.06.2021

Date of receipt: 31.05.2021; date of publication: 30.06.2021