Вестник Томского государственного университета. Филология. 2022. № 80. С. 5–29
Tomsk State University Journal of Philology. 2022. 80. pp. 5–29

ЛИНГВИСТИКА

Научная статья
УДК 81-13
doi: 10.17223/19986645/80/1

Где живут чудовища? Корпусный метод обнаружения англицизмов и их производных в русскоязычном Интернете

Юлия Матвеевна Алюнина[1]

[1] *Российский университет дружбы народов, Москва, Россия, aliunina-yum@rudn.ru*

Аннотация. Описан метод автоматизированного обнаружения английских заимствований и их производных с помощью менеджера корпусов Sketch Engine и его инструмента Keyword, работающего на основе принципа TF-IDF. Пилотное исследование было проведено на материале небольшого количества блоговых текстов о моде (174 213 словоупотреблений – 218 091 токен) с сайта LiveJournal, в которых благодаря применению функции Keyword было выявлено 84 заимствования в сфере моды (4 506 вхождений) и 32 производных (1 194 вхождения).

Ключевые слова: англицизмы, заимствования, поиск англицизмов, корпусная лингвистика, методы корпусного анализа

Источник финансирования: исследование выполнено при поддержке темы № 056118-0-000 на базе Института современных языков, межкультурной коммуникации и миграций филологического факультета Российского университета дружбы народов.

Для цитирования: Алюнина Ю.М. Где живут чудовища? Корпусный метод обнаружения англицизмов и их производных в русскоязычном Интернете // Вестник Томского государственного университета. Филология. 2022. № 80. С. 5–29. doi: 10.17223/19986645/80/1

Original article
doi: 10.17223/19986645/80/1

Where do the wild things live? Corpus method to detect anglicisms and their derivatives on Russian Internet

Yulia M. Alyunina[1]

[1] *Peoples' Friendship University of Russia, Moscow, Russian Federation, aliunina-yum@rudn.ru*

Abstract. Many articles show the results of the study of anglicisms, and we must assume that as long as languages accept anglicisms, their study will remain topical. Nowadays, more and more attention is being paid to the issue of automated detection of English loanwords and their derivatives in different languages. This article de-

scribes the corpus method of detecting English loanwords and their derivatives in Russian fashion blogs by means of corpus manager *Sketch Engine* and its tool *Keyword*, which operates on TF-IDF principle. The relevance of the study is related to the following objectives: to detect the newest anglicisms that have no lexicographic fixation and to determine their number and frequency; to optimize the search of anglicisms and their derivatives; to reduce the human factor in the search of anglicisms and their derivatives. The structure of this article includes, *first*, an explanation of the terms *anglicism* and *derivative* and ways of anglicisms adaptation to Russian; *second*, a description of existing software methods for detecting anglicisms on the Internet (based on neural networks training and the use of *AntConc* corpus manager); *third*, a description of corpus method to detect anglicisms with *Sketch Engine*, which has not been used to search for anglicisms on Russian Internet, and an explanation of key terms necessary to understand the mechanism of the described method. A pilot research was conducted on a small number of fashion blog posts (174,213 words – 218,091 tokens) from *LiveJournal*, in which 84 fashion loanwords (4,506 occurrences) and 32 derivatives (1,194 occurrences) were detected using the *Keyword* function: *bini, bodi, nyud, skini, slipon; zamiksovat', kezhual'shchik, nyudovyy etc.* The pilot study has shown that the use of the *Sketch Engine* contributes to solving the problems of automating the search of anglicisms and their derivatives on Russian Internet. The implementation of the proposed method requires the preliminary preparation of a focus corpus and subsequent keyword analysis. A preliminary preparation implies: (1) selection of texts united by a common topic; (2) manual removal of hidden hyperlinks in the texts if the corpus is not compiled by crawling Internet pages, but by loading texts independently copied from Internet pages; (3) selection of a suitable reference corpus reflecting the colloquial language. Subsequent keyword analysis involves: (1) excluding irrelevant lexical units from the list of keywords; (2) lemmatising anglicisms and their derivatives and lemmatising individual word forms to lemmas where necessary. The proposed method can be applied not only to the search of English loanwords on Russian Internet but also to the texts in other languages covered by *Sketch Engine*. The prospect of further exploration of this method consists in studying the specifics of its use to search for anglicisms and their derivatives in other languages, other thematic areas and also on a larger array of texts.

Keywords: anglicisms, loanwords, automatic detections, corpus linguistics

Financial Support: The study was carried out at the Institute of Modern Languages, Intercultural Communication and Migration, Faculty of Philology, Peoples' Friendship University of Russia, Topic No. 056118-0-000.

For citation: Alyunina, Yu.M. (2022) Where do the wild things live? Corpus method to detect anglicisms and their derivatives on Russian Internet. *Vestnik Tomskogo gosudarstvennogo universiteta. Filologiya – Tomsk State University Journal of Philology.* 80. pp. 5–29. (In Russian). doi: 10.17223/19986645/80/1

Введение

Вопросу изучения английских заимствований в разных языках посвящено множество исследований, и, надо полагать, пока языки осваивают англицизмы, их изучение будет оставаться актуальным. Так, на сегодняшний день увеличивается внимание к проблеме автоматизированного выявления английских заимствований и их производных в интернет-текстах на разных языках [1–3].

Интерес к автоматизированному обнаружению англицизмов и их производных, метафорически именуемых нами чудовищами, объясняется их активным проникновением в язык-реципиент в разных сферах коммуникации, что нередко воспринимается носителями как угроза принимающему языку. Так, в научных и научно-популярных исследованиях сегодня можно встретить мнения о том, что англицизмы есть деградация [4. P. 38], «болезнь, разрушение и упадок» языка-реципиента («sickness, destruction, and demise» [5. P. 34–35]), которые ведут к его «порче» и засорению [6. С. 17], «подмене понятий и потере национального самоопределения» [7]. Популярным стало мнение о необходимости и возможности заменять заимствования существующими в принимающем языке лексическими единицами: «иностранное слово можно, нужно и вовсе не трудно заменять русским» [8. С. 68].

Объективизация представления о современной тенденции к освоению английских слов языком невозможна без комплексного изучения англицизмов и их производных, для чего требуется их массовое выявление, которое закономерным образом ставит такие задачи, как:

– обнаружение новейших англицизмов, не имеющих лексикографической фиксации, и определение их количества и частотности;

– оптимизация процесса поиска англицизмов и их дериватов (увеличение объёмов анализируемых источников и сокращение времени на ознакомление с ними);

– уменьшение влияния человеческого фактора на процесс обнаружения англицизмов и их дериватов (усталость и снижение концентрации внимания при ручном поиске иноязычных слов в больших массивах текстов).

Эти и другие причины стимулируют поиск новых методов и способов автоматизированного выявления слов от английских этимонов[1] в принимающем языке.

В настоящей статье на примере русского языка предлагается описание корпусного метода обнаружения англицизмов и их производных в блогах о моде с помощью корпусного менеджера *Sketch Engine* [11] и его инструмента *Keyword*.

Структура настоящей статьи предполагает, во-первых, пояснение понятий *англицизм* и *производное* и способов адаптации англицизмов к русскому языку; во-вторых, описание существующих программных методов обнаружения англицизмов в Интернете (основанных на обучении нейронных сетей и применении корпусных технологий); в-третьих, описание корпусного метода обнаружения англицизмов, который ранее не применялся для поиска английских заимствований в русскоязычном Интернете, а также пояснение ключевых терминов, необходимых для понимания механизма рассматриваемого метода.

[1] Здесь и далее термин *этимон* употребляется нами вслед за А.И. Дьяковым [9. С. 6] и Е.В. Мариновой [10. С. 213] для обозначения английского слова (*box*), от которого в языке-реципиенте появился англицизм (*бокс*).

Кто такие чудовища?

В отечественном языкознании хрестоматийной считается дефиниция, закреплённая в *Лингвистическом энциклопедическом словаре*, в котором под заимствованием понимается «элемент чужого языка (слово, морфема, синтаксическая конструкция и т.п.), перенесённый из одного языка в другой в результате контактов языковых, а также сам процесс» такого перехода [12. С. 158]. Заимствования из английского языка называются англицизмами (англ. *casual* → *кэжуал;* англ. *shopping* → *шоппинг*), а слова, созданные на их основе с использованием аффиксов языка-реципиента, – производными или дериватами (*кэжуальность, кэжуальщик; шопиться*). В настоящей статье для обобщённого названия англицизмов и их производных будет применяться термин *слова от английских этимонов*.

При заимствовании иноязычные слова адаптируются к принимающему языку. В русском языке выделяют четыре основные формы освоения заимствований: фонетическую, графическую, грамматическую, семантическую [13. С. 102]. Понимание этих форм освоения важно при разработке автоматизированного способа обнаружения английских заимствований в русском языке.

Фонетическое освоение заимствований представляет собой изменение звуковой оболочки иноязычных слов в соответствии с произносительными нормами принимающего языка [14. С. 223] и отсутствие вариативности в их произношении, которое может быть свойственно некоторым словам даже после их вхождения в узус: англ. *jeans* → *джинсы / джинцы;* англ. *discourse* → *дискурс / дискурс*.

Графическое освоение состоит в передаче иностранного слова на письме алфавитными символами языка-реципиента [13. С. 104]. Так, заимствования в русском языке передаются с помощью кириллицы: англ. *casual* → *кэжуал;* англ. *lookbook* → *лукбук*. Неустойчивость графической формы может говорить о том, что заимствование является новым или не до конца освоенным: англ. *total look* → *тотал **лук**, тотал-**лук***; англ. *second-hand* → *секонд**хэнд**, секонд-**хенд***.

Грамматическое освоение заключается в приспособлении иноязычных лексических единиц к грамматике принимающего языка вне зависимости от наличия определённых грамматических категорий в языке-доноре [15. С. 105]. Например, от формы мн.ч. англ. *boot**s*** с финалией *-s* в русский язык заимствована лексема *бут**с***. Последняя в русском языке является формой ед.ч. с нулевым окончанием[1], а форма мн.ч. образуется от неё с добавлением окончания *-ы* – *бутс**ы***. То есть в русском *бут**-с-ы*** формально представлены две формы множественного числа, одна из которых этимологически восходит к англ. *boot**s*** (*-с-* в *бут**с**ы*), а вторая является приобре-

[1] Пример из газетного подкорпуса Национального корпуса русского языка (НКРЯ): *Нападающий «Реала» Криштиану Роналду в четвертый раз стал обладателем приза лучшему бомбардиру европейских чемпионатов «Золотая **бутса**»* [16].

тённой (*-ы* в *бутс**ы***) в результате приспособления англицизма к парадигме словоизменения в русском языке.

Семантическое освоение представляет собой «процесс, в результате которого иноязычное слово входит в систему понятий заимствующего языка» [17]. Это названия понятий и объектов внеязыковой действительности, которые вошли в жизнь носителей языка-реципиента (из англ. *процессор, букмекер, каршеринг* и др.), а также слова, выражающие дополнительные смысловые оттенки имеющихся в принимающем языке эквивалентов (*комфортный – удобный*; *гуглить – искать; бежевый – айвори – нюд*).

В контексте автоматизированного обнаружения англицизмов и их производных в интернет-текстах на русском языке наиболее значимым, с нашей точки зрения, является аспект графического освоения иноязычной лексемы, поскольку рассматриваемые далее механизмы предполагают машинное распознавание слов от английских этимонов по их формальному признаку – графической форме. Таким признаком является, как правило, нетипичное для языка-реципиента сочетание графем в лексической единице, которое отличает новое заимствование (*оверсайз*) или его дериват (*оверсайзный*[1], *оверсайзность*[2]) от исконных слов принимающего языка или давно освоенных заимствований, приспособившихся к парадигме словоизменения языка-реципиента (*свитер, шорты, пуловер, кардиган*). В некоторых случаях на графическую адаптацию большое влияние оказывает фонетическое приспособление нового слова к произносительной норме в принимающем языке. Если в произношении прослеживается вариативность, то и на письме она может сохраняться в виде нестабильности графической формы, повторяющей фонологическую оболочку лексемы (англ. *loafers* → *л**о**феры, л**оу**феры*). Тем не менее общность лексического значения таких вариантов слов, а также общность обозначаемых ими денотатов свидетельствуют о том, что перед нами одна и та же лексическая единица.

Понимание правил грамматического приспособления слов от английских этимонов, в особенности к правилам русского словоизменения, необходимо при ручной проверке результатов их автоматизированного поиска. Эта необходимость связана с тем, что некоторые механизмы автоматизированного обнаружения слов от английских этимонов не предполагают их лемматизации (например, механизм лемматизации отсутствует в менеджере корпусов *AntConc* [19]): частотность словоформ ошибочно вычисляется как частотность отдельных лемм (*лукбук, лукбуке, лукбуках* и др.), что приводит к неверному определению количества и частотности заимствований.

---

[1] *Так вот, чтобы добиться этого трендового эффекта в аутфите, нам вполне подойдут и обычные **оверсайзные** рубашки в клетку виши* [18] (здесь и далее в примерах выделения наши. – *Ю.А.*).

[2] *Я даже не могу толком конкретизировать модель, но отличительными чертами мастхэва являются растянутость, видимая **оверсайзность** и вот эти непонятные принты «как в детстве»* [18].

Возможная сложность машинной лемматизации таких лексических единиц объясняется их нетипичным сочетанием графем.

Понимание правил семантической адаптации английских заимствований и их производных тоже необходимо при ручной проверке результатов автоматизированного поиска новых слов для исключения ошибочной омонимии, которая может сказаться на частотности англицизма. Например, лексема *лук*, заимствованная от англ. *look* в значении *внешность, внешний вид*, является омонимом слову *лук*, которое употребляется в русском языке в значении *овощ*, и слову *лук*, которое употребляется в значении *стрелковое оружие*. То есть абсолютная частотность слова *лук* в любом корпусе без семантической разметки будет вычислена корпусом как частотность графемы *лук* во всех присущих ей формах и значениях.

Таким образом, в контексте проблематики автоматизированного поиска англицизмов и их производных в текстах на русском языке в центре внимания оказывается графическая адаптация слов от английских этимонов, которая сопряжена с фонетическим, грамматическим и семантическим аспектами освоения новых лексем. Понимание механизмов освоения иноязычной лексики русским языком позволяет снизить вероятность ошибочных результатов их автоматизированного поиска. При поиске англицизмов автоматизированными методами в других языках важными могут оказаться другие особенности приспособления новой лексики к языку-реципиенту.

Автоматизированные методы обнаружения англицизмов и их производных в Интернете

Для понимания механизмов работы автоматизированных методов выявления слов от английских этимонов в Интернете требуется краткое введение в их механику. Под автоматизированными методами в данной статье понимаются не приёмы, позволяющие полностью автоматизировать процесс выявления англицизмов и их производных в интернет-текстах, а методы, помогающие ускорить этот процесс благодаря обращению к возможностям корпусной и компьютерной лингвистики. В рамках настоящей статьи мы опишем два метода, которые уже были использованы их авторами для поиска англицизмов и их дериватов: метод, основанный на обучении нейронных сетей на материале русского языка [1–2], и метод [3] с использованием менеджера корпусов *AntConc* [19] для поиска англицизмов в текстах на испанском и датском языках. Хотя корпусные и компьютерные технологии на сегодняшний день имеют широкое применение в анализе естественного языка (см., например, [27, 34]), в основе предлагаемого обзора лежат лишь три публикации [1–3], поскольку в ходе анализа существующих корпусных и компьютерных методов автоматизированного поиска англицизмов были обнаружены только эти исследования.

Перейдём к рассмотрению названных методов.

Нейронные сети: программирование и машинное обучение. Авторы этого метода предлагают алгоритм автоматизированного поиска англицизмов и их производных, который был апробирован на материале 10 млн текстов на русском языке с сайта *LiveJournal* [1. P. 34]. В результате анализа было обнаружено 4 300 слов, из которых примерно 1 150 не имело лексикографической фиксации на момент проведения исследования (2016 г.) [1. P. 36].

Алгоритм обнаружения англицизмов и их производных в русскоязычном Интернете, использованный авторами, не предполагает предварительной ручной обработки текстов [1. P. 32]. Метод реализуется с помощью кода, написанного на Python, и нейросети, обученной на материале словарей англицизмов и русских грамматик. В основе метода лежит гипотеза о том, что большинство англицизмов в текстах на русском языке транслитерируется, сохраняя в определённой степени фонологическую оболочку [1. P. 32; 2. P. 68].

Суть рассматриваемого метода состоит в следующем.

С помощью кода, написанного на языке Python, были проанализированы блоговые статьи *LiveJournal* на русском (10 млн текстов) и на английском (10 млн текстов) языках и выявлены лексические единицы, «одинаковые» в текстах на двух языках [1. P. 3; 2. P. 70]. Под одинаковыми авторы описываемого метода понимают слова, написанные в текстах на английском языке латиницей и встречающиеся в транслитерированном виде в текстах на русском языке на кириллице (*complex – комплекс, module – модуль, bowl – боул*). Правила транслитерации были прописаны в коде на основе ГОСТов [1. P. 33]. Этим правилам была обучена нейронная сеть, которая осуществляла поиск англицизмов и их производных. В результате такого поиска сформировался список «одинаковых» слов.

После обнаружения таких слов имеющиеся тексты на русском языке были исследованы для выявления в них производных от англицизмов. Этот этап анализа осуществлялся с помощью рекуррентной нейронной сети, обученной на алгоритмах CBoW и Skip-Gram правилам словообразования на материале 97 000 слов из словаря WikiDictionary [1. P. 33]. В результате нейросеть выявила не только англицизмы (*контраст, клик* и др.), но и их дериваты, восходящие к общим этимонам (*контрастность, кликнуть* и др.). Полученный список лексических единиц нейросеть сопоставила со словами, зафиксированными в словарях англицизмов [20, 21], что позволило выявить 1 150 лексем из 4 300 обнаруженных слов, не имеющих лексикографической фиксации, но использующихся в блогах.

Авторы отмечают, что достоинством их метода является обнаружение сложных слов (*киберспорт* от англ. *cyber sport*), производных от англицизмов (*ретвитнуть* от англ. *retweet*) и лексем, не зафиксированных словарями [1. P. 36; 2. P. 70]. Основным недостатком авторы метода называют временны́е затраты на его реализацию. Также к недостаткам отнесено иногда неверное соотнесение английского этимона с заимствованием (*клип* якобы от англ. *creep*). Эту ошибку иллюстрируют множественные примеры,

обнаруженные нами после изучения списка англицизмов и их производных, опубликованного авторами на GitHub [1. P. 36; 22]: *Берлин* и *берлинский* (якобы от англ. *Berlin*), *Прага* (якобы от англ. *Prague*), *водка* (якобы от англ. *vodka*), *мыловарня* (якобы от англ. *meal*), *это* (якобы от англ. *at*), *балалайка* (якобы от англ. *balalaika*). Эти слова, безусловно, не являются англицизмами. Данная ошибка означает, что ручная проверка результатов автоматизированного поиска слов от английских этимонов всё же требуется, что может стать весьма трудоёмкой работой, учитывая количество лексических единиц (в данном случае 4 300), которое необходимо проверить вручную.

Рассмотрим ещё один способ выявления английских заимствований в интернет-текстах, основанный на корпусных технологиях.

Корпусный метод: ключевые слова в менеджере корпуса. Один из способов автоматизированного обнаружения слов от английских этимонов связан с использованием менеджера корпусов. Менеджером корпуса называется «программа, предназначенная для управления корпусами текстов: создания корпусов, их редактирования, аннотирования, осуществления поиска в них и т.д.» («a program used to manage text corpora, i.e. to build, edit, annotate and search corpora») [23]. С помощью менеджера корпусов можно создать свой собственный корпус текстов, который отвечает методическим задачам (корпус текстов по специальности для «составления профессиональных лексических минимумов» [24. С. 44]) или исследовательским (корпус на базе сборника *Карелы: модели языковой мобилизации. Сборник материалов и документов* для изучения особенностей языкового регулирования в Карелии [25. P. 52]).

В рамках настоящей статьи внимание обращается на два менеджера корпусов – *AntConc* [19] и *Sketch Engine* [11], которые рассматриваются в аспекте их использования для автоматизации обнаружения слов от английских этимонов с помощью функции *Keyword,* позволяющей выявить ключевые слова в корпусе текстов.

Ключевыми словами в корпусной лингвистике называются слова, которые «чаще встречаются в фокусном корпусе, чем в референтном корпусе» [23]. Это значит, что свойство «быть ключевым» относится не к языку вообще, а к конкретному массиву текстов, в котором ключевые слова выделяются на основании законов математической статистики [24. С. 45] и действуют при сопоставлении фокусного корпуса с референтным. Фокусным корпусом называется корпус, с которым работает исследователь, или корпус, в котором осуществляется поиск. Референтный корпус[1] – это корпус, с которым сопоставляется фокусный корпус для выявления ключевых слов в последнем [23]. Помимо наиболее частотных лексических единиц, к ключевым словам также относятся лексемы, которые встречаются только в фокусном корпусе и не повторяются в референтном. Как правило, объём референтного корпуса больше, чем фокусного, или сопоставим с ним («Typically,

[1] Иногда в русскоязычной научной литературе референтный корпус называется справочным или опорным [24. P. 46].

a reference corpus is larger than or similar in size to the corpus of interest <…>» [26. P. 81]). Больший объём референтного корпуса позволяет ему стать «достоверным образцом того языка, на котором написан изучаемый текст» [24. С. 46], и тем самым исключить из списка ключевых слов лексические единицы, которые относятся к числу наиболее общих в языке, как лексемы *новый, нравиться, люди, быть*, а также служебные слова, местоимения и междометия.

Функция *Keyword* «сравнивает содержание фокусного корпуса с референтным корпусом и определяет, какие слова и фразы являются значимыми для первого на основе их частотности» [27. P. 193]. В менеджере корпуса эта функция работает на основе принципа «TF.IDF (term frequency by inverse document frequency)» [28. P. 240], согласно которому каждому слову в документе присваивается числовое значение или вес («score» или «keyness score» [29]), вычисляемый как отношение частоты слова в фокусном корпусе к обратной частоте в референтном корпусе [26. P. 85; 30. С. 12] по формуле, интегрированной в работу корпусного менеджера [29]:

$$Score = \frac{fpm\ focus + N}{fpm\ ref + N},$$

где *fpm focus* – относительная частотность слова (*frequency per million*) в фокусном (*focus*) корпусе; *fpm ref* – относительная частотность слова в референтном (*ref*) корпусе; *N* – сглаживающий коэффициент («smoothing parameter» [29]), равный единице и необходимый, чтобы избежать деления на ноль, когда рассматриваемое слово не встречается в референтном корпусе, то есть его частотность равна нулю [26. P. 85].

Вес лексем (score), которые часто встречаются в тексте вне зависимости от его тематической или жанровой принадлежности (например, служебные слова, местоимения и др.), приближается к нулю, поскольку такие лексические единицы, как правило, не являются специфическими для определённого типа текстов [28. P. 240]. Высокий вес слова говорит о его специфичности в референтном корпусе, то есть делает его ключевым («words displaying a higher score would be considered more specialized than those associated to a lower or even negative value» [31. P. 91]).

Для выявления ключевых слов в корпусе текстов необходимо априорное представление о том, какие лексемы могут являться в нём ключевыми [26. P. 82], поскольку «свойство слова быть ключевым является текстуальной характеристикой» [24. С. 48]. Это значит, что лексические единицы, попавшие в список ключевых, «являются важными в тексте, так как в них отражена главная идея» [24. С. 48–49], а одним из показателей её значимости оказывается высокая частотность соответствующих лексем. Так, если фокусный корпус состоит из текстов о погоде на русском языке, а референтный корпус представляет русский язык во всём его многообразии, то скорее всего, ключевыми словами в первом будут *дождь, ветер, снег, температура, облачность, солнечно, опускаться, подниматься* и т.п. Также должно быть представление о предполагаемом изменении состава клю-

чевых слов в фокусном корпусе при изменении референтного корпуса [26. P. 82]. Например, если фокусный корпус состоит из русскоязычных текстов о погоде на Аляске, а референтный корпус включает тексты о погоде в Бразилии, то в список ключевых слов фокусного корпуса наверняка попадут лексемы *снег, ветер, субарктический, снежная буря* и не попадут такие единицы, как *тропический ливень, жара, засуха, песчаная буря*.

Возможность использования функции *Keyword* для выявления англицизмов в корпусе текстов обусловлена тем, что они являются новыми единицами языка, которые, как правило, заимствуются для их использования в определённой сфере коммуникации. Это значит, что их частотность в рамках соответствующей сферы значительно выше, чем за её пределами. Так, если фокусный корпус состоит из текстов по экономике, содержащих английские заимствования, а тексты референтного корпуса представляют язык во всём его многообразии, то при применении функции *Keyword* к первому в результате поиска попадут все лексические единицы, которые не встречаются во втором корпусе или имеют в нём низкую частотность, в том числе англицизмы, например: *фьючерс, депорт, консигнация, овердрафт, форфейтинг, трейдинг* [32. С. 68].

Единственный пример использования функции *Keyword* для поиска английских заимствований, обнаруженный нами в ходе изучения вопроса, представлен в тезисах конференции *2015 IEEE International Professional Communication Conference* [33]. Авторы тезисов [3] кратко описывают методологию работы с бесплатным менеджером корпуса *AntConc* [19] для автоматизированного поиска англицизмов в текстах, взятых из финансовых блогов на датском и испанском языках. В рамках их исследования выявление англицизмов происходило в два этапа [3. P. 3]:

1) применение функции *Keyword* к текстам финансовых блогов на датском языке и выявление в результатах поиска заимствований (*daytrading, earnings, gearing, price* и др.);

2) ручной поиск «датских» англицизмов в корпусе текстов на испанском языке для выявления одинаковых заимствований в двух языках (в дат. *cash flow* и в исп. *cash flow* от англ. *cash flow*).

Авторы исследования не уточняют объём корпусов датского и испанского языков, с которыми они работали, не поясняют, какой корпус являлся референтным, не указывают количество обнаруженных англицизмов, но приводят скриншоты, подтверждающие работу с *AntConc* [3. P. 6–7]. Наличие этих скриншотов в публикации и понимание механизма работы функции *Keyword* позволяет сделать вывод о том, что её использование способствует автоматизации обнаружения английских заимствований, но требует ручной проверки результатов машинного поиска.

Описанный способ автоматизированного обнаружения английских заимствований с помощью функции *Keyword* в *AntConc* является более простым в исполнении, чем метод на основе обучения нейронных сетей, поскольку не требует навыков программирования, а только понимания механики работы менеджера корпусов. Однако описание процедуры использова-

ния данного метода, представленное авторами тезисов [3], с нашей точки зрения, является недостаточным для понимания принципов его работы: отсутствие информации о количественных параметрах фокусного и референтного корпусов, о процедуре и критериях выбора текстов для фокусного корпуса, о количестве обнаруженных англицизмов и их производных. Отсутствие этих сведений в статье не позволяет читателю сделать объективных выводов об эффективности корпусного метода выявления англицизмов.

Таким образом, на настоящий момент удалось обнаружить два способа [1–3], которые позволяют автоматизировать процесс обнаружения слов от английских этимонов в принимающем языке. С нашей точки зрения, описание метода обнаружения англицизмов с использованием корпусного менеджера [3], применённого к датскому и испанскому языкам, требует уточнения. Метод с применением приёмов машинного обучения нейронных сетей, выполненный на материале русского языка, видится весьма трудозатратным и времяёмким. При этом ни один их описанных методов не является полностью автоматизированным: *один* требует предварительного написания кода и последующего тщательного анализа результатов поиска, *второй* – предварительной выборки текстов и последующей ручной сортировки ключевых слов.

Учитывая видимые достоинства и недостатки существующих методов автоматизированного обнаружения английских заимствований, а также с опорой на теорию языкового заимствования, в настоящей статье предлагается описание процедуры пилотного исследования для демонстрации возможностей корпусного менеджера *Sketch Engine* выявлять слова от английских этимонов в собственном исследовательском корпусе на русском языке.

Где живут чудовища?
Поиск англицизмов и их производных с помощью *Sketch Engine*

В нашем пилотном исследовании для поиска слов от английских этимонов в русскоязычных интернет-текстах был выбран корпусный менеджер *Sketch Engine* [11], одним из преимуществ которого перед корпусным менеджером *AntConc* является наличие механизма лемматизации. В менеджере корпуса этот механизм необходим для вычисления частотности леммы, а не каждой отдельной словоформы.

Чтобы осуществить автоматизированный поиск англицизмов и их производных в текстах при помощи *Sketch Engine* через функцию *Keyword*, в менеджере корпусов необходимо создать собственный корпус. Для этого существует два способа:

1) загрузка заранее подготовленных текстов, собранных вручную;

2) автоматизированный сбор текстов менеджером корпуса:

– по заданным вручную ключевым словам[1] (минимум трём);

[1] В данном случае ключевыми словами называются слова и словосочетания, которые характеризуют определённую сферу коммуникации или тему. Например, если тре-

– по выбранным URL-адресам (корпус генерируется из текстов, находящихся на выбранных интернет-страницах);

– по сайтам (корпус генерируется из текстов, находящихся на выбранных сайтах).

Второй способ называется краулингом или веб-краулингом (*crawling, web crawling* [26. P. 18; 34. P. 340]) интернет-страниц от англ. *crawling* – *сканирование* или *сбор данных в Интернете* [35]. В ходе создания корпуса методом краулинга из формирующегося корпуса исключаются тексты рекламы, размещённой на интернет-страницах или сайтах, тексты интерфейса (*Домашняя страница, Меню, Личный кабинет* и т.п.), тексты гиперссылок, а также повторяющиеся фрагменты текстов, которые иногда встречаются на разных сайтах.

В нашем случае для автоматизированного обнаружения англицизмов и их производных в русском языке была использована собственная коллекция блоговых текстов о моде, собранных вручную на русскоязычной версии сайта *LiveJournal* [36] и загруженная в *Sketch Engine* в формате doc. Объём фокусного корпуса составляет 174 213 словоупотреблений (218 091 токен) и включает тексты шести блогеров [18, 37–41], написанных в 2014–2018 гг. В *Sketch Engine* наш исследовательский корпус получил название *RuFashBlog*.

В качестве референтного корпуса был использован существующий на *Sketch Engine* корпус *ruTenTen 2011* (14 553 856 113 словоупотреблений – 18 280 486 876 токенов), созданный разработчиками *Sketch Engine* методом краулинга интернет-страниц. Объём референтного корпуса значительно превышает объём фокусного корпуса, что отвечает одному из требований выбора сопоставляемых корпусов для вычленения ключевых слов [26. P. 81].

Источником текстов в сопоставляемых корпусах служит Интернет. Фокусный корпус является тематическим, поскольку его тексты посвящены моде, а референтный корпус представляет общеразговорный язык («general language corpus»[1]), так как в него вошли тексты, не ограниченные с точки зрения принадлежности к определённой теме или сфере коммуникации.

Поскольку корпус *RuFashBlog* содержит тексты о моде, ожидается, что ключевыми словами в нём будут лексические единицы, являющиеся номинациями предметов одежды (*юбка, платье, блузка*), обуви (*туфли, сапоги*), аксессуаров (*сумка, шляпа*), материалов (*замша, замшевый, шёлк, шёлковый*) и др. на русском языке. Также ожидается, что среди этих лексем будут английские заимствования и их производные, которые являются либо высокочастотными в текстах о моде (например, наиболее распространённые названия предметов одежды, как *свитер, кардиган* и *джинсы*), либо но-

буется составить корпус текстов о кулинарии, ключевыми словами могут быть *еда, кулинария, рецепт*.

[1] «A general language corpus is a sample of language taken from a very large population – in the case of a general corpus the population consists of all of the language that people produce during a certain period of time» [26. P. 15].

выми номинациями в области моды и потому употребляются преимущественно в данной сфере коммуникации, редко встречаясь в языке повседневного общения (*аутфит, лоферы, оверсайз, слипоны, тотал-лук* и др.).

После создания корпуса *RuFashBlog* к нему была применена функция *Keyword*, которая в *Sketch Engine* по умолчанию формирует список из 1 000 ключевых слов в фокусном корпусе, сопоставляя его с референтным. Количество единиц в списке ключевых слов можно регулировать вручную, но в рамках настоящего пилотного исследования мы не пользовались этой возможностью.

На рис. 1 представлена первая из двадцати страница результатов поиска после применения функции *Keyword* к *RuFashBlog*.

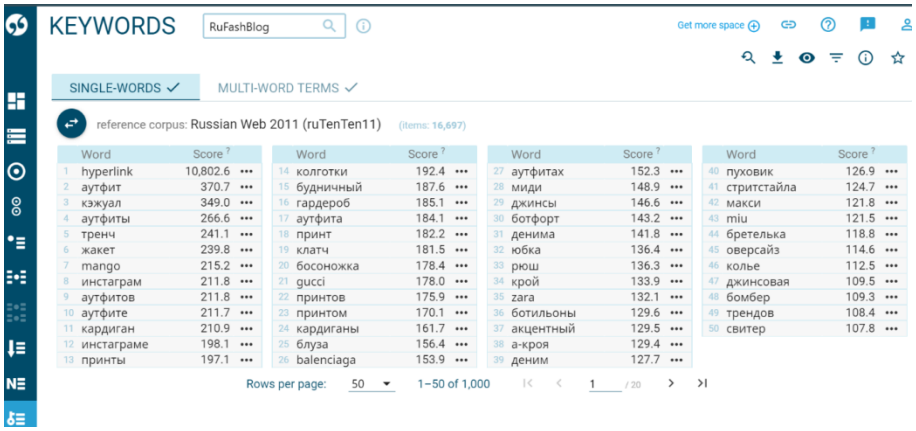


Рис. 1. Скриншот первой страницы результатов поиска ключевых слов в *RuFashBlog*

В полученных результатах внимание привлекают три особенности:

Во-первых, согласно приведённому изображению (рис. 1), самым частотным словом в *RuFashBlog* является *hyperlink* (10 802,6[1]), что связано с технической особенностью распознавания токенов корпусным менеджером: если к слову привязана скрытая гиперссылка, то она распознаётся как отдельный токен *hyperlink*. Это значит, что перед загрузкой собственноручно собранной коллекции текстов в *Sketch Engine* необходимо удалить из текстов все скрытые гиперссылки. Это можно сделать одномоментно во всём документе, применив к нему специальное сочетание клавиш, напри-

[1] Здесь и далее числовое значение, приводимое нами рядом с лексической единицей, означает её вес (score) в *RuFashBlog*, на основании значения которого в *Sketch Engine* формируется список ключевых слов.

Как показано на рис. 1, в *Sketch Engine* десятичные доли отделяются точками, а тысячи – запятыми. В настоящей статье написание чисел адаптировано к системе, принятой в России, то есть в качестве десятичного разделителя используется запятая (1,5 – одна целая пять десятых), а в качестве разделителя для групп разрядов (тысячи, десятки тысяч и т.д.) – пробел (1 000 – одна тысяча).

мер Ctrl+Shift+F9 (на разных устройствах сочетание клавиш для удаления скрытых гиперссылок может отличаться). Такая особенность неверного или нежелательного распознавания знака корпусом создаёт эффект, называемый в корпусной лингвистике шумом (от англ. *noise*) ([42] см. разд. *Лексико-грамматический поиск*).

Во-вторых, в результатах поиска (см. рис. 1) присутствуют нелемматизированные словоформы (*аутфит, аутфиты, аутфитов, аутфите, аутфита; инстаграм, инстаграме* и др.), которых не должно быть, потому что *Sketch Engine* поддерживает автоматическую лемматизацию. Наличие словоформ в нашем списке ключевых слов объясняется тем, что большинство их них – англицизмы и их производные, т.е. лексические единицы, новые для русского языка. Сочетание графем в этих лексических единицах является нетипичным для русского слова, поэтому практически каждая иноязычная словоформа ошибочно распознаётся как самостоятельная лексема. То есть в данном случае в список из 1 000 ключевых слов входят не только слова, но и словоформы. Аналогичная сложность в лемматизации характерна для сложных слов, не являющихся англицизмами. В *RuFashBlog* это, например, словоформы *платье-ночнушку* (14,7)*, платья-ночнушки* (14,7) и *юбка-карандаш* (22,0)*, юбками-карандаш* (14,7)*, юбками-платьями* (14,7)*, юбки-карандаш* (23,5)*, юбкой-карандаш* (14,5), которые распознаются как самостоятельные леммы. Такие слова попали в список ключевых, поскольку в общеразговорном языке, представленном в нашем случае референтным корпусом *RuTenTen 2011*, они практически не встречаются.

В-третьих, в списке ключевых слов (см. рис. 1) обнаруживаются лексемы, написанные на латинице: *mango* (215,2), *gucci* (178,0), *balenciaga* (153,9), *zara* (132,1), *miu* (121,5). Появление этих лексических единиц в списке ключевых тоже объясняется высокой частотностью наименований брендов в блогах о моде в сравнении с их частотностью в общеразговорном языке и их нетипичным для русского языка сочетанием графем – они написаны латиницей, а не кириллицей.

Таким образом, первая из перечисленных особенностей создаёт шум в результатах поиска, вторая и третья, помимо создания шума, свидетельствуют о необходимости ручной проверки результатов поиска для выявления искомых англицизмов и их производных.

Ручная проверка первой страницы (рис. 1) ключевых слов позволяет увидеть, что в список искомых англицизмов в сфере моды и их производных попало 16 из 33[1] лексем на русском языке: *аутфит* (от англ. *outfit*), *бомбер* (от англ. *bomber*), *деним* (от англ. *denim*), *джинсы, джинсовая* (от англ. *jeans*), *кардиган* (от англ. *cardigan*), *клатч* (от англ. *clutch*), *кэжуал* (от англ. *casual*), *макси* (от англ. *maxi*), *миди* (от англ. *midi*), *оверсайз* (от англ. *oversize*), *принт* (от англ. *print*), *свитер* (от англ. *sweater*), *стритстайл* (от англ. *street style*), *тренд* (от англ. *trend*), *тренч* (от англ.

[1] В данном случае имеются в виду 33 лексемы на первой странице результатов поиска, а не 50 ключевых слов и словоформ, которые представлены на рис. 1.

trench). То есть практически половина ключевых лексем на первой странице поиска являются англицизмами и их дериватами.

Как было сказано ранее, в *Sketch Engine* список ключевых слов по умолчанию формируется из 1 000 единиц, что усложняет их ручную проверку и сортировку на сайте менеджера корпусов. Для ускорения и частичной автоматизации ручной проверки дальнейшую работу по обнаружению англицизмов и их производных необходимо выполнять в *Excel*, скачав сформировавшийся список ключевых слов в соответствующем формате (эта возможность предусмотрена *Sketch Engine*).

Для выявления англицизмов в сфере моды и их производных в списке ключевых слов в *Excel* требуется выполнить следующий алгоритм действий:

1. Применить функцию *Таблица* к списку ключевых слов с их числовыми значениями для удобства синхронизированной сортировки строк.

2. Отсортировать строки в алфавитном порядке, в результате чего верхние строки будут заняты словами или словоформами, написанными латиницей, что позволит единовременно их исключить из списка ключевых слов. В нашем случае было исключено 229 слов, написанных латиницей (*adidas, armani, asos, chanel, chloe, cors, dutti, fendi, kari, kenzo, lakbi, lamoda, moschino, prada, valentino, wildberries, zara* и др.).

3. Исключить слова и словоформы, которые не отвечают цели поиска. В нашем случае такими единицами стали:

– англицизмы и их производные, которые не относятся к сфере моды (*кликабельный, лайфхак, лого, экспресс-пост, экспресс-текст* и др.);

– слова и словоформы, которые относятся к сфере моды, но не являются англицизмами и их производными (*балетками, балетки, балеток, бант, воротник, вязаный, капюшон* и др.);

– слова, не имеющие отношения к сфере моды (*как-будто, любимчик, любительница, цейлонский* и др.);

– просторечия и сленгизмы (*адски, ботан, капец, крайняк* и др.);

– имена собственные и их производные (*инстаграм, инста* и др.; *Айплатов* – фамилия дизайнера; *Винтур* – фамилия главного редактора американского издания журнала *Vogue*; *Честейн* – фамилия актрисы Джессики Честейн).

После исключения нерелевантных слов и словоформ из результатов поиска в нашем списке осталось 250 слов (*джинсовый, тренч, шоппер*) и словоформ (*джинсовая, джинсовой, джинсовую; тренча, тренчами, тренчей; шоппера, шопперы*), которые являются англицизмами и их производными в сфере моды.

4. Лемматизировать словоформы для удобства подсчёта лексем-англицизмов и лексем-производных от англицизмов. Частичное автоматизирование лемматизации возможно с помощью инструментария *Excel*, а именно его функции *Найти и заменить*, в которой в поисковой строке *Найти* необходимо указать основу лексемы со знаком * вместо окончания (*аутфит**), а в окне *Заменить* указать лемму (*аутфит*), на которую необ-

ходимо заменить все словоформы[1]. При такой замене все словоформы в таблице (*аутфит, аутфитом, аутфита, аутфиты*) будут автоматически заменены на словарную форму слова (*аутфит*). Более наглядный пример реализации этой функции представлен в прил. I.

В результате лемматизации и подсчёта выявленных лемм список англицизмов в сфере моды сформировали 84 лексические единицы с количеством вхождений в *RuFashBlog* 4 506 (*бини, боди, бойфренды, нюд, оверсайз, свитшот, скини, слипон, стиль, тренд, тренч, тренчкот, чиносы, шоппер* и др.) и 32 производных с количеством вхождений 1 194 (*замиксовать, кэжуальщик, кэжуальность, миксовать, нюдовый, стилевой, топовый, трендовый* и др.) (рис. 2). Полный список обнаруженных слов от английских этимонов приведён в прил. II.

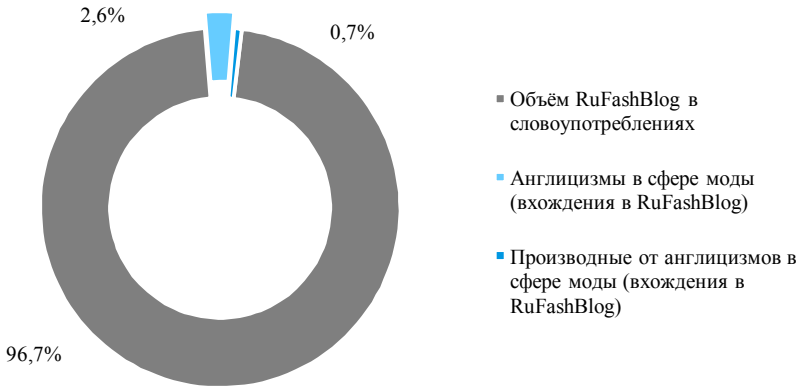


Рис. 2. Процентное соотношение англицизмов и их производных в сфере моды в корпусе RuFashBlog

Описанный способ автоматизировать процедуру обнаружения англицизмов и их производных в русскоязычном Интернете с помощью *Sketch Engine* имеет ряд преимуществ перед описанными ранее методами на основе обучения нейронных сетей и с использованием менеджера *AntConc*. С нашей точки зрения, к таким преимуществам относятся:

– отсутствие в результатах поиска (в списке ключевых слов) ошибочных англицизмов и их производных, таких, например, как *Берлин* и *берлинский*, *балалайка* и *мыловарня*;

– небольшие временные затраты, которые зависят от объёма обрабатываемых текстов (в рамках работы над данной статьёй с момента загрузки текстов в менеджер корпуса до составления финального списка англицизмов в сфере моды и их производных прошло примерно 4 часа);

– обнаружение новейших англицизмов, некоторые из которых не зафиксированы НКРЯ и не имеют лексикографической фиксации в искомом

[1] Знак (*) является регулярным выражением, который на языке поисковых запросов означает «Ни одного или несколько любых символов» [43].

значении (в нашем случае – наименование объекта моды), например *гёрлфренды* и *бойфренды* (фасоны джинсов);

– наличие механизма лемматизации в *Sketch Engine*, который если и не работает в случае с англицизмами и их производными, но всё же объединяет все словоформы лексической единицы (*юбкой, юбки* и др.) в лемму (*юбка*) и тем самым сокращает излишний шум в результатах поиска.

Безусловно, можно выделить и некоторые недостатки в процедуре поиска слов от английских этимонов с помощью *Sketch Engine*, которыми являются:

– возможность выявления англицизмов и их производных только в текстах, принадлежащих определённой тематике, что обусловливает необходимость предварительной подготовки фокусного корпуса;

– потеря англицизмов, которые являются омонимами лексических единиц, частотных в общеразговорном языке, и которые можно обнаружить только в ходе прочтения текстов из фокусного корпуса, например лексемы *лук* (в общеразговорном языке – овощ, стрелковое оружие, как англицизм – внешний вид), *бокс* (вид спорта и маленькая женская сумочка), *финиш* (конечный пункт дистанции и финальное покрытие обуви), *челси* (спортивный клуб и вид обуви).

Названные достоинства и недостатки описанного метода автоматизированного выявления англицизмов и их производных с помощью *Sketch Engine*, с нашей точки зрения, говорят о том, что на настоящий момент ни один из способов не является полностью автоматизированным, потому что требуется либо предварительная ручная работа, либо последующий ручной анализ результатов поиска, либо и то и другое. Тем не менее метод с использованием корпусного менеджера *Sketch Engine* позволяет нивелировать некоторые недостатки описанных способов автоматизации процесса поиска английских заимствований и их производных, в частности в русском языке.

Заключение

Таким образом, выполненное пилотное исследование на материале небольшого корпуса блоговых текстов о моде показало, что использование менеджера корпусов *Sketch Engine* способствует решению задач по автоматизации поиска англицизмов и их производных в русскоязычном Интернете. Для реализации описанного метода требуется предварительная подготовка корпуса текстов и последующий анализ ключевых слов.

Предварительная подготовка корпуса текстов предполагает:

– выбор текстов, объединённых общей тематикой (мода), что, во-первых, позволяет предвидеть возможные ключевые слова (*боди, деним, нюд, оксфорды, ретро*) или их рубрики (названия стилей, цветов, предметов одежды и обуви) и, во-вторых, обеспечивает высокую частотность лексем в рамках текстов на выбранную тематику, что делает их ключевыми в фокусном корпусе;

– ручное удаление гиперссылок в текстах, если корпус составляется не методом краулинга интернет-страниц, а посредством загрузки текстов, самостоятельно скопированных с интернет-страниц;

– выбор подходящего референтного корпуса, отражающего общеразговорный язык, современный языку текстов в фокусном корпусе.

Последующий анализ ключевых слов включает:

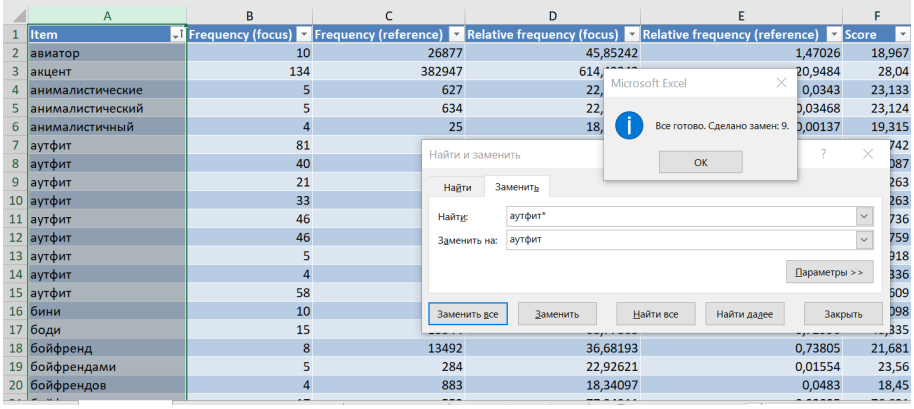
– исключение нерелевантных лексических единиц из списка ключевых слов;

– лемматизацию слов от английских этимонов и сведение к лемме отдельных словоформ в тех случаях, где это необходимо.

Описанный метод может быть использован не только для поиска английских заимствований в русскоязычном Интернете, но и в текстах на других языка, работа с которыми предусмотрена возможностями *Sketch Engine*. Изучение специфики применения данного метода для поиска англицизмов и их производных в других языках, других тематических сферах, а также на большем массиве текстов составляет перспективу его дальнейшего исследования.

Приложение I

Пример использования *Excel* для автоматической лемматизации



	A	B	C	D	E	F
1	Item	Frequency (focus)	Frequency (reference)	Relative frequency (focus)	Relative frequency (reference)	Score
2	авиатор	10	26877	45,85242	1,47026	18,967
3	акцент	134	382947	614,	20,9484	28,04
4	анималистические	5	627	22,	0,0343	23,133
5	анималистический	5	634	22,	0,03468	23,124
6	анималистичный	4	25	18,	0,00137	19,315
7	аутфит	81				742
8	аутфит	40				087
9	аутфит	21				263
10	аутфит	33				263
11	аутфит	46				736
12	аутфит	46				759
13	аутфит	5				918
14	аутфит	4				836
15	аутфит	58				509
16	бини	10				098
17	боди	15				335
18	бойфренд	8	13492	36,68193	0,73805	21,681
19	бойфрендами	5	284	22,92621	0,01554	23,56
20	бойфрендов	4	883	18,34097	0,0483	18,45

Приложение II

Список выявленных англицизмов в сфере моды и их производных

Англицизмы			Производные
1. Авиаторы	33. Мартенсы	65. Топ	1. Акцентирование
2. Акцент	34. Мастхэв	66. Тотал	2. Акцентный
3. Анималистичный[1]	35. Металлик	67. Тотал-лук	3. Базовость

[1] Лексема *анималистичный* рассматривается нами как англицизм с морфологическим оформлением, а не как производное, поскольку для формирования производного от англицизма необходима производящая основа – английское заимствование. В русском языке такая производящая основа отсутствует (ею могло бы быть слово *анимал* от

Англицизмы			Производные
4. Аутфит	36. Миди[1]	68. Тренд	4. Базовый
5. Бини	37. Микс	69. Тренч	5. Богемный
6. Боди	38. Милитари	70. Тренчкот	6. Джинсовый
7. Бойфренды	39. Минимализм	71. Угги	7. Замиксовать
8. Бомбер	40. Мюли	72. Фэшениста	8. Капроновый
9. Бохо	41. Нейтральный	73. Фэшн	9. Ковбойский
10. Броги	42. Нюд	74. Хаки	10. Коктейльный
11. Гёрлфренды	43. Оверсайз	75. Хантеры	11. Контрастность
12. Гикшик	44. Оксфорды	76. Хит	12. Кроссовки
13. Гламур	45. Пейсли	77. Хобо	13. Кэжуальный
14. Деконструктивизм	46. Пижама	78. Худи	14. Кэжуально
15. Деним	47. Преппи	79. Чиносы	15. Кэжуальность
16. Джемпер	48. Принт	80. Чокер	16. Кэжуальщик
17. Джинсы	49. Ретро	81. Шопер	17. Миксовать
18. Дресс-код	50. Свитер	82. Шопинг	18. Минималистичный
19. Капрон	51. Свитшот	83. Шоппер	19. Нейтральный
20. Кардиган	52. Сет	84. Шорты	20. Некэжуальный
21. Каффы	53. Скини		21. Нюдовый
22. Кеды	54. Слиперы		22. Оверсайзный
23. Клатч	55. Слипоны		23. Оверсайзность
24. Контраст	56. Смартвотч		24. Пижамный
25. Кроп-топ	57. Снуд		25. Стилистический
26. Кросс-боди	58. Стайлинг		26. Стильно
27. Кэжуал	59. Стилизация		27. Стильный
28. Леггинсы	60. Стилист		28. Тимберы
29. Логомания	61. Стилистика		29. Топовый
30. Лоуферы	62. Стиль		30. Трендыбренды
31. Лукбук	63. Стристайл		31. Трендовый
32. Макси[2]	64. Тимберленды		32. Футболка

англ. *animal*, но такого заимствования в русском языке нет), а этимоном лексемы *анималистичный* является английское слово *animal*.

[1] В данном случае имеется в виду употребление лексемы *миди* в блогах в значении, заимствованном из английского языка [44] – *длина* или *одежда длины миди*, например:

• *К **миди** можно подобрать туфли или же снова ботильоны с коротким голенищем* [37].

• *Носите эту вещь с летящими платьями длины **миди** и высокими сапогами, брючными костюмами, джинсами клеш и вельветовыми брюками* [38].

[2] В данном случае имеется в виду употребление лексемы *макси* в блогах значении, заимствованном из английского языка [44] – *длина* или *одежда длины макси*, например:

• *Мы поломали голову, но нашли идеальный фасон платья: **макси** на запах* [18].

• *В случае с плиссе мы ушли от **макси** и дошли до длины миди, складки стали резче и жестче, в цветах появился металлик, а также глубокие насыщенные оттенки, а также появились новые варианты образов и вещей, которые мы носим с такими юбками* [37].

Список источников

1. *Fenogenova A., Kazorin V., Karpov I.* A General Method Applicable to the Search for Anglicisms in Russian Social Network Texts // Proceedings of the Artificial Intelligence and Natural Language AINL FRUCT 2016 Conference. Saint-Petersbourg, 2016. P. 31–36. URL: https://publications.hse.ru/en/chapters/194779964 (дата обращения: 03.01.2022).

2. *Fenogenova A.S., Karpov I., Kazorin V., Lebedev I.V.* Comparative Analysis of Anglicism Distribution in Russian Social Network Texts // Материалы Международной конференции по компьютерной лингвистике и интеллектуальным технологиям «Диалог 2017» : в 2 Т. М. : Изд-во РГГУ. 2017. Т. 1. С. 65–74. URL: https://publications.hse.ru/books/206282438 (дата обращения: 03.01.2022).

3. *Laursen A.L., Mousten B.* Tracking Anglicisms in Domains by the Corpus-Linguistic Method – A Case Study of Financial Language in Stock Blogs and Stock Analyses // 2015 IEEE International Professional Communication Conference (IPCC). Limerick, 2015. P. 1–7. URL: https://ieeexplore.ieee.org/document/7235806?reload=true (дата обращения: 03.01.2022).

4. *Scherling J.* Holistic loanword integration and loanword acceptance. A comparative study of anglicisms in German and Japanese // AAA – Arbeiten aus Anglistik und Amerikanistik. 2013. Vol. 1 (38). P. 37–51.

5. *Onysko A.* Exploring discourse on globalizing English // English Today. 2009. Vol. 25 (1). P. 25–36. URL: https://www.cambridge.org/core/journals/english-today/article/abs/exploring-discourse-on-globalizing-english/F0F61668C8BE8866C857AB45B11991FB (дата обращения: 04.01.2022).

6. *Елифёрова М.* Панталоныфракжилет. М. : Альпина Диджитал, 2020. 157 с.

7. *Артамонов А.* Татьяна Миронова: Переживать надо, когда лингвистика служит сокрытию деяний. URL: https://omiliya.org/article/tatyana-mironova-perezhivat-nado-kogda-lingvistika-sluzhit-sokrytiyu-deyaniy (дата обращения: 03.07.2020).

8. *Галь Н.* Куда же идёт язык? // Слово живое и мёртвое. М. : АСТ, 2017. С. 65–79.

9. *Дьяков А.И.* Словарь английских заимствований русского языка. Новосибирск : Новосибирское книжное издательство, 2010. 588 с.

10. *Маринова Е.В.* Иноязычная лексика современного русского языка. М. : ФЛИНТА : НАУКА, 2012. 288 с.

11. *Sketch Engine*. URL: https://www.sketchengine.eu/ (дата обращения: 04.03.2020).

12. *Лингвистический* энциклопедический словарь / под ред. В.Н. Ярцевой. М. : Советская энциклопедия, 1990. 685 с.

13. *Рахманова Л.И., Суздальцева В.Н.* Современный русский язык : учеб. пособие. М. : Изд-во МГУ, ЧеРо, 1997. 480 с.

14. *Кожевникова Е.И.* Фонетическая и грамматическая ассимиляция галлицизмов в современном английском языке // Известия Уральского государственного университета. Серия 1. Проблемы образования, науки и культуры. 2010. Т. 5 (84). С. 222–225. URL: https://elar.urfu.ru/handle/10995/18868 (дата обращения: 04.01.2022).

15. *Володарская Э.Ф.* Заимствование как отражение русско-английских контактов // Вопросы языкознания. 2002. № 4. С. 96–118. URL: https://vja.ruslang.ru/ru/archive/2002-4/96-118 (дата обращения: 04.01.2022).

16. *Национальный* корпус русского языка. URL: http://www.ruscorpora.ru/new/ (дата обращения: 20.11.2021).

17. *Семантическое* освоение заимствованных слов в русском языке. URL: http://www.textologia.ru/russkiy/leksikologia/slovo-proishozhdenie/semanticheskoe-osvoenie-zaimstvovannih-slov-v-russkom-yazike/1224/?q=463&n=1224 (дата обращения: 14.01.2020).

18. *7 одёжек.* Свой гардероб – свои правила. URL: https://7odezhek.livejournal.com/ (дата обращения: 14.01.2019).

19. *AntConc*. URL: https://www.laurenceanthony.net/software/antconc/ (дата обращения: 19.10.2021).

20. *Дьяков А.И.* Словарь англицизмов русского языка. URL: http://anglicismdictionary.ru/ (дата обращения: 01.05.2022).

21. *Словарь* молодёжного сленга. URL: https://teenslang.su/ (дата обращения: 07.01.2019).

22. *lab533/Anglicisms*. URL: https://github.com/lab533/Anglicisms (дата обращения: 14.01.2020).

23. *Glossary*. Sketch Engine. URL: https://www.sketchengine.eu/guide/glossary/ (дата обращения: 19.04.2021).

24. *Горина О.Г.* Методика и математика ключевых слов // Открытое и дистанционное образование. 2017. Т. 2 (66). С. 44–51. URL: http://journals.tsu.ru//ou/&journal_page=archive&id=1579&article_id=35320 (дата обращения: 23.11.2021).

25. *Moskvitcheva S.* Prototypical Notions of Minority Languages in the Soviet Union and Russia: “Native Language” (rodnoj âzyk) and “National Language” (nacional’nij âzyk) // Minority Languages from Western Europe and Russia: Comparative Approaches and Categorical Configurations / ed. by S. Moskvitcheva, A. Viaut. Cham : Springer International Publishing, 2019. P. 49–67. URL: https://doi.org/10.1007/978-3-030-24340-1_5 (дата обращения: 23.11.2021).

26. *Brezina V.* Statistics in Corpus Linguistics: A Practical Guide. Cambridge : Cambridge University Press, 2018. 314 p.

27. *Thomas J.* Discovering English with Sketch Engine. 2nd ed. New Delhi : Versatile, 2017. 229 p.

28. *Kilgarriff A.* Comparing corpora // International journal of corpus linguistics. 2001. Vol. 6 (1). P. 97–133.

29. *Simple maths.* URL: https://www.sketchengine.eu/documentation/simple-maths/ (дата обращения: 27.06.2021).

30. *Белоусов К.И., Баранов Д.А., Зелянская Н.Л., Пономарёв Н.Ф., Рябинин К.В.* Когнитивно-информационное моделирование социальной реальности: концепты, события, приоритеты // Вестник Томского государственного университета. Филология. 2021. № 72. С. 5–26.

31. *Pérez M.J.M.* Measuring the degree of specialisation of sub-technical legal terms through corpus comparison. A domain-independent method // Terminology. 2016. Vol. 1 (22). P. 80–102.

32. *Яхина Р.Р., Ильдуганова Г.М.* Особенности модификации заимствований англоязычного происхождения на материале экономической и финансовой терминологии // Вестник Вятского государственного университета. 2017. № 5. С. 67–71.

33. 2015 IEEE International Professional Communication Conference (IPCC). URL: https://ieeexplore.ieee.org/xpl/conhome/7210374/proceeding (дата обращения: 29.09.2021).

34. *A Practical Handbook of Corpus Linguistics* / ed. by Paquot M., Gries S.Th. Cham : Springer, 2020. 686 p.

35. *Multitran*. URL: https://www.multitran.com (дата обращения: 20.10.2021).

36. *LiveJournal*. URL: https://www.livejournal.com/ (дата обращения: 11.11.2021).

37. *Стильные заметки, блог о стиле и моде* // LiveJournal. URL: https://upryamka.livejournal.com/ (дата обращения: 03.01.2022).

38. *Lena View* // LiveJournal. URL: https://lena-view.livejournal.com/profile (дата обращения: 03.01.2022).

39. *Блог* визуальных осколков. Иллюстрированный журнал Алексея Наседкина // LiveJournal. URL: https://nasedkin.livejournal.com/ (дата обращения: 03.01.2022).

40. *Дневник* очаровательной киберледи // LiveJournal. URL: https://kibernetika.livejournal.com/367565.html?media (дата обращения: 03.01.2022).

41. *Anything* for a quiet life // LiveJournal. URL: https://olga-srb.livejournal.com/ (дата обращения: 03.01.2022).

42. *Инструкция* для пользователя Национальным корпусом русского языка // Studiorum: Образовательный портал НКРЯ. URL: https://studiorum-ruscorpora.ru/manual/basic/ (дата обращения: 22.03.2021).

43. *VBA Excel.* Регулярные выражения (объекты, свойства, методы) // Время не ждёт. URL: https://vremya-ne-zhdet.ru/vba-excel/regulyarnyye-vyrazheniya/ (дата обращения: 08.01.2022).

44. *Merriam-Webster* Dictionary. URL: https://www.merriam-webster.com/ (дата обращения: 15.09.2021).

References

1. Fenogenova, A., Kazorin, V. & Karpov, I. (2016) A General Method Applicable to the Search for Anglicisms in Russian Social Network Textsю *Proceedings of the Artificial Intelligence and Natural Language AINL FRUCT 2016 Conference*. Saint Petersburg. 10–12 November 2016. Saint Petersburg. pp. 31–36. [Online] Available from: https://publicationshse.ru/en/chapters/194779964 (Accessed: 03.01.2022).

2. Fenogenova, A.S. et al. (2017) Comparative Analysis of Anglicism Distribution in Russian Social Network Texts. Materialy Mezhdunarodnoy konferentsii po komp'yuternoy lingvistike i intellektual'nym tekhnologiyam "Dialog 2017": v 2 t. [Computational Linguistics and Intellectual Technologies. Proceedings of the Annual International Conference "Dialogue" (2017): in 2 vols]. Vol. 1. Moscow: RSUH. pp. 65–74. [Online] Available from: https://publications.hse.ru/books/206282438 (Accessed: 03.01.2022).

3. Laursen, A.L. & Mousten, B. (2015) Tracking Anglicisms in Domains by the Corpus-Linguistic Method – A Case Study of Financial Language in Stock Blogs and Stock Analyses. *2015 IEEE International Professional Communication Conference (IPCC)*. Limerick. pp. 1–7. [Online] Available from: https://ieeexplore.ieee.org/document/7235806?reload=true (Accessed: 03.01.2022).

4. Scherling, J. (2013) Holistic loanword integration and loanword acceptance. A comparative study of anglicisms in German and Japanese. *AAA – Arbeiten aus Anglistik und Amerikanistik*. 1 (38). pp. 37–51.

5. Onysko, A. (2009) Exploring discourse on globalizing English. *English Today.* 25 (1). pp. 25–36. [Online] Available from: https://www.cambridge.org/core/journals/english-today/article/abs/exploring-discourse-on-globalizing-english/F0F61668C8BE8866C857AB45B11991FB (Accessed: 04.01.2022).

6. Eliferova, M. (2020) *Pantalonyfrakzhilet* [Pantaloonsfrockcoatvest]. Moscow: Al'pina Didzhital.

7. Artamonov, A. (2020) *Tat'yana Mironova: Perezhivat' nado, kogda lingvistika sluzhit sokrytiyu deyaniy* [Tatyana Mironova: We Should Worry When Linguistics Serves to Conceal Deeds]. [Online] Available from: https://omiliya.org/article/tatyana-mironova-perezhivat-nado-kogda-lingvistika-sluzhit-sokrytiyu-deyaniy (Accessed: 03.07.2020).

8. Gal', N. (2017) Kuda zhe idet yazyk? [Where Does the Language Go?]. In: *Slovo zhivoe i mertvoe* [A Word Alive and Dead]. Moscow: AST. pp. 65–79.

9. D'yakov, A.I. (2010) *Slovar' angliyskikh zaimstvovaniy russkogo yazyka* [The Dictionary of English Loanwords in the Russian Language]. Novosibirsk: Novosibirskoe knizhnoe izdatel'stvo.

10. Marinova, E.V. (2012) *Inoyazychnaya leksika sovremennogo russkogo yazyka* [Foreign Words in Contemporary Russian]. Moscow: FLINTA: NAUKA.

11. *Sketch Engine*. [Online] Available from: https://www.sketchengine.eu/ (Accessed: 04.03.2020).

12. Yartseva, V.N. (ed.) (1990) *Lingvisticheskiy entsiklopedicheskiy slovar'* [Linguistic Encyclopedic Dictionary]. Moscow: Sovetskaya entsiklopediya.

13. Rakhmanova, L.I. & Suzdal'tseva, V.N. (1997) *Sovremennyy russkiy yazyk: ucheb. posobie* [Modern Russian Language: Students' Book]. Moscow: Moscow State University; CheRo.

14. Kozhevnikova, E.I. (2010) Foneticheskaya i grammaticheskaya assimilyatsiya gallitsizmov v sovremennom angliyskom yazyke [Phonetic and grammar adaptation of gallicisms in modern English]. *Izvestiya Ural'skogo gosudarstvennogo universiteta. Seriya 1. Problemy obrazovaniya, nauki i kul'tury*. 5 (84). pp. 222–225. [Online] Available from: https://elar.urfu.ru/handle/10995/18868 (Accessed: 04.01.2022).

15. Volodarskaya, E.F. (2002) Zaimstvovanie kak otrazhenie russko-angliyskikh kontaktov [Loanwords as a Reflection of Russian and English Contacts]. *Voprosy yazykoznaniya*. 4. pp. 96–118. [Online] Available from: https://vja.ruslang.ru/ru/archive/2002-4/96-118 (Accessed: 04.01.2022).

16. *Russian National Corpus*. [Online] Available from: http://www.ruscorpora.ru/new/ (Accessed: 20.11.2021). (In Russian).

17. Rakhmanova, L.I. & Suzdal'tseva, V.N. (1997) *Semanticheskoe osvoenie zaimstvovannykh slov v russkom yazyke* [Semantic Adaptation of Loanwords in the Russian Language]. [Online] Available from: http://www.textologia.ru/russkiy/leksikologia/slovo-proishozhdenie/semanticheskoe-osvoenie-zaimstvovannih-slov-v-russkom-yazike/1224/?q=463&n=1224 (Accessed: 14.01.2020).

18. 7 odezhek. (2019) Svoy garderob – svoi pravila [Your wardrobe – Your Rules]. *LiveJourna*. [Online] Available from: https://7odezhek.livejournal.com/ (Accessed: 14.01.2019).

19. *AntConc*. [Online] Available from: https://www.laurenceanthony.net/software/antconc/ (Accessed: 19.10.2021).

20. D'yakov, A.I. (2022) *Slovar' anglitsizmov russkogo yazyka* [Dictionary of Anglicisms in the Russian Language]. [Online] Available from: http://anglicismdictionary.ru/ (Accessed: 01.05.2022).

21. Anon. (2019) *Slovar' molodezhnogo slenga* [Dictionary of Youth Slang]. [Online] Available from: https://teenslang.su/ (Accessed: 07.01.2019).

22. *lab533/Anglicisms*. [Online] Available from: https://github.com/lab533/Anglicisms (Accessed: 14.01.2020).

23. Sketch Engine. (2021) *Glossary*. [Online] Available from: https://www.sketchengine.eu/guide/glossary/ (Accessed: 19.04.2021).

24. Gorina, O.G. (2017) Methodology and mathematics of key words. *Otkrytoe i distantsionnoe obrazovanie*. 2 (66). pp. 44–51. [Online] Available from: http://journals.tsu.ru//ou/&journal_page= archive&id=1579&article_id=35320 (Accessed: 23.11.2021). (In Russian).

25. Moskvitcheva, S. (2019) Prototypical Notions of Minority Languages in the Soviet Union and Russia: "Native Language" (rodnoj âzyk) and "National Language" (nacional'nij âzyk). In: *Minority Languages from Western Europe and Russia: Comparative Approaches and Categorical Configurations*. Cham: Springer International Publishing. pp. 49–67. [Online] Available from: https://doi.org/10.1007/978-3-030-24340-1_5 (Accessed: 23.11.2021).

26. Brezina, V. (2018) *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge: Cambridge University Press.

27. Thomas, J. (2017) *Discovering English with Sketch Engine*. 2nd ed. New Delhi: Versatile.

28. Kilgarriff, A. (2001) Comparing corpora. *International Journal of Corpus Linguistics*. 6 (1). pp. 97–133.

29. Sketch Engine. (2021) *Simple maths*. [Online] Available from: https://www.sketchengine.eu/documentation/simple-maths/ (Accessed: 27.06.2021).

30. Belousov, K.I. et al. (2021) Cognitive-information modeling of social reality: Concepts, events, priorities. *Vestnik Tomskogo gosudarstvennogo universiteta. Filologiya – Tomsk State University Journal of Philology*. 72. pp. 5–26. (In Russian). DOI: 10.17223/19986645/72/1

31. Pérez, M.J.M. (2016) Measuring the degree of specialisation of sub-technical legal terms through corpus comparison. A domain-independent method. *Terminology*. 1 (22). pp. 80–102.

32. Yakhina, R.R. & Il'duganova, G.M. (2017) Modification features of english origin borrowings in the material of economic and financial terminology. *Vestnik Vyatskogo gosudarstvennogo universiteta.* 5. pp. 67–71.

33. IPCC. (2015) *2015 IEEE International Professional Communication Conference (IPCC)*. [Online] Available from: https://ieeexplore.ieee.org/xpl/conhome/7210374/proceeding (Accessed: 29.09.2021).

34. Paquot, M. & Gries, S.Th. (eds) (2020) *A Practical Handbook of Corpus Linguistics.* Cham: Springer.

35. *Multitran*. [Online] Available from: https://www.multitran.com (Accessed: 20.10.2021).

36. *LiveJournal.* [Online] Available from: https://www.livejournal.com/ (Accessed: 11.11.2021).

37. LiveJournal. (2022) *Stil'nye zametki, blog o stile i mode* [Stylish Notes, Blog on Style and Fashion]. [Online] Available from: https://upryamka.livejournal.com/ (Accessed: 03.01.2022).

38. LiveJournal. (2022) *Lena View*. [Online] Available from: https://lena-view.livejournal.com/profile (Accessed: 03.01.2022).

39. LiveJournal. (2022) *Blog vizual'nykh oskolkov. Illyustrirovannyy zhurnal Alekseya Nasedkina* [Blog of Visual Fractions. Illustrated Journal of Aleksey Nasedkin]. [Online] Available from: https://nasedkin.livejournal.com/ (Accessed: 03.01.2022).

40. LiveJournal. (2022) *Dnevnik ocharovatel'noy kiberledi* [Diary of Charming Cyberlady]. [Online] Available from: https://kibernetika. livejournal.com/367565.html?media (Accessed: 03.01.2022).

41. LiveJournal. (2020) *Anything for a quiet life.* [Online] Available from: https://olga-srb.livejournal.com/ (Accessed: 03.01.2022).

42. Studiorum. (2022) *Instruktsiya dlya pol'zovatelya Natsional'nym korpusom russkogo yazyka* [Instruction for Russian National Corpus Users]. [Online] Available from: https://studiorum-ruscorpora.ru/manual/basic/ (Accessed: 22.03.2021).

43. Vremya ne zhdet. (2022) *VBA Excel. Regulyarnye vyrazheniya (ob"ekty, svoystva, metody)* [VBA Excel. Regular Expressions (Objects, Properties, Methods)]. [Online] Available from: https://vremya-ne-zhdet.ru/vba-excel/regulyarnyye-vyrazheniya/ (Accessed: 08.01.2022).

44. *Merriam-Webster Dictionary*. [Online] Available from: https://www.merriam-webster.com/ (Accessed: 15.09.2021).

Информация об авторе:

Алюнина Ю.М. – канд. филол. наук, Ph.D. in Lexicology and Multilingual Terminology and Translation, ассистент кафедры иностранных языков филологического факультета, научный сотрудник Научно-образовательного Института современных языков, межкультурной коммуникации и миграций Российского университета дружбы народов (Москва, Россия). E-mail: aliunina-yum@rudn.ru

Автор заявляет об отсутствии конфликта интересов.

Information about the author:
Yu.M. Alyunina, Cand. Sci. (Philology), Ph.D. in Lexicology and Multilingual Terminology and Translation, teaching assistant at the Department of Foreign Languages of Faculty of Philology, research fellow at the Research and Academic Institute of Modern Languages, Intercultural Communication and Migration of Peoples' Friendship University of Russia (Moscow, Russian Federation).

The author declares no conflicts of interests.

Статья поступила в редакцию 14.02.2022;
одобрена после рецензирования 26.06.2022; принята к публикации 16.11.2022.

The article was submitted 14.02.2022;
approved after reviewing 26.06.2022; accepted for publication 16.11.2022.