28 Feb 2021

Lacunae Matrices, Web Scraping. Theses.

Irina V. Ivliyeva
Missouri University of Science and Technology, ivliyeva@mst.edu

Perry Koob

Ивлиева, И.В., Куб, Перри. **Метод веб-извлечения парадигм русских глаголов из электронных словарей и баз данных. Матричная организация лакун, их кодификация и классификация (на материале глаголов звучания).** Рабочие тезисы. Февраль 2021.

I. Вводный тезис. **Система глагольных форм** в русском языке имеет **матричную структуру** (существуют как полные парадигмы глаголов, так и дефектные - типа «*побежу*», «*пропылесошу*»). Для чего необходимы таблицы глагольных форм? Прежде всего, чтобы заполнить «матрицы» (таблицы) каждого глагола всеми имеющимися в русском языке актуальными словоформами.

II. Лакуны и **псевдолакуны.** В русском языке существуют как кодифицированные единицы, так и «невидимые» элементы языковой системы - некодифицированные глаголы. Кодификация может быть частичной, например, когда словари содержат только исходные формы глагола (инфинитив), а приставочные образования данного глагола – дериваты-модификации, в словарях отсутствуют.

Лакунами считаем реально пустующие клетки глагольной парадигмы. Это, по разным причинам, невозможные образования в языке, связанные с различными запретами, а *псевдолакунами* мы именуем «невидимые» элементы языковой системы – это некодифицированные единицы, но имеющие потенциал стать объектом семантизации.

III. **Где находится информация о формах глаголов**. Исходная форма (инфинитив) находится в толковых, орфографических, словообразовательных и грамматических словарях. *Формы глагола* находим в словаре Зализняка и частично в толковых словарях в виде глагольных окончаний: *проахаю, -**ем, -ешь, -ете.***

IV. ЧТО СДЕЛАНО. Таблица 1. Подробрная информация о графах, о строении, краткие выводы, и связь с национальным корпусом рассматриваестя в следующих публикациях:

Koob, P., Ivliyeva, I. *An Assessment of the Propagation of Contemporary Russian Mechanical Phonation Verbs to Online Dictionaries.* Web scraping project by Missouri S&T, IT and ALP departments. [Electronic resource]. January 10, 2020. – Access mode: https://scholarsmine.mst.edu/artlan_phil_facwork/157/

Koob, P., Ivliyeva, I. *Russian National Corpus Web Scraping Project 2019-2020*. [Electronic resource]. February 1, 2020. – Access mode: https://scholarsmine.mst.edu/artlan_phil_facwork/158/

Где опубликованы результаты исследования:

Особенности семантики лакун-модификаций. К идее создания словаря лакун (на материале глаголов звучания) // Интерактивная наука. – 2020. – № 6 (52). – С. 57-70. – ISSN 2414-9411. Access mode: https://scholarsmine.mst.edu/artlan_phil_facwork/163/

V. ЧТО ПЛАНИРУЕМ СДЕЛАТЬ.

1. Разработать **продолжение Таблицы 1.** Какие графы вводим и почему (см. ниже).
2. **Привести доказательную базу** (примеры) того, что неразработанные в толковых словарях (некодифицированные) приставочные глаголы звучания активно используются в речи (деловой и разговорной), художественной литературе и публицистике. То есть слова, которые не отражены в словарях, и недоступны изучающим русский язык, на самом деле фактически, активно живут в речи носителей языка, используются в литературе и других источниках. Кроме того, «невидимые» в лексикографии слова вынужденно не участвуют и в научных разработках, так как большинство научных исследований ведется на базе существующих словарей.

Планируем доказать, что необходимо ввести отсутствующий корпус слов в лексикон, научный оборот и систему русского языка, а также дать ему лексикографическое описание.

В словарной традиции принято считать отсутствующей ту форму глагола, которая **не имеет** зафиксированных примеров своего использования (в словарях, в речи и литературе и т.п.). Данные единицы считаются несуществующими ни в литературном языке, ни в узусе.

Наше исследование ставит целью поиск и дальнейшую семантизацию (лексикографическое описание) подобных глаголов.

VI. **Национальный корпус русского языка** (НКРЯ) **– это основная база** и **источник** поиска примеров, подтверждающих наличие той или иной искомой словоформы (для того чтобы заполнить пустующие клетки матрицы глагола).

VII. Наша задача **максимум** (и **цель** создания программы) – **найти в узусе** (в частности, в базе НКРЯ) недостающие *действующие формы* глаголов звучания и заполнить матричную таблицу по каждому глаголу, выявив тем самым и **систему лакун**, которая является одной из важных составляющих матрицы.

VIII. Для достижения поставленной цели требуется решить две взаимосвязанные задачи:

1. Составить наиболее полную **таблицу** приставочных форм глаголов звучания, включая все времена (наст., буд., прош.), наклонения (изъявительное, повелительное), учитывая дифференциацию по родам (м., ж., ср.), числам (ед., мн.) и лицам (1-е лицо, 2-е лицо, 3-е лицо) и

2. Создать программу (разработать алгоритм) по заполнению **Таблицы 2** формами *будущего времени*, *повелительного наклонения*, *причастными* и *деепричастными* формами глаголов звучания.

IX. **Составление таблицы**. Таблица заполняется всеми актуальными для данного глагола словоформами в двух наклонениях: изъявительном и повелительном (сослагательное наклонение не учитываем, причины см. ниже).

Изъявительное наклонение. Таблица включает формы - ***Время; Число***; ***Род***; ***Лицо***.

Прошедшее время: Я пел, она пела, оно пело, они пели.

Настоящее время: Я пою, ты поешь, он поет, мы поем, вы поете, они поют.

Будущее время: Я (буду петь) *пропою*, ты (будешь петь) *пропоешь*, он (будет петь) *пропоет*, мы (будем петь) *пропоем*, вы (будете петь) *пропоете*, они (будут петь) *пропоют*.

Аналитические формы будущего времени (*буду петь* и т.п.) в таблицу не вносим, так как они повторяют исходную форму глагола (инфинитив), в данном примере – **петь.**

Повелительное наклонение. В таблицу вносим формы – ***Число; Лицо***.

1-е лицо *мн*. числа: давайте **мы** споем, споемте;

2-е лицо *ед*. и *мн*. числа: **ты** спой, **вы** спойте;

3-е лицо *ед*. и *мн*. числа: пусть **он** споет, пусть **они** споют.

В таблицу **НЕ вносим формы сослагательного наклонения:** (число; род): Единственное число: *Он спел бы, она спела бы, оно пелось бы*; множественное число: *Они спели бы*), так как данные формы повторяют формы глаголов прошедшего времени, они уже ранее внесены в Таблицу 1.

Информация по разработке заполнения Таблицы 1 формами прошедшего и настоящего времени глаголов звучания представлена в программе (Май 2020 г.).

«Таблица 1», «Таблица 2» - это условное (рабочее) название таблиц.

X. Информацию по заполнению матричных форм будущего времени и повелительного наклонения, а также причастными и деепричастными формами получаем с помощью базы данных морфологического словаря русского языка https://morfologija.ru/, основанного на материале грамматического словаря русского языка А. А. Зализняка.

В данном морфологическом словаре содержится более 170 тыс. слов и более 5 млн. словоформ грамматического словаря А.А. Зализняка. Грамматический словарь русского языка (словарь Зализняка́) содержит приблизительно 100 тыс. слов русского языка с их полным морфологическим описанием.

Словарь А.А. Зализняка считается словарем полных акцентуированных парадигм, он относится к справочной литературе и предназначен для того, чтобы отразить современное русское словообразование и словоизменение.

Для каждого слова, расположенного в словаре, приводится информация о том, как именно и в каких окончаниях оно может меняться, склоняться или спрягаться, и может ли меняться вообще. Так, например, акцентуированная парадигма глаголов *ахать* и *проахать* в словаре Зализняка выглядит следующим образом:

АХАТЬ

а'хать, а'хаю, а'хаем, а'хаешь, а'хаете, а'хает, а'хают, а'хая, а'хал, а'хала, а'хало, а'хали, а'хай, а'хайте, а'хающий, а'хающая, а'хающее, а'хающие, а'хающего, а'хающей, а'хающего, а'хающих, а'хающему, а'хающей, а'хающему, а'хающим, а'хающий, а'хающую, а'хающее, а'хающие, а'хающего, а'хающую, а'хающее, а'хающих, а'хающим, а'хающей, а'хающею, а'хающим, а'хающими, а'хающем, а'хающей, а'хающем, а'хающих, а'хавший, а'хавшая, а'хавшее, а'хавшие, а'хавшего, а'хавшей, а'хавшего, а'хавших, а'хавшему, а'хавшей, а'хавшему, а'хавшим, а'хавший, а'хавшую, а'хавшее, а'хавшие, а'хавшего, а'хавшую, а'хавшее, а'хавших, а'хавшим, а'хавшей, а'хавшею, а'хавшим, а'хавшими, а'хавшем, а'хавшей, а'хавшем, а'хавших.

ПРОАХАТЬ

проа'хать, проа'хаю, проа'хаем, проа'хаешь, проа'хаете, проа'хает, проа'хают, проа'хая, проа'хал, проа'хала, проа'хало, проа'хали, проа'хай, проа'хайте, проа'хавший, проа'хавшая, проа'хавшее, проа'хавшие, проа'хавшего, проа'хавшей, проа'хавшего, проа'хавших, проа'хавшему, проа'хавшей, проа'хавшему, проа'хавшим, проа'хавший, проа'хавшую, проа'хавшее, проа'хавшие, проа'хавшего, проа'хавшую, проа'хавшее, проа'хавших, проа'хавшим, проа'хавшей, проа'хавшею, проа'хавшим, проа'хавшими, проа'хавшем, проа'хавшей, проа'хавшем, проа'хавших

Каждый глагол русского языка имеет парадигму, аналогичную приведенным выше парадигмам глаголов *ахать* и *проахать*, с той лишь разницей, что какие-то глаголы имеют более полную парадигму своих грамматических форм, а какие-то – неполную, дефектную парадигму - с недостающими элементами.

XI. Наша задача: **извлечь из базы грамматического словаря** А.А. Зализняка те словоформы, которые относятся к глаголам звучания: исходные формы и производные (дериваты), включая все словоформы приставочных и суффиксальных производных (учитывая окончания всех времен, наклонений, родов, лиц) и создать таблицу, которая будет отражать полный набор теоретически возможных форм каждого глагола звучания.

Для поиска словоформ глаголов звучания в словаре Зализняка мы составили исходный список глаголов (см. Таблицу 1, май 2020), который включает примерно 800 единиц. Средняя парадигма русского глагола, по словарю Зализняка, составляет 80-120 единиц (см. вышеприведенную парадигму глаголов *ахать, проахать*). Следовательно, поиск должен вестись теоретически примерно по 96 000 (девяносто шести тысячам) форм $(800\times 120 = 96000)$.

Очевидно, что поиск такого количества слов в ручном режиме сопряжен с огромными затратами времени и не является целесообразным.

С целью автоматизации данного процесса нами (**Ivliyeva, Koob 2020**) была разработана универсальная поисковая программа по методологии Web Scraping (**Приложение 2**), которая позволяет осуществить поиск необходимых сведений в словаре А.А. Зализняка по любому заданному списку слов, любой частеречной принадлежности (глагол, существительное, наречие и т.п.) и самой различной семантики (лексико-семантические группы *движения, чувства, цвета, созидания* и т.п.). Программа предполагает также автоматический перенос данных в **Таблицу** в формате Excel. Процедура перевода результатов поиска из таблицы в формате Excel в формат Microsoft Word описана **в Приложении 3.**

В нашем случае – в Таблицу 2 осуществляется перенос недостающих словоформ будущего времени (табл. 2-2), повелительного наклонения (табл. 2-3), и деепричастные (табл. 2-1) формы глаголов звучания.

ПРИМЕЧАНИЕ: В таблицах 2-1, 2-2, 2-3 серым маркером отмечены **не найденные** формы глаголов-дериватов в инфинитиве (графа word). Они отсутствуют как в словарях, так и в НКРЯ. Серым цветом с двумя гризонтальными штрихами (--) выделены незаполненные клетки таблиц, когда устанавливается отсутствие и словоформ слова (word), которое использовалось при поиске.

Результаты компьютерного поиска на материле изучаемых глаголов **выборочно** проверены дополнительно, вручную. Пробелов и несоответсвий системного характера в методологии и в статистическом представлении результатов не обнаружено.

Табл. 2-1 Деепричастия *сов*. и *несов*. вида

№ п/п	root	word	Деепричастие		Примеры НКРЯ
			сов.вид	**несов. вид**	
0	ахать		ахав, ахавши	ахая	ахав, ахавши – не обнаружено *Лэри, **ахая** и постанывая, свесил со своей полки ноги*
1	ахать	ахнуть	ахнув; ахнувши	--	*В голос **ахнув**, Мира замерла и молча вытянула руку.*
1	ахать	взахать	--	--	--
1	ахать	заахать	заахав; заахавши	--	--
1	ахать	поахать	поахав; поахавши	--	***Поахав** и поохав, и почесав затылки, крестьяне сделали тут же из ветвей деревьев носилки* Поахавши –не обнаружено

1	ахать	поахивать	--	--	--
1	ахать	проахать	проахав; проахавши	--	--
1	ахать	разахаться	разахавшись	--	*Василий Дергушин,* ***разахавшись****, мазал словами и эдак, и так его.*

Результаты поиска словоформ деепричастия в полном объёме представлены **в приложениях 6 (Excel) и 6А (Microsoft Word).**

Таблица 2-2 Формы будущего времени глаголов звучания

№	root	word	ед.число			мн.число		
			1-е л. я	2-е л. ты	3-е л. он/она/оно	1-е л. мы	2-е л. вы	3-е л. они
1	ахать	ахнуть	ахну	ахнешь	ахнет	ахнем	ахнете	ахнут
1	ахать	взахать						
1	ахать	заахать	заахаю	заахаешь	заахает	заахаем	заахаете	заахают
1	ахать	поахать	поахаю	поахаешь	поахает	поахаем	поахаете	поахают
1	ахать	поахивать --	--	--	--	--	--	--
1	ахать	проахать	проахаю	проахаешь	проахает	проахаем	проахаете	проахают
1	ахать	разахаться	разахаюсь	разахаешься	разахается	разахаемся	разахаетесь	разахаются

Результаты поиска словоформ будущего времени в полном объёме представлены **в приложениях 4 (Excel) и 4А (Microsoft Word).**

Таблица 2-3 Формы повелительного наклонения глаголов звучания

№	root	word	Ед. число			Мн. число		
			1-е л. Я	2-е л. Ты	3-е л. Он/она/ оно	1-е л. Мы	2-е л. Вы	3-е л. Они
0	**ахать**	**--**	**--**	**--**	**--**	**--**	**--**	**--**
1	ахать	ахнуть	--	ахни	--	ахнем; ахнемте	ахните	--
1	ахать	взахать	--	--	--	--	--	--
1	ахать	заахать	--	заахай	--	заахаем; заахаемте	заахайте	--
1	ахать	поахать	--	поахай	--	поахаем; поахаемте	поахайте	--
1	ахать	поахивать	--	--	--	--	--	--
1	ахать	проахать	--	проахай	--	проахаем; проахаемте	проахайте	--
1	ахать	разахаться	--	разахайся	--	разахаемся; разахаемтесь	разахайтесь	--

Результаты поиска словоформ повелительного наклонения в полном объёме представлены **в приложениях 5 (Excel) и 5А (Microsoft Word).**

XII. Несмотря на то, что информация, приведенная в словаре А.А. Зализняка, считается в русской лексикографии наиболее полной и убедительной, база словаря так же, как и другие словарные источники, не дает полной картины парадигм многих приставочных глаголов.

Объясняется это тем, что словник словаря А.А. Зализняка восходит (хоть и со значительными изменениями) к «Обратному словарю русского языка» (М.: Советская энциклопедия, 1974), составленному на основе лексики из четырех фундаментальных толковых словарей советского периода (словари Ушакова и Ожегова, а также малый и большой академические словари (МАС и БАС). И, как уже отмечалось выше, современные толковые словари русского языка и словари советского периода не отражают полный состав русских глаголов.

Так, например, в словаре Зализняка не находим таких приставочных дериватов как *покукарекать* и *раскукарекаться*.

XIII. В подобных случаях (когда слово не найдено) дополнительную информацию необходимо получить из базы НКРЯ. И **это следующий этап разработки** поисковой программы, призванной связать базы нескольких сайтов.

XIV. В базе НКРЯ можем найти те недостающие формы, которые примерами доказывают «реальность» псевдолакун. Так, например, найдены нашей программой в НКРЯ глаголы *раскукарекаться* и *покукарекать* в различных формах будущего времени и причастия, что в очередной раз доказывает то, что данные дериваты у глагола *кукарекать* есть и они прекрасно «живут» и используются в языке:

Даже две самые чокнутые, самые прилежные прихожанки не могли бы слушать какой-нибудь магнификат с тем самозабвением, с каким эти две курицы внимали ***раскукарекавшемуся*** *в моей комнате романтическому петушку.* Марина Палей, Дань саламандре (2008);

- *Протянул Кязым песню и замолк, прислушиваясь к тишине. Чегемские петухи вовсю* ***раскукарекались****.* Фазиль Искандер, Сандро из Чегема (1989);
- *Ну, думаю, жаба, ты у меня* ***покукарекаешь*** *еще.* Алексей Иванов, Географ глобус пропил (2002);
- *— Теперь мы вас, ребята, надолго закроем! Похлебаете баланды,* ***покукарекаете*** *«петушками»! Ивана подняли с земли, поволокли к машине.* Андрей Житков, Супермаркет (2000);
- *А просто так петь ради искусства — ему неинтересно. Ему интереснее* ***покукарекать.*** *У Алены естьзнакомый Вова — сосед четырех лет*. Виктория Токарева, Тайна Земли (1964-1994);
- *— Все равно, пусть* ***покукарекают*** *каратели, — отозвался Сергей Ломов.* К. С. Бадигин, Секретгосударственной важности (1974).

XV. В тех случаях, когда искомые формы глагола все же не найдены ни в одной словарной базе и базах цитирования, а языковое чутье носителя языка подсказывает, что форма существует, ее следует искать вручную (пока не разработаны другие алгоритмы поиска) через известные поисковики Яндекс, Google и т.п.

Так, например, в научной записи фольклорной экспедиции вручную был найден глагол *поахивать,* образованный от *ахать* по продуктивной модели (и по одной из модификационных схем синтеза: **ахать** + ***по-…-ива***):

Травникова-то жона говорит,
Жубрикова-то жона говорит:
Спит мой муж во дому,
Спит мой муж на боку,
С перепою он ***поахивает****.* З. И. Власова, Скоморохи и фольклор.

Данный глагол **отсутствовал во всех** приведенных выше таблицах и словарях.

SELECTED RESOURCES

2010. Statistical Parsing of Morphologically Rich Languages (SPMRL) What, How and Whither.

2013. Parsing Morphologically Rich Languages: Introduction to the Special Issue.

2016. The Research and Innovation Action “Quality Translation 21 (QT21)”. This project has received funding from the European Union’s Horizon 2020 program for ICT under grant agreement no. 645452.

2020. Koob, Perry B., and Ivliyeva, Irina V. "Russian National Corpus Web Scraping Project 2019-2020." [Electronic resource]. February 1, 2020. – Access mode: https://scholarsmine.mst.edu/artlan_phil_facwork/158/

2020. Ивлиева, И.В. Особенности семантики лакун-модификаций. К идее создания словаря лакун (на материале глаголов звучания) // Интерактивная наука. – 2020. – № 6 (52). – С. 57-70. – ISSN 2414-9411. Access mode: https://interactive-science.media/ru/article/551671/discussion_platform

DICTIONARIES AND CORPORA

Грамматичческий словарь А.А. Зализняка. The grammatical dictionary of the Russian language A.A. Zalyzniak. Access mode: https://morfologija.ru/.

Национальный корпус русского языка. The Russian National Corpus. Access mode: https://ruscorpora.ru/new/index.html

Национальный корпус русского языка (параллельный перевод, 03.02.2021). Access mode: https://ruscorpora.ru/new/search-multi.html