29 Apr 2024

Russian Verbs of Sound's Web-Scraping Results from the A.A. Zalizniak Grammatical Dictionary and the Russian National Corpus. Multi-Dimensional Scaling Techniques and Visualization Strategies

John Simmons

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

Russian Verbs of Sound's Web-Scraping Results from the A.A. Zalizniak Grammatical Dictionary and the Russian National Corpus. Multi-Dimensional Scaling Techniques and Visualization Strategies.

John Simmons, Computer Science

Advisor: Irina Ivliyeva, Ph.D. Curators' Distinguished Teaching Professor of Russian

The project is funded by the Intelligent Systems Center. Thank you!

I. Terminology (in the order of appearance)

1. Information Extraction, Retrieval
2. Semantic Modification
3. Lacuna
4. Lexical Semantics
5. Linguistic Resources
6. Phonology, Morphology
7. Statistical and Knowledge Based Methods
8. Machine Translation
9. Text Mining

II. Challenges to a typical approach to linguistic data processing

1. What is processing?
2. Historical, or Traditional (manual)
3. Space and time challenges
4. Difficult to organize and systematize

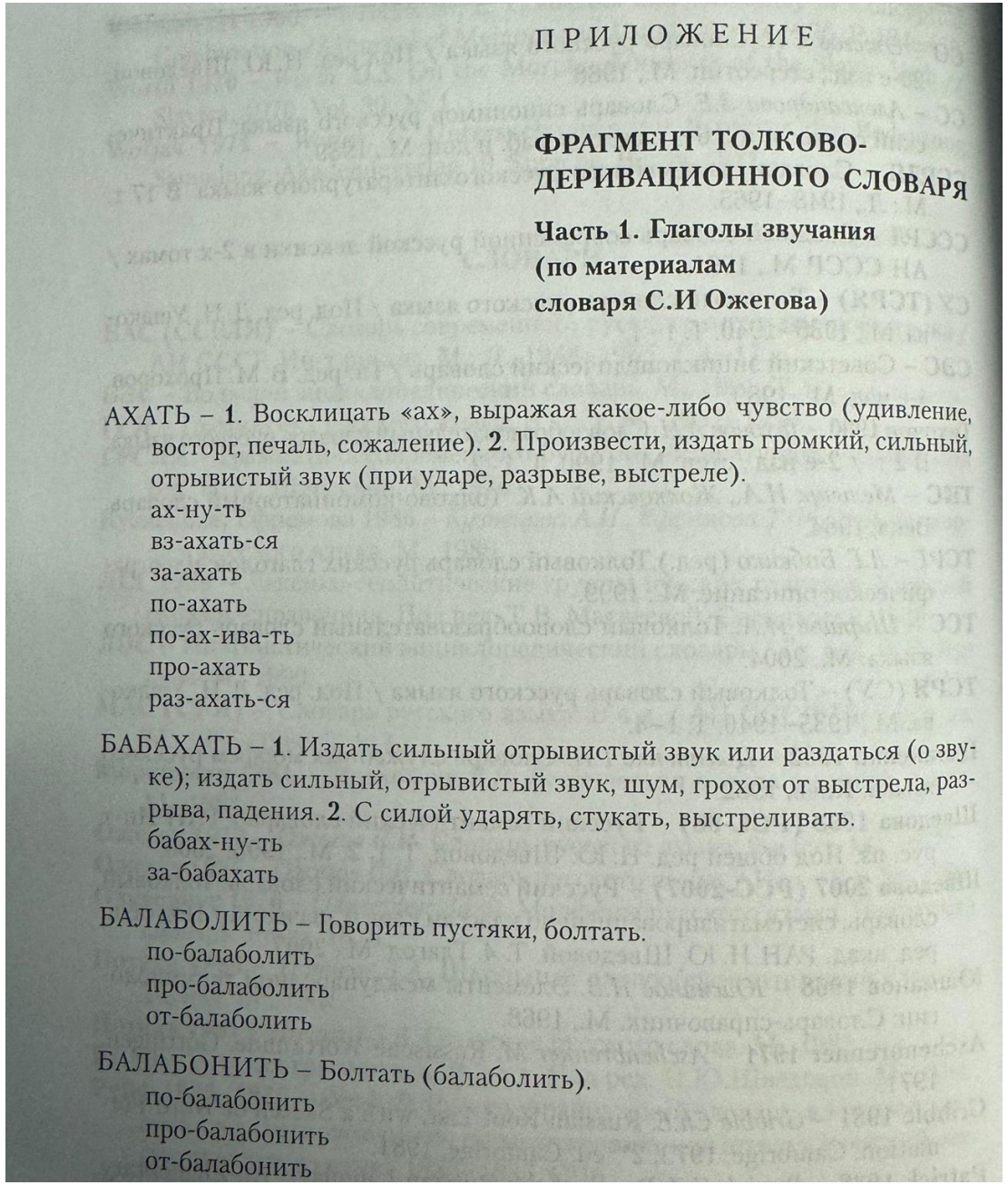
An example page 230 from:

И.В. Ивлиева. Русские глагольные модификации. Опыт составления словаря. Издательство «Элпис».

Москва, 2008. 275 стр.

ISBN 5-902872-24-5

[Translation: Irina V. Ivliyeva. Russian Verbal Modifications. An Experience of a Dictionary Creating. Elpis Press, Moscow, 2008. 275 pages.]



ПРИЛОЖЕНИЕ

ФРАГМЕНТ ТОЛКОВО-ДЕРИВАЦИОННОГО СЛОВАРЯ

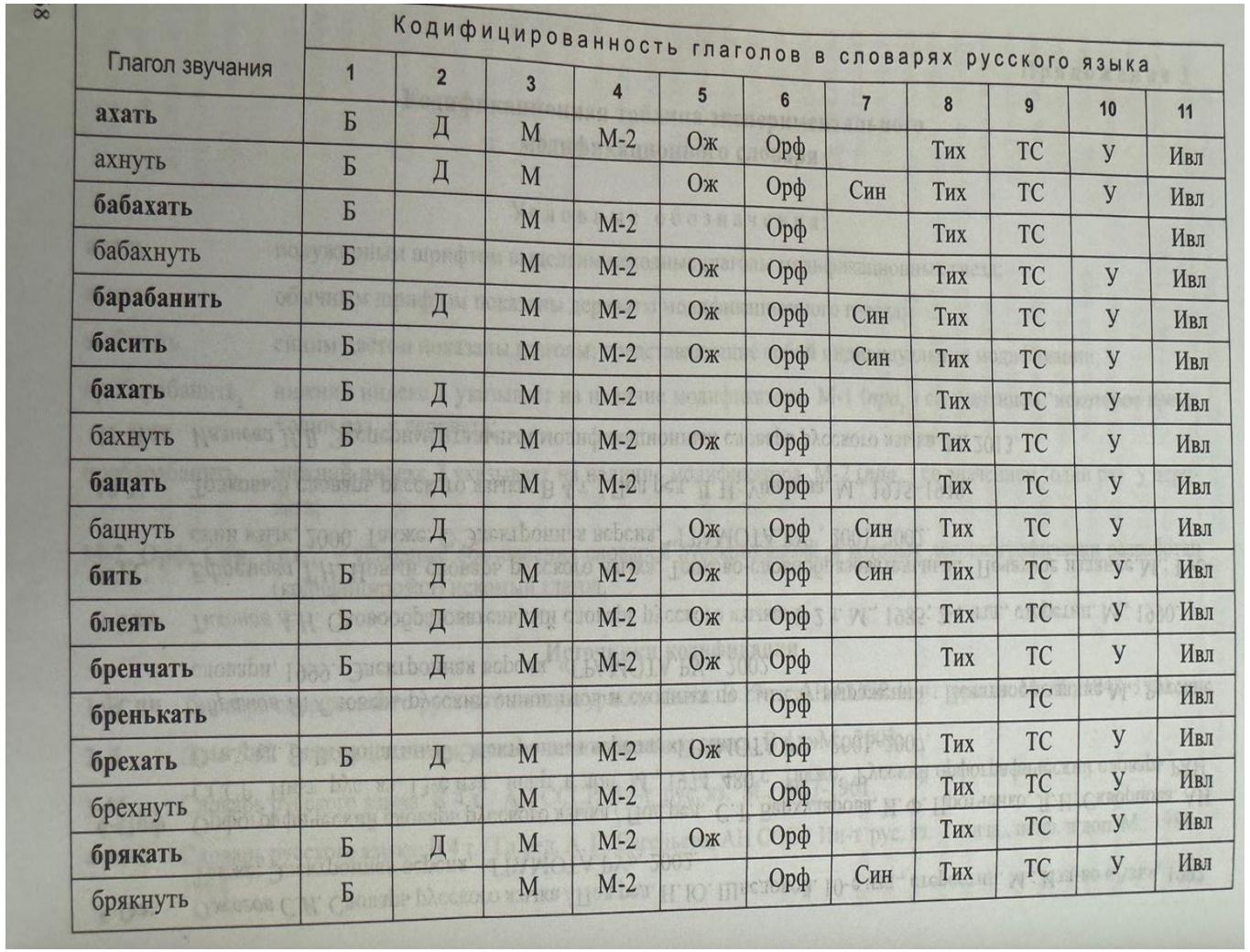
Часть 1. Глаголы звучания (по материалам словаря С.И Ожегова)

АХАТЬ – **1**. Восклицать «ах», выражая какое-либо чувство (удивление, восторг, печаль, сожаление). **2**. Произвести, издать громкий, сильный, отрывистый звук (при ударе, разрыве, выстреле).
ах-ну-ть
вз-ахать-ся
за-ахать
по-ахать
по-ах-ива-ть
про-ахать
раз-ахать-ся

БАБАХАТЬ – **1**. Издать сильный отрывистый звук или раздаться (о звуке); издать сильный, отрывистый звук, шум, грохот от выстрела, разрыва, падения. **2**. С силой ударять, стукать, выстреливать.
бабах-ну-ть
за-бабахать

БАЛАБОЛИТЬ – Говорить пустяки, болтать.
по-балаболить
про-балаболить
от-балаболить

БАЛАБОНИТЬ – Болтать (балаболить).
по-балабонить
про-балабонить
от-балабонить



Глагол звучания	Кодифицированность глаголов в словарях русского языка										
	1	2	3	4	5	6	7	8	9	10	11
ахать	Б	Д	М	М-2	Ож	Орф		Тих	ТС	У	Ивл
ахнуть	Б	Д	М		Ож	Орф	Син	Тих	ТС	У	Ивл
бабахать	Б		М	М-2		Орф		Тих	ТС		Ивл
бабахнуть	Б		М	М-2	Ож	Орф		Тих	ТС	У	Ивл
барабанить	Б	Д	М	М-2	Ож	Орф	Син	Тих	ТС	У	Ивл
басить	Б		М	М-2	Ож	Орф	Син	Тих	ТС	У	Ивл
бахать	Б	Д	М	М-2		Орф		Тих	ТС	У	Ивл
бахнуть	Б	Д	М	М-2	Ож	Орф		Тих	ТС	У	Ивл
бацать		Д	М	М-2		Орф		Тих	ТС	У	Ивл
бацнуть		Д			Ож	Орф	Син	Тих	ТС	У	Ивл
бить	Б	Д	М	М-2	Ож	Орф	Син	Тих	ТС	У	Ивл
блеять	Б	Д	М	М-2	Ож	Орф		Тих	ТС	У	Ивл
бренчать	Б	Д	М	М-2	Ож	Орф		Тих	ТС	У	Ивл
бренькать						Орф			ТС		Ивл
брехать	Б	Д	М	М-2	Ож	Орф		Тих	ТС	У	Ивл
брехнуть			М	М-2		Орф		Тих	ТС	У	Ивл
брякать	Б	Д	М	М-2	Ож	Орф		Тих	ТС	У	Ивл
брякнуть	Б		М	М-2		Орф	Син	Тих	ТС	У	Ивл

An example page 368 from:

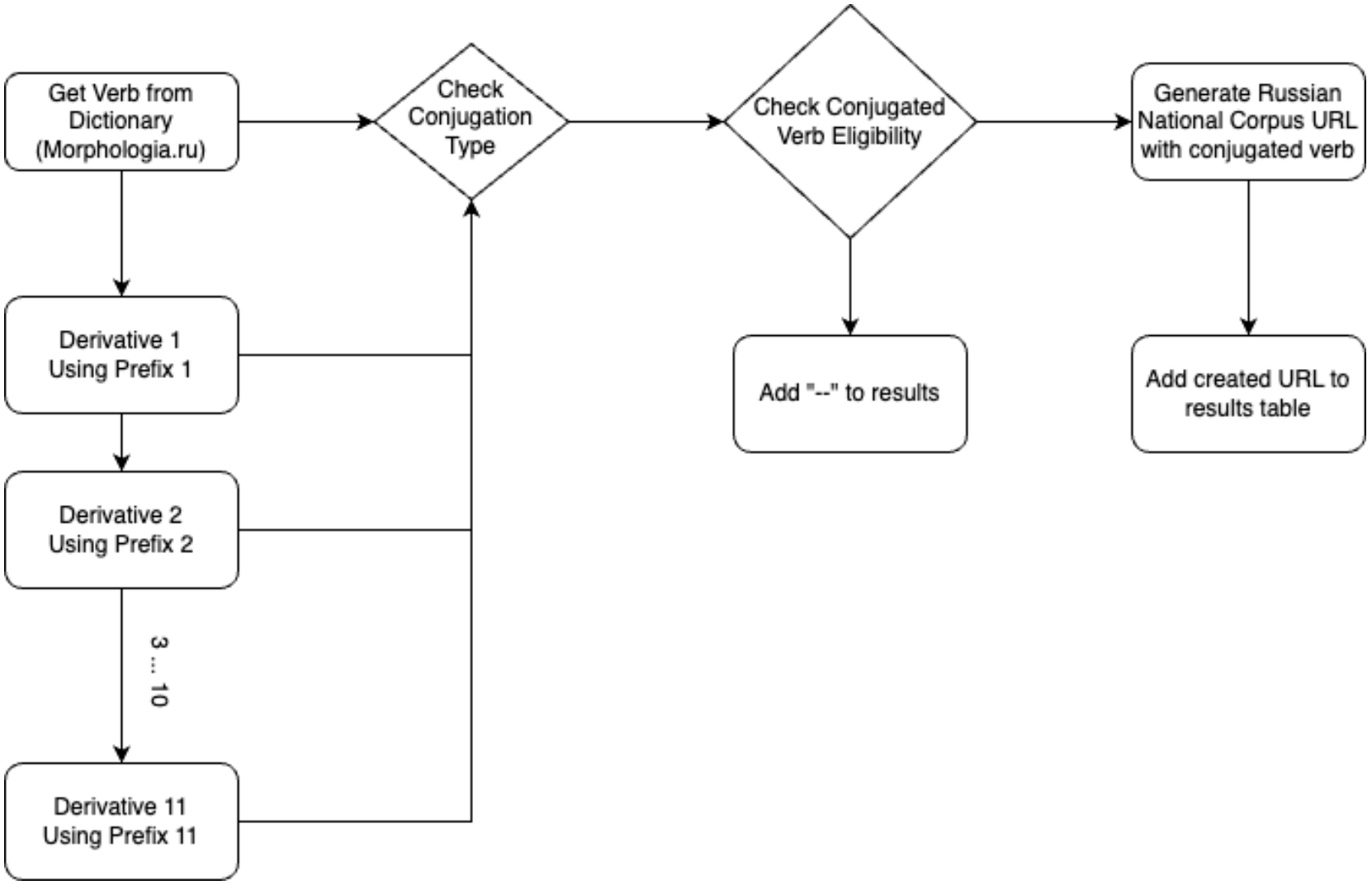
И.В. Ивлиева. [Экспериментальный модификационный словарь русского языка](). Издательство «Азбуковник». Москва, 2013. 467 стр. ISBN 978-5-91172-080-3 [Translation: Irina V. Ivliyeva. [Experimental Russian Verbal Modification Dictionary](). Azbukovnik Publishing House, Moscow, 2013. 467 pages.]

Digital Humanities: the new frontiers

III. Digital Humanities

1. What is digital humanities?
2. Advantages of Digital Humanities approaches
   a. Automated usage of online resources
   b. More accurate results
   c. Access to Technology, Equity
   d. Time Factor
   e. Harnessing the power of Large Language Models (Claude, ChatGPT, etc.)

IV. Result Generation

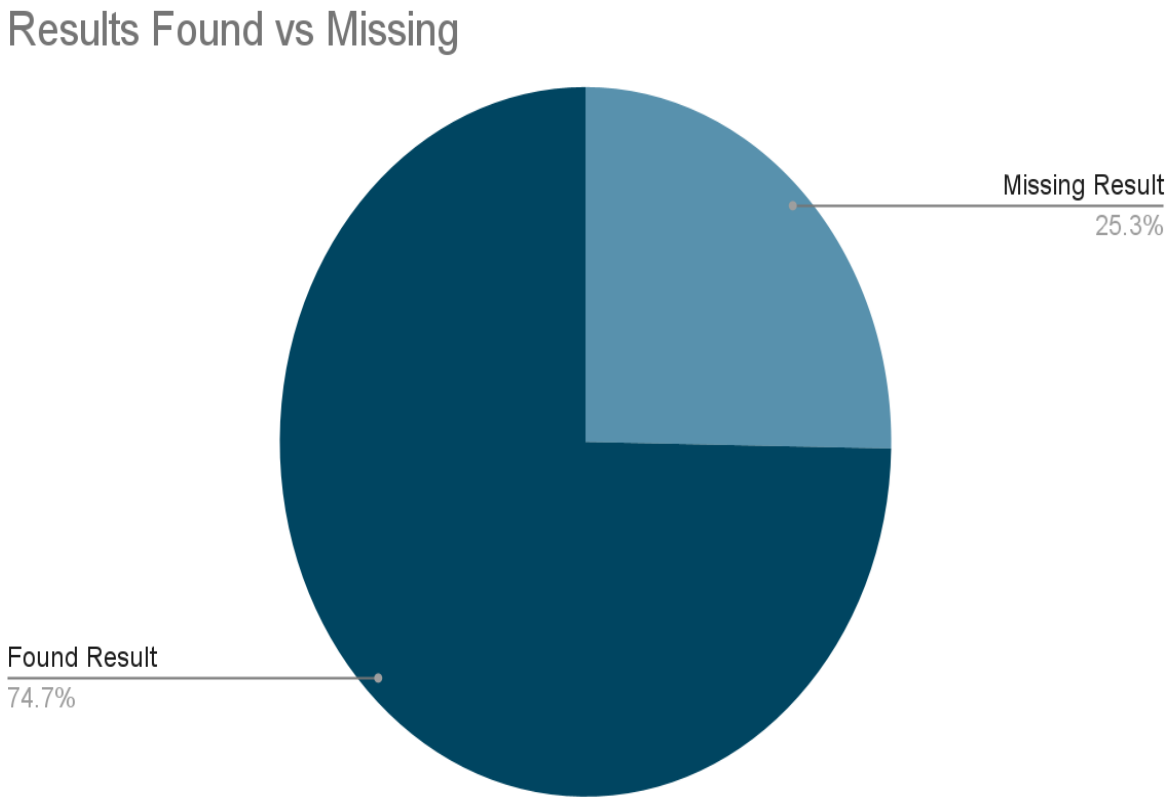


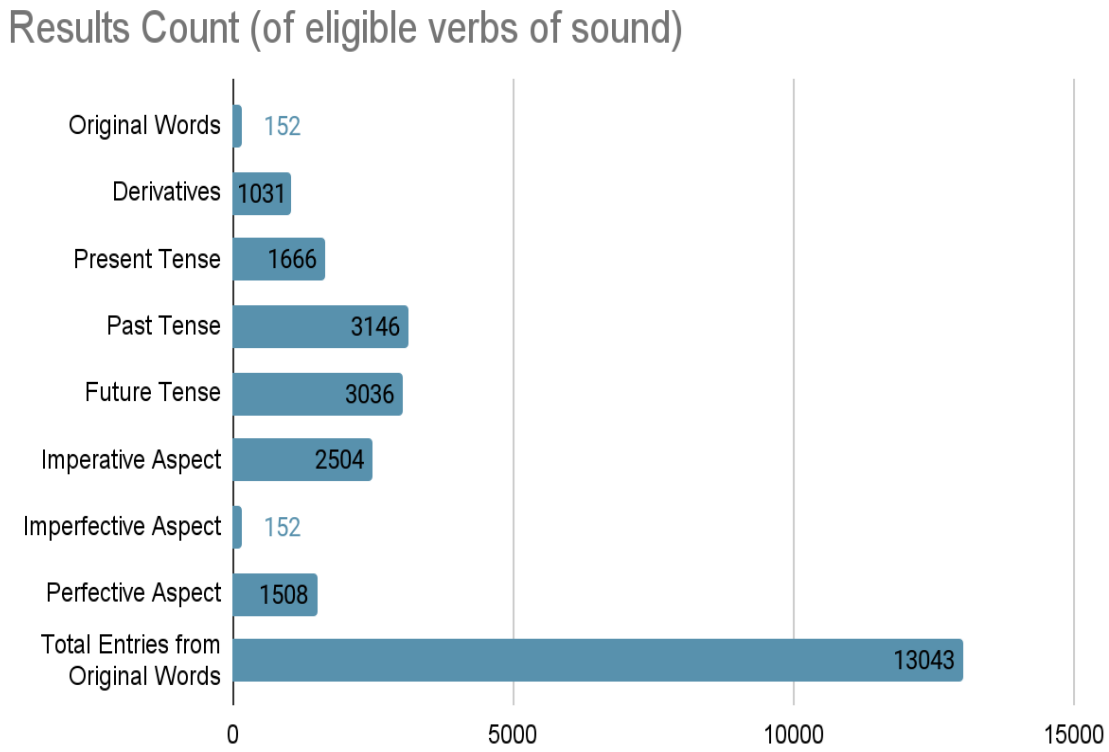
V. Results Table Sample

Number	Original Word	Derivative	Pronoun	Present Tense	Past Tense	Future Tense	Imperative Aspect	Imperfective Aspect	Perfective Aspect
1.0.0.0	ахать	ахать	--	--	--	--	--	ахая	--
1.0.0.1	ахать	ахать	--	--	--	--	--	--	ахав
1.0.0.2	ахать	ахать	--	--	--	--	--	--	ахавши
1.0.1.1	ахать	ахать	Я	ахаю	ахал	--	--	--	--
1.0.1.2	ахать	ахать	Я	ахаю	ахала	--	--	--	--
1.0.1.3	ахать	ахать	Я	ахаю	ахало	--	--	--	--
1.0.2.1	ахать	ахать	Ты	ахаешь	ахал	--	ахай	--	--
1.0.2.2	ахать	ахать	Ты	ахаешь	ахала	--	--	--	--
1.0.2.3	ахать	ахать	Ты	ахаешь	ахало	--	--	--	--
1.0.3.1	ахать	ахать	Он	ахает	ахал	--	--	--	--
1.0.3.2	ахать	ахать	Она	ахает	ахала	--	--	--	--
1.0.3.3	ахать	ахать	Оно	ахает	ахало	--	--	--	--
1.0.4.0	ахать	ахать	Мы	ахаем	ахали	--	--	--	--
1.0.5.0	ахать	ахать	Вы	ахаете	ахали	--	ахайте	--	--
1.0.6.0	ахать	ахать	Они	ахают	ахали	--	--	--	--
1.1.0.1	ахать	ахнуть	--	--	--	--	--	--	ахнув
1.1.0.2	ахать	ахнуть	--	--	--	--	--	--	ахнувши
1.1.1.1	ахать	ахнуть	Я	--	ахнул	ахну	--	--	--
1.1.1.2	ахать	ахнуть	Я	--	ахнула	ахну	--	--	--
1.1.1.3	ахать	ахнуть	Я	--	ахнуло	ахну	--	--	--
1.1.2.1	ахать	ахнуть	Ты	--	ахнул	ахнешь	ахни	--	--
1.1.2.2	ахать	ахнуть	Ты	--	ахнула	ахнешь	--	--	--

I**vliyeva, Irina and Koob, Perry,** "Multi-dimensional scaling of web-scraping results from the A.A Zalizniak Grammatical Dictionary and the Russian National Corpus. Creating a corpus fragment of all possible word-forms of modified Russian sound verbs using web-scraping methodology. Compilation of a summary table for the present tense, past tense, future tense, imperative, imperfective and perfective gerund forms" (2024). *Research Data*. 12. https://scholarsmine.mst.edu/research_data/12

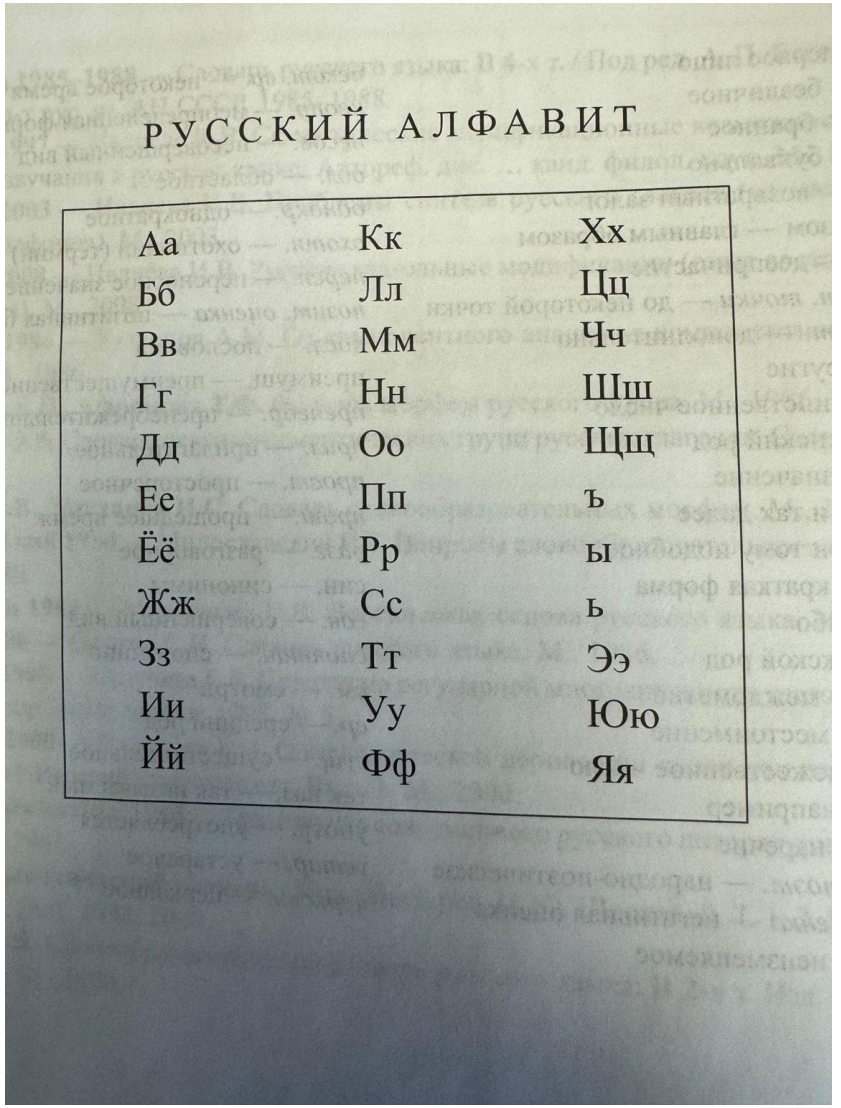
Results at a glance





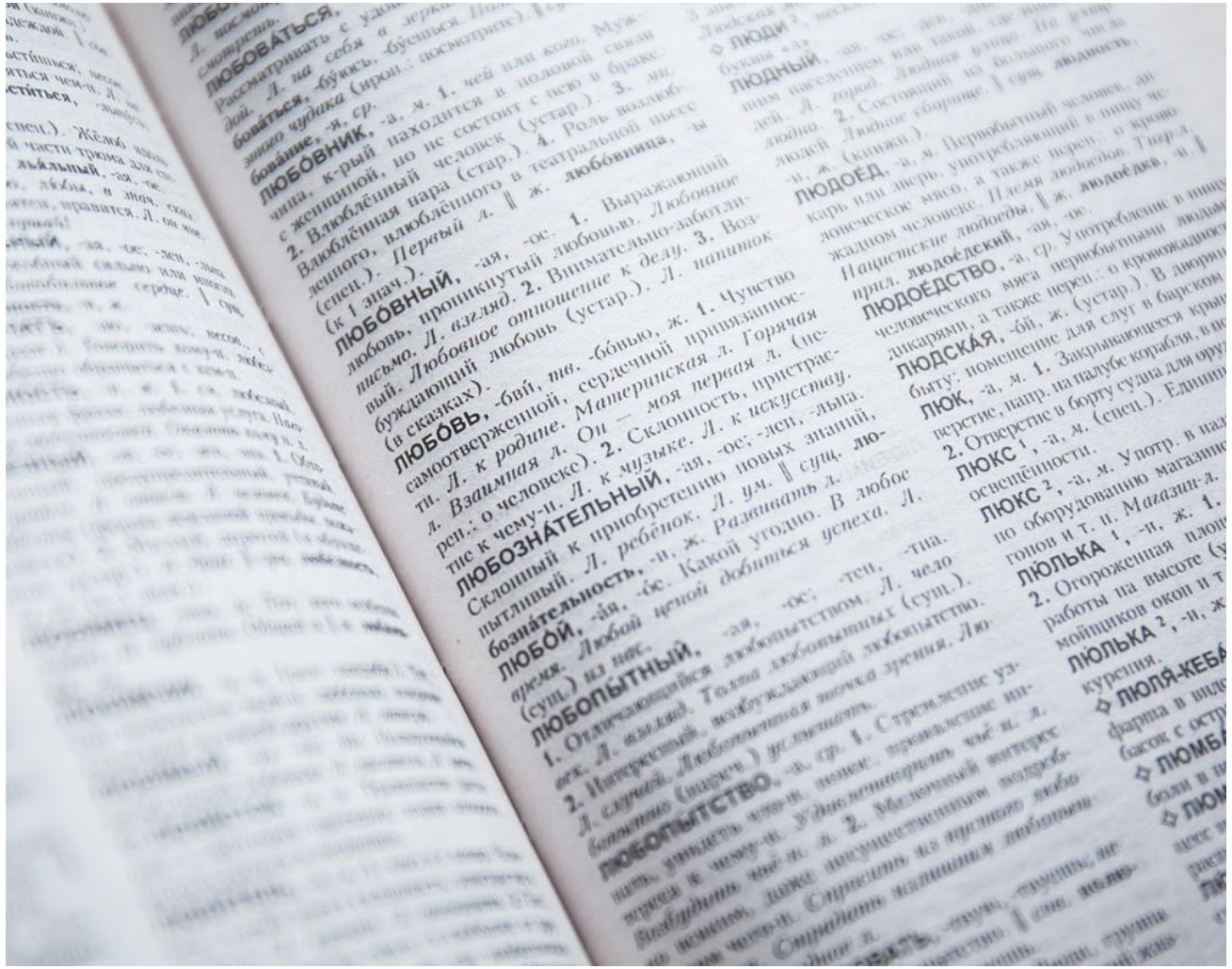
VI. Specific Challenges

1. Cryllic Alphabet
2. Access Stability
    a. Web resource availability
3. Unstructured data
4. Visualizing the data
    a. Over 12000 entries, from just 152 input verbs (Ivliyeva, 2013)



VII. Future Development

1. Processing optimization
2. Adding data from other digital dictionaries
3. Improved aggregation of the results
4. Enhanced dissemination of the methods and data



VIII. Bibliography
Methodology and Techniques

Darmawan, Irfan, et al. “Evaluating Web Scraping Performance Using XPath, CSS Selector, Regular Expression, and HTML DOM with Multiprocessing Technical Applications.” *JOIV : International Journal on Informatics Visualization*, vol. 6, no. 4, 31 Dec. 2022, https://doi.org/10.30630/joiv.6.4.1525

Luscombe, Alex, et al. “Algorithmic Thinking in the Public Interest: Navigating Technical, Legal, and Ethical Hurdles to Web Scraping in the Social Sciences.” Quality Quantity, no. 56, 24 May 2021, https://doi.org/10.1007/s11135-021-01164-0

P. Matta, S. Sharma and N. Uniyal "Comparative Study Of Various Scraping Tools: Pros And Cons," 2022 IEEE Delhi Section Conference (DELCON), New Delhi, India, 2022, doi: 10.1109/DELCON54057.2022.9753358

R. R. N. R, N. R. S and V. M. "Web Scrapping Tools and Techniques: A Brief Survey," *2023 4th International Conference on Innovative Trends in Information Technology (ICITIIT)*, Kottayam, India, 2023, doi: 10.1109/ICITIIT57246.2023.10068666

VIII. Bibliography. Foundational publications by Irina Ivliyeva

MONOGRAPHS

Экспериментальный модификационный словарь русского языка. Издательство «Азбуковник». Москва, 2013. 467 стр. ISBN 978-5-91172-080-3 [Translation: Experimental Russian Verbal Modification Dictionary. Azbukovnik Publishing House, Moscow, 2013. 467 pages.]

Русские глагольные модификации. Опыт составления словаря. Издательство «Элпис». Москва, 2008. 275 стр. ISBN 5-902872-24-5 [Translation: Russian Verbal Modifications. An Experience of a Dictionary Creating. Elpis Press, Moscow, 2008. 275 pages.]

Проблемы синтеза русского глагола (на материале антропофонов). Москва, 2003. Издательство Российского университета дружбы народов. 120 страниц. ISBN 5-209- 02286-2. [Translation: Problems in Synthesis of Russian Verbs, RUDN Press Moscow, 2003. 120 pages.]

VIII. Bibliography
Digital Lexicographic Resources

Грамматический словарь А.А. Зализняка. The grammatical dictionary of the Russian language by A.A. Zalyzniak, https://morfologija.ru/. Accessed February 11, 2024.

Национальный корпус русского языка. The Russian National Corpus, https://ruscorpora.ru/. Accessed February 11, 2024.

VIII. Bibliography
Recent Publications by Research Team Members

I**vliyeva, Irina and Koob, Perry,** "Multi-dimensional scaling of web-scraping results from the A.A Zalizniak Grammatical Dictionary and the Russian National Corpus. Creating a corpus fragment of all possible word-forms of modified Russian sound verbs using web-scraping methodology. Compilation of a summary table for the present tense, past tense, future tense, imperative, imperfective and perfective gerund forms" (2024). *Research Data*. 12. https://scholarsmine.mst.edu/research_data/12

Ivliyeva, Irina V., Koob, Perry. Experimental multi-dimensional scaling of web-scraping results from the A.A Zalizniak Grammatical Dictionary and the Russian National Corpus. Creating a corpus fragment of all possible word-forms of modified Russian sound verbs using web-scraping methodology. Compilation of a summary table for the present tense, future tense, imperative, imperfective, and perfective gerund forms. April 2022 – May 2023. [Electronic resource]. Working theses. Research Data. June 5, 2023. – Access mode: https://scholarsmine.mst.edu/research_data/11/

Ивлиева И.В. Принципы лексикографического описания терминологии словообразовательного синтеза (на материале глаголов звучания русского языка) // Тема номера. *Интерактивная наука*. – 2023. – 6 (82). – С. 7-15. – ISSN 2414-9411. – DOI 10.21661/r-560185. [Translation: Principles of lexicographic description for word formation synthesis terminology (on the material of verbs of sound in the Russian language)]. [Electronic resource]. Feature article. *Interactive science*, 2023, 6 (82). P. 7-15. Access mode: https://interactive-science.media/ru/article/560185/discussion_platform?utm_source=ticket&utm_medium=email&utm_campaign=request_onsite&utm_term=ru&utm_content=discussion_platform