14 Nov 2024

Multi-Dimensional Visualization Strategies using Web Scraping Tools: A Word Formation Synthesis Case Study for Russian Verbs of Sound

John Simmons

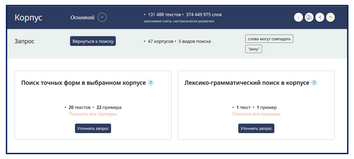
Multi-Dimensional Visualization Strategies using Web Scraping Tools: A Word Formation Synthesis Case Study for Russian Verbs of Sound

Student: John Simmons, Computer Science

Faculty Advisor(s): Irina Ivliyeva PhD, ALP & Perry Koob MS, IT S&T

I. Project Objectives

1. Expand the initial dataset of verbs of sound (**V of S**).
2. Calibrate web scraping (**WS**) algorithms through cyclical testing.
3. Refine visualization strategies (**VS**) to showcase results in two-dimensional summary tables.
4. Normalize the web-scraping outputs.
5. Transition 2D Visualization technique to 3D format by adding hyperlinks to Russian National Corpus (**RNC**).



Key Words: Web Scraping (**WS**),Semantic Modification (**SM**), Word Formation Synthesis (**WFS**), Verbs of Sound (**V of S**), corpus linguistics, Russian National Corpus (**RNC**).

II. Background

1. Word Formation Synthesis is the process of creating new meanings from original units.
2. The Russian Verb morphology follows a systematic, matrix-like structure, yet a single dictionary does not reflect the full lexicographic profiles of words in all their forms.
3. Emergence of digital lexicographic resources creates an opportunity to more accurately inventory existing units of meaning, while at the same identifying duplicates, explain lexical gaps (lacunae), and document novel units.
4. The research methods and the **WS** data obtained allow for improvements to the overall completeness of digital dictionaries and accuracy of digital dictionaries at large.

III. Approach

1. Create a web extraction mechanism and data processing methods to index the Russian **V of S** and their derivatives.
2. Compile the **WS** results in a comprehensive Excel table.

A	B	C	D	E	F	G	H	I	J
number	root	word	pronoun	present	past	future	imperative	Imp. gerund	Perf. gerund
1.0.0.0	ахать	ахать	--	--	--	--	--	ахая	--

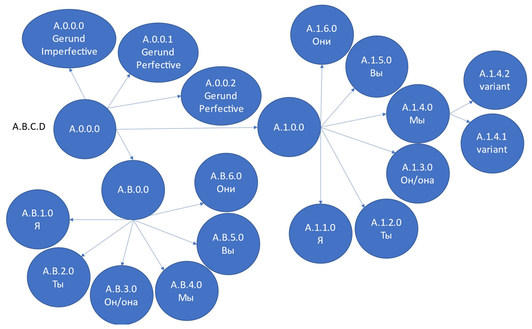
3. Record each individual instance of output using a four digit index number to allow for searching, comparing, and contrasting results that may suggest further applications.
4. During **WS** search, connect each index with **RNC** to bring in third dimension.

A	B	C	D	E	F	G
Number	Root	Word	Pronoun	Present	Past	Future
1.1.1.1	ахать	ахнуть	Я	--	ахнул	ахну
1.1.1.2	ахать	ахнуть	Я	--	ахнула	ахну

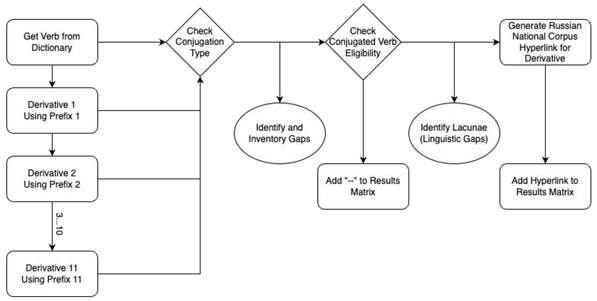
Russian National Corpus Hyperlinking:
ахну =HYPERLINK("https://ruscorpora.ru/explore?req=ахну","ахну")

IV. Results

The significant result of this research is in establishing the interactive connections in the Matrix system of **V of S** forms (around 12000 units) and with the Russian National Corpus (around two billion words).



"RNC Text Mining Output 23 Apr 2019 Excel Spreadsheet" was the most popular paper in September 2021 in the Scholars' Mine, remained the second most popular item in Scholars' Mine in May 2022, and as of January 2024 has over 22,162 downloads.



V. Discussion

1. Cyrillic Alphabet: challenges for coding; knowledge of Russian grammatical system by all researchers.
2. Reliable and safe access to various Russian lexicographic web resources outside the US: Covid, geopolitical restrictions for accessing by a program vs. manual access.
3. Integration of the new data with data from prior research "on paper": a set of verbal forms and their modifications, a system of modifiers of the sound verbs, and a scheme of synthesis (algorithm) of various lexical groups; a set of information about lacunae and codified derivatives (modifications), as well as a lot of other information that is not currently present in linguistic dictionaries.
4. Linguistic Challenges: how to determine the semantic potential of a particular verb sound, using information about lacunae and codification principles; how to determine the presence or absence of a particular verbal modification in the language, if initially only the original form of the verb is known.

VI. Concluding Remarks

1. Finalize a set of verbal forms and their modifications, a system of modifiers of the sound verbs, and a scheme of synthesis (algorithm) of various lexical groups; a set of information about lacunae and codified derivatives (modifications), as well as a lot of other information that is not currently present in linguistic dictionaries.
2. Illustrate how closed and open digital linguistic databases can be linked together in a multidimensional fashion.
3. Help the researcher and the language learner to spend as little time as possible on routine tasks to find the necessary information, while focusing on the main tasks, e.g., analyzing and visualizing data

VII. Future Works

1. Provide new perspectives on the asymmetries in the behavior of Russian affixes in morpho-syntactical context and their status in modern theories of grammar, taking into account the main tasks of the Russian verb synthesis.
2. Create the terminological dictionary that aims at describing and presenting the terminology of word-formation synthesis as a dynamic system,
3. Enhance the dissemination of results (e.g., Scholars Mine, GitHub, scientific journals).
4. Find future collaborators and funds.

VIII. Acknowledgements

- Intelligent Systems Center (ISC)
- Arts Language Philosophy Department (ALP)
- Advisors: Irina Ivliyeva PhD, Professor of Russian
  Perry Koob MS, Chief Security Officer S&T

Sponsored by the Intelligent Systems Center

MISSOURI S&T