PRC-Saltillo Project: Codifying the Russian Language

Disciplines:
Computer Science
Slavic Linguistics

Gillian Gregory ‘19, Erin Tupman ‘19, Harry Dunham ‘20, and David Sasu ‘19 (Ashesi)
Advisors: Prof. Nathan Sommer, Prof. Tatiana Filimonova, Prof. Michael Furman

AMRE COLLEGE OF WOOSTER

About the Client

PRC Saltillo

- PRC-Saltillo is a software and medical device company based in Wooster, OH. It designs and manufactures speech-processing software which allows those who cannot speak to communicate.
- Their products support English, Spanish, German, and others. The company is looking to expand into Slavic languages - specifically Croatian.

About the AMRE Team

Computer Science
- Harry Dunham ’20
- David Sasu ‘19 (Ashesi University)

Russian Studies
- Gillian Gregory ’19
- Erin Tupman ‘19

About the Project

PRC-Saltillo tasked us to identify the morphological rules of the Russian language, then use these rules to create software in Python which, when given an input word form, would produce the proper output forms.

As our client’s target language was Croatian, a secondary goal of the project was to examine the similarities between Russian and Croatian, identify rules if there was time, and determine what parts of the Russian code could be repurposed for the future Croatian code.

Our Approach

- Writing/Research Phase: “divide and conquer.”
  - Linguists: take one part of speech (or subset therein) and define its rules
  - Programmers: transfer the rules for different parts of speech into code
  - Both: brainstorming for highly-complicated topics
- Testing Phase:
  - Both: check accuracy of output words, correct and add morphological rules and REGEXs

Morphological Rules

- **Morphemes** = the smallest units of a language that carry meaning.
  - Ex: the suffix {-ed} in English, when attached to verb roots, indicates the past tense.
- Morphemes can be set endings, like {-ed}, or can mutate when paired with other morphemes.
- We defined rules for declension/conjugation morphemes and mutations across various parts of speech.

Morphological Rules of Adjectives
Adjectives must agree in gender, number, and case with the noun they are describing. Adjectives come before the noun they are describing. Animacy is relevant to the accusative case declensions, like with nouns.

I. Nominative
  a. Masculine (This nominative, masculine, singular form of the adjective would be the input)
    i. This is the dictionary form of the adjective. Other cases will be derived from these endings.
    ii. The 7-Letter Spelling Rule
      1. Instead of -ы, write -и after the letters [к,г,х,ш,щ,ж,ч].
      2. Example: хороший
    iii. Hard Stem Ending: -ый
      1. Identify the gender, case and number of the noun the adjective is describing (in this case it should be: masculine, nominative, singular)
      2. Example: новый
    iv. Hard Stem Ending: -ой
      1. Adjectives with stress on the ending use -ой, not -ый/-ий in nominative masculine singular case
      2. Identify the gender, case and number of the noun the adjective is describing (in this case it should be: masculine, nominative, singular)
      3. Example: молодой
    v. Soft Stem Ending: -ий
      1. Identify the gender, case and number of the noun the adjective is describing (in this case it should be: masculine, nominative, singular)
      2. Example: синий
      3. Example: молодой

Example from the Adjective Rules

Workflow and Priorities
1. Identified parts of speech (PoS)
2. Sorted PoS by utility and relevance
3. High Priority: nouns and verbs
4. Lower-priority: adjectives, adverbs, particles, prepositions, special modifiers, conjunctions, etc.

Exception Tables

- Some words did not follow the morphological rules that we defined
- To compensate, we declined or conjugated the word into all of its possible forms and place it into an exceptions table.
- The Engine then referenced the tables, retrieving the proper word form

			1st person singular Я			2nd person singular Ты		
Root	Infinitive	Aspect	Past	Present	Future	Past	Present	Future
ед	есть	imperfective	ел/ела	ем	буду есть	ел/ела	ешь	будешь есть
ед	съесть	perfective	съел/съела	-	съем	съел/съела	-	съешь
хот	хотеть	imperfective	хотел/хотела	хочу	буду хотеть	хотел/хотела	хочешь	будешь хотеть

Excerpt from the Verbs Exceptions Table

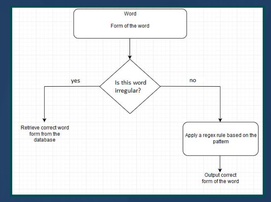
Russian vs Croatian

Russian
- East Slavic language
- Cyrillic Alphabet (33 letters)
- 6 Cases
- 3 Tenses
- Perfective vs Imperfective
- 3 Genders
- Animate vs Inanimate
- 2 Forms of Adjectives
- Exceptions to morphological gender markers

Croatian
- South Slavic language
- Latin Alphabet (30 letters)
- 7 Cases
- 7 Tenses
- Perfective vs Imperfective
- 3 Genders
- Animate vs Inanimate
- 3 Forms of Adjectives
- Exceptions to morphological gender markers

The Engine

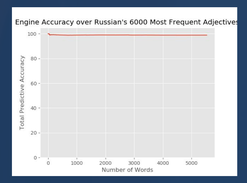
- The engine is composed of 6 basic functions: decline_noun, decline_adjective, prefix, make_ordinal, decline_numeral, and conjugate.
- Each function takes in a word and a desired conjugation or declension form
- Regular Expressions account for morphological transformations, substitution, and mutations.



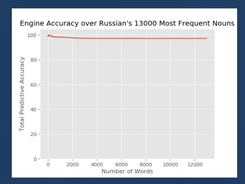
Regular Expressions

	Infinitive/Root	Conjugated	REGEX Match	REGEX Sub
Past Tense in English	Walk	Walked	^.*[^ey]$	\g<0>ed
Past Tense (feminine) in Russian:	Ходить	Ходила	(^.*)ть$	\g<1>ла

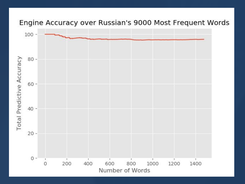
Accuracy



- 98.81% adjective accuracy
- 1.19% error stems from short form declensions



- 97.21% noun accuracy
- 2.79% error stems from irregular words



- 95.83% verb accuracy
- 4.17% error results from irregular verbs

Final Product

- A 144 page report, including sociolinguistic commentary, key grammatical differences, programming notes, and all morphological rules identified during the project
- A Python code consisting of 1,185 regular expressions
- A database of about 800 words which the program can process correctly (including exceptions tables)