Paradigms: cognitive plausibility and pedagogical application

Laura A. Janda, UiT The Arctic University of Norway
Francis M. Tyers, Indiana University and Higher School of Economics, Moscow
Why Russian?

• Morphologically relatively complex
  • relatively large numbers of forms in paradigms
  • relatively numerous inflectional classes
  • high proportion of irregular and suppletive word forms

• Full paradigms are not cognitively plausible
• We propose a usage-based model of Russian inflection and a pedagogical application
Overview

1. Distributional Facts about Russian Paradigms

2. Computational Experiment on Learning of Paradigms

3. Introducing the SMARTool
   http://uit-no.github.io/smartool/
1. Distributional Facts about Russian Paradigms

**Word frequency** is inversely proportional to frequency rank (Zipf’s Law)

**Zipfian distribution:**
- Few words of high frequency
- Sharp decline
- Scales up infinitely

George K. Zipf
## Zipf’s Law Applies to Wordforms

<table>
<thead>
<tr>
<th>Language &amp; Corpus Name</th>
<th>Corpus Size</th>
<th>Paradigm Size</th>
<th>Total Lexemes</th>
<th>Lexemes with full Paradigm</th>
<th>% Lexemes with full Paradigm</th>
</tr>
</thead>
<tbody>
<tr>
<td>English Web Treebank</td>
<td>254,830</td>
<td>2</td>
<td>6,369</td>
<td>1,524</td>
<td>23.92%</td>
</tr>
<tr>
<td>Norwegian Dependency Treebank</td>
<td>311,277</td>
<td>4</td>
<td>12,587</td>
<td>393</td>
<td>3.12%</td>
</tr>
<tr>
<td>Russian SynTagRus</td>
<td>1,032,644</td>
<td>12</td>
<td>21,945</td>
<td>13</td>
<td>0.06%</td>
</tr>
<tr>
<td>Czech Prague Dependency Treebank</td>
<td>1,509,242</td>
<td>14</td>
<td>17,904</td>
<td>3</td>
<td>0.02%</td>
</tr>
<tr>
<td>Estonian ArborEst</td>
<td>234,351</td>
<td>28</td>
<td>14,075</td>
<td>0</td>
<td>0%</td>
</tr>
</tbody>
</table>
Zipf’s Law Applies to Wordforms

<table>
<thead>
<tr>
<th>Language &amp; Corpus Name</th>
<th>Corpus Size</th>
<th>Paradigm Size</th>
<th>Total Lexemes</th>
<th>Lexemes with full Paradigm</th>
<th>% Lexemes with full Paradigm</th>
</tr>
</thead>
<tbody>
<tr>
<td>English Web Treebank</td>
<td>254,830</td>
<td>2</td>
<td>6,369</td>
<td>1,524</td>
<td>23.92%</td>
</tr>
<tr>
<td>Norwegian Dependency Treebank</td>
<td>311,277</td>
<td>4</td>
<td>12,587</td>
<td>393</td>
<td>3.12%</td>
</tr>
<tr>
<td>Russian SynTagRus</td>
<td>1,032,644</td>
<td>12</td>
<td>21,945</td>
<td>13</td>
<td>0.06%</td>
</tr>
<tr>
<td>Czech Prague Dependency Treebank</td>
<td>1,509,242</td>
<td>14</td>
<td>17,904</td>
<td>3</td>
<td>0.02%</td>
</tr>
<tr>
<td>Estonian ArborEst</td>
<td>234,351</td>
<td>28</td>
<td>14,078</td>
<td>0</td>
<td>0%</td>
</tr>
</tbody>
</table>

Because Zipf’s Law scales up, these numbers will never change substantially, no matter how large the corpus is.
Language Exposure as a Big Corpus

• A large corpus is a close approximation to the **lifetime linguistic input** for a native speaker, estimated at about 5-10 million words per year.

• Zipfian distributions **remain the same** even for very large corpora, like those that approximate a speaker’s exposure to their native language.

• A native speaker of Russian encounters all twelve paradigm forms of **less than 0.1% of nouns** that they are exposed to in a lifetime.

• The portion of adjectives and verbs attested in all paradigm forms is **virtually zero**.
Paradigm Cell Filling Problem
(Ackerman et al. 2009)

Native **speakers** of languages with complex **inflectional morphology** routinely **recognize** and **produce** forms that they have **never** encountered.

**Q:** How is this possible?

**A:** Inflectional morphology is mastered through exposure to **partially overlapping subsets of paradigms** arranged according to Zipf’s Law, a **cognitively plausible** model.
Paradigm Cell Filling Problem
(Ackerman et al. 2009)

Native speakers of languages with complex inflectional morphology routinely recognize and produce forms that they have never encountered.

Q: How is this possible?
A: Inflectional morphology is mastered through exposure to partially overlapping subsets of paradigms arranged according to Zipf's Law, a cognitively plausible model.

This also explains why native speakers have the intuition that full paradigms exist.
Illustration of Overlapping Subsets of Paradigms

Frequency distributions of wordforms of 982 noun lexemes

All lexemes with frequency ≥ 50 in SynTagRus representing five paradigm types:

- masculine inanimate (312 lexemes)
- masculine animate (95 lexemes)
- neuter inanimate (238 lexemes)
- feminine inanimate II (ending in -а/-я, 261 lexemes)
- feminine inanimate III (ending in -ъ, 75 lexemes)
## Frequency Distributions of Some High-frequency Russian Nouns

<table>
<thead>
<tr>
<th>Case</th>
<th>Noun</th>
<th>Gender</th>
<th>Noun</th>
<th>Gender</th>
<th>Noun</th>
<th>Gender</th>
<th>Noun</th>
<th>Gender</th>
<th>Noun</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nsg</td>
<td>страх</td>
<td>мужчина</td>
<td>солдат</td>
<td>мужчина</td>
<td>отделение</td>
<td>русский</td>
<td>концепция</td>
<td>русский</td>
<td>память</td>
<td>русский</td>
</tr>
<tr>
<td>Gsg</td>
<td>страх</td>
<td>мужчина</td>
<td>солдата</td>
<td>женщина</td>
<td>отделения</td>
<td>русские</td>
<td>концепции</td>
<td>русские</td>
<td>памяти</td>
<td>русские</td>
</tr>
<tr>
<td>Dsg</td>
<td>страху</td>
<td>мужчина</td>
<td>солдату</td>
<td>женщина</td>
<td>отделению</td>
<td>русском</td>
<td>концепция</td>
<td>русской</td>
<td>память</td>
<td>русской</td>
</tr>
<tr>
<td>Asg</td>
<td>страх</td>
<td>мужчина</td>
<td>солдата</td>
<td>женщина</td>
<td>отделение</td>
<td>русском</td>
<td>концепции</td>
<td>русской</td>
<td>память</td>
<td>русской</td>
</tr>
<tr>
<td>Lsg</td>
<td>страхе</td>
<td>мужчина</td>
<td>отделении</td>
<td>русском</td>
<td>концепции</td>
<td>русской</td>
<td>памяти</td>
<td>русской</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Npl</td>
<td>страхи</td>
<td>мужчины</td>
<td>отделения</td>
<td>русскими</td>
<td>концепция</td>
<td>русскими</td>
<td>память</td>
<td>русскими</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gpl</td>
<td>страхов</td>
<td>мужчины</td>
<td>солдат</td>
<td>русскими</td>
<td>отделений</td>
<td>русскими</td>
<td>концепцией</td>
<td>русскими</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dpl</td>
<td></td>
<td></td>
<td>отделения</td>
<td>русского</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Apl</td>
<td>страхи</td>
<td>мужчины</td>
<td>солдат</td>
<td>русскими</td>
<td>отделения</td>
<td>русскими</td>
<td>концепции</td>
<td>русскими</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ipl</td>
<td>страхами</td>
<td>мужчины</td>
<td>отделениями</td>
<td>русскими</td>
<td>концепциями</td>
<td>русскими</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lpl</td>
<td>страхах</td>
<td>мужчины</td>
<td>солдатах</td>
<td>русскими</td>
<td>отделениях</td>
<td>русскими</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Key:** bold >20%, plain >10%, grey 1-9%, (blank) unattested
Typically a lexeme is found in only 1-3 wordforms
Any single lexeme gives exposure to only a subset of the paradigm. Each lexeme has a different subset of most typical forms. Collectively they populate the entire “space” of case/number combinations.
Distributional Facts about Russian Paradigms: Summary

• Even a small vocabulary can have >100,000 wordforms
• <10% of wordforms are frequent, others rare or unencountered
• Each lexeme is common in a subset of wordforms
• The wordforms common to a lexeme are motivated by typical collocations and grammatical constructions
• Overlapping subsets of wordforms create the illusion of a full paradigm, make it possible for native speakers to comprehend and produce wordforms they have never encountered
2. Computational Learning Experiment

- Based on an ordered list of the most frequent forms for nouns, verbs, and adjectives in SynTagRus

- Machine learning:
  - Given the 100 most frequent forms, predict the next 100 most frequent forms
  - Given the 200 most frequent forms, predict the next 100 most frequent forms
  - Given the 300 most frequent forms, predict the next 100 most frequent forms
  - Given the 400 most frequent forms, predict the next 100 most frequent forms
  - Given the 500 most frequent forms, predict the next 100 most frequent forms
  - ... until 5400, when SynTagRus runs out of data
Comparison of accuracy training on individual forms and full paradigms

1800-5400: Single forms model outperforms full paradigms
After 11 iterations, the errors committed by the single forms model are consistently smaller.
Computational Learning Experiment: Summary

• Learning is potentially enhanced by focus only on the most typical wordforms attested for each lexeme: accuracy increases and severity of errors decreases

• This finding is consistent with a usage-based cognitively plausible model
3. Introducing the SMARTool

- **Strategic Mastery of Russian Tool**
- The user can browse over 3000 Russian words according to proficiency level, topic, and grammatical categories.
- For each word, the SMARTool provides the three most common forms, plus example sentences that show typical collocations and grammatical constructions.
- The SMARTool provides audio and translations
Members of the SMARTool team

Radovan Bast
Tore Nesset
Svetlana Sokolova
Mikhail Kopotev
Francis Tyers
Ekaterina Rakhilina
Olga Lyashevskaya
Valentina Zhukova
James McDonald
Evgeniia Sudarikova
Vocabulary Selection from 5 Textbooks and Лексический минимум; Balanced for Nouns, Verbs, Adjs (RNC ratio)

<table>
<thead>
<tr>
<th>CEFR Level</th>
<th>ACTFL Equivalent</th>
<th>Russian Equivalent</th>
<th>SMARTool number of lexemes</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1 “Beginner”</td>
<td>Novice Low-Mid</td>
<td>ТЭУ Элементарный уровень</td>
<td>500</td>
</tr>
<tr>
<td>A2 “Elementary”</td>
<td>Novice High</td>
<td>ТБУ Базовый уровень</td>
<td>500</td>
</tr>
<tr>
<td>B1 “Intermediate”</td>
<td>Intermediate Low-Mid</td>
<td>ТРКИ-1 Сертификационный уровень</td>
<td>1,000</td>
</tr>
<tr>
<td>B2 “Upper Intermediate”</td>
<td>Intermediate High-Advanced Low</td>
<td>ТРКИ-2</td>
<td>1,000</td>
</tr>
</tbody>
</table>
Typical Contexts Illustrated by Examples

• For each lexeme we identify 3 most common wordforms and **most typical grammatical constructions** and **lexical collocations**, and provide corpus-inspired **example sentences**

• Based on queries:
  • SynTagRus Corpus
  • the Russian National Corpus ([http://ruscorpora.ru](http://ruscorpora.ru))
  • the Collocations Colligations Corpora ([http://cococo.cosyco.ru/](http://cococo.cosyco.ru/))
  • the Russian Constructicon ([https://spraakbanken.gu.se/karp/#?mode=konstruktikon-rus](https://spraakbanken.gu.se/karp/#?mode=konstruktikon-rus))
SMARTool Filters

• **Select a level:** A1, A2, B1, B2, all levels
  • **Search by topic:** внутренний мир, время, еда, животные/растения, жильё, здоровье, люди, магазин, мера, общение, одежда, описание, погода, политика, путешествие, свободное время, транспорт, учёба/работка
  • **Search by analysis:** Select grammatical features, such as “Ins.Sing”
  • **Search by dictionary**

• **Translations and audio** available for all example sentences
1) Select a Level
2) Search by topic, analysis, dictionary
Level: A1

Topic:
- Внутренний мир (mental experience)
- Время (time)
- Еда (food)
- Животные/растения (animals/plants)
- Жилье (home)
- Здоровье (health)
- Люди (people)
- Магазин (shopping)
- Мера (measurement)
- Общение (communication)
- Одежда (clothing)
- Описание (description)
- Погода (weather)
- Политика (politics)
- Путешествие (travel)
- Свободное время (leisure)
- Транспорт (transportation)
- Учёба/работа (study/work)
бизнесмен

Новый закон защищает интересы бизнесменов. (Gen.Plur ?)

Бизнесмен должен быть честным. (Nom.Sing ?)

Российские бизнесмены протестуют против повышения налогов. (Nom.Plur ?)
**бирнешмен**

Новый закон защищает интересы бизнесменов. (Gen.Plur)  
*The new law protects the interests of businessmen.*

Бизнесмен должен быть честным. (Nom.Sing)  
*A businessman has to be honest.*

Российские бизнесмены протестуют против повышения налогов. (Nom.Plur)  
*Russian businessmen are protesting against a tax increase.*
Я ухожу навсегда! Прощай! (Imperf.Imper.Sing.2 -question)
I'm leaving forever! Goodbye!

Слушай учительницу внимательно! Не отвлекайся! (Imperf.Imper.Sing.2 -question)
Listen carefully to the teacher! Don't get distracted!

Не волнуйся, ты сдашь экзамен! (Imperf.Imper.Sing.2 -question)
Don't worry, you'll pass that exam!

Одевайся, мы уже выходим. (Imperf.Imper.Sing.2 -question)
Get dressed, we're leaving.
Москва (Moscow)
автобусный (bus)
автомат (machine)
автомобиль (car)
азиатский (Asian)
академия (academy)
альбом (album)
американец (American)
американка (American)
американский (American)
анализ (analysis)
анализировать (analyze)
ангина (sore throat)
ансамбль (ensemble)
армия (army)
архитектурный (architectural)
атмосфера (atmosphere)
атомный (atomic)
афиша (poster)
балерина (ballerina)
Anna Pavlova dreamed of becoming a ballerina.

As a child Anna Pavlova dreamed of becoming a ballerina.
The SMARTool:

- **Inspired by research** on the distribution and simulated learning of Russian wordforms (**cognitively plausible**)
- Strategic focus on the **highest frequency wordforms and contexts** that motivate their use (**usage-based**)
- **Reduces the task** of learning a basic vocabulary of about 3,000 lexemes **by over 90%**
- Can be **continuously updated** and custom-tailored
- Potentially **portable to other languages** with rich inflectional morphology
References


