Predicting Russian Aspect: A corpus study and an experiment

Laura A. Janda
UiT The Arctic University of Norway

with: Hanne M. Eckhoff, Olga Lyashevskaya and Robert J. Reynolds
How learnable is Russian aspect?

- Use and meaning of Russian aspect is topic of long-standing debate (cf. Janda 2004 and Janda et al. 2013 and references therein)
- It is unclear how children acquire Russian aspect in L1
  - Generativist theory would assume that aspect is part of UG
  - Gvozdev (1961), based on his diary of son Ženja, claimed Russian aspect was fully acquired early on, but re-analysis of his and other data (Stoll 2001, Gagarina 2004) has shown that L1 acquisition is far from complete even at age 6
- It is clear that L2 learners struggle with Russian aspect
  - Russian aspect is considered the most difficult grammatical feature for L1 English speakers (Offord 1996, Andrews et al. 1997, Cubberly 2002); it is not clear how L2 acquisition takes place (Martelle 2011)
  - “Rules” offered in textbooks for when to use perfective vs. imperfective are relevant for only 2% of verb forms in a corpus (Reynolds 2016)
Aspect in Russian (a crash course)

- All forms of all verbs obligatorily express **perfective** vs. **imperfective** aspect
- **Perfective aspect**: unique, complete events with crisp boundaries
  - *Pisatel’ na-pisal/na-pišet roman* ‘The writer has written/will write a novel’
- **Imperfective aspect**: ongoing or repeated events without crisp boundaries
  - *Pisatel’ pisl/pišet roman* ‘The writer was writing/is writing a novel’

- Morphological marking is very helpful, but not entirely reliable:
  - **bare verb**: usually **imperfective** (*pisat’* ‘write’), some biaspectral (*ženit’sja* ‘marry’), a few perfective (*dat’* ‘give’)
  - **prefix + verb**: usually **perfective** (*pere-pisat’* ‘rewrite’), some imperfective (*pre-obladat’* ‘prevail’, *pere-xodit’* ‘walk across’)
  - **prefix + verb + suffix**: **imperfective** (*pere-pis-yva-t’* ‘rewrite’)

Cognitive Linguistics: Empirical Approaches to Russian
Where the aspects do and do not compete

• Paradigmatically competing:
  Non-past (future if perfective, present if imperfective)
  Past
  Imperative
  Infinitive

• Paradigmatically non-competing:
  Past gerunds and participles are perfective
  Present gerunds and participles are imperfective

• Syntagmatically competing:
  In some contexts, either aspect is grammatical

• Syntagmatically non-competing:
  In some contexts only one aspect is allowed

Study 1 takes a paradigmatic perspective
Study 2 takes a syntagmatic perspective
Research Questions

Study 1 Paradigmatic Perspective:
To what extent can the aspect of a verb be figured out based on the distribution of its grammatical forms (grammatical profiling)?
Can this type of learning be modeled by means of corpus data?

Study 2 Syntagmatic Perspective:
To what extent can the aspect of a verb be figured out based on the context in which it appears?
Can this type of learning be modeled by means of experiments?
Study 1 Paradigmatic Perspective: Aspect via Grammatical Profiles

- Janda & Lyashevskaya (2011) showed that, for paired verbs, **perfective and imperfective verbs have** in aggregate **different grammatical profiles**
  - This was a **top-down approach** (we started out by segregating perfective from imperfective verbs) and was **limited to paired verbs**
- Can aspect be approached **bottom-up**?
- Is it possible to figure out the aspect of **individual verbs of all types** (not just paired verbs) **based only on the distribution of their grammatical forms in a corpus**?
- Goldberg (2006) gives evidence that children are **sensitive to statistical tendencies** in L1 acquisition
- Could children learn to distinguish between perfective and imperfective verbs **based solely on the distributions** of their forms?
What is a grammatical profile?

Verbs have different forms:

- *eat* 749 M
- *eats* 121 M
- *eating* 514 M
- *eaten* 88.8 M
- *ate* 258 M

The grammatical profile of *eat*
Janda & Lyashevskaya 2011
Grammatical Profiles of Russian Verbs Top-Down

<table>
<thead>
<tr>
<th></th>
<th>Nonpast</th>
<th>Past</th>
<th>Infinitive</th>
<th>Imperative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Imperfective</td>
<td>1,330,016</td>
<td>915,374</td>
<td>482,860</td>
<td>75,717</td>
</tr>
<tr>
<td>Perfective</td>
<td>375,170</td>
<td>1,972,287</td>
<td>688,317</td>
<td>111,509</td>
</tr>
</tbody>
</table>

chi-squared  = 947756
df = 3
p-value < 2.2e-16
effect size
(Cramer’s V)
= 0.399
(medium-large)
Janda & Lyashevskaya 2011
Grammatical Profiles of Russian Verbs **Top-Down**

<table>
<thead>
<tr>
<th></th>
<th>Nonpast</th>
<th>Past</th>
<th>Infinitive</th>
<th>Imperative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Imperfective</td>
<td>1,330,016</td>
<td>915,374</td>
<td>482,860</td>
<td>75,717</td>
</tr>
<tr>
<td>Perfective</td>
<td>375,170</td>
<td>1,972,287</td>
<td>688,317</td>
<td>111,509</td>
</tr>
</tbody>
</table>

Can we turn this upside-down and go **Bottom-Up**?

**CLEAR**
Cognitive Linguistics: Empirical Approaches to Russian
Grammatical Profiles of Russian Verbs **Bottom-Up**

Data extracted from the manually disambiguated Morphological Standard of the Russian National Corpus (approx. 6M words), 1991-2012
Stratified by genre, 0.4M word sample for each

<table>
<thead>
<tr>
<th>Genre</th>
<th># Verb Tokens</th>
<th># Verb Lemmas</th>
<th># Verb Lemmas Frequency &gt;50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Journalistic</td>
<td>52 716</td>
<td>5 940</td>
<td>185</td>
</tr>
<tr>
<td>Scientific-technical</td>
<td>43 528</td>
<td>4 494</td>
<td>174</td>
</tr>
<tr>
<td>Fiction</td>
<td>78 084</td>
<td>8 665</td>
<td>225</td>
</tr>
</tbody>
</table>

Study 1 focuses on **Journalistic** data
Correspondence Analysis of Journalistic Data

**Input:** 185 vectors (1 for each verb) of frequencies for verb forms

Each vector tells how many forms were found for each verbal category: indicative non-past, indicative past, indicative future, imperative, infinitive, non-past gerund, past gerund, non-past participle, past participle

**rows are verbs, columns are verbal categories**

**Process:**

Matrices of distances are calculated for rows and columns and represented in a multidimensional space defined by factors that are mathematical constructs. Factor 1 is the mathematical dimension that accounts for the largest amount of variance in the data, followed by Factor 2, etc.

Plot of the first two (most significant) Factors, with Factor 1 as x-axis and Factor 2 as the y-axis

You can think of **Factor 1** as the **strongest parameter** that splits the data into **two groups** (negative vs. positive values on the x-axis)
On the Following Slide...

- Results of correspondence analysis for Journalistic data
- Perfective verbs represented as “p”
- Imperfective verbs represented as “i”
- Remember that the program was not told the aspect of the verbs
- All it was told was the frequency distributions of grammatical forms
- All it was asked to do was to construct the strongest mathematical Factor that separates the data along a continuum from negative to positive (x-axis)
Factor 1 looks like aspect

Factor 1 (39.1 %)

Factor 2 (19.7 %)

Perfective

Imperfective
Factor 1 correctly predicts aspect 91.5% (negative = perfective vs. positive = imperfective)

Of the 185 verbs:

- **87 perfectives**
  - 84 negative values, 3 positive values, so 96.6% correct
    - 3 deviations are: *obojtis* ‘make do without’, *smoč* ‘manage’, *prijtis* ‘be necessary’

- **96 imperfectives**
  - 83 positive values, 13 negative values, so 86.5% correct

- **2 biaspectuals** with low negative values
  - *obеščat* ‘promise’, *ispol’zovat* ‘use’
Interpretation of 0 (zero) Value for Factor 1 in Correspondence Analysis

Kamphuis 2016: 78-79:
“The first thing we notice is that the verbs that are shown in the scatter plot and in the corresponding table (Eckhoff & Janda 2014: 240-241) show a continuum and not a clear division into two groups. This makes the vertical line drawn at 0 (zero), dividing the lefties and the righties, look arbitrary. The zero does not have a clear meaning, nor does it seem to be a natural boundary in the scatter plot.”

Kamphuis 2016: 152:
“The lines are just as arbitrary as the line that Eckhoff & Janda (2014: 238) draw at zero and have no consequences for the final assessment of the aspect of the verbs in the group.”

What is the basis for these statements?
“the loadings on axis/factors/principal components/corresp. analysis dimensions are some form of correlation, that tell you to what extent your original variable is correlated with your new axis. So if you have a loading of zero, it means the original factor does not predict your current axis, they are "uncorrelated". A large positive loading indicates the original predictor axis and the new axis are very similar (highly correlated, small angle between their vectors), and a large negative loading indicates they point in opposite directions.”
“The origin (zero point) represents the centre of gravity…, or, in other words, the centroid (average) of the column and row profiles. The further a point representing a row or a column from the origin, the greater the chi-squared distance from it to the centroid and the more it should contribute to the inertia (the CA parlance for variance). So, the zero point is by no means arbitrary or meaningless!”
Stefan Th. Gries

“the 0-point is not arbitrary”
Aspectual morphology and Factor 1 values for Journalistic prose
Perfective verbs (aside from *smoč´, prijti´s, obojti´s*) are well-behaved.
Biaspectuals and indeterminate motion verbs lie near the line, but with perfectives
Other imperfectives are mostly well-behaved.
Summary of Study 1: Paradigmatic Perspective

- When we look at the distribution of verb forms, **aspect** (or a close approximation) **emerges as the most important factor** distinguishing verbs.
- It is **possible to sort** high-frequency **verbs as perfective vs. imperfective** based only on the distribution of their forms with about 93% accuracy.
- **Individual verbs can deviate strongly** from overall patterns.
- May have implications for L1 acquisition and machine learning.
Study 2 Syntagmatic Perspective: Aspect via Context

- This study is still underway!!
- Given contexts where both aspects are morphologically possible, what happens when you offer a choice of a perfective vs. an imperfective form to:
  - L1 native speakers of Russians

- Source material: six texts representing three written genres (journalistic, scientific-technical, fiction) and two spoken genres (monologue, dialogue)
- All texts represent authentic Russian (produced by native speakers) and plenty of context (1100-1700 words)
Examples of triggers cited in Russian textbooks
(Only available 2% of the time but 96% reliable)

<table>
<thead>
<tr>
<th>Adverbials</th>
<th>Complements of verbs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Perfective</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Imperfective</strong></td>
<td>Phasal verbs: stat’ ‘start’, načat’/načinat’ ‘begin’, prodolžit’/prodolžat’ ‘continue’, končit’/končat’ ‘stop’</td>
</tr>
</tbody>
</table>
The contexts where both aspects are morphologically possible in Russian: example verb *na-pisat’*(p) vs. *pisat’*(i) ‘write’

<table>
<thead>
<tr>
<th></th>
<th>Perfective</th>
<th>Imperfective</th>
</tr>
</thead>
<tbody>
<tr>
<td>Past</td>
<td><em>na-pisal</em> ‘he wrote’</td>
<td><em>pisal</em> ‘he wrote’</td>
</tr>
<tr>
<td>Future</td>
<td><em>na-pišet</em> ‘s/he will write’</td>
<td>*budet pisat’‘s/he will write’</td>
</tr>
<tr>
<td>Infinitive</td>
<td><em>na-pisat’</em> ‘write’</td>
<td><em>pisat’</em> ‘write’</td>
</tr>
<tr>
<td>Imperative</td>
<td><em>na-piši</em> ‘write!’</td>
<td><em>piši</em> ‘write!’</td>
</tr>
</tbody>
</table>

Excluded:

- Present tense (imperfective only)
- Gerunds & participles (specific to one aspect or the other)
- Biaspectual verbs
- Verbs not paired for aspect (Aktionsarten, -*sja* passives)
- Forms of verb *byt’* ‘be’
## The texts in the experiment

<table>
<thead>
<tr>
<th>Text</th>
<th>Genre</th>
<th>Source</th>
<th># words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Беспризорник Жук</td>
<td>Fiction</td>
<td>© 2015 Финеева Елизавета (<a href="mailto:laure.lenberg@gmail.com">laure.lenberg@gmail.com</a>) Библиотека Максима Мошкова</td>
<td>1459</td>
</tr>
<tr>
<td>История о том</td>
<td>Spoken Narrative</td>
<td>Из корпуса «РАССКАЗЫ О СНОВИДЕНИЯХ И ДРУГИЕ КОРПУСА ЗВУЧАЩЕЙ РЕЧИ» (А. А. Кибрик et al.) © 2016</td>
<td>1275</td>
</tr>
<tr>
<td>МГЛУ</td>
<td>Spoken Narrative</td>
<td>The Multimodal Communication and Cognition Laboratory at Moscow State Linguistic University (Alan Cienki, Olga Iriskhanova) © 2014</td>
<td>1617</td>
</tr>
<tr>
<td>Нефтяной саммит</td>
<td>Journalistic Prose</td>
<td>Михаил Крутихин Московский центр Карнеги © 2016</td>
<td>1116</td>
</tr>
<tr>
<td>Выяснили ученые</td>
<td>Scientific-Technical Prose</td>
<td>Александр Марков Элементы.ру 18.04.16</td>
<td>1558</td>
</tr>
<tr>
<td>Иван Дмитриевич</td>
<td>Spoken Interview</td>
<td>ГТРК «Липецк». Передача цикла «Встречи», ноябрь 2004 г.</td>
<td>1468</td>
</tr>
</tbody>
</table>
Distribution of aspect across the subparadigms in the original texts

<table>
<thead>
<tr>
<th></th>
<th>past</th>
<th>future</th>
<th>infinitive</th>
<th>imperative</th>
</tr>
</thead>
<tbody>
<tr>
<td>perfective</td>
<td>291</td>
<td>51</td>
<td>104</td>
<td>32</td>
</tr>
<tr>
<td>imperfective</td>
<td>133</td>
<td>6</td>
<td>66</td>
<td>10</td>
</tr>
</tbody>
</table>
A sample text (Journalistic prose)

В Соединенных Штатах наготове дожидается своего часа огромное число буровых установок и комплектов оборудования для гидроразрыва пласта, чтобы [возобновить / возобновлять] работу, как только оставленные на время промыслы [выйдут / будут выходить] на уровень рентабельности.

In the USA there stands at the ready an enormous number of drills and sets of equipment for the fracking of rock layers in order [RESUME] work as soon as the temporarily held up businesses [ACHIEVE] the level of profitability.


Thousands of holes have been bored where there [REMAIN] only [INITiate] periodic fracking operations.
В квадратных скобках всегда предлагается два глагола на выбор. Для каждого глагола выберите уровень приемлемости ("Отлично", "Допустимо" или "Невозможно") в данном контексте. Оба глагола в квадратных скобках могут получить одинаковую оценку. Выбор уровня приемлемости зависит от вашего личного восприятия.

По завершении задания вы получите лотерейный код. Не оставляйте никаких глаголов на голубом фоне и не отвечайте случайным образом, иначе ваш код будет исключен из лотереи.

Участие в эксперименте добровольно. Вы можете прервать эксперимент в любой момент.
Согласно условиям эксперимента, вы понимаете и принимаете условия эксперимента.

Право выбора жизненного пути - большая подарок судьбы. У Василия этого права не было. Он безропотно [принял / принимал] выбор, который за него [сделала / делала] судьба, и это был великий шаг.

Василий родился в семье зажиточного кулака в Брянской области. Когда мальчику [исполнилось / исполнилось] три года, его отец [сожился / сходился] с работницей. Мать Василия была из семьи священника, и ничего мужу не [сказала / говорила], по-христиански [потерпела / терпела] все
<table>
<thead>
<tr>
<th>Text</th>
<th># items</th>
<th># verb pairs</th>
<th># respondents</th>
<th># outliers</th>
<th># respondents - outliers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Беспризорник Жук</td>
<td>300</td>
<td>150</td>
<td>83</td>
<td>1</td>
<td>82</td>
</tr>
<tr>
<td>История о том</td>
<td>160</td>
<td>80</td>
<td>99</td>
<td>4</td>
<td>95</td>
</tr>
<tr>
<td>МГЛУ</td>
<td>278</td>
<td>139</td>
<td>78</td>
<td>2</td>
<td>76</td>
</tr>
<tr>
<td>Нефтяной саммит</td>
<td>166</td>
<td>83</td>
<td>84</td>
<td>4</td>
<td>80</td>
</tr>
<tr>
<td>Выяснили ученые</td>
<td>198</td>
<td>99</td>
<td>72</td>
<td>1</td>
<td>71</td>
</tr>
<tr>
<td>Иван Дмитриевич</td>
<td>244</td>
<td>122</td>
<td>85</td>
<td>3</td>
<td>82</td>
</tr>
<tr>
<td>Totals</td>
<td>1346</td>
<td>673</td>
<td>501</td>
<td>15</td>
<td>486</td>
</tr>
</tbody>
</table>
(1) Categorical negation: here according to “objective” criteria, only imperfective should be possible. "ženščina nikogda ne obrugala / *rugala ego... ‘the woman never yelled at him’

<table>
<thead>
<tr>
<th></th>
<th>impossible</th>
<th>possible</th>
<th>excellent</th>
</tr>
</thead>
<tbody>
<tr>
<td>perfective</td>
<td>79</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>imperfective</td>
<td>0</td>
<td>0</td>
<td>82</td>
</tr>
</tbody>
</table>

(2) No “objective” criterion for choosing aspect, but native speakers consistently choose imperfective. "Pokazalos´ / *Kazalos´, čto ego mat´... byla dlja nego angelom xranitelem. ‘It seemed that his mother was his guardian angel’

<table>
<thead>
<tr>
<th></th>
<th>impossible</th>
<th>possible</th>
<th>excellent</th>
</tr>
</thead>
<tbody>
<tr>
<td>perfective</td>
<td>80</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>imperfective</td>
<td>0</td>
<td>1</td>
<td>81</td>
</tr>
</tbody>
</table>

(3) No “objective” criteria, and in this case native speakers accept both aspects. "Deti u mačexi Vasilija [ *pošli / šli ] odin za drugim. ‘Vasilij’s stepmother had (lit. ‘went’) one child after another’

<table>
<thead>
<tr>
<th></th>
<th>excellent</th>
<th>possible</th>
<th>impossible</th>
</tr>
</thead>
<tbody>
<tr>
<td>perfective</td>
<td>7</td>
<td>26</td>
<td>49</td>
</tr>
<tr>
<td>imperfective</td>
<td>11</td>
<td>47</td>
<td>24</td>
</tr>
</tbody>
</table>
Sometimes respondents were very undecided:

Fagov [ podvergli / *podvergali ] polnogenomnomu sekvenirovaniju...
‘The phages underwent full gene sequencing...’

<table>
<thead>
<tr>
<th></th>
<th>impossible</th>
<th>possible</th>
<th>excellent</th>
</tr>
</thead>
<tbody>
<tr>
<td>perfective</td>
<td>24</td>
<td>24</td>
<td>23</td>
</tr>
<tr>
<td>imperfective</td>
<td>2</td>
<td>13</td>
<td>56</td>
</tr>
</tbody>
</table>

So what do you expect in the distribution?
Any breaks or gaps between items that are categorical and those that are not?
The mode (the most popular choice) and what percentage of people chose it.
Weighted averages:
невозможно ‘impossible’ = 0
dопустимо ‘possible’ = 1
отлично ‘excellent’ = 2
Along the x-axis are the 1346 items, sorted according to their average score from 0 to 1.
Along the x-axis are the 1346 items, sorted according to their average score from 0 to 1, but divided into two groups: original (black) are items that match the original text, non-original (grey) are items that conflict with the original text.

**Binary data:**
невозможно ‘impossible’ = 0
dопустимо ‘possible’ or отлично ‘excellent’ = 1
Binary data:
невозможно ‘impossible’ = 0
dопустимо ‘possible’ or отлично ‘excellent’ = 1

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>mean</th>
<th>median</th>
<th>1st quartile</th>
</tr>
</thead>
<tbody>
<tr>
<td>original</td>
<td>673</td>
<td>0.967</td>
<td>1.000</td>
<td>0.974</td>
</tr>
<tr>
<td>non-original</td>
<td>673</td>
<td>0.376</td>
<td>0.316</td>
<td>0.085</td>
</tr>
</tbody>
</table>
Along the x-axis are the 673 pairs, sorted according to the absolute difference between their average scores from 0 (indicating uncertainty) to 1 (indicating certainty).
Summary of Study 2: Syntagmatic Perspective

• For 2% of verb forms in a corpus the choice of aspect is clearly marked by a “trigger” in the context: here everyone (L1, L2, machine learning) should know what to do and be correct 96% of the time
  – But what about the rest? How do native speakers know which aspect is most felicitous?
  – What about the examples where there is variation? How do they differ from the ones that are clear to native speakers?
  – Why are native speakers so good at rating the original aspect and so bad at rating the non-original aspect?
Conclusions

• In the many cases, the aspect of a verb can be determined either solely on the basis of the distribution of forms, or solely on the basis of context.
• It is likely that L1 learners use both cues in acquisition.
• But we don’t know enough about the cues.
• More study could tell us more about the role of construal in language.
• And we could learn things that can be applied to pedagogy.
References


