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 (Martella 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’ pisal/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 biaspectual (ženit’ ‘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’)

Where the aspects do and do not compete
- Paradigmatically competing:
  - Non-past (future if perfective, present if imperfective)
  - Past
  - Imperative
  - Infinitive
- Syntagmatically competing:
  - In some contexts, either aspect is grammatical
- Paradigmatically non-competing:
  - Present gerunds and participles are perfective
  - Past gerunds and participles are imperfective
- Syntagmatically non-competing:
  - In some contexts only one aspect is allowed

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)

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

Janda & Lyashevskaya 2011
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>180</td>
</tr>
<tr>
<td>Scientific-technical</td>
<td>43,528</td>
<td>4,494</td>
<td></td>
</tr>
<tr>
<td>Fiction</td>
<td>78,084</td>
<td>6,656</td>
<td></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 2 (19.7\%) has a larger inertia origin, the greater the chi-squared representing a row or a column from the row profiles. The further a point is from the centre of gravity, the greater the deviation from the mean, 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!

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?

R. Harald Baayen

"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."

Natalia Levshina

"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"
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)
Items and respondents (13.-20.09.2016)

<table>
<thead>
<tr>
<th>Text</th>
<th># items</th>
<th># verb pairs</th>
<th># respondents</th>
<th># outliers</th>
<th># respondents</th>
<th>- outliers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Беспризорник</td>
<td>300</td>
<td>150</td>
<td>83</td>
<td>1</td>
<td>82</td>
<td></td>
</tr>
<tr>
<td>Жук</td>
<td>199</td>
<td>99</td>
<td>76</td>
<td>2</td>
<td>74</td>
<td></td>
</tr>
<tr>
<td>История</td>
<td>156</td>
<td>81</td>
<td>84</td>
<td>4</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>Нефтяной саммит</td>
<td>100</td>
<td>50</td>
<td>72</td>
<td>1</td>
<td>71</td>
<td></td>
</tr>
<tr>
<td>Московские унiversity</td>
<td>140</td>
<td>70</td>
<td>85</td>
<td>3</td>
<td>82</td>
<td></td>
</tr>
<tr>
<td>Миха Днепровский</td>
<td>143</td>
<td>71</td>
<td>85</td>
<td>3</td>
<td>82</td>
<td></td>
</tr>
<tr>
<td>Totals</td>
<td>1,346</td>
<td>673</td>
<td>107</td>
<td>15</td>
<td>486</td>
<td></td>
</tr>
</tbody>
</table>

(1) Categorical negation: here according to "objective" criteria, only imperfective should be possible. Женщина никогда не [обругала / *ругала] его...

- The woman never yelled at him

(2) No "objective" criterion for choosing aspect, but native speakers consistently choose imperfective ([Показалось / *Казалось], ч то его мат... byla dla nego angelom xranitelem

- It seemed that his mother was his guardian angel

(3) No "objective" criteria, and in this case native speakers accept both aspects.

- Дети у матери Василия [*пошли / шли] одн за другим. Василия’s stepmother had (lit. ‘went’) one child after another

Sometimes respondents were very undecided:

<table>
<thead>
<tr>
<th>Mode</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>10</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?

Weighted averages:

<table>
<thead>
<tr>
<th>Category</th>
<th>Weighted avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>невозможное 'impossible'</td>
<td>0</td>
</tr>
<tr>
<td>допустимо 'possible'</td>
<td>1</td>
</tr>
<tr>
<td>отлично 'excellent'</td>
<td>2</td>
</tr>
</tbody>
</table>

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 (gray) are items that conflict with the original text

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