Automating the identification of meter and rhyme in Russian verse

David J. Birnbaum
djbpitt@gmail.com
http://www.obdurodon.org

LAUDATIO – Mini-Workshop
2013-05-07

Outline

- The nature of Russian verse
- Quantitative metrics and the study of Russian verse
- The task: automate identification of meter and rhyme
- Why it's worth doing
- Why it's hard
- Why it's nonetheless possible
- How to make it easier
- Why it's still hard
  … and therefore interesting

The nature of Russian verse

- Classical Russian poetry is *syllabotonic*
  - Organized by *number of syllables* and *place of stress*
  - Like English
- Poems are divided into *lines* (sometimes *stanzas*)
- Lines are divided metrically into *feet*
  - Irrespectively of *word divisions*
- Lines with phonetically similar endings *rhyme*
- Metrical and rhyming *patterns* recur
- Meter and rhyme create poetic *cadence* and *structure*

Example (iambic pentameter)

No longer mourn for me when I am dead
Than you shall hear the surly sullen bell
Give warning to the world that I am fled
From this vile world with vilest worms to dwell:

Meter: o x | o x | o x | o x | o x
Iambic = o x
Pentameter = five feet (iambs)
Rhyme: abab

(William Shakespeare, Sonnet 71, lines 1–4)

Lexical stress vs metrical ictus

No longer mourn for me when I am dead
Than you shall hear the surly sullen bell
Give warning to the world that I am fled
From this vile world with vilest worms to dwell:

o x | o x | o o | o o | o x
o x | o x | o x | o x | o x
o x | o o | o x | o o | o x
o x | o x | o x | o x | o x

Lexical stress vs metrical ictus

- Pyrrhic (o o), spondee (x x), trochaic (x o) substitutions in iambic verse
- Metrical variation
  - Preserves meter, while preventing poetry from becoming "sing-song"
  - Establishes associations among words
  - Draws attention to important moments
  - Adapts international meter to local linguistic properties (stress system, word length)

Meter and language

- Stress
  - Long English words often have secondary stress
  - Secondary stress occurs in Russian only in compound words: trëxètažnyj 'three-story'
  - Otherwise Russian words, no matter how long, have only primary stress: dostoprimečatel'nost'
- Word length
  - Average word length in Shakespeare Sonnet 71 is 3.8 letters = *lots of short words*
  - Average word length in first stanza of Pushkin's *Eugene Onegin* in Russian is 9.5 letters = *lots of long words*
  - Neither English nor Russian fits iambic meter naturally

Quantitative metrics

- Andrej Belyj, Boris Tomaševskij, Viktor Žirmunskij
- Kiril Taranovski, *Russian binary meters,* 1953
- Mixail Gasparov, 1960+
- James Bailey, Nila Friedberg, Emily Klenin, Barry Scherr, J. Thomas Shaw, Marina Tarlinskaja
- Morris Halle, Bruce Hayes, Paul Kiparsky

What quantitative metrics tells us about Russian verse

- Final stress must always be realized
- "Law of regressive accentual dissimilation" (Taranovski)
  - Pre-final foot is weakest
  - Iambic tetrameter: 2 3 1 4
  - Iambic pentameter: 3 2 4 1 5
- Pattern holds over 18th, 19th, 20th centuries (Friedberg), but with changes
- No such regularity in English (Tarlinskaja)

But we already knew that!

- How can we be confident? Can we prove it?
  - Humans are not uniformly attentive to details
  - We notice the presence of oddities easily
  - We don't notice absences as easily
- Accuracy and consistency
- Reproducible results
- Examine and critique the evidence
  - Selection of texts
  - Method of counting (stress is not really binary)

Quantitative metrics is

- Labor-intensive
- Error-prone
- Hard to maintain (oops! I should have counted …)
- Hard to control (ictus is not really binary)
- Hard to validate

- Can it be automated?

Why it's hard

- Native Russian orthography almost never marks stress
- Input must be in native Russian orthography
- Meter
  - Meter depends on stress
- Rhyme
  - Rhyme depends on pronunciation
  - Pronunciation can be inferred from orthography only if stress is also known
- But if we can add determine stress automatically …

Why it's possible: meter

- Russian classical verse is overwhelmingly
  - Regular and syllabotonic
  - Binary or ternary
- Every vowel letter is syllabic
  - We can count syllables by counting vowel letters
- If we know which vowels are stressed, we can infer meter

Ambient meter

- Binary or ternary
  - Tabulate all distances from stress to stress
  - What percentage is divisible by 2? By 3?
  - (Use interstress distance instead of stress position to avoid anacrusis/catalexis distortions)
- Number of feet predictable from number of syllables and binary/ternary value
- Subcategory
  - Last stress is obligatory
  - Number of syllables before final stress distinguishes iamb ~ trochee and anapest ~ dactyl (possible anacrusis complications?)

Why it's possible: rhyme

- Russian spelling is broadly phonetic
  - Except that it doesn't mark stress
- If we know the place of stress, we can convert automatically form orthography to broad phonetic transcription

What we lose

- e [e] ~ ë [o]
  - vse [fs'e]
  - vsë [fs'o]
  - tetja [t'ét'i]
  - telo [t'élə]
- Degrees of "vowel reduction"
  - goroda ('city' Gsg) [górədə]
  - goroda ('city' NApl) [gərʌdá]
- Idiosyncrasies
  - solnce [sóntse]
  - solnečnyj [sóln'ičnij]
- Clitic stress
  - zá ruku ('by the hand' [Asg])
  - né bylo ('there wasn't' [past Nsg])

Vowel reduction depends on stress

- Five vowel phonemes: /a e i o u/
- Unstressed /o/, /a/ after nonpalatalized C > [ə]
- Unstressed /o/, /a/ after palatalized C > [i]
- Unstressed e > [i]

- We ignore unstressed /a/ > [ə ~ ʌ]
- We ignore unstressed /e/ > [e]

Classical Russian rhyme

- Final stressed vowel and everything following it must match
- Open masculine rhyme requires a preceding "supporting" consonant
- I think that I shall never *see*
  A poem lovely as a *tree*
  - Not a rhyme in Russian

("Trees," Joyce Kilmer)

Taking stock

- We can count syllables by counting vowels
- If we know the place of stress
  - We get meter
  - We almost get pronunciation, and therefore rhyme
- If we also know e ~ ë
  - We also get rhyme

How to make it easy

- Figure out the place of stress in every word
- Input: Russian verse in normal orthography
- Output: Russian verse with stress marked
  - Format can be adjusted to subsequent processing requirements

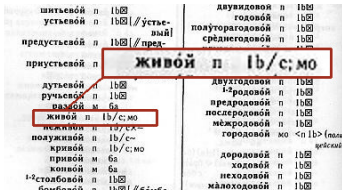
Why that's possible

- Russian stress is lexical and morphological
- Lexical: stress pattern of a word is not reliably predictable from anything else
- Morphological: stress may occur in different places in the same lexeme under inflection
  - ruká 'hand' Nsg, rúku Asg, rukí Gsg, rúki Napl
- Stress shifts under inflection are constrained
- Stress pattern is part of inflectional paradigm
- Stress patterns are numerous, but not infinite

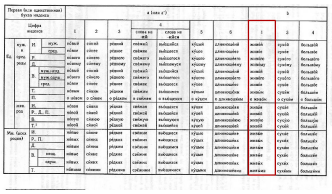
How to get stress for free

- Take all of the lexemes in Russian
- Tag each according to its inflection paradigm, including stress pattern
- Automatically generate all inflected forms, with stress
- Use a database to pass in a form without stress and return all stress possibilities
- Use a database to pass in a form without stress and return all stress possibilities
- (Resolve lacunae and ambiguities later …)

The Zaliznjak grammatical dictionary



Paradigm from dictionary



Keypunch file

```
XIVODERSTVO          0101 XIVOD<ÐRSTVO S 1ª
XIVODERSTVOVAT+ 0101 XIVOD<ÐRSTVOVAT+ NSV NP 2ª
XIVOJ                0101DXIV<OJ P 1V/S
XIVOJ                0101BXIV<OJ MO <P 1V/S>
XIVOKOST+            0101 X<IVOKOST+ X 8A (_RASTENIE_)
XIVOPISANIE          0101 XIVOPIS<ANIE S 7ª
XIVOPISAT+           0101 XIVOPIS<AT+ NSV 1A? @ _NAST._ XIVOPIS<UÐ, -ET
```

XML lexicon

```
<item>
   <unstressed>живой</unstressed>
   <line-total>01</line-total>
   <stressed>жив<stress>о</stress>й</stressed>
   <pos>n</pos>
   <index>1b/c</index>
</item>
```

(simplified)

XML paradigm

```
<paradigm index="1b/c" truncate="2">
   <!-- sample: живой -->
   <form case="N" gender="m" number="sg"><stress>о</stress>й</form>
   <form case="N" gender="n" number="sg"><stress>о</stress>е</form>
   … other long form endings …
   <form case="A" animacy="a" number="pl"><stress>ы</stress>х</form>
   <form case="I" number="pl"><stress>ы</stress>ми</form>
   <form type="short" gender="m" number="sg"/>
   <form type="short" gender="f" number="sg"><stress>а</stress></form>
   <form type="short" gender="n" number="sg">о</form>
   <form type="short" gender="mfn" number="pl">ы</form>
   <form type="comp"><stress>е</stress>е</form>
</paradigm>
```

Components: input

- Input: natural Russian orthography
- Dictionary: add lexical stress (includes lacunae, ambiguities)
- Interim output: text with stresses identified
  - Format to be determined

Components: poetic analysis

- Retrieve stress from dictionary
- Metrical analysis 1: determine ambient meter
- Stress disambiguation: use ambient meter to resolve lexical stress ambiguities and lacunae
  - Add lacunae to dictionary for future use
  - Fails where multiple resolutions are possible
- Metrical analysis 2: reevaluate meter, including metrical variation
- Rhyme analysis 1: determine rhyme scheme
- Rhyme analysis 2: identify non-canonic rhyme

Example: binary or ternary

- Mój djádja sámyx čéstnyx právil
  Kogdá ne v šútku zanemóg
  Ón uvažát' sebjá zastávil
  I lúčše výdumat' ne móg
- x x ○ x ○ x ○ x ○ (1,2,2,2)
  ○ x ○ x ○ ○ ○ x (2,4)
  x ○ ○ x ○ x ○ x ○ (3,2,2)
  ○ x ○ x ○ ○ ○ x (2,4)
- Syllables per line: 9-8-9-8
  Interstress distances: 1, 2, 2, 2; 2, 4; 3, 2, 2; 2, 4 (total: 11)
  Interstress distances divisible by 2: 9/11 = 82%
  Interstress distances divisible by 3: 1/11 = 9%

→ Binary meter

Example: metrical type

- Mój djádja sámyx čéstnyx právil
  Kogdá ne v šútku zanemóg
  Ón uvažát' sebjá zastávil
  I lúčše výdumat' ne móg
- x x | ○ x | ○ x | ○ x (○)
  ○ x | ○ x | ○ ○ | ○ x
  x ○ | ○ x | ○ x | ○ x (○)
  ○ x | ○ x | ○ ○ | ○ x
- Binary meter
  Last stressed syllable of every line is 8 (even)

→ Iambic tetrameter
  Hypercatalectic first and third lines

Example: Reconciling ictus and stress

- Moj djádja sámyx čéstnyx právil
  Kogdá ne v šútku zanemóg
  On uvažát' sebjá zastávil
  I lúčše výdumat' ne móg
- ○ x | ○ x | ○ x | ○ x (○)
  ○ x | ○ x | ○ ○ | ○ x
  ○ ○ | ○ x | ○ x | ○ x (○)
  ○ x | ○ x | ○ ○ | ○ x
- Iambic tetrameter
  Pronouns stressed only in strong metrical positions (Friedberg, p. 34)

→ Three pyrrhic feet
  Distribution follows regressive accentual dissimilation

Example: rhyme

- Moj djádja sámyx čéstnyx právil
  Kogdá ne v šútku zanemóg
  On uvažát' sebjá zastávil
  I lúčše výdumat' ne móg.
- moj DáDi sámix čéstnix práVil
  kagdá NiSútku zaNimók
  on uvažáT SiBá zastáVil
  ilúčši vídumaT Nimók

→ abab rhyme
  Rhyme of lines 2 and 4 is richer than minimum

Components: output

- Poem with stress and meter
- Metrical report (ambient meter, metrical variation)
- Rhyme report (rhyme scheme, non-canonic or enriched rhyme)
- Enriched dictionary

Ah, and about the dictionary …

- Includes
  - Lexeme
  - Part of speech
  - Inflectional class
  - Fully inflected paradigm
    - Including grammatical categories
    - Including stress
- Annotation possibilities
  - Lemmatization
  - Part of speech tagging
  - Morphological annotation
- Quantitative poetic analysis according to the preceding

Thank you!

David J. Birnbaum
http://www.obdurodon.org
djbpitt@gmail.com

Assisted by: Sam Depretis, Erin Harrington, Elise Thorsen