Markup and meter

Using XML tools to teach a computer to think about versification

David J. Birnbaum and Elise Thorsen
http://poetry.obdurodon.org

Balisage 2015: The markup conference, 2015-08-11

About this presentation

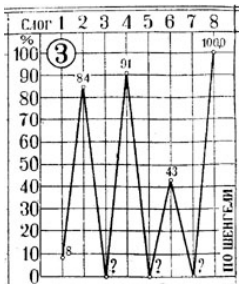
- Poetry scholars care about form
  - Meter, rhyme etc.
  - Can we identify form without massive human effort?
- What else
  - Possibly a different way to think about poetry
  - And because this is Balisage:
    - Possibly a different way to think about overlap

Outline

- Verse scholarship
  - Research goals and the research context
  - Stress, meter, and rhyme
- Markup
  - Explicit and implicit markup
  - Explicit and implicit overlap
  - Elements without borders
  - Mixed content as a type of overlap

Research framework

- Input
  - Russian plain-text verse corpus in normal orthography
- Output
  - Meter, rhyme, and other formal features
  - Text- and corpus-level analysis



Why it's hard intellectually

- Written text is accessible
  - But it lacks information about stress and pronunciation
- Meter and rhyme depend on the place of stress
  - The place of stress is not marked
  - The place of stress is not predictable without linguistic knowledge
  - Metrical and linguistic emphasis may differ
- Rhyme further depends on pronunciation

Why it's hard technically

- Overlapping hierarchies
- Mixed content

Assumptions

- Russian quantitative verse studies are worth doing
  - Andrej Belyj, 1910, *Simvolizm*; Jurij Tynjanov, 1924, *Problema stixotvornogo jazyka*; Viktor Žirmunskij 1925, *Vvedenie v metriku. Teorija stixa*; Kiril Taranovski, 1953, *Ruski dvodelni ritmovi*; Boris Ėjxenbaum, 1969, *O poèzii*; Mixail Gasparov, 1984, *Očerk istorii russkogo stixa. Metrika, ritmika, rifma, strofika*
  - Vladimir Nabokov, 1964, *Notes on prosody*; J. Thomas Shaw, 1993, *Pushkin's poetics of the unexpected: The nonrhymed lines in the rhymed poetry and the rhymed lines in the nonrhymed poetry*, Ian K. Lilly, 1995, *The dynamics of Russian verse*
  - Handbooks and textbooks: Boris Unbegaun 1956, Barry Scherr 1986, Michael Wachtel 2004
  - Generative poetics: Morris Halle, Bruce Hayes, Paul Kiparsky
  - Names to watch: James Bailey, Nila Friedberg, Emily Klenin, Barry Scherr, J. Thomas Shaw, Marina Tarlinskaja
- Target corpus is generally regular syllabotonic verse: stanzas, lines, feet, meter, rhyme

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:
[Shakespeare, Sonnet 71, iambic pentameter]

○ x | ○ x | ○ ○ | ○ ○ | ○ x
○ x | ○ x | ○ x | ○ x | ○ x
○ x | ○ ○ | ○ x | ○ ○ | ○ x
○ x | x x | ○ x | ○ x | ○ x

Lexical stress vs metrical ictus

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

Meter and language: orthography

- In English
  - The relationship between vowel letters and vowel sounds (syllables) is not one to one
- In Russian
  - Every vowel is syllabic
  - No silent vowels (cf. English Adelaide)
  - No representation of single vowel sounds by sequences of vowel letters (cf. Eng. Adelaide)
- Which means
  - Vowel *letters* in Russian are surrogates for *syllables*

Meter and language: stress

- English
  - Long words often have secondary stress
- Russian
  - Secondary stress only in compound words: трёхэтажный trëxètažnyj 'three-story'
  - Otherwise Russian words, no matter how long, have only primary stress: достопримечательность dostoprimečatel'nost' '(tourist) attraction'
  - So what happens with long words in binary meter?

Implicit meter and actual stress

Meter and language: 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 binary meter naturally

Meter and language: verse convention

- Russian
  - Strong sense of line
  - Strong sense of foot
  - Strong syllabotonic orientation
- English
  - Stronger role for tonic organization

"The old woman of Berkeley"

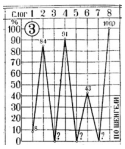
Robert Southey (1774–1842; 1799)

The ra|ven croaked | as she sate | at her meal, |	2 2 3 3	a
And the Old | Woman knew | what he said; |	3 3 3	b
And she | grew pale | at the Ra|ven's tale, |	2 2 3 2	c
And sick|ened, and went | to her bed. |	2 3 3	b

Vasilij Andreevič Žukovskij (1783–1852; 1814/1831)

На кро|вле во|рон ди|ко про|кричал —|	2 2 2 2 2	a
Стару|шка слы|шит и | бледнеет.	2 2 2 2+	B
Понят|но ей, | что во|рон тот | сказал: |	2 2 2 2 2	a
Слегла | в постель, | дрожит, | хладеет.	2 2 2 2+	B

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)

[image from: N. V. Lapšina, I. K. Romanovič, and B. I. Jarxo, *Metričeskij spravočnik k stixotvorenijam A. S. Puškina*, Moscow: Academia, 1934, p. 134bis. http://feb-web.ru/feb/pushkin/critics/jar/jar-005-.htm]

What the system should tell us

- For individual poems:
  - Identify which syllables are stressed linguistically
  - Identify metrical structures and ambient meter
  - Identify deviations from the ambient meter
  - Identify rhyme schemes
  - (Other formal regularities?)
- Corpus level
  - Historical patterns (authors, periods, movements)
  - Relationships between form and meaning
    - E.g., semantic halo

TEI

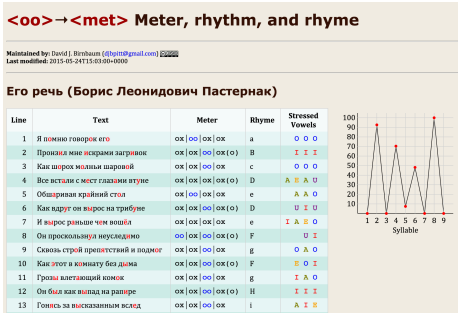
<div type="book" n="1"
met="-+|-+|-+|-+|-+/" rhyme="aa">

- Inline vs standoff
  - Meter and rhyme stand apart from text
- Data vs metadata
  - The human analyzes the line and writes the result
- Tagging the text vs (meta)tagging the poem

Tagging the text vs the poem

```
<l>
    <w>
        <orth>Я</orth>
        <str>я</str>
    </w>
    <w>
        <orth>помню,</orth>
        <str>п<stress>о</stress>мню</str>
    </w>
    <w>
        <orth>говорок</orth>
        <str>говор<stress>о</stress>к</str>
    </w>
    <w>
        <orth>его</orth>
        <str>ег<stress>о</stress></str>
    </w>
</l>
```

Sample browsing output



From plain text input to rich output

- Input must be in native Russian orthography
  - Native Russian orthography almost never marks stress
- 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 determine stress automatically …

Procedure

1. Make stress explicit (dictionary lookup)
2. Metrical valence
3. Strong and weak position
4. Metrical type (binary ~ ternary)
5. Metrical subtype (foot type)
6. Line length
7. Catalexis and hypermetricality
8. (Rhyme)

[All processing is XQuery and XSLT]

1. Dictionary lookup

- Input is word in normal orthography
  - Mixed case, punctuation, no stress
  - Dictionary contains word forms with stress and morphological information
    - Morphological information is irrelevant for our purposes
- Eventual output has all vowels tagged
  - Stressed
  - Unstressed
  - Unknown
    - Not in dictionary
    - Dictionary evidence is contradictory

Dictionary content (eXist)

```
<item>
    <unstressed>карий</unstressed>
    <stressed>к<stress>а</stress>рий</stressed>
    <pos>n</pos>
    <form>
      <categories>
        <category case="N" gender="m" number="sg"/>
        <category case="A" gender="m" animacy="i" number="sg"/>
      </categories>
      <content>к<stress>а</stress>рий</content>
    </form>
    <!-- other <form> elements -->
</item>
```

Final dictionary output

к
<vowel stress="1">а</vowel>
р
<vowel stress="-1">и</vowel>
й

2. Metrical valence

- For each vocalic position in the line
  - Ignore text node children of <line>
  - Line as sequence of <vowel> element
- Metrical valence
  - stressed / (stressed + unstressed)
    - Ignore unknowns
  - Varies between 0 and 1
  - Sample output: 0 0.5 0 1 0 0 0 1 0

3. Strong and weak position

- Compare valence of each position to preceding and following
  - Assume a 0 value if preceding or following is missing, i.e., at beginning or end of a line
- If target value is higher than both neighbors: strong
- If target value is lower than both neighbors: weak
- Otherwise: weak
  - Provisional; adjacent strong positions do not occur in common Russian meter

4. Metrical type (binary ~ ternary)

- Calculate how often the strong ~ weak property of a syllable matches the property two (resp. three) syllables earlier
- Count both strong/strong and weak/weak matches
- The greater number of matches determines the type
- Resolve tagging ambiguities according to positional valence (where possible)
- Retag all vowels with @stress values of 0 or 1
  - Represents strong ~ weak (not necessarily stress)

5. Metrical subtype (foot type)

- Having determined metrical type (binary ~ ternary)
- Subtype is based on last foot
  - Last stress is the only obligatory one
  - Iamb ~ trochee
  - Dactyl ~ anapest ~ amphibrach

6. Line length (number of feet)

- Number of strong positions = number of feet
- May be global or line-specific

Во всем мне хочется дойти	ox | ox | ox | ox
До самой сути.	ox | ox (o)
В работе, в поисках пути,	ox | ox | oo | ox
В сердечной смуте.	ox | ox (o)

[Pasternak 1956]

7 Catalexis and hypermetricality

- Catalexis: Number of syllables
  - Is sufficient for the number of feet
  - Is not sufficient for the number of complete feet
- Hypermetricality
  - Syllables after the final stress are easily identified
  - Hypermetrical caesura: Demarcate feet based on strong position

Как ветер мокрый, ты бьешься в ставни,	ox|ox(o) ‖ ox|ox(o)
Как ветер черный, поешь: ты мой!	ox|ox(o) ‖ ox|ox
Я древний хаос, я друг твой давний,	ox|ox(o) ‖ ox|ox(o)
Твой друг единый,- открой, открой!	ox|ox(o) ‖ ox|ox

[Gippius, Neljubov', 1907]

Rhyme

- Meter vs rhyme
  - Meter can be identified through vowel *letters*
  - Rhyme requires *sounds*
    - род ~ рот [ˈrot]
    - ноги [naˈgʲi] ~ ноги [ˈnogʲi]
  - Rhyme runs from last stressed vowel to end of line
    - Exception: open masculine rhyme: *tree ~ see*
- Russian orthography can be mapped to coarse phonetics algorithmically, except that
  - Stress must be known
  - e ~ ë must be resolved
  - Both are accessible in dictionary

Meter ~ rhyme

- Meter: privilege markup
  - Distinguish line/vowel from line/text()
- Rhyme: privilege text
  - Convert mixed content to text
  - String matching
- Imperfect (slant) rhyme: privilege markup
  - Decompose segments into distinctive features
  - Identify rhyme scheme on the basis of exact rhyme
  - Infer imperfect rhyme on the basis of the ambient rhyme (cf. ambient meter)
  - Characterize slant rhyme by neutralized features

Taking stock

- We can count syllables by counting vowel letters
- If we know the place of stress
  - We get meter
  - We get most pronunciation …
    … and therefore most rhyme
- If we also know e ~ ë
  - We get the rest of pronunciation
  - We also get rhyme

Overlapping hierarchies: caesura

- Свои|ми кольцами ‖ она, | упорная,
- Iambic tetrameter, dactylic caesura and clausula
- Feet: o x|o x(o o)‖o x|o x(o o)
- Words: o x o|x o o ‖o x|o x o o

```
<word>
   св
   <vowel stress="0">о</vowel>
   <vowel stress="1">и</vowel>
   м
   <vowel stress="0">и</vowel>
</word>
```

Implicit and explicit markup

- P5: "A text is not an undifferentiated sequence of words, much less of bytes."
  - Quotation marks delimit quotations
  - Space characters delimit words
  - New line characters delimit lines of poetry
  - Multiple new line characters delimit paragraphs of prose
  - Asterisks or underscores delimit emphasized text,
- Metrical foot and word hierarchies overlap
- Can we use a combination of explicit and implicit markup to represent the *logical* overlap without *syntactic* overlap?
  - Is there pseudo-markup of the metrical hierarchy that we can use?

Implicit and explicit markup

- Words are tagged explicitly as <word>
- Feet are implicit
  - Represented at level of line/descendant::vowel
  - Ignore everything else

```
let $vowelCount := min(//line/count(descendant::vowel))
let $midPoint := $vowelCount div 2
let $targets := //line/descendant::vowel[position() eq $midPoint]
return
   if ($targets/following-sibling::vowel)
   then
      'no caesura'
   else
      'caesura'
```

Q: So where are the foot boundaries?

<word>сво<stress>и</stress>ми</word>
<word>к<stress>о<stress/>льцами</word>
<word>он<stress>а</stress></word>
<word>уп<stress>о</stress>рная</word>

- *A: We don't care!*

- If there is a word boundary between feet
  - Word boundary = foot boundary (potential caesura)
- Otherwise
  - We don't need to locate the foot boundary
- Not only do we not need to tag the foot boundaries, but we don't even need to be able to find them

What if the words weren't tagged?

<w>Св<v>о</v><v>и<v/>м<v>и</v></w>
<w>к<v>о</v>льц<v>а</v>м<v>и</v></w>
<w><v>о</v>н<v>а</v></w>
<w><v>у</v>п<v>о</v>рн<v>а<v/><v>я</v></w>

Сво<str>и</str>ми
к<str>о</str>льцами
он<str>а</str>
уп<str>о</str>рная

What if the words weren't tagged?

- The data
  - Feet are still implicit
  - Now words are encoded with pseudo-markup white space
- Processing
  - What we want to do is tokenize(line)
    - Oops!
  - Convert markup to text
  - Convert text to markup
    - Replace white space in line/text() with <w/> tags
    - Convert milestones to wrappers (<xsl:for-each-group>)

What if it isn't just words?

<lpa>съ грѣкы. w̃пусти слы w<marginalia>даривъ.
<lb/> ско</marginalia>рою</lpa>

Not generally identified as overlap

- Syntactically it isn't
- Logically it is
- White space as pseudo-markup
- It may raise the same processing challenges as traditional types of overlap

So what have we learned?

- We can identify an "element" without start or end tags
  - And without knowing where it starts and ends
  - Foot boundaries are not only untagged, but also unknown
- Overlapping hierarchies may hide in plain sight
  - An overlapping hierarchy may be encoded through plain text pseudo-markup (e.g., white space)
  - Avoidance of overlap is only apparent, and may vanish when the implicit hierarchy needs to be processed
- Needing to treat words during processing as though they had been marked up individually is as much overlap as when the markup is overt
  - White space as milestone

So what have we learned?

- With absent and implicit markup, as with other strategies for avoiding syntactic overlap
  - No well-formedness errors
  - Similar processing challenges as soon as the researcher needs to engage with them explicitly

Thank you!

Elise Thorsen (enthorsen@gmail.com)
David J. Birnbaum (djbpitt@gmail.com)
http://poetry.obdurodon.org

Assisted by: Sam Depretis, Erin Harrington
Thanks to: Elisa Beshero-Bondar, Sibelan Forrester

Birnbaum, David J., and Elise Thorsen. "Markup and meter: Using XML tools to teach a computer to think about versification." Presented at Balisage: The Markup Conference 2015, Washington, DC, August 11–14, 2015. In *Proceedings of Balisage: The Markup Conference 2015.* Balisage Series on markup technologies, vol. 15 (2015). doi:10.4242/BalisageVol15.Birnbaum01. http://www.balisage.net/Proceedings/vol15/html/Birnbaum01/BalisageVol15-Birnbaum01.html