The beginning of a beautiful friendship: Rule-based and statistical analysis of Middle Russian

Aleksandrs Berdicevskis (UiT), Hanne Eckhoff (UiT) and Tatiana Gavrilova (HSE)

CLEAR
Cognitive Linguistics: Empirical Approaches to Russian

Dialogue, 04.06.2016
Moscow
Introduction

- Historical texts pose particular problems to automatic analysis
  - no standard variant
  - no standard orthography (or many standards)
  - small pool of texts
  - lack of tools for analysis
Aims of this paper

• We describe two tools that provide lemmatisation, part-of-speech and morphology tagging for Middle Russian
  • “TOROT”: a practical statistical procedure for pre-processing texts for the Tromsø Old Russian and OCS Treebank (disambiguating)
  • “RNC”: a rule-based analyser developed for annotating parts of the historical subcorpus of the RNC (no disambiguation)

• Prediction: TOROT will perform better since it does disambiguation
• How good is RNC without disambiguation?
• Can they enhance each other’s performance?
The RNC analyser

- Designed at the School of Linguistics, NRU “Higher School of Economics” for annotating the Middle Russian corpus – a part of the historical subcorpus of the RNC.
- Based on UniParser (Arkhangelsky 2012, Arkhangelskiy, Belyaev and Vydrin 2012) – a rule-based, dictionary-based language-independent morphological analyzer
- Does NOT resolve ambiguity
- UniParser requires:
  - Description of the language’s grammar (a dictionary of inflections)
  - Grammatical dictionary of lexemes
- A module for dealing with spelling variability was developed
The RNC analyser: a dictionary of inflections

- Inflectional class, annotation: \(<\text{number of stem}>\). grammeme
- Created manually, based on:
  - Poljakov’s inflection tables for OCS (Poljakov 2014)
  - Zalizniak’s inflectional tables of Old Russian

- Diachronic rules are applied to the dictionary:
  - Loss of palatalization
  - Palatalisation of velars
    
    други – други – друзья
    N1g, Nom.pl: \(<0>.ы – <0>.и – <1>.и\)
The RNC analyser: a dictionary lexemes

• Every lexeme is annotated by POS, inflectional class and possible stems
• Based on:
  – Poljakov’s OCS grammatical dictionary (Poljakov 2014)
  – Pronouns added manually
• Stems generated automatically for which inflectional class using rules бра-ти – бер-у, окн-о – окон-ъ / окен-ъ
• Diachronical rules are applied to the dictionary:
  – new inflectional classes were added
  – ἰ-declension, masculine → ὀ-declension (гость)
  – ръ → е
• Some regular differences between OSC and Old Russian were taken into account
  – едîн – один
  – одежда - одежа
The RNC analyser: a module for dealing with spelling variability

- **Absolute approach - add stems in the dictionary:**
  - кели-я → кель-я
  - княгин-я → княин-я
  - Verb prefixes

- **Relative approach – preprocessing of the text and the dictionaries:**
  - Capital letters → lower case
  - Letters, corresponding to the same sound reduced
  - Consonant clusters simplified (переводшик - переводчик)
  - Jers between consonants deleted
  - Some other combinations of letters changed (жы → жи)
The Tromsø Old Russian and OCS Treebank

- Ca. 180,000 word tokens of Old and Middle Russian (browse at https://nestor.uit.no, downloadable versioned data releases at torottreebank.github.io, syntactic queries at http://clarino.uib.no/iness)
  - Lemmatisation
  - Part-of-speech tags
  - Fine-grained morphology (10-place positional tag)
  - Enriched dependency grammar analysis
- Large base of form, lemma and tag correspondences
- Possible to train successful morphological taggers for Old and Middle Russian (either separately or taken as a single stage)
- We combine statistical tagging with direct lookups in the database
Statistical tagging

- We use Trigrams’nTags (TnT, Brants 2000), which uses trigrams and word-final character sequences.
- To optimise annotation, we normalise both the training set and the new text to be tagged:
  - all diacritics stripped off
  - all capital letters replaced with lower-case
  - all ligatures resolved
  - all variant representations of a single sound are reduced to one (including juses and jat)
- Note that we do not normalise the forms stored in the treebank database, they can be maintained in a manuscript-near form.
TOROT preprocessing procedure

• Is the form present in the database?
  – YES: assign the most frequent analysis (lemma, POS and morphology)
  – NO: normalise the form, is it present in the database?
    • YES: assign the most frequent analysis
    • NO: assign the TnT POS + morphology, normalise form to lemma standard
  • Is there a lemma matching the normalised form and the TnT POS?
    – YES: assign that lemma
    – NO: chop off the final character from the lemma and check it against the opening strings of all lemmas with the correct POS, is there a match?
      » YES: assign that lemma
      » NO: repeat unless the form < 4 characters
    – STILL NO MATCH? Assign dummy lemma “FIXME”
da po истинъ велико то иес намъ Ḥ бґа дарованіе дарова сѧ. дивлю же сѧ о сем. како толико лъть минуло. а житиє ег не писано. о семъ съжалих сѧ сѣло. како оубо такови старецъ пречудныи і предобрыи. Ḥнелъже престави сѧ. кѣ лът преиде. никтоже не
The experiment

• The TnT tagger was trained on the full Old and Middle Russian data set (166,183 word tokens, 10,603 lemmata at the time)
• The text (1710 tokens) was preprocessed with the TOROT tagger (> TOROT)
• The preprocessed text was manually annotated by one annotator and proofread by another in the TOROT webapp (> Gold)
• A normalised version of the text was lemmatised and tagged with the RNC tagger (> RNC)
• Gold was compared with TOROT (directly) and with RNC (via harmonisations)
POS tag harmonisation

- The TOROT POS inventory is more fine-grained than the RNC one, and the correspondences are complicated.
- We deemed an RNC POS tag to be correct if it corresponded to any of a list of possible TOROT tags.
- Extreme example: pronouns
  - A-PRO => A-, Pd, Pi, Pk, Pp, Pr, Ps, Pt, Px
  - N-PRO => Pp, Pk, Pi, Px
Morphological tag harmonisation

- Morphology: RNC tags for most of the same features as TOROT, but the configurations differ per part of speech
- RNC has extra features (transitivity, aspect, reflexivity, animacy), which were dropped
- TOROT tags converted to RNC tags stripped of extra features and features problematic due to POS differences
- Long form / short form dropped due to considerable differences in definition of adjectives

- дивлю: 1spia----i > indic, praes, sg, 1p
- старецъ: -s---mn--i > m, sg, nom
Lemma harmonisation

- TOROT lemma orthography is much more archaic than RNC’s
- Gold lemmas were harmonised with the RNC lemmas:
  - Havlik vocalisation routine (all strong jers vocalised, all other jers deleted)
  - б > е
  - кы/гы/хы > ки/гу/хи
  - зс > сс
  - double consonants shortened to one
  - o removed from во- and co- in the beginning of the word longer than four letters
  - жде > же
  - Ad hoc rules for three frequent lemmata: сии, тыи, пьсати
- reduces number of RNC lemma guesses unjustly labeled as wrong to 10
TOROT accuracy, lemmatisation and POS

<table>
<thead>
<tr>
<th>Metric</th>
<th>Lemma + POS, %</th>
<th>POS only, %</th>
<th>Number of tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>69.8</td>
<td>89.5</td>
<td>1710</td>
</tr>
<tr>
<td>Accuracy (not “FIXME”)</td>
<td>88.5</td>
<td>93.9</td>
<td>1348</td>
</tr>
<tr>
<td>Accuracy (no RNC guess)</td>
<td>42.5</td>
<td>78.9</td>
<td>327</td>
</tr>
</tbody>
</table>

- POS accuracy is much better than lemmatisation
- If there is a lemma guess, it’s 88.5 % correct
- When RNC fails to make a guess, lemmatisation accuracy is low in TOROT as well
## RNC accuracy, lemmatisation and POS

<table>
<thead>
<tr>
<th>Metric</th>
<th>Lemma + POS, %</th>
<th>POS only, %</th>
<th>Number of tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy (exact)</td>
<td>47.3</td>
<td>54.2</td>
<td>1710</td>
</tr>
<tr>
<td>Accuracy (fuzzy)</td>
<td>74.3</td>
<td>77.0</td>
<td>1710</td>
</tr>
<tr>
<td>Accuracy (exact, when guess)</td>
<td>58.5</td>
<td>67.0</td>
<td>1383</td>
</tr>
<tr>
<td>Accuracy (fuzzy, when guess)</td>
<td>91.9</td>
<td>95.2</td>
<td>1383</td>
</tr>
</tbody>
</table>

- TOROT always provides a guess, RNC only provides a guess for 1383 of the tokens
- Exact accuracy: RNC provides a single correct guess
- Fuzzy accuracy: RNC provides several guesses, at least one of which is correct
- RNC fuzzy accuracy is higher than TOROT’s accuracy: disambiguation missing
Morphological tags, compared accuracy

<table>
<thead>
<tr>
<th></th>
<th>Accuracy, %</th>
<th>Number of tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOROT</td>
<td>81.5</td>
<td>1710</td>
</tr>
<tr>
<td>RNC (exact)</td>
<td>16.6</td>
<td>1710</td>
</tr>
<tr>
<td>RNC (fuzzy)</td>
<td>70.2</td>
<td>1710</td>
</tr>
<tr>
<td>RNC (exact, when guess)</td>
<td>20.5</td>
<td>1383</td>
</tr>
<tr>
<td>RNC (fuzzy, when guess)</td>
<td>86.8</td>
<td>1383</td>
</tr>
</tbody>
</table>

- RNCs exact accuracy is considerably worse than for lemma/POS
- But again, the fuzzy accuracy is better than TOROT's in the cases where RNC does provide a guess
- Disambiguation missing
TOROT morphology tags, Hamming distances

<table>
<thead>
<tr>
<th>Hamming distance</th>
<th>count</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1393</td>
<td>81.5</td>
</tr>
<tr>
<td>1</td>
<td>128</td>
<td>7.5</td>
</tr>
<tr>
<td>2</td>
<td>57</td>
<td>3.3</td>
</tr>
<tr>
<td>3</td>
<td>14</td>
<td>0.8</td>
</tr>
<tr>
<td>4</td>
<td>38</td>
<td>2.2</td>
</tr>
<tr>
<td>5</td>
<td>14</td>
<td>0.8</td>
</tr>
<tr>
<td>6</td>
<td>26</td>
<td>1.5</td>
</tr>
<tr>
<td>7</td>
<td>10</td>
<td>0.6</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>0.5</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>0.2</td>
</tr>
</tbody>
</table>
Boosting TOROT lemmatisation accuracy

- Can the two analysers help each other out?
- Seems unlikely for POS and morphology at the current stage: TOROT does best
- But we can use RNC lemma guesses to boost TOROT's lemmatisation performance
- We take every token lemmatised as “FIXME” by TOROT and which has one or multiple RNC guess(es)
- For each token, we go through every RNC guess, checking them against the (harmonised) TOROT lemma list
- If there is a match, we assign the lemma and its POS tag from the TOROT lemma list
- We will only find lemmas that are already in the TOROT lemma list
Lemmatisation boosting results

<table>
<thead>
<tr>
<th></th>
<th>Lemma + POS</th>
<th>POS only</th>
<th>Number of tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Success rate when fixing “FIXME”</td>
<td>90.3</td>
<td>92.7</td>
<td>165</td>
</tr>
<tr>
<td>TOROT accuracy</td>
<td>69.5</td>
<td>89.5</td>
<td>1710</td>
</tr>
<tr>
<td>Boosted TOROT accuracy</td>
<td>78.5</td>
<td>91.4</td>
<td>1710</td>
</tr>
</tbody>
</table>

- The booster attempts to provide lemma guesses for 165 tokens, and gets it right in 149 cases.
- The TOROT lemmatisation accuracy is considerably improved.
- There is also a slight improvement in POS tagging
Conclusions

• TOROT vastly outperforms RNC in lemmatisation, POS assignment and morphological analysis, for several reasons
• RNC does not disambiguate (but is slightly better at providing guesses)
• The characteristics of the text favours a statistical analyser trained on a large and varied training set: considerable variability in orthography and morphology, unresolved abbreviations (typical of the era)
• RNC is lemma-oriented and will only provide a guess if it can find a lemma, while TOROT guesses morphology and POS with no reference to lemmatisation
• A future dream analyser for Middle Russian: RNC with a good disambiguator
• Now: The best option is to use TOROT with the RNC lemmatisation booster
• RNC’s Middle Russian subcorpus holds more than 7 million word tokens, 80% success for lemmatisation and POS/morphology is a great practical gain
References


