Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2016"

Moscow, June 1–4, 2016

THE BEGINNING OF A BEAUTIFUL FRIENDSHIP: RULE-BASED AND STATISTICAL ANALYSIS OF MIDDLE RUSSIAN

Berdičevskis A. (aleksandrs.berdicevskis@uit.no),
Eckhoff H. (hanne.m.eckhoff@uit.no)

UiT The Arctic University of Norway, Tromsø, Norway

Gavrilova T. (tanya96gavrilova@gmail.com)

The National Research University "Higher School of Economics", Moscow, Russia

We describe and compare two tools for processing Middle Russian texts. Both tools provide lemmatization, part-of-speech and morphological annotation. One ("RNC") was developed for annotating texts in the Russian National Corpus and is rule-based. The other one ("TOROT") is being used for annotating the eponymous corpus and is statistical. We apply the two analyzers to the same Middle Russian text and then compare their outputs with high-quality manual annotation. Since the analyzers use different annotation schemes and spelling principles, we have to harmonize their outputs before we can compare them. The comparison shows that TOROT performs considerably better than RNC (lemmatization 69.8% vs. 47.3%, part of speech 89.5% vs. 54.2%, morphology 81.5% vs. 16.7%). If, however, we limit the evaluation set only to those tokens for which the analyzers provide a guess and in addition consider the RNC response correct if one of the multiple guesses it provides is correct, the numbers become comparable (88.5% vs. 91.9%, 93.9% vs. 95.2%, 81.5% vs. 86.8%). We develop a simple procedure which boosts TOROT lemmatization accuracy by 8.7% by using RNC lemma guesses when TOROT fails to provide one and matching them against the existing TOROT lemma database. We conclude that a statistical analyzer (trained on a large material) can deal with non-standardised historical texts better than a rule-based one. Still, it is possible to make the analyzers collaborate, boosting the performance of the superior one.

Key words: Old Russian; Middle Russian; morphological tagging; lemmatization; rule-based approach; statistical approach

НАЧАЛО ПРЕКРАСНОЙ ДРУЖБЫ: ПРАВИЛОВЫЙ И СТАТИСТИЧЕСКИЙ АНАЛИЗ СТАРОРУССКОГО ЯЗЫКА

Бердичевский А. (aleksandrs.berdicevskis@uit.no),
Экхофф Х. (hanne.m.eckhoff@uit.no)
Университет Тромсё — Норвежский арктический университет, Тромсё, Норвегия

Гаврилова Т. (tanya96gavrilova@gmail.com)
Национальный исследовательский университет «Высшая школа экономики», Москва, Россия

Ключевые слова: древнерусский; старорусский; автоматическая морфологическая разметка; лемматизация; правиловый подход; статистический подход

0. Introduction

Apart from the usual challenges for NLP, processing of historical texts faces a number of additional ones, such as absence of a standard variant, absence of a standardized orthography and smaller resources, both in terms of existing tools and available texts (Piotrowski 2012). In this paper, we describe and compare two tools for processing Old/Middle Russian[1] texts. Both tools provide lemmatization, part of speech (POS) and morphological annotation. One analyzer (labeled "RNC"), described in Section 1, was developed for annotating parts of the historical subcorpus of the Russian National Corpus, and is rule-based. The other one (labeled "TOROT"), described in Section 2, is statistical, and is used for pre-annotating the eponymous corpus.

Since the analyzers were developed independently, and since they employ two different approaches, it is particularly interesting to compare their performance. Our expectation is that TOROT will perform better, since RNC does not perform disambiguation when several guesses are possible. Apart from testing this expectation empirically, we are also interested in checking whether it is possible to boost the TOROT performance by making the analyzers collaborate.

[1] For the purposes of this article we do not distinguish between Middle Russian "proper" and Church Slavic of the Russian recension, since both models deal well with both types of text, and since many texts are mixed. The text chosen for our performance test (see section 3.1) is a (late) Church Slavic text of the Russian recension.

1. The RNC analyzer

The "RNC analyzer" is a morphological analyzer for Middle Russian designed at Higher School of Economics (Moscow) for annotating the Middle Russian corpus (a part of the Russian National Corpus, ruscorpora.ru). The analyzer is based on Uniparser (Arxangelsky 2012, Arkhangelskiy, Belyaev and Vydrin 2012), which can give grammatical annotation to a text in any language provided that there exists a description of the language's grammar (a dictionary of inflections) and a grammatical dictionary of lexemes. The Uni-parser splits a word in all possible ways and looks for its parts in the description of the grammar and the grammatical dictionary. If one part of the word can be found in the dictionary of inflections, the other one in the dictionary of lexemes, and these parts are marked with the same inflectional class, then the word gets an analysis. There can be several possible analyses for one word. The parser does not create hypotheses for words which cannot be found in the dictionary and does not resolve ambiguity. The Uni-parser is intended for working on modern languages, so a module for dealing with spelling variability was developed by the third author of this paper. All letters which correspond to the same sound are reduced to one letter, geminate consonants are reduced to one letter, all jers between consonants are deleted and so on. Overall more than fifteen rules apply to a wordform before it is processed via Uni-parser.

The description of Middle Russian grammar was created manually. Due to the lack of a grammatical dictionary of Middle Russian, a grammatical dictionary of Old Church Slavic[2] (Poljakov 2014) was used. The dictionary was automatically adapted to Middle Russian: new inflectional classes were added, some regular differences between Old Church Slavic and Old Russian were taken into account. As far as Middle Russian contains both archaic and innovative forms, diachronic rules were applied to the dictionary. As a result, words which changed their inflectional class historically got two classes in the dictionary: the old one and the new one. Some word classes which are missing in the Old Church Slavic dictionary were added manually, e.g. pronouns and pronominal adjectives. The Uni-parser format requires information about all possible stems, so they were created automatically for each lexeme depending on its inflectional class. Different spelling variants were also added in the dictionary. For example, the lexeme "*княгиня*" 'princess' has two stems—"*княгин*" and "*княин*". The second one is a possible spelling variant with loss of the intervocalic *г*.

A lexical entry can contain several paradigms and several stems for each of them. For example the lexeme "*премощи*" 'overcome' has four stems in the dictionary (*премо*, *премог*, *премож*, *премоз*). There can be up to fifteen stems for one lexeme.

[2] http://feb-web.ru/febupd/slavonic/dicgram/

2. The TOROT analyzer

The Tromsø Old Russian and OCS Treebank (TOROT, nestor.uit.no, see Eckhoff & Berdičevskis 2015) contains approximately 175,000 word tokens of annotated Old Russian and Middle Russian text (15th–17th century), fairly equally distributed between the two periods. The texts are all lemmatised and have fine-grained part-of-speech and morphology tags, in addition to syntactic annotation, yielding a large database of form, lemma and tag correspondences. This database is used systematically for linguistic preprocessing of texts: lemmatisation, part-of-speech assignment and morphological tagging. With a training set of this size, it is possible to train very successful statistical morphological taggers for these language stages, either separately or taken as a single stage. For this purpose, the TnT tagger (Trigrams 'n Tags, as described in Brants 2000), a statistical morphological tagger which takes trigrams and word-final letter sequences as its input is used (for the motivation behind this choice, see Skjærholt 2011).

To improve the performance of the tagger, both the training data and the new text to be tagged in the process are normalized. The normalisation consists in considerable orthographical simplification. All diacritics are stripped off, all capital letters are replaced with lower-case letters, all ligatures are resolved (e.g., *ѿ* to *от*), all variant representation of single sounds are reduced to one (all *о* variants are reduced to *о* and all *і* variants are reduced to *и*, for instance).[3] The juses are simplified to *я* and *у* (*ю*), and the jat to *е*.

When preprocessing a text, the tagger output is used in combination with direct lookups in the database.[4] For each word token in the text, it is checked whether that form is present in the database already, first as it is, then again with different kinds of orthographic simplifications. If one or more matches are found, the most frequent analysis (lemma + part of speech + morphology) is assigned. If the form is not found in the base, the TnT part-of-speech and morphology tag are assigned, and an attempt is made to find a suitable lemma in the database. If the word form (normalized to the lemma orthography style) matches a lemma with the part-of-speech tag the TnT tagger assigned, that lemma is assigned. If not, letters from the end of the word form are dropped one by one, the remainder checked again against the opening strings of lemmata of the correct part of speech. If no matches are found, a dummy lemma ("FIXME") is assigned, and the annotators will have to assign a lemma manually. This process is represented as a simplified flowchart on Figure 1.

---

[3] Supplementary materials can be found in the TROLLing data repository at http://hdl.handle.net/10037.1/10303. They include the normalization routine, the harmonization and comparison scripts (Section 3), and more detailed comparison results (Section 4).

[4] We are indebted to Professor Dag Haug at the University of Oslo for writing procedures for Latin and Greek, which we have modified for Slavic.

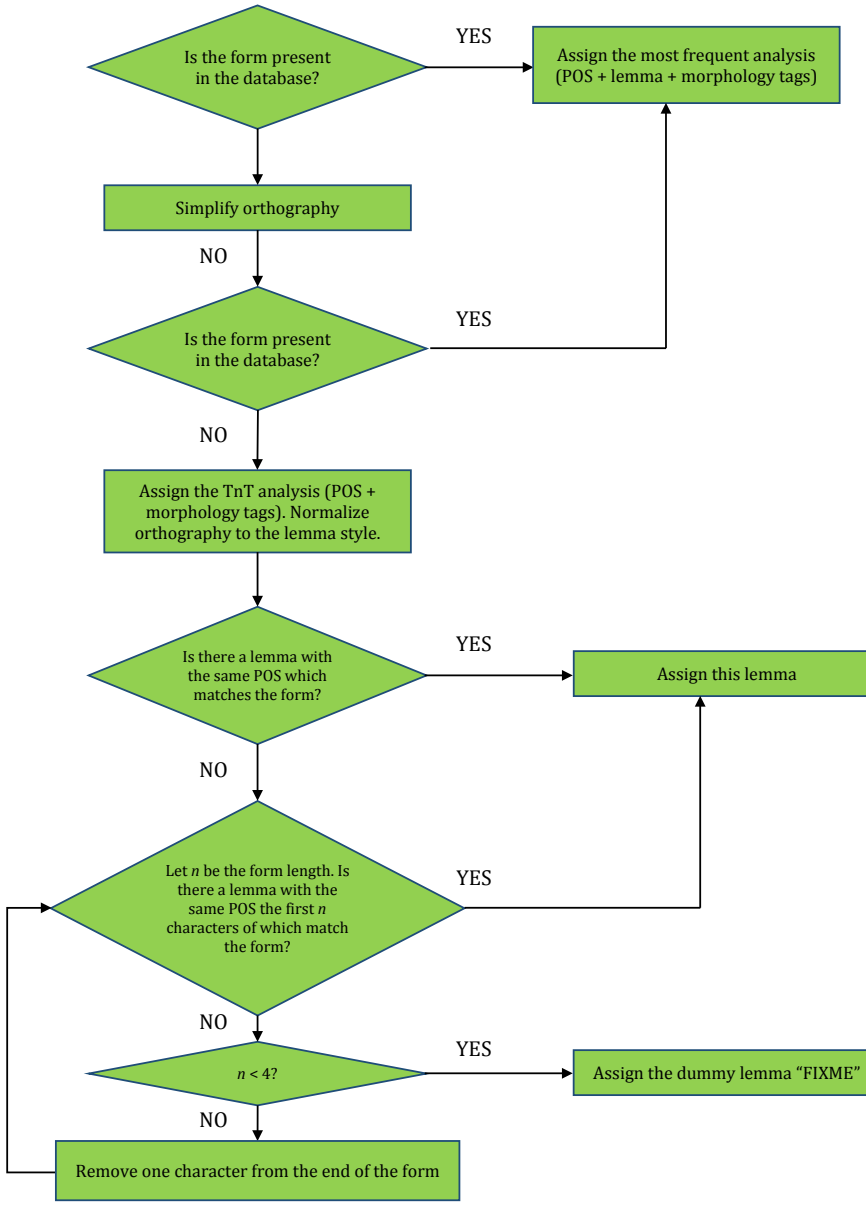

Figure 1. The TOROT automatic pre-annotation technique

3. Comparison

3.1. Test set and preprocessing

As a test set, we chose the preface to the "Life of Sergij of Radonezh" (1696 words in the unprocessed text), an early 15th century Russian Church Slavic text digitized after a late 16th century manuscript.[5] The text is late enough for the RNC analyzer (which is unlikely to perform optimally with earlier texts), but is still within the period which is of interest for TOROT. We normalized orthography (see Section 2) and ran both the RNC and TOROT analyzers on the otherwise unprocessed text.[6]

Since there were no discrete releases of the TOROT corpus at the time of the experiment (it is being expanded and corrected continuously), we preserved the training data as they were before the "Life of Sergij" was added to the corpus. That includes the whole set of Old and Middle Russian data that the TnT tagger was trained on (166,183 word tokens), and the full lemmata list that the TOROT analyzer used for lemma guessing (10,603 lemmata).

3.2. Gold standard and alignment

After preprocessing the annotation was manually corrected by a human expert, the annotation of every sentence was subsequently reviewed by at least one another expert. The resulting annotation was used as the gold standard.

The TOROT text import module assigns an id to every word in a text, and these ids are not normally changed by the annotators. These ids are used to align gold with the output of the TOROT analyzer. TOROT and RNC, in turn, are aligned using the source document: for every word in it, the corresponding TOROT guess and RNC guess (or a set of several guesses) are found. Note that TOROT always provides a guess (it may use the dummy lemma "FIXME", but the POS and morphology will still be provided), while RNC does not, which means that sometimes the TOROT guess will correspond to a blank.

Importantly, annotators can sometimes change tokenization, splitting an existing word token into two (14 cases, e.g. *неписано* > *не* 'not' and *писано* 'written'). This creates extra tokens (which are present in gold, but not in RNC or TOROT), so the total token count goes up to 1,710. Alternatively, the annotators can merge two tokens into one (5 cases, e.g. *во истинну* > *воистину* 'indeed'). This results in some tokens (*истинну*) existing only in RNC and TOROT, but not in gold. In both cases, there

---

[5] The text was digitised by Catherine Sykes and Hanne Eckhoff after the illuminated late 16th century manuscript of the Trinity Lavra of St. Sergius, available in facsimile online at http://old.stsl.ru/manuscripts/book.php?col=3&manuscript=001. The TOROT version is available at https://nestor.uit.no/sources/215.

[6] Note that the TOROT analyzer takes non-normalized text as input and uses both normalized and non-normalized tokens in the lookup process. We did an RNC analyzer test run with non-normalized input, the results were nearly the same.

always is at least one token matched against a blank, something which will be counted as an error for both analyzers.

3.3. Harmonization of the analyzers

We were interested in comparing the accuracy of POS tagging, full morphological tagging and lemmatization of the RNC and TOROT analyzers. Gold and TOROT, obviously, share the annotation format, but RNC uses a different one (different POS and morphological tags, lemmatization principles and orthography). In order to make the analyzers comparable, gold/TOROT and RNC have to be harmonized first. Some information is lost in the harmonization process, especially for the morphological tags.

3.3.1. Harmonization of the POS tags

The correspondences between RNC and TOROT POS classes are complicated. For every RNC tag, Table 1 lists all TOROT tags that can potentially correspond to it and were considered a correct match.

Table 1. POS tag correspondences

RNC POS tag	TOROT POS tag(s)
A	A-
A-PRO	A-, Pd, Pi, Pk, Pp, Pr, Ps, Pt, Px
ADV	Df
ADV-PRO	Df, Du
CONJ	C-, G-, Df
CONJ/PART	Df, G-
INTJ	I-
N	Nb, Ne
N-PRO	Pp, Pk, Pi, Px
NUM	Ma
PART	Df
PREP	R-
V	V-

If a token with a tag from the "RNC" column had one of the corresponding tags from the "TOROT" column in gold, the annotation was considered correct. TOROT tags: A- — adjective, Pd — demonstrative pronoun, Pi — interrogative pronoun, Pk — personal reflexive pronoun, Pp — personal pronoun, Pr — relative pronoun, Ps — possessive pronouns, Pt — possessive reflexive pronoun, Px — indefinite pronoun, Df — adverb, Du — interrogative adverb, C- — (coordinating) conjunction, G- — subjunction, Ma — cardinal numeral, I- — interjection, Nb — common noun, Ne — proper noun, R- — preposition, V- — verb. RNC tags: A — adjective, A-PRO — adjective pronoun, ADV — adverb, ADV-PRO — pronominal/interrogative adverb, CONJ — conjunction, CONJ/PART — a special tag for *да* 'so as / let", INTJ — interjection, N — noun, NUM — cardinal numeral, PART — particle, PREP — preposition, V — verb

3.3.2. Harmonization of the lemmatization

We consider lemmatization of a token correct if and only if both the lemma itself and the POS tag match the gold standard. There are numerous discrepancies in the spelling of the lemmata. TOROT consistently uses conservative orthography, largely following Sreznevskij (1895–1902) for the sake of better comparability of earlier and later texts. RNC focuses on the Middle Russian period and uses less archaic orthography. After manually analyzing the discrepancies, the following harmonization procedure was implemented. In gold lemmas (which are spelled according to the TOROT principles) all jers that are strong according to Havlik's law and the СъRC-rule were vocalized. Jers in the clusters *чьск* and *чьст* were vocalized, too. All remaining jers were deleted; yat was replaced by *е*; *кы/гы/хы* were changed to *ки/ги/хи*; double consonants were shortened to one. In RNC lemmas all jers were deleted; *зс* was changed to *сс*; double consonants were shortened to one; *о* was removed from *во-* and *со-* in the beginning of the word longer than four letters (this *о* is almost always a reflex of a jer in a prefix which gets missed by the vocalization rule applied to the gold lemmata); *жде* was changed to *же*. Ad hoc rules were created for three frequent lemmata: pronouns *сеи* and *тои* (changed into resp. *сии* and *тыи*) and verb *писати* (changed into *пьсати*).

After this procedure, the number of cases when a RNC lemmatization guess is unjustly labeled as wrong (some cases of "unexpected" jer vocalization; inconsistencies to the tagging of participles; some other spelling discrepancies) is reduced to 10, which we deem acceptable.

3.3.3. Harmonization of the morphology tags

The two morphological tag sets are not entirely compatible either. The RNC analyzer tags for a number of features that the TOROT analyzer ignores, namely transitivity (intr, tr), aspect (pf, ipf), reflexivity (med) and animacy (inan, anim).[7] In the comparison, these features are dropped. Both analyzers tag for long form/short form, but this is relevant for adjectives and participles only, and not adjectival pronouns. There are, however, considerable differences between the formats as to what is considered a pronoun and what an adjective. We therefore disregard this feature in the comparison. For the same reason, we ignore degree of comparison for adjectives and adverbs. Table 2 shows the harmonized tags per TOROT part of speech.

[7] We have nonetheless used the animacy tags to control for genitive-accusatives: TOROT tags these as genitives, RNC as accusatives. RNC masculine singular animate accusatives are thus considered matches of gold masculine singular genitives.

Table 2. Harmonized morphological tags used for comparison between the TOROT and RNC analyzers. The original RNC tags are in this format already, but are stripped of the features we chose to exclude (see main text). TOROT tags are converted into the simplified RNC format[8]

TOROT POS tag	Subcategory	Tagged for	Example of a possible harmonized tag
V-	l-participle	tense	perf
V-	participle	mood	participle[8]
V-	indicative	mood, tense, number, person	indic, praes, sg, 3p
V-	no mood feature	inflection	noninfl
V-	other	mood	inf
Nb, Ne	none	gender, number, case	f, sg, acc
A-, Pd, Pr, Ps, Pt	none	number, gender, case	sg, f, acc
Px, Ma, Mo	none	number, case	sg, acc
Pk, Pp, Pi	none	case	acc
Df	none	inflection	noninfl
Other	non-inflecting	inflection	noninfl

4. Results and performance boost

4.1. Results

The accuracy of lemmatization and POS tagging for TOROT and RNC are provided in resp. Tables 3 and 4. For RNC, we measure both "exact" (there is only one guess, and it is correct) and "fuzzy" (there are several guesses, and one of them correct) accuracy. Consider, for example, the form *paðu*. The RNC analyzer at its current stage will always assign three analyses to this form: the preposition *paðu* 'for the purpose of'; the verb *paðumu* 'take care' (2/3 person aorist singular); the adjective *paðъ* 'glad' (strong plural masculine nominative). Obviously, the RNC guess for *paðu* will never be an exact match. If, however, at least one of the three analyses correct, it will be considered a fuzzy match.

[8] In the vast majority of cases, the RNC analyzer is unable to provide a guess for participles, since the necessary rules have not been implemented yet. If it does hazard a guess, it is mostly erroneous. This tag is therefore simplified.

Table 3. Accuracy of the lemmatization and POS tagging by the TOROT analyzer

Metric	Lemma +POS, %	POS only, %	Number of tokens
Accuracy	69.8	89.5	1,710
Accuracy (when lemma is not "FIXME")	88.5	93.9	1,348
Accuracy (when RNC does not have a guess)	42.5	78.9	327

TOROT performs better on both accounts. Unsurprisingly, the numbers go up considerably for both analyzers if we take into account only those tokens for which they had a guess. RNC has a guess for 1383 tokens out of 1710 (81%). TOROT has a lemma guess for 1348 tokens (79%), a POS guess is always provided.

For RNC, fuzzy accuracy is much higher than exact one. When we are dealing only with tokens which have a guess, fuzzy accuracy is even higher than that of TOROT. Interestingly, if we limit ourselves to the tokens for which RNC failed to provide a guess, TOROT accuracy decreases noticeably. In other words, what is unsurmountable for RNC, is difficult for TOROT, too.

Table 4. Accuracy of the lemmatization and POS tagging by the RNC analyzer. "Exact" means that the analyzer provided a correct guess and nothing else; "fuzzy" means that there were several guesses, only one of each was correct

Metric	Lemma +POS, %	POS only, %	Number of tokens
Accuracy (exact)	47.3	54.2	1,710
Accuracy (fuzzy)	74.3	77.0	1,710
Accuracy (exact, when there is a guess)	58.5	67.0	1,383
Accuracy (fuzzy, when there is a guess)	91.9	95.2	1,383

A comparison of the morphological annotations is found in Table 5.

Table 5. Performance of the TOROT and RNC analyzers on morphological tags

	Accuracy, %	Number of tokens
TOROT	81.5	1,710
RNC (exact)	16.6	1,710
RNC (fuzzy)	70.2	1,710
RNC (exact, when there is a guess)	20.5	1,383
RNC (fuzzy, when there is a guess)	86.8	1,383

It should also be noted that a good number of the TOROT guesses are off by only one or two tags, as seen in Table 6. Since the TOROT morphological tags are 10-place positional tags, this can be measured by Hamming distances (the distance shows how many features got an incorrect tag).

The off-by-one errors are typically ambiguous forms such as *домъ* "house", which could be either nominative singular or accusative singular. It could also be a genitive plural, which might lead to a off-by-two error. Such morphological guesses are still of great practical use to the TOROT annotators, who will only have to make one or two corrections in the morphological tag, rather than providing a full new analysis.

Table 6. Hamming distances between gold tags and TOROT guess tags (10-place positional tag)

Hamming distance	count	%
0	16393	81.5
1	128	7.5
2	57	3.3
3	14	0.8
4	38	2.2
5	14	0.8
6	26	1.5
7	10	0.6
8	8	0.5
9	3	0.2
no tag	19	1.1

4.2. Boosting TOROT lemmatization accuracy

A question of practical importance is whether the analyzers are able to cooperate, helping each other out. Differences between the annotation formats, however, represent an important problem here. While we managed to harmonize the analyzers' outputs, some information got lost in the process. It does not seem realistic to do anything with morphological and POS tags, at least not without a more sophisticated harmonization. In addition, considering TOROT's better results, using it to boost RNC performance might be more complicated than simply using TOROT.

A promising avenue is to use RNC lemma guesses when TOROT fails to find one and resorts to "FIXME". We experiment with the following boosting procedure. For every token which is lemmatized as "FIXME" by TOROT and which has a RNC guess (either single or multiple), we go through all RNC lemma guesses. We harmonize the lemma and try to find a match in the (harmonized) TOROT lemma list (described in Section 3.1). If there is a match, the POS tag of the lemma guess and the potential match are compared, and if they are the same, the (non-harmonized version of the) lemma is taken as a guess,[9] otherwise the booster proceeds to the next RNC guess, if there is one. Obviously, this simple method can only work for lemmata which were

[9] Note that this can potentially result in a POS tag change due to the complex many-to-many corrrespondences used for the harmonization (see Table 1).

already in the TOROT list, but were not identified by the guesser described in Section 2. It transpires that even this can give performance a significant boost, see Table 7.

Table 7. Boosting TOROT lemmatization accuracy using RNC guesses

Metric	Lemma+ POS	POS only	Number of tokens
Success rate when fixing "FIXME"	90.3	92.7	165
Boosted TOROT accuracy	78.5	91.4	1,710

The booster attempts to provide lemma guesses for 165 tokens and gets it right in 149 cases. This increases TOROT lemmatization accuracy to 78.5% from 69.8% (see table 3). In addition, there is a slight improvement in POS tagging: 91.4% instead of 89.5%.

5. Conclusion

As was expected, the TOROT analyzer outperformed the RNC analyzer on all three accounts. There are several reasons for that.

The most prominent one is RNC's inability to disambiguate if there are several possible analyses. In addition, the selected text is non-standardized and displays considerable morphological variability and, even when consistent, idiosyncratic morphological endings, both in choice of form and orthography. The text also has numerous unresolved abbreviations. While our findings do not necessarily generalize to any historical text, these features are entirely typical of the texts of this era, and it seems reasonable to conclude that they strongly favour a statistical analyzer (trained on a large material) rather than a rule-based one. Furthermore, the RNC POS and morphological guesses are dependent on the analyzer's ability to come up with a lemma guess, whereas the TOROT analyzer guesses morphology with no reference to lemmatization. Finally, the RNC analyzer systematically misses a number of words altogether, such as all words with a titlo and most participles.

If we relax the evaluation criteria, requesting only the presence of a correct guess (not its uniqueness) and limit the evaluation set to those tokens for which RNC produces a guess, then RNC performs slightly better than TOROT. In other words, the analyzers are almost equally good at producing a guess, but differ in their ability to distinguish between several candidates. This finding shows that RNC has large potential, but one would have to develop a disambiguating technique in order to make this potential practically applicable, and this is a very time-consuming task.[10] At the current stage, the most practical thing to do if one wants to pre-annotate a Middle Russian text would be to use TOROT with the RNC lemmatization booster.

[10] Two anonymous reviewers asked how the RNC performance could be increased. Our answer is that the most important thing to do would be to implement disambiguation, but this task is far beyond the scope of this paper.

As described in 4.2, RNC can help TOROT out when it fails to provide a lemma guess. It is possible to check RNC lemma suggestions against the TOROT lemma list, and, if a match is discovered, use it as a lemma guess. This simple procedure boosts TOROT lemmatization accuracy by 8.7%, and POS tagging accuracy by 1.9%. For lemmatization, the difference is significant ($\chi^2(1) = 33.39$, $p < 0.001$), the effect size is small (Cohen's $h = 0.20$). For POS, the difference is not significant, the effect size is negligible ($\chi^2(1) = 3.46$, $p = 0.062$, $h = 0.07$). Thus, although a statistical model seems best for POS and morphological tagging, a rule-based model may considerably aid lemmatization.

Further work will no doubt result in better analyzers for Old and Middle Russian. However, the current approach is of great practical use. Especially for Middle Russian, there is a vast bulk of text available that could provide very interesting data for linguistic studies: the RNC Middle Russian subcorpus holds more than 7 million word tokens. Needless to say, the cost of manually analysis of all this text would be very high. On this background, an analyzer with around 80% success rate for both lemmatization and morphological annotation is a considerable gain, especially taking into consideration the unruly and unnormalized nature of these texts.

References

1. *Arxangelsky, T.* (2012), Principles of Morphological Parser Construction for Multi-structural Languages [Principy postroenija morfologičeskogo parsera dlja raznostrukturnyx jazykov]. PhD dissertation, Moscow State University.
2. *Arkhangelskiy T., Belyaev O., Vydrin A.* (2012), The creation of large-scaled annotated corpora of minority languages using UniParser and the EANC platform. Proceedings of COLING 2012: Posters. Mumbai, pp. 83–91.
3. *Brants, T.* (2000), TnT: a statistical part-of-speech tagger. In S. Nirenburg (ed.): Proceedings of the sixth conference on applied natural language processing 3, ANLC '00. Stroudsburg: Association for Computational Linguistics, pp. 224–231.
4. *Eckhoff, H. M., Berdičevskis, A.* (2015), Linguistics vs. digital editions: The Tromsø Old Russian and OCS Treebank. Scripta & e-Scripta Vol. 14–15.
5. *Piotrowski, M.* (2012): Natural Language Processing for Historical Texts. Morgan & Claypool Publishers.
6. *Poljakov, A.* (2014), Church Slavonic corpus: spelling and grammar problems [Корпус церковнославянских текстов: проблемы орфографии и грамматики]. Przegląd wschodnioeuropejski Vol. 5 (1): 245–254.
7. *Skjærholt, A.* (2011), More, faster: Accelerated corpus annotation with statistical taggers. Journal for Language Technology and Computational Linguistics, Vol. 26:2.
8. *Sreznevskij, I.* (1895–1902), Materials for a Dictionary of the Old Russian Language [Materialy dlja slovarja drevnerusskogo jazyka]. Tipografija Imperatorskoj Akademii Nauk, St. Petersburg.