# From Diachronic Treebank to Dictionary Resource: the Varangian Rus Project

**Hanne Martine Eckhoff, Aleksandrs Berdičevskis**

UiT The Arctic University of Norway
e-mail: hanne.m.eckhoff@uit.no, aleksandrs.berdicevskis@uit.no

## Abstract

In this paper we present the Varangian Rus' dictionary resource, which is based on the Tromsø Old Russian and OCS[1] Treebank (TOROT),[2] a diachronic treebank of Russian containing a balanced selection of 11th–17th century Old East Slavic and Middle Russian texts. The treebank is lemmatised and has rich morphological and syntactic annotation. With simple glossing of the word meanings found in the treebank, we are able to generate a dictionary resource with rich grammatical information. The dictionary work also enables us to improve the lemmatisation of the treebank considerably, as well as annotation at other levels. The dictionary resource will be published as a searchable online tool towards the end of 2016 and is intended for students and scholars alike.
**Keywords:** treebank; historical dictionary; Russian

## 1 The TOROT

The Tromsø Old Russian and OCS Treebank (TOROT, nestor.uit.no)[3] is the only existing treebank of Old East Slavic and Middle Russian texts. There are other tagged resources, such as the Old Russian subcorpus of the Russian National Corpus[4] and the Manuscript corpus,[5] but none of them, to our knowledge, currently provides syntactic annotation. The TOROT presently contains approximately 89,000 word tokens of fully lemmatised and morphosyntactically annotated Kiev-era Old East Slavic (11th–14th centuries), and 87,000 word tokens of fully lemmatised and morphosyntactically annotated 15th–17th-century Middle Russian. The TOROT is a part of a larger family of treebanks of ancient languages, originating in the PROIEL project.[6] The PROIEL project developed an enriched dependency grammar scheme, a scheme for detailed morphological tagging and various other annotation schemes, such as ones for information structure and semantics. Accompanying web annotation, browsing and query tools were also created. These schemes and tools are tailored especially for the structures typically found in early attestations of Indo-European languages (rich case and verbal inflection systems, word order driven by information structure). They are therefore obviously useful beyond the original PROIEL languages (Ancient Greek, Latin,

[1] Old Church Slavonic is abbreviated OCS throughout.
[2] The TOROT annotation and browsing interface is found at https://nestor.uit.no. Versioned releases of annotated texts in XML format can be accessed via http://torottreebank.github.io/
[3] The TOROT is being developed as part of the project "Birds and Beasts: Shaping Events in Old Russian" at UiT The Arctic University of Norway. The dictionary resource is a part of its companion project "The Varangian Rus' Digital Environment", a cooperation between UiT and two Russian partners: The Higher School of Economics and Moscow State University. The project also aims at creating other electronic resources for learners of Old East Slavic and Middle Russian, such as online exercises and reading tools with text commentary.
[4] http://www.ruscorpora.ru/help-old_rus.html, the subcorpus is the work of the Russian Language Institute, Russian Academy of sciences, see also http://www.lrc-lib.ru/index.php?id=5
[5] http://mns.udsu.ru/mns/portal.main?p1=1&p_lid=2&p_sid=1
[6] "Pragmatic Resources in Old Indo-European Languages", University of Oslo 2008–2012, principal investigator: Dag Haug. For further details on the PROIEL treebank, see Haug et al. 2009. The Old Church Slavonic (OCS) part of the TOROT is an expansion of the OCS part of the PROIEL parallel corpus.

Classical Armenian, Gothic and OCS), and have been taken in use by a number of other projects, including the TOROT.

The PROIEL tools and schemes are developed for and by linguists, and yield data with rich morphological and syntactic information, as well as lemmatisation. We currently have a lemma base of over 11,000 entries. The TOROT thus contains a lot of lexicographically interesting information, which can easily be developed into a learner's dictionary resource.[7] In this article we discuss the lemmatisation work, glossing and types of information that can be generated for the dictionary resource entries.

## 2 The Corpus Builder as Lexicographer

The TOROT's lemmatisation scheme requires sophisticated decisions on the part of the corpus annotator. However, these decisions are often difficult to make, due to the lack of complete dictionaries for Old East Slavic and Middle Russian.

In the corpus annotation process, every word token is assigned to a lemma – if the lemma does not already exist, the annotator automatically creates it in the annotation process. Each lemma is tagged with ISO code and a part-of-speech tag. Identical-looking lemmata with different part-of-speech tags[8] are stored as different lemmata. We also have the option to assign variant numbers to lemmata with identical form and part of speech, which we use to distinguish homographic lemmata: lemmata deemed to be semantically different (not just polysemous) and lemmata which do not belong to the same morphological paradigm. For the lemma pair *sъpasti*#1 'save' and *sъpasti*#2 'fall down', both of these considerations are relevant.

Even when there is no doubt about the part of speech or semantics, lemmatisation is often not at all straightforward. The lemmatisation process can be represented as consisting of two tasks: grouping together word forms that belong to the same lexeme and choosing a label (headword) for this lexeme. Ideally, this is not a task that should be done by corpus linguists, but by lexicographers: the corpus linguist should ideally rely on already available dictionary resources. However, in the case of Old East Slavic and Middle Russian, the existing dictionary resources are unsatisfactory. The only dictionary that covers the whole alphabet, Sreznevskij 1893–1912, is based on a relatively small selection of texts, and also leaves out large classes of words, such as ethnonyms. It is also a relatively unsystematic piece of work both when it comes to choice and form of headword, and organisation and form of definitions. The more comprehensive and structured dictionaries, SRJa XI–XVII and SDJa XI–XIV, are still incomplete. Also, SRJa XI–XVII has chosen a relatively modern orthography for their headwords, to fit their newest source texts. This is not an option for us: since we want our OCS and East Slavic lemma inventories to be maximally comparable, the natural choice is to give headwords in archaic form.

Our choice is therefore to follow Sreznevskij 1893–1912 as much as possible when choosing labels. When this is impossible (when Sreznevskij does not list a word; or gives several variants of equal status; or his solution is unacceptable for some reason), the annotators should try to follow the so-called "etymological" principle, i.e. choose a maximally conservative label. When multiple forms are found in the sources, Sreznevskij often provides double labels, such as *zvati=zъvati* 'call, shout'. In this case, the annotator should prefer the second option as it preserves more etymological information.

Grouping the word forms is a more crucial task which presents more difficulties. Some of these problems are systematic and require a certain general policy. These include South Slavic–East Slavic variation and orthographic-or-nearly-orthographic variation. Old Russian texts abound with variation such as *volodějutь* 'rule' (East) vs. *vladějutь* 'rule' (South), or *reči* 'say' (East) vs. *rešti*

[7] For other lexicographical projects exploiting treebank data, see e.g. the Perseus Dynamic Lexicon project (http://nlp.perseus.tufts.edu/lexicon/) and the Hinoki Treebank and Sensebank (Bond et al. 2008).

[8] For a full part-of-speech tag inventory and a detailed description of the TOROT, see Eckhoff and Berdičevskis (2015).

'say' (South) etc. In addition, there are also cases of variation with do not seem to be related to the South–East differences: *pomagai* vs. *pomogai* 'help!'; *xrstilъ* vs. *krstilъ* 'baptised' etc. Sreznevskij 1893–1912 usually lumps together variants like *reči* and *rešti* 'say', but keeps separate lemmata for cases like *pomogati* and *pomagati* 'help'. In the case of both types of variation, we assume that in many cases it does not really affect the lexical level. We therefore try to ascribe the variants to the same lemma wherever possible. The "etymological" variant is chosen as a label for the quasiorthographic cases, the South Slavic variant for the other ones (to ensure maximal comparability with the OCS texts). In both cases, forms that are different from the label get a special additional tag in order to preserve information about variation.

## 3 Glossing the TOROT Lemma Base

The TOROT dictionary resource is not a full-fledged dictionary, and provides simple glosses rather than structured and ranked definitions. Glosses are given in English and Russian, and aim at covering all and only the meanings attested in the TOROT treebank. Below we discuss some of our glossing principles.

Glosses should be as simple as possible. If one word is deemed sufficient, no more glosses are added. When several glosses are needed, they are separated by commas. Clarifications are given in parentheses. In some cases it is not obvious whether to include or exclude a gloss. Should, for instance, *velikyj* 'great, big' have the gloss 'loud' since it can occur in the expression *velikъ golosъ* 'loud voice'? The glossers are asked to consult dictionaries in such cases, and also to apply a transparency principle: If the meaning can easily and unambiguously be deduced from the immediate context ('big shoulders' = 'broad shoulders'), glossers are told not to posit an additional meaning. If the meaning cannot easily be deduced (does *big voice* mean 'magnificent voice' or just 'loud voice'), they are told to posit an additional meaning.

Since we gloss the lemmata in both English and Russian, we aim for the two sets of glosses to be as similar as possible. One language should only have more glosses than the other if it is semantically justified. For instance, *gora* is glossed with 'gora' in Russian, but with 'hill, mountain' in English, since the attestations in the TOROT include examples also from an English perspective.

In order to make sure that only actually occurring meanings are included in the gloss, the glossers are asked to go through all occurrences of every lemma. This policy sometimes make us exclude "obvious" meanings, such as 'main' for the adjective *glavьnyi*, since that adjective has just a single attestation in TOROT. The single attestation is in the collocation *vlasi glavnii* 'hair of the head'. The dictionary lookup should thus only hold the meaning 'of the head'.

Reflexive verb forms are a special challenge. In the treebank, the reflexive clitic is always analysed as a separate syntactic word, and accordingly, headwords are always listed as non-reflexive, even in the case of reflexiva tantum. If a (sub)meaning applies only to the reflexive version of a verb, that meaning is tagged refl. We do not posit separate meanings for true reflexives (*myti sja* 'wash oneself') and passives (*sъzьdati sja* 'be created').

## 4 Generating Dictionary Entries

The lemmatisation, naturally, makes it possible to provide a listing of all sentences containing a given lemma. In addition, the TOROT has detailed morphological and syntactic annotation. The morphological annotation means that the dictionary entries can also include a full listing of all attested forms of each lemma, sorted by grammatical features, i.e. we can provide paradigms insofar as they are attested in the TOROT. As far as possible, we are using digital texts that are maximally close to a single manuscript. This means that we are in position to give a rich survey of the occurring forms in each paradigm. The 47 attestations of *varjagъ* 'Varangian' in the TOROT, for instance, have the distribution of unique forms seen in Table 1.

| | singular | dual | plural |
|---|---|---|---|
| nominative | варѧгъ | варѧга | варѧзи |
| accusative | | варѧга, варѭга | варѧги, вариги, варѧгы, вѧрѧгы |
| genitive | | | варѧгъ |
| dative | | | варѧгомъ |
| instrumental | | | варѧги, варѧгы |

Table 1: Attested paradigm in the TOROT for varjagъ - 'Varangian'.

The syntactic annotation can also be mined for lexicographically interesting information. In particular, we are able to provide rich valency information for prepositions and verbs. For prepositions, we can give a full listing (including frequencies) of all cases that occur with the preposition in the TOROT.

| | |
|---|---|
| dative | 841 |
| locative | 344 |
| accusative | 163 |
| genitive or genitive-accusative | 22 |

Table 2: Case of complements of the preposition *po* 'along, after' in the TOROT

For verbs, we can list and give frequencies of all attested argument structure frames, exemplified in Table 3 with the argument frames of *zъvati* 'call' (49 occurrences), ranked by frequency (subjects not included).

| | |
|---|---|
| direct object | 10 |
| instrumental predicative complement | 8 |
| reflexive/passive, nominative predicative complement | 7 |
| reflexive/passive, no arguments | 5 |
| active, no arguments | 3 |
| direct object + oblique argument headed by *kъ* 'to' | 2 |
| direct object + oblique argument headed by *na* 'on' | 2 |
| direct object + instrumental predicative complement | 2 |
| oblique argument headed by *vъ* 'into' | 2 |
| oblique argument headed by *kъ* 'to' | 2 |
| direct object + nominative predicative complement | 1 |
| direct object + other predicative complement | 1 |
| passive, instrumental passive agent | 1 |
| passive, oblique argument headed by *na* 'on' + instrumental passive agent | 1 |
| reflexive/passive, instrumental predicative complement | 1 |
| passive, nominative predicative complement + PP passive agent | 1 |

Table 3: Attested argument frames with *zъvati* 'call, shout' in the TOROT.

The dictionary resource is diachronic in nature, which makes it necessary to give indications of diachronic variation in the use of words. Since the TOROT also contains metadata about the text sources, we can give precise indications of a lemma's distribution across sources and time periods. We are able to do the same with our generated paradigms and valency frames, thus tracking morphological and syntactic change in the dictionary entries.

The generated dictionary entries also serve as a useful error detection tool for the treebank, since the systematic representation easily reveals many types of annotation errors, such as morphological

misclassification or faulty lemmatisation. The work on the dictionary resource is therefore directly beneficial to the treebank.

## 5 Conclusion

This paper describes the development of a diachronic Old East Slavic and Middle Russian dictionary resource on the basis of the TOROT treebank. By adding simple glossing to the treebank's lemma inventory, we are able to generate a rich dictionary resource that fills a gap in the available lexicographical tools for historical Russian. We are not only able to provide every example of a given lexeme, the dictionary can also provide paradigms of attested forms, frequency information and valency frames. The same information can also be sorted by their appearance in particular texts, and therefore be given a diachronic dimension. The dictionary resource will be published as a searchable web tool towards the end of 2016.

## 6 References

Bond, F., Fujita, S. & Tanaka, T. (2008). The Hinoki syntactic and semantic treebank of Japanese. In *Language Resources and Evaluation* 42(2), pp. 243–251.

Eckhoff, H. & Berdičevskis A. (2015). Linguistics vs. digital editions: The Tromsø Old Russian and OCS Treebank. In *Scripta & e-Scripta* 14–15.

Haug, D.T.T., Jøhndal, M, Eckhoff, H.M., Welo, E., Hertzenberg, M.J.B. & Müth, A. (2009). Computational and Linguistic Issues in Designing a Syntactically Annotated Parallel Corpus of Indo-European Languages. In *Traitement Automatique des Langues* 50.

Sreznevskij, I.I. (1893–1912). *Materialy dlja slovarja drevnerusskogo jazyka.* I–III. Saint Petersburg.

SRJa XI–XVII = Slovar' russkogo jazyka XI–XVII vv. (1995–). Moscow: Nauka.

SDJa XI–XIV = Slovar' drevnerusskogo jazyka XI–XIV vv. (1988–). Moscow: Institut russkogo jazyka AN.