SCRIPTA
& e-SCRIPTA

14–15 / 2015

e-SCRIPTA

Linguistics vs. digital editions: The Tromsø Old Russian and OCS Treebank

Hanne Eckhoff, Aleksandrs Berdičevskis

1. Introduction

The Tromsø Old Russian and OCS Treebank (TOROT, nestor.uit.no)[1] is, along with its parent treebank, the PROIEL corpus (foni.uio.no), the only existing treebank of Old Church Slavonic (OCS), Old East Slavic and Middle Russian texts. There are other tagged resources, such as the Old Russian subcorpus of the Russian National Corpus[2] and the Manuskript corpus,[3] but none of them, to our knowledge, currently provide syntactic annotation.

The TOROT presently contains approximately 160,000 word tokens of fully annotated OCS (Codex Marianus[4] and Codex Suprasliensis), 85,000 word tokens of fully annotated Kiev-era Old East Slavic, and 60,000 word tokens of fully annotated 15th–17th-century Middle Russian. In addition, it contains the Codex Zographensis with automatic and partially hand-corrected morphological annotation and lemmatisation (sections of the Gospels missing in the Codex Marianus also have full syntactic annotation), and the PROIEL version of the Greek Gospels, with which the Codex Marianus and the Codex Zographensis are both aligned at token level (automatically, then hand-corrected).

The TOROT is a part of a larger family of treebanks of ancient languages, originating in the PROIEL project.[5] A central aim of the PROIEL project was building a parallel treebank of old Indo-European languages:

[1] 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.

[2] 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

[3] http://mns.udsu.ru/mns/portal.main?p1=1&p_lid=2&p_sid=1

[4] The Codex Marianus was annotated as a part of the PROIEL project, and is found both at the PROIEL corpus site and the TOROT corpus site.

[5] *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 original Greek New Testament and its translations into Latin (Vulgate), Gothic (Wulfila), Classical Armenian and Old Church Slavonic (Codex Marianus).[6] An open-source customised web-based annotation tool was developed for the purposes of the project by Marius L. Jøhndal,[7] written in Ruby on Rails with a relational database backend (MySQL, PostgreSQL). The application serves as a tool to apply PROIEL's enriched dependency grammar annotation scheme, described in detail in the PROIEL guidelines for syntactic annotation.[8] The project also developed an annotation scheme and interface for annotating information status and anaphoric relationships, see Haug et al. 2014.

The PROIEL annotation schemes and web tools are thus tailored for the structures typically found in old 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, and have been taken in use by a number of other projects. The PROIEL treebank itself has been expanded with more Classical Greek and Latin (Herodotus; Caesar, Cicero, Plautus, Terence) as well as more recent Greek and Latin (Byzantine chronicles; Peregrinatio Aetheriae). It also hosts the compatible treebanks of three other projects: the Old English, Old French, Old Spanish and Old Portuguese treebanks of the ISWOC project (Information Structure and Word Order Change in Germanic and Romance Languages, University of Oslo), the Old Norwegian treebank of the Menotec project and the Old Icelandic treebank (Eddic poems) of the Greinir Skáldskapar project.[9] The TOROT has its own treebank web site.

The treebanks have all been developed in close collaboration, which means that guidelines for individual languages are based on the PROIEL guidelines, with adjustments arrived at through communal discussion. The analyses are therefore immediately compatible and comparable. The projects also have in common that they are developed for and by linguists. The data produced are specifically tailored for a linguist's needs, not those of a traditional philologist or textologist. However, several of the projects include cooperations with scholars working on electronic text editions, and we believe this to be the ideal situation: edition philologists should produce text with all the necessary care and detail, linguists should produce maximally refined, many-layered linguistic annotation, and the results of both groups should be combined in interactive digital editions.

In this article we discuss principles and selected problems at several levels of analysis in the TOROT, and then return to a brief discussion of the division of labour between linguists and edition philologists.

[6] foni.uio.no:3000

[7] https://github.com/mlj/proiel-webapp

[8] http://folk.uio.no/daghaug/syntactic_guidelines.pdf

[9] http://bragi.info/greinir/

2. Text selection and preprocessing

As far as possible, the TOROT aims to use already existing high-quality text digitisations. For our annotation of the Codex Suprasliensis we are in an ideal position: We cooperate with the Suprasliensis project at the Bulgarian Academy of Sciences (Anisava Miltenova, David Birnbaum) and have the permission to use and publish the manuscript transcriptions used in their digital edition[10] with annotations. Work is under way to enrich the edition with our annotations, which can thus serve both as high-quality linguistic data and a practical reading aid, and reach a much wider audience than the treebank can alone. Our annotation of the Primary Chronicle (Codex Laurentianus) uses the text of the e-PVL,[11] and a similar collaboration with David Birnbaum to enrich the electronic edition with linguistic annotation is planned.

We also have text collaborations with the Regensburg Russian Diachronic Corpus and the Russian Language Institute at the Russian Academy of Sciences. In all our text collaborations, we offer to coindex our annotated text with the text of our collaboration partner, for maximum ease of annotation transfer.

To some extent it has been necessary for the TOROT team to do text digitisations, either because no digital text was available to us, or because the available text had been normalised to such an extent that it could affect linguistic analysis. For example, the Life of Avvakum and our texts from the Uspenskij sbornik were digitised by our team members. In these cases we have made digitisations of a single high-quality manuscript, sticking closely to the text and ignoring editorial corrections and insertions in available editions. We have reproduced the orthography as far as Unicode allows, but have largely ignored diacritics that are not abbreviation marks. We have taken down all supralinear letters, but never expand abbreviations. As far as possible, we have relied on texts with available manuscript facsimiles. Our tokenisation has been guided by already available text editions, but the tokenisation in the editions is sometimes overridden due to general TOROT principles. For instance, we always treat the reflexive marker сѩ as a separate token, and the relative pronoun иже as a single token. All our text downloads contain a metadata header that describes the editorial work. We release these texts under a Creative Commons Attribution licence (CC BY 4.0) for the use of other scholars.

In order to import the texts into the web application, they must be converted into the PROIEL xml format,[12] which is also used for exports. The texts are divided into chapter divisions if appropriate (for instance, each year entry in the Primary Chronicle is a separate chapter division), and the word tokens are preliminarily organised into sentences on the basis of punctuation. Since early Slavic punctuation virtual-

[10] http://suprasliensis.obdurodon.org/

[11] http://pvl.obdurodon.org/

[12] https://nestor.uit.no/exports/proiel.xsd

ly always indicates smaller units than sentences, sentence boundaries are manually adjusted by annotators in the web interface as part of the annotation work flow.

One of TOROT's major assets is its large database of form, lemma and tag correspondences. We are able to use this database for linguistic preprocessing of texts, which increases precision and speeds up the annotation process considerably.

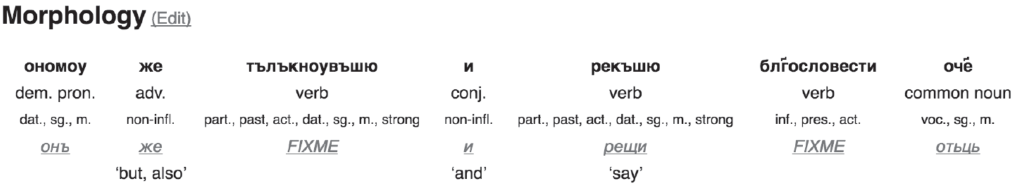
With approximately 160,000 tokens of annotated OCS, and approximately 145,000 tokens of Old East Slavic and Middle Russian, we are able to train very successful statistical morphological taggers for these language stages.[13] For this purpose, we use the TnT tagger (Trigrams 'n Tags, as described in Brants 2000), a statistical morphological tagger that looks at trigrams and word-final letter sequences (for the motivation behind this choice, see Skjærholt 2011).

To improve the performance of the tagger, we normalise both the training data and the new text to be tagged in the process. The normalisation consists in considerable orthographical simplification. For Old East Slavic, 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 ï variants are reduced to и, for instance). The juses are simplified to я and у (ю), and the jat to е. Note that this normalisation takes place behind the scenes, as it were, no text in the treebank is normalised in this way, only transformed text files used in the tagging process.

When preprocessing a text, we use the tagger output in combination with direct lookups in the database. For each word token in the text, we check whether we have that form in the database already, first as it is, then again with different kinds of orthographic simplifications. If the form is not found in the base, we assign the TnT morphtag and try to find a suitable lemma in the database. If the word form (normalised to our lemma orthography style) matches a lemma with the part-of-speech tag the TnT tagger assigned, then we assign that lemma. If not, we drop letters from the end of the word form one by one and check again against the opening strings of lemmata of the correct part of speech. If we get no matches, we assign a dummy lemma ("FIXME"), and the annotators will have to assign a lemma manually. The method is crude, but quite successful – with the current size of the database, we get 70–90% of the lemmata right, depending on the subject matter.

Figure 1. Example of automatic lemmatisation and morphological tagging from the Life of Feodosij Pečerskij (Uspenskij sbornik). Tagger trained on Old East Slavic data only

[13] In the treebank, both Old East Slavic and Middle Russian are organised under a single ISO code, orv (Old Russian). Experiments show that we do best with a tagger trained on Old East Slavic data alone for the Kiev-era texts, but that later texts are better tagged with a tagger trained on data from both periods.



Morphology (Edit)

ономоу	**же**	**тълъкноувъшю**	**и**	**рекъшю**	**блг҇ословести**	**оч҇е**
dem. pron.	adv.	verb	conj.	verb	verb	common noun
dat., sg., m.	non-infl.	part., past, act., dat., sg., m., strong	non-infl.	part., past, act., dat., sg., m., strong	inf., pres., act.	voc., sg., m.
онъ	*же*	*FIXME*	*и*	*рещи*	*FIXME*	*отьць*
	'but, also'		'and'	'say'		

The pre-tagging is not good enough to serve directly as linguistic data, but gives excellent annotation support and increases precision and annotation speed. However, in the case of very close textual variants, we can do very successful automatic lemmatisation and morphological analysis. For the purposes of Eckhoff and Haug 2015, we did automatic tagging of the Codex Zographensis on the basis of the Codex Marianus analysis. We were then able to get viable verb data from the Zographensis with only a brief round of corrections: assistants completed the lemmatisations and checked the morphological analyses of all verbs. We were then also able to automatically align the Codex Zographensis with the Greek Gospels, and then hand-corrected the alignment. Since the Marianus is already aligned with the Greek, this also linked the two OCS gospel variants to each other on token level. Similar automatic analysis could very profitably be applied e.g. to the manuscript variants of the Primary Chronicle, since TOROT already has a full analysis of the Laurentian manuscript.

3. Lemmatisation

Lemmata are stored in a separate table in the database backend of the annotation web application. Each lemma has a language tag and a part-of-speech tag: identical-looking lemmata with different language tags are stored as different lemmata. Likewise, identical-looking lemmata with different part-of-speech tags (table 1) are also 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 paradigm. For the lemma pair *сънасти*#1 'save' and *сънасти*#2 'fall down', both of these considerations are relevant. As a consequence, some multifunctional items are assigned to multiple lemmata. For example, OCS has four lemmata on the form ꙗко: a subjunction, a relative adverb and two variant regular adverbs ("as, approximately" vs. introductory "for"). For an example of how the choice of part-of-speech tags interact with the choice of syntactic analysis, see section 5.

A-	adjective	*Mo*	ordinal numeral
Df	adverb	*Pp*	personal pronoun
S-	article	*Pk*	personal reflexive pronoun
Ma	cardinal numeral	*Ps*	possessive pronoun
Nb	common noun	*Pt*	possessive reflexive pronoun
C-	conjunction	*R-*	adposition
Pd	demonstrative pronoun	*Ne*	proper noun
F-	foreign word	*Py*	quantifier
Px	indefinite pronoun	*Pc*	reciprocal pronoun
N-	infinitive marker	*Dq*	relative adverb
I-	interjection	*Pr*	relative pronoun
Du	interrogative adverb	*G-*	subjunction
Pi	interrogative pronoun	*V-*	verb

Table 1. Part-of-speech tag inventory.

Even when there is no doubt about the part of speech or semantics, lemmatisation is often not at all straightforward. The main principles that we follow are the same as elsewhere in TOROT: first, we want to preserve useful linguistic information; second, we want to make it easily retrievable; third, we want to make different language stages in the corpus maximally comparable (which usually means that we want to be conservative). Interestingly, these principles do not always point into the same direction.

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. While a more correct usage is to reserve the word *lemma* only for the label, it is a widespread practice in corpus linguistics to use it both for the label and the lexeme, i.e. the whole set of words (Knowles and Mohd Don 2004), and for simplicity's sake we will stick to this tradition.

Obviously enough, choosing a label is a less challenging task, although it is still important for retrievability and users' convenience. For Old Russian (Old East Slavic + Middle Russian) we try to follow Sreznevskij's *Materialy dlja slovarja drevnerusskogo jazyka* as much as possible when choosing labels.[14] When this is impossible

[14] We choose Sreznevskij for two main reasons: It is currently the only complete dictionary of Old Russian in the sense that it covers the whole alphabet, and it generally strives to give etymologically correct spellings of the lemmata.

(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 most conservative (within reason) label. Sreznevskij's entry for the verb *clamare*, for instance, looks like звати=зъвати. In this case, the annotator should prefer the second option as preserving more etymological information.

Grouping the word forms is a more crucial task which presents more difficulties. We will illustrate some of them on the example of Old Russian verb lemmata. One of the most prominent issues is to decide to which lemma a token should belong in some notorious verb families. This concerns, for instance, families like *имати*; *имѣти*; *емати*; *яти* and their numerous derivatives, or *вѣдати*; *вѣдѣти*/*вѣсти* (though we do not attempt to separate the latter two). Another (somewhat opposite) case includes diverging (or emerging) lemmata, for instance, the pair *въпити* and *въпияти*. While it seems safe to assume that all forms in the older texts (*воньеть*, *впити*, *вопьюще* etc.[15]) belong to one lemma (with the former label), in Middle Russian texts both forms like *вопиют* (Domostroj) and *вопят* (The Life of Avvakum) are attested. On the one hand, assigning these forms to the same paradigm contradicts linguistic intuition. On the other hand, it is not quite clear whether the existence of the lemma *въпияти* should already be assumed for Domostroj. A similar case is represented by the verb *болѣти* (diverges into *болѣти-болѣю* and *болѣти-болю*).

The solutions for these problems lemma convergence and divergence are mostly based on thorough analyses of individual cases, but there are also systematic problems that require a certain general policy. These include South Slavic–East Slavic variation and orthographic-or-nearly-orthographic variation.

As regards the former, Old Russian texts, obviously, abound with variation like *володѣють* vs. *владѣють*, or *речи* vs. *рещи*, or *могуче* vs. *могуще*. (present active participle) etc. In addition, there are also cases of variation with do not seem to be related to the South–East differences: *помагаи* vs. *помогаи*; *хрстилъ* vs. *крстилъ*; *ядѧт* vs. *едят* etc. Sreznevskij usually lumps together variants like *речи* and *рещи*, but keeps separate lemmata for cases like помогати and помагати. The latter type of variation is difficult to classify. It is not purely orthographic variation (the differences are not merely orthographic, cf. *хрстилъ* and *крстилъ*). On the other hand, it is not random, of course, that it is written counterparts of similar (and not just any) sounds that occur in the same place, thus creating variation. "Quasiorthographic" might capture the nature of the phenomenon. In any case, a further nice illustration can be offered by the following set of tokens found in TOROT: ıсповѣдыват, ıсповѣдующе, ıсповѣдуем.[16] Sreznevskij

[15] The example tokens are given exactly as they are represented in TOROT.

[16] *-ся* is never part of a verb lemma, since it is always split off, i.e. we do not distinguish between reflexive and non-reflexive verbs at this level.

has the lemma ИСПОВЕДОВАТИ, but should ıсповѣдыват be part of it (*Slovar' russkogo jazyka XI–XVII vv.* lists both *исповѣдыватися and исповѣдоватися*)? And is it another case of quasiorthographic variation, or do we witness a suffix *-ыва-* here? In the latter case, the form should be lemmatised separately from the other two, in the former case the solution is less clear.

The currently accepted policies (at the moment of submission being refined and implemented) are similar for both types of variation. Assuming that in many cases the variation, whatever its nature is, does not really affect the lexical level, i.e. does not create different words, we try to ascribe the variants to the same lemma, wherever possible. From the examples above, *речи* and *рещи*; *могуче* and *могуще*; *помогати* and *помагати* will get lumped together (o/a being a widespread variation, clearly caused by the similarity of the two vowels), the others will not. The "etymological" variant will be 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 will get a special additional tag in order to preserve information about variation.

4. Morphology

The TOROT offers detailed morphological analysis. The morphology is stored in the database in a ten-place positional tag with the following features:[17]

1. Person	1, 2, 3, x (uncertain)
2. Number	s (singular), d (dual), p (plural), x (uncertain number)
3. Tense	p (present), i (imperfect), r (perfect), s (resultative, i.e. l-form), a(aorist), u (past), l (pluperfect), f (future), t (future perfect), x (uncertain)
4. Mood (combined moo and finiteness)	i (indicative), s (subjunctive), m (imperative), o (optative), n (infinitive), p (participle), d (gerund), g (gerundive), u (supine), x (uncertain mood)
5. Voice	a (active), m (middle), p (passive), e (middle or passive)
6. Gender	m (masculine), f (feminine), n (neuter), p (masculine or feminine), o (masculine or neuter), r (feminine or neuter), q (masculine, feminine or neuter), x (uncertain gender)
7. Case	n (nominative), a (accusative), o (oblique), g (genitive), c (genitive or dative), d (dative), b (ablative), i (instrumental), l (locative), v (vocative), x (uncertain case)

[17] Not all these features are in use in the analysis of Slavic.

8. Degree	p (positive), c (comparative), s (superlative), x (uncertain degree)
9. Strength (long form / short form)	w (weak, i.e. long form), s (strong, i.e. short form), t (weak or strong)
10. Inflection	n (non-inflecting), i (inflecting)

Table 2. Morphological tags

Note especially that there is no separate case tag for the genitive-accusative, or one for animacy. In OCS and Old East Slavic the variability between genitive-accusative and nominative-accusative is still so great that we cannot tell whether, for example, a genitive human object dependent on a negated verb is a genitive-accusative or a genitive proper. We therefore rely on the interaction between the syntactic analysis (direct objects get the tag "OBJ" regardless of their case) and the morphology (genitive-like forms get the case value tag "g" regardless of their syntactic role) to identify the "real" genitive-accusatives for us (see Eckhoff 2015 for a study of OCS genitive-accusatives using such data).

An important issue in a diachronic corpus is how to deal with morphological change. A case in point is the rise of the gerunds in the history of Russian, i.e. non-inflecting varieties of the present active and past active participles in the adverbial usages that regular inflecting participles had in the first place ("conjunct participles", "converbs"). Even in the earliest Old East Slavic (and OCS) texts, we see signs of agreement failures between converb usages of participles and their agreement controller (i.e. their external subject). In example (1) we see a singular noun controlling an apparently plural conjunct participle.

(1) **оувѣдѣвше же се ѡканьныи с҃тополкъ яко еще дышеть. посла два варѧга прикончатъ ѥго.**

"but having realised that he was still breathing, the cursed Svjatopolk sent two Varyags to finish him off" (PVL 134.16–18)

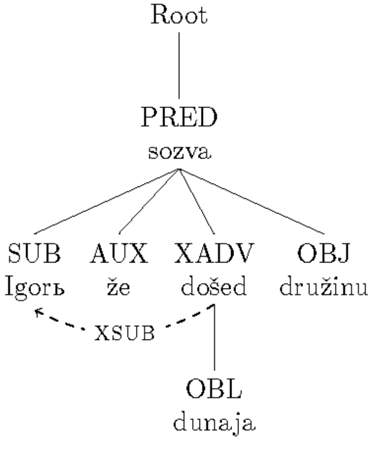
However, the mismatching form of the participle, naturally, always matches some case/number/gender form from its paradigm, even in cases where it is not the expected one. We also see that the mismatching form is not always the same, we find examples of apparent masculine nominative singulars, feminine nominative singulars and masculine nominative plurals (as in example 1) in mismatching positions. We therefore have essentially three choices: 1) the conservative solution: we analyse the participle as a participle with the face-value morphology; 2) the moderate solution: we analyse the form as a non-inflecting participle; 3) the radical solution: we analyse the form as a non-inflecting gerund.

We have opted for the conservative solution. Since the syntactic annotation scheme encodes the relationship between a converb and its external subject (see section 5), we are able to retrieve and judge the agreement match, thus providing sophisticated data for diachronic studies of the development. The solution is, naturally, a better fit for OCS and Old East Slavic than for Middle Russian, but there is still variation in our Middle Russian data.

5. Syntax

TOROT uses the PROIEL dependency grammar annotation scheme,[18] an enriched dependency grammar inspired by and convertible to Lexical-Functional Grammar's F-structure (the component of LFG that handles grammatical functions). It differs from classical dependency grammar in several respects. Empty verb and conjunction nodes are systematically employed to model ellipsis, null copulae, gapping and asyndetic coordination. This makes the annotated data less useful as training data for a syntactic parser, but on the other hand, it preserves structural information that would have been lost in a model without empty nodes. It also employs secondary dependencies to indicate external dependencies, for instance the external subjects of conjunct participles (cf. section 3).

Figure 2: Syntactic analysis of example (2).



(2) игорь же дошед дуная. созва дружину
"having reached the Danube, Igor' summoned his retinue" (PVL 45.29)

The scheme also has a richer set of syntactic relation labels than is found in e.g. the Prague Dependency Treebank (table 3).

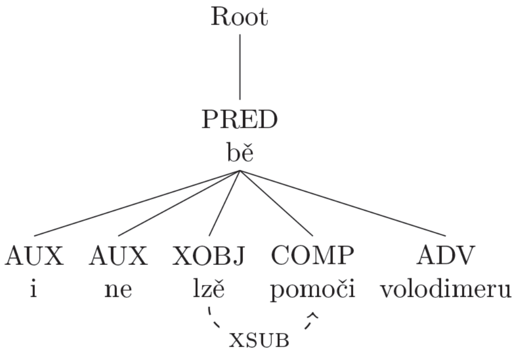
[18] For an exhaustive description of how the scheme is applied in Slavic, see the TOROT guidelines for syntactic annotation, http://folk.uio.no/hanneme/torot.pdf

adnom	adnominal (supertag*)	*obl*	oblique argument
adv	adverbial	*parpred*	parenthetical predication
ag	passive agent	*part*	partitive
apos	apposition	*per*	peripheral (supertag)
arg	argument (supertag)	*pred*	predicate
atr	attribute	*rel*	relative clause (supertag)
aux	auxiliary	*sub*	subject
comp	complement	*voc*	vocative
expl	expletive	*xadv*	adverbial with external subject
narg	adnominal argument	*xobj*	argument with external subject
nonsub	non-subject (supertag)	***pid***	predicate identity (secondary)
obj	direct object	***xsub***	external subject (secondary)

* Supertags are tags to be used by annotators in cases of doubt: if it is not clear whether an adnominal dependent is an atr, apos, narg or part, adnom can be used.

Table 3. Syntactic relation label inventory.

Figure 3. Syntactic analysis of example (3)

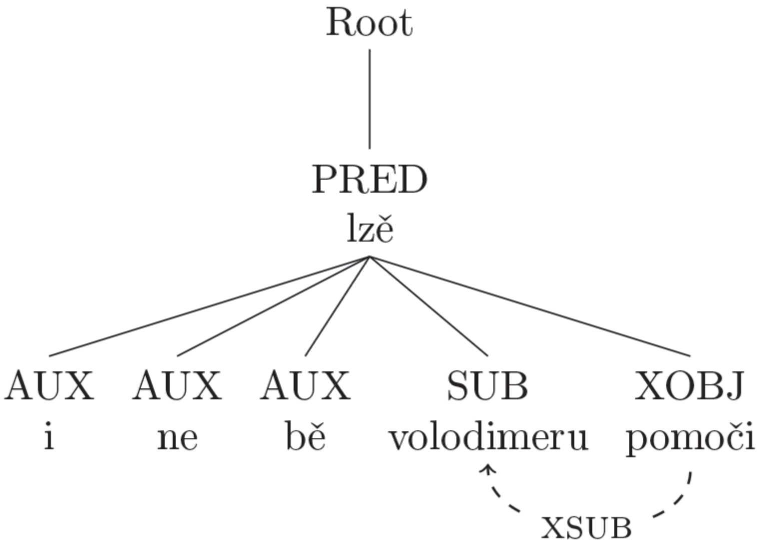


(3) и нє бѣ лꙃѣ володимєрꙋ помочи.
“and it was impossible for Volodimer to help” (PVL 127.16)

In the later stages, an analysis of such words as independent predicates with dative subjects and an auxiliary verb to indicate tense is probably a better linguistic analysis, as shown in the tree below, but it is not obvious when to switch to that analysis.

As in the morphological analysis and lemmatisation, one of the challenges in the syntactic analysis is how to deal with changing phenomena. Often we follow a similar approach to that taken for the emerging gerunds: We make very conservative assumptions about structure. A case in point is the syntactic history of the so-called “predicative adverbs”, such as *нельзя* ‘(it is) impossible, not allowed’. It seems clear that this class of words display increasingly predicate-like and even verb-like behaviour in the history of Russian. However, for the sake of easy retrievability, we analyse them in the same way in all periods: the “predicative adverb” is taken to be a predicative complement (XOBJ) dependent on a (null) copula, with the infinitive as its external subject (analysed as COMP due to a convention for infinitival and clausal arguments).

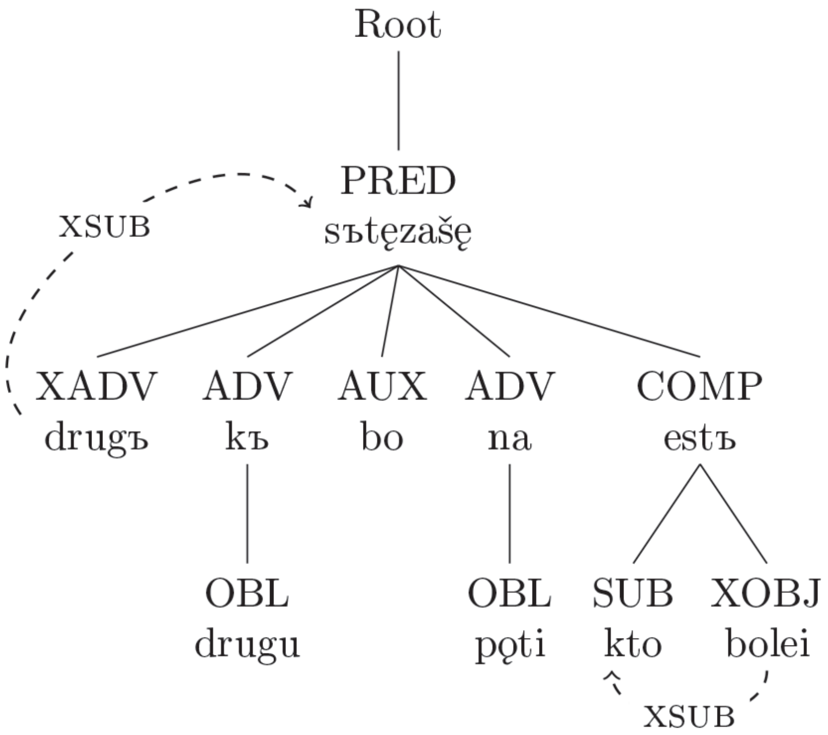
Figure 4. Alternative syntactic analysis of example (3)



In other diachronic processes, we have no choice: it is necessary to change the analysis as the change progresses. A case in point is the emergence of relative pronouns from interrogative pronouns, such as *которыи* ‘which’ *къто* ‘who’ and *чьто* ‘what’. In OCS, there is no need to have a separate relative pronoun lemma *къто.* All non-indefinite occurrences of *къто* in dependent clauses can plausibly be analysed as interrogative pronouns in indirect questions.[19] Indirect questions are virtually always complement clauses (COMP) dependent on speech, thought or emotion verbs.

[19] Note that *къто* can also be an indefinite pronoun in certain types of dependent clauses. We will not discuss these examples here, since they provide considerable additional complications.

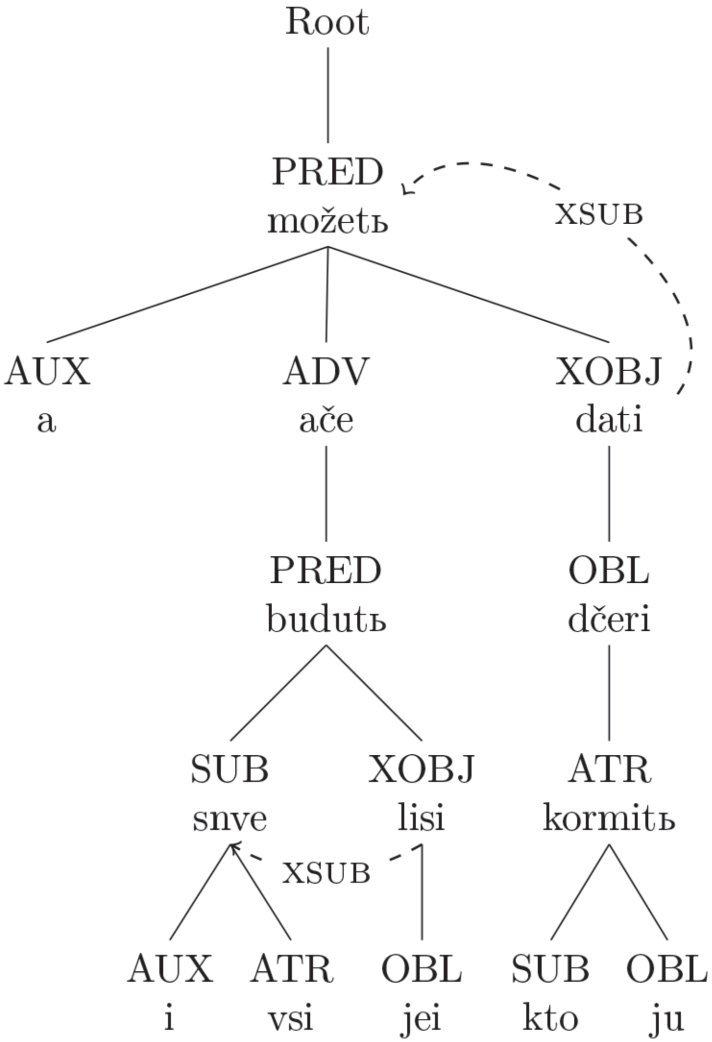
Figure 5. Syntactic analysis of example (4)



(4) Дроугъ къ дроугоу бо сътѧѕашѧ на пѫти. кто естъ болеи.
“For they had discussed with one another about who was the greatest” (Codex Marianus, Mark 9.34)

In Old East Slavic and Middle Russian, on the other hand, we see that this single analysis is no longer adequate. Examples crop up where the dependent clause containing *къто* is unambiguously a relative clause, dependent on a nominal head, as seen in (5). Such clauses must be analysed as attributes (ATR) on their nominal heads, and we must lemmatise *къто* as a relative pronoun to signal that this is a relative clause, so that it can be retrieved along with all other relative clauses.

Figure 6. Syntactic analysis of example (5)



(5) аче и вси сн҃ве ѥи будуть лиси. а дчери мѡжеть дати. кто ю кормить

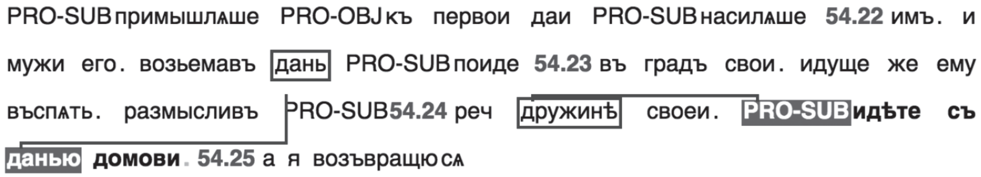
“if all her sons are mean to her, then she can give (her property) to a daughter who feeds her” (Russkaja pravda 106)

As soon as we open for a relative-clause analysis, however, a lot of examples become ambiguous: it is not always clear whether something is a headless relative clause argument or a complement clause. In such a scenario, we continue to take all dependent clauses containing къто as complement clauses if they are dependent on speech, thought or emotion verbs. Other argument dependent clauses containing *къто* are taken to be relative clauses, къто is lemmatised as a relative pronoun, and the dependent clause is given the relevant argument tag (SUB, OBJ, OBL …).

6. Expanding the TOROT

The TOROT treebank is currently under expansion in two ways. On the one hand, we are taking advantage of the corpus architecture to add additional layers of annotation. On the other hand, we are expanding the text base: we are steadily adding more Old East Slavic and Middle Russian text, and we have added a modern Russian stage by converting the SynTagRus treebank into the PROIEL dependency format.

As mentioned in section 1, the annotation application has a separate annotation interface for information status annotation and anaphoric links. For the OCS Gospels, such annotation can be transferred via the token alignment links from the Greek Gospels, which were annotated in full during the PROIEL project. There are also annotated passages in the Codex Suprasliensis and the Primary Chronicle, which can easily be expanded.



PRO-SUB примышлѧше PRO-OBJ къ первои даи PRO-SUB насилѧше **54.22** имъ. и мужи его. возьемавъ дань PRO-SUB поиде **54.23** въ градъ свои. идуще же ему въспѧть. размысливъ PRO-SUB **54.24** реч дружинѣ своеи. **PRO-SUB идѣте съ данью домови**. **54.25** а я возъвращю сѧ

Figure 7. Information status annotation. Old referents are marked in red, anaphoric links in blue. Highlighted sentence: "Go home with the tribute" (PVL 54.24)

There is also the possibility to add customised tags at sentence, lemma and token level. For instance, OCS verbs have lemma-level tags indicating their stem, prefix (if any) and suffix (if any), and nouns and denominal adjectives have lemma-level tags indicating animacy. The same type of tagging will be applied to the Old and Middle Russian data.

In many diachronic studies it is important to be able to compare the historical data with modern data, in our case, data from Modern Russian. In order to have truly comparable data, we have converted the SynTagRus treebank[20] (approximately 800,000 word tokens) to the PROIEL dependency format (see Berdičevskis and Eckhoff 2014 and 2015), and the converted treebank will be published on the TOROT corpus website. In effect, we have a Slavic diachronic treebank spanning over 1000 years.

[20] SynTagRus (found at http://ruscorpora.ru/search-syntax.html) was developed by the Laboratory of Computational Linguistics at the Institute for Information Transmission Problems, who have kindly granted us access to the offline version of the treebank and allowed us to publish the converted treebank data, for which we are very grateful.

7. Linguistics and philology: some perspectives

One of TOROT's strengths is that it is designed and made for and by linguists. The linguistic annotation is based on modern linguistic theory, but is still mostly recognisable from a more traditional point of view. The fact that it is a treebank, providing syntactic analyses of every sentence, is a great advantage both for the part-of-speech assignment and the morphological analysis. Any manual morphological analysis will necessarily be based on implicit syntactic analyses. In TOROT, we make these analyses explicit, and are also able to encode examples where the syntax and morphology are at odds. Since the data is constantly in use as data for various linguistic studies by the corpus builders, the treebank is under continual targeted correction: every linguistic study yields more precise data. Additional layers of tagging are also typically added in the course of particular research projects. For example, the verb affix and stem tagging in OCS was added for the purposes of Eckhoff and Haug 2015. We encourage all scholars using our data to add their personal classifications to our treebank.

However, linguists are not necessarily good at text editions, since their view of the text is shaped by their linguistic interests. Linguists focus on larger units, such as words, clauses and sentences: they are keen on segmentation, but inclined to ignore non-linguistic (and even prosodic) information. In general they strive to choose a single interpretation, which is not necessarily fair to the text.

Therefore, collaboration is the ideal situation. Treebank builders can create very rich and sophisticated linguistic data. This data should ideally be incorporated as one viewing level in a rich and flexible digital editions created by textologists. We are eager to offer up our treebank data for further collaborations of this kind.

REFERENCES

Berdičevskis and Eckhoff 2015: Berdičevskis, Aleksandrs and Hanne Eckhoff. "Automatic identification of shared arguments in verbal coordinations". *Computational Linguistics and Intellectual Technologies. International Conference Dialogue 2015 Proceedings 14(21), 33–43*. Moscow 2015.

Berdičevskis and Eckhoff 2014: Berdičevskis, Aleksandrs and Hanne Eckhoff. "Verbal constructional profiles: reliability, distinction power and practical applications". *Proceedings of the Thirteenth International Workshop on Treebanks and Linguistic Theories* (TLT13), Tübingen 2014.

Eckhoff 2015: Eckhoff, Hanne Martine. "Animacy and differential object marking in Old Church Slavonic". *Russian Linguistics* 39/2 (2015), 233–254.

Eckhoff and Haug 2015: Eckhoff, Hanne and Dag Haug. "Aspect and prefixation in Old Church Slavonic". *Diachronica* 32:2 (2015), 186–230.

Brants 2000: Brants, Thorsten "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, 2000, 224–231.

Haug, Eckhoff and Welo 2014: Haug, Dag, Hanne Eckhoff and Eirik Welo. "The theoretical foundations of givenness annotation". In Kristin Bech and Kristine Eide (eds.): *Information Structure and Syntactic Change in Germanic and Romance Languages.* Amsterdam: John Benjamins 2014.

Haug et al. 2009: Haug, Dag Trygve Truslew, Marius Jøhndal, Hanne Martine Eckhoff, Eirik Welo, Mari Johanne Bordal Hertzenberg and Angelika Müth. "Computational and Linguistic Issues in Designing a Syntactically Annotated Parallel Corpus of Indo-European Languages". *Traitement Automatique des Langues 50 (2009).*

Knowles and Mohd Don 2004: Knowles, Gerry, and Zuraidah Mohd Don. "The Notion of a 'lemma': Headwords, Roots and Lexical Sets". *International Journal of Corpus Linguistics* 9:2 (2004), 69–81.

Russian Academy of Sciences 1975–: Словарь русского языка XI–XVII вв. Москва: Наука (1975–).

Skjærholt 2011: Skjærholt, Arne. "More, faster: Accelerated corpus annotation with statistical taggers". *Journal for Language Technology and Computational Linguistics*, 26: 2 (2011), 151–163.

Sreznevskij 1893–1903: Срезневский, Измаил И. Материалы для словаря древнерусского языка. Москва 1893–1903/1958.

About the authors…

Hanne Eckhoff is a researcher at UiT The Arctic University of Norway. Her chief research interests are Slavic historical corpus linguistics, semantics, syntax and information structure. She is the leader of the team building and maintaining the Tromsø Old Russian and OCS Treebank (TOROT), and has published work on Old Church Slavonic and Old Russian possessive constructions, Old Church Slavonic aspect and animacy, and contrastive work on Indo-European prepositional semantics. She also has several publications on more technical issues in corpus annotation and use.

Aleksandrs Berdičevskis is a postdoctoral fellow at UiT The Arctic University of Norway. His research interests lie within language change, empirical methods and Russian linguistics. He has published on language change in modern Russian, language complexity, experimental approaches to language evolution, and corpus linguistics. He is a member of the team building and maintaining the Tromsø Old Russian and OCS Treebank (TOROT).