TOROT: The Tromsø Old Russian and OCS Treebank

Hanne Eckhoff

UiT Arctic University of Norway

April 21, 2015

Birds and Beasts and the TOROT

- Birds and Beasts: Shaping Events in Old Russian (2013–2016)

Birds and Beasts and the TOROT

- Birds and Beasts: Shaping Events in Old Russian (2013–2016)
- Two main purposes
  - Study Russian verbal prefixation patterns diachronically and contrastively
  - Build a treebank of OCS, Old and Middle Russian (goal: 220 000 (new) word tokens)

Birds and Beasts and the TOROT

- Birds and Beasts: Shaping Events in Old Russian (2013–2016)
- Two main purposes
  - Study Russian verbal prefixation patterns diachronically and contrastively
  - Build a treebank of OCS, Old and Middle Russian (goal: 220 000 (new) word tokens)
- TOROT: Tromsø Old Russian and OCS Treebank at nestor.uit.no
- No treebank is perfect, but ours should now be ready to use

PROIEL

- Point of departure: the OCS part of the PROIEL corpus

PROIEL

- Point of departure: the OCS part of the PROIEL corpus
- PROIEL: Pragmatic Resources in Old Indo-European Languages

PROIEL

- Point of departure: the OCS part of the PROIEL corpus
- PROIEL: Pragmatic Resources in Old Indo-European Languages
- By what linguistic means do these languages express pragmatics and information structure?

PROIEL

- Point of departure: the OCS part of the PROIEL corpus
- PROIEL: Pragmatic Resources in Old Indo-European Languages
- By what linguistic means do these languages express pragmatics and information structure?
- Word order, anaphoric expressions, definiteness, participles (background events), discourse particles

PROIEL

- Point of departure: the OCS part of the PROIEL corpus
- PROIEL: Pragmatic Resources in Old Indo-European Languages
- By what linguistic means do these languages express pragmatics and information structure?
- Word order, anaphoric expressions, definiteness, participles (background events), discourse particles
- Centrepiece: A parallel corpus of old Indo-European New Testament texts (Greek, Latin, Gothic, Classical Armenian and OCS)

PROIEL

- Point of departure: the OCS part of the PROIEL corpus
- PROIEL: Pragmatic Resources in Old Indo-European Languages
- By what linguistic means do these languages express pragmatics and information structure?
- Word order, anaphoric expressions, definiteness, participles (background events), discourse particles
- Centrepiece: A parallel corpus of old Indo-European New Testament texts (Greek, Latin, Gothic, Classical Armenian and OCS)
- Focus on making the most of a limited dataset by in-depth manual annotation on many levels

A family of treebanks for ancient languages

- Classical Latin and Ancient Greek: expansions of the PROIEL corpus
- Byzantine Greek hosted by PROIEL
- Germanic and Romance: ISWOC
- Old Norse: Menotec and Greinir Skáldskapar (and experimental work on Old Swedish in Gothenburg)

A family of treebanks for ancient languages

- Classical Latin and Ancient Greek: expansions of the PROIEL corpus
- Byzantine Greek hosted by PROIEL
- Germanic and Romance: ISWOC
- Old Norse: Menotec and Greinir Skáldskapar (and experimental work on Old Swedish in Gothenburg)
- All share an open-source annotation tool custom-built for ancient Indo-European languages (the PROIEL webapp)

A family of treebanks for ancient languages

- Classical Latin and Ancient Greek: expansions of the PROIEL corpus
- Byzantine Greek hosted by PROIEL
- Germanic and Romance: ISWOC
- Old Norse: Menotec and Greinir Skáldskapar (and experimental work on Old Swedish in Gothenburg)
- All share an open-source annotation tool custom-built for ancient Indo-European languages (the PROIEL webapp)
- All share guidelines for syntactic (and information structure) annotation

A family of treebanks for ancient languages

- Classical Latin and Ancient Greek: expansions of the PROIEL corpus
- Byzantine Greek hosted by PROIEL
- Germanic and Romance: ISWOC
- Old Norse: Menotec and Greinir Skáldskapar (and experimental work on Old Swedish in Gothenburg)
- All share an open-source annotation tool custom-built for ancient Indo-European languages (the PROIEL webapp)
- All share guidelines for syntactic (and information structure) annotation
- Both annotation tool and guidelines were developed through practical annotation and custom-made for the old Indo-European languages (rich morphology, free word order)

A family of treebanks for ancient languages

- Classical Latin and Ancient Greek: expansions of the PROIEL corpus
- Byzantine Greek hosted by PROIEL
- Germanic and Romance: ISWOC
- Old Norse: Menotec and Greinir Skáldskapar (and experimental work on Old Swedish in Gothenburg)
- All share an open-source annotation tool custom-built for ancient Indo-European languages (the PROIEL webapp)
- All share guidelines for syntactic (and information structure) annotation
- Both annotation tool and guidelines were developed through practical annotation and custom-made for the old Indo-European languages (rich morphology, free word order)
- Corpus builders are also corpus users; linguist's needs in focus

A family of treebanks for ancient languages

- Classical Latin and Ancient Greek: expansions of the PROIEL corpus
- Byzantine Greek hosted by PROIEL
- Germanic and Romance: ISWOC
- Old Norse: Menotec and Greinir Skáldskapar (and experimental work on Old Swedish in Gothenburg)
- All share an open-source annotation tool custom-built for ancient Indo-European languages (the PROIEL webapp)
- All share guidelines for syntactic (and information structure) annotation
- Both annotation tool and guidelines were developed through practical annotation and custom-made for the old Indo-European languages (rich morphology, free word order)
- Corpus builders are also corpus users; linguist's needs in focus
- Advantages to TOROT: established annotation practice for early Slavic; lemma/form base

Texts

- The goal is to amass *linguistic* knowledge, not to represent manuscripts as such

Texts

- The goal is to amass *linguistic* knowledge, not to represent manuscripts as such
- Corpus texts are at best a tertiary source, users must refer to good editions

Texts

- The goal is to amass *linguistic* knowledge, not to represent manuscripts as such
- Corpus texts are at best a tertiary source, users must refer to good editions
- Manuscript corrections and interpolations are nonetheless problematic

Texts

- The goal is to amass *linguistic* knowledge, not to represent manuscripts as such
- Corpus texts are at best a tertiary source, users must refer to good editions
- Manuscript corrections and interpolations are nonetheless problematic
- Linguists need philologists and textologists!

Texts

- The goal is to amass *linguistic* knowledge, not to represent manuscripts as such
- Corpus texts are at best a tertiary source, users must refer to good editions
- Manuscript corrections and interpolations are nonetheless problematic
- Linguists need philologists and textologists!
- Ideal collaboration: Textologists carefully prepare texts with all necessary detail, linguists provide linguistic annotation, information may be merged in an electronic edition

Texts

- The goal is to amass *linguistic* knowledge, not to represent manuscripts as such
- Corpus texts are at best a tertiary source, users must refer to good editions
- Manuscript corrections and interpolations are nonetheless problematic
- Linguists need philologists and textologists!
- Ideal collaboration: Textologists carefully prepare texts with all necessary detail, linguists provide linguistic annotation, information may be merged in an electronic edition
- Advantages to text contributors: Indexing of your choice for easy transfer of annotation

Text collaborations

- The Suprasliensis project (BAS; Anisava Miltenova and David Birnbaum): TOROT lemmatisation, morphology (and syntax?) can be integrated into the electronic edition

Text collaborations

- The Suprasliensis project (BAS; Anisava Miltenova and David Birnbaum): TOROT lemmatisation, morphology (and syntax?) can be integrated into the electronic edition
- The e-PVL (David Birnbaum)

Text collaborations

- The Suprasliensis project (BAS; Anisava Miltenova and David Birnbaum): TOROT lemmatisation, morphology (and syntax?) can be integrated into the electronic edition
- The e-PVL (David Birnbaum)
- Texts from the RRuDi (Roland Meyer)

Text collaborations

- The Suprasliensis project (BAS; Anisava Miltenova and David Birnbaum): TOROT lemmatisation, morphology (and syntax?) can be integrated into the electronic edition
- The e-PVL (David Birnbaum)
- Texts from the RRuDi (Roland Meyer)
- Middle Russian texts from the Institut russkogo jazyka

TOROT digitisations

- Project members have (reluctantly) digitised several manuscripts that were unavailable or unavailable in sufficient detail
- *Russkaja pravda, Life of Avvakum, Life of Feodosij Pečerskij*, some letters and legal acts
- Principle: always stick to a single good manuscript
- Retain original orthography as far as possible
- Consult manuscript facsimile when possible
- Base tokenisation on existing editions
- Release digitised text freely

Goals and results

text	morph	syntax	reviewed	goal
OCS	207 893	157 726	121 577	150 000
Old Russian	–	74 156	69 489	100 000
Middle Russian	–	48 097	47 403	50 000

Text inventory

text	morph	syntax	reviewed
Codex Marianus	–	57577	57554
Codex Suprasliensis	–	98077	63042
Codex Zographensis	52181	2072	981
Codex Laurentianus	–	55368	55013
Mstislav's letter	–	159	0
Russkaja pravda	–	4021	3928
Statute of Prince Vladimir	–	650	0
Uspenskij sbornik	–	13818	10548
Varlaam's donation charter	–	140	0
Domostroj	–	22662	22640
The Life of Avvakum	–	22210	22205
The Tale of Luka Koločskij	–	897	281
The taking of Pskov	–	2328	2277

Preprocessing: statistical morphological tagging

- One of TOROT's major assets is the large database of form, lemma and tag correspondences

Preprocessing: statistical morphological tagging

- One of TOROT's major assets is the large database of form, lemma and tag correspondences
- 157 000 annotated OCS tokens, 121 000 annotated Old/Middle Russian tokens is enough for good results with a statistical tagger

Preprocessing: statistical morphological tagging

- One of TOROT's major assets is the large database of form, lemma and tag correspondences
- 157 000 annotated OCS tokens, 121 000 annotated Old/Middle Russian tokens is enough for good results with a statistical tagger
- TnT tagger: Statistical morphological tagger that looks at trigrams and word-final letter sequences

Preprocessing: statistical morphological tagging

- One of TOROT's major assets is the large database of form, lemma and tag correspondences
- 157 000 annotated OCS tokens, 121 000 annotated Old/Middle Russian tokens is enough for good results with a statistical tagger
- TnT tagger: Statistical morphological tagger that looks at trigrams and word-final letter sequences
- To optimise the results, we normalise both the training data and the new data (simplified orthography)

Preprocessing: statistical morphological tagging

- One of TOROT's major assets is the large database of form, lemma and tag correspondences
- 157 000 annotated OCS tokens, 121 000 annotated Old/Middle Russian tokens is enough for good results with a statistical tagger
- TnT tagger: Statistical morphological tagger that looks at trigrams and word-final letter sequences
- To optimise the results, we normalise both the training data and the new data (simplified orthography)
- We can do a good deal of lemmatisation with a combination of lookups in the database and guessing (several layers of normalisation)

Preprocessing: statistical morphological tagging

- One of TOROT's major assets is the large database of form, lemma and tag correspondences
- 157 000 annotated OCS tokens, 121 000 annotated Old/Middle Russian tokens is enough for good results with a statistical tagger
- TnT tagger: Statistical morphological tagger that looks at trigrams and word-final letter sequences
- To optimise the results, we normalise both the training data and the new data (simplified orthography)
- We can do a good deal of lemmatisation with a combination of lookups in the database and guessing (several layers of normalisation)
- The pre-tagging is not good enough to serve directly as data, but gives good annotation support

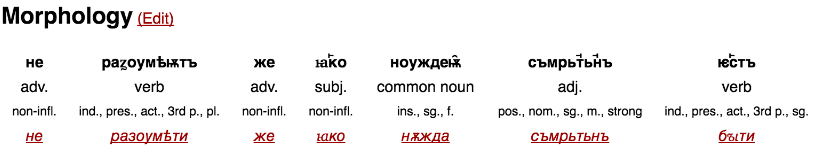
Preprocessing: statistical morphological tagging

- One of TOROT's major assets is the large database of form, lemma and tag correspondences
- 157 000 annotated OCS tokens, 121 000 annotated Old/Middle Russian tokens is enough for good results with a statistical tagger
- TnT tagger: Statistical morphological tagger that looks at trigrams and word-final letter sequences
- To optimise the results, we normalise both the training data and the new data (simplified orthography)
- We can do a good deal of lemmatisation with a combination of lookups in the database and guessing (several layers of normalisation)
- The pre-tagging is not good enough to serve directly as data, but gives good annotation support
- Feasible: autotagging of very close text variants (Codex Zographensis)

Preprocessing: statistical morphological tagging

- One of TOROT's major assets is the large database of form, lemma and tag correspondences
- 157 000 annotated OCS tokens, 121 000 annotated Old/Middle Russian tokens is enough for good results with a statistical tagger
- TnT tagger: Statistical morphological tagger that looks at trigrams and word-final letter sequences
- To optimise the results, we normalise both the training data and the new data (simplified orthography)
- We can do a good deal of lemmatisation with a combination of lookups in the database and guessing (several layers of normalisation)
- The pre-tagging is not good enough to serve directly as data, but gives good annotation support
- Feasible: autotagging of very close text variants (Codex Zographensis)
- Auto-tag other PVL manuscripts and align?

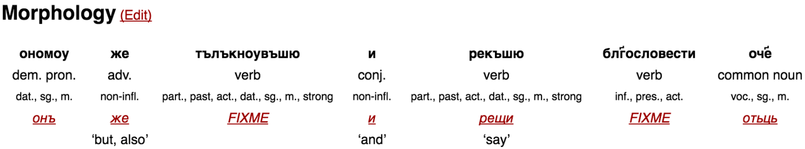
Auto-tagged Suprasliensis



Morphology (Edit)

не	раꙁоумѣѭтъ	же	ꙗ҄ко	ноуждеѭ̑	съмрь҄ть҄нъ	ѥ҄стъ
adv.	verb	adv.	subj.	common noun	adj.	verb
non-infl.	ind., pres., act., 3rd p., pl.	non-infl.	non-infl.	ins., sg., f.	pos., nom., sg., m., strong	ind., pres., act., 3rd p., sg.
не	разоумѣти	же	ıако	нѫжда	съмрьтьнъ	бꙑти

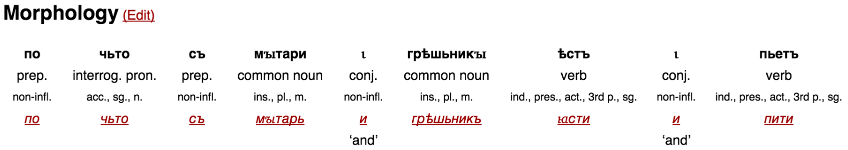
Auto-tagged Feodosij Pečerskij



Morphology (Edit)

ономоу	же	тълъкноувъшю	и	рекъшю	блг҃ословести	оч҃е
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’		

Auto-tagged Zographensis with some corrections



Morphology (Edit)

по	чьто	съ	мꙑтари	ι	грѣшьникꙑ	ѣстъ	ι	пьетъ
prep.	interrog. pron.	prep.	common noun	conj.	common noun	verb	conj.	verb
non-infl.	acc., sg., n.	non-infl.	ins., pl., m.	non-infl.	ins., pl., m.	ind., pres., act., 3rd p., sg.	non-infl.	ind., pres., act., 3rd p., sg.
по	чьто	съ	мꙑтарь	и ‘and’	грѣшьникъ	ꙗсти	и ‘and’	пити

Annotation work flow

- International team of annotators working online

Annotation work flow

- International team of annotators working online
- Check sentence and word division

Annotation work flow

- International team of annotators working online
- Check sentence and word division
- Correct morphology and lemmatisation

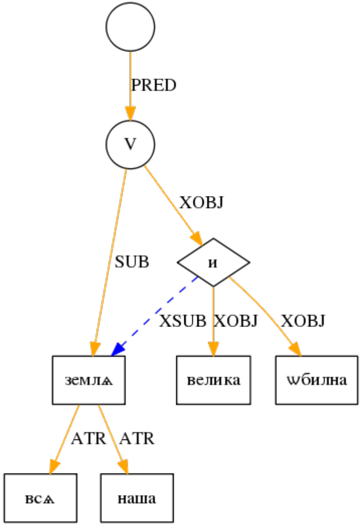
Annotation work flow

- International team of annotators working online
- Check sentence and word division
- Correct morphology and lemmatisation
- Give syntactic analysis (enriched dependency grammar) guided by rule-based guesses

Annotation work flow

- International team of annotators working online
- Check sentence and word division
- Correct morphology and lemmatisation
- Give syntactic analysis (enriched dependency grammar) guided by rule-based guesses
- Future: Experiment with syntactic parsing and pre-tagging?

Syntactic analysis



Extra layers

- Separate layer for annotating information status and anaphoric relations

Extra layers

- Separate layer for annotating information status and anaphoric relations
- All NT texts are aligned with the Greek text at token level

Extra layers

- Separate layer for annotating information status and anaphoric relations
- All NT texts are aligned with the Greek text at token level
- Customised tagging available at token, lemma and sentence level

Extra layers

- Separate layer for annotating information status and anaphoric relations
- All NT texts are aligned with the Greek text at token level
- Customised tagging available at token, lemma and sentence level
- OCS: nouns are annotated for animacy, verbs are annotated for prefixation, suffixation and stem

Availability

- All sentences are (to be) checked by a reviewer

Availability

- All sentences are (to be) checked by a reviewer
- We do consistency checks continually

Availability

- All sentences are (to be) checked by a reviewer
- We do consistency checks continually
- Anyone can register and use the corpus

Availability

- All sentences are (to be) checked by a reviewer
- We do consistency checks continually
- Anyone can register and use the corpus
- Simple query interface that allows combinations of features

Availability

- All sentences are (to be) checked by a reviewer
- We do consistency checks continually
- Anyone can register and use the corpus
- Simple query interface that allows combinations of features
- For syntactic queries, TOROT is available in the INESS treebank facility at http://clarino.uib.no/iness

Availability

- All sentences are (to be) checked by a reviewer
- We do consistency checks continually
- Anyone can register and use the corpus
- Simple query interface that allows combinations of features
- For syntactic queries, TOROT is available in the INESS treebank facility at http://clarino.uib.no/iness
- Annotated data may also be downloaded in several formats, including ones that can serve as input to syntactic query facilities (TigerXML, CoNLL)

Availability

- All sentences are (to be) checked by a reviewer
- We do consistency checks continually
- Anyone can register and use the corpus
- Simple query interface that allows combinations of features
- For syntactic queries, TOROT is available in the INESS treebank facility at http://clarino.uib.no/iness
- Annotated data may also be downloaded in several formats, including ones that can serve as input to syntactic query facilities (TigerXML, CoNLL)
- The data are released under a Creative Commons Attribution-NonCommercial-ShareAlike license

Availability

- All sentences are (to be) checked by a reviewer
- We do consistency checks continually
- Anyone can register and use the corpus
- Simple query interface that allows combinations of features
- For syntactic queries, TOROT is available in the INESS treebank facility at http://clarino.uib.no/iness
- Annotated data may also be downloaded in several formats, including ones that can serve as input to syntactic query facilities (TigerXML, CoNLL)
- The data are released under a Creative Commons Attribution-NonCommercial-ShareAlike license
- For demonstrations of the query options: demo session!

A user-built corpus

- A full-coverage corpus will have less bias than a database collected and annotated for a specific study

A user-built corpus

- A full-coverage corpus will have less bias than a database collected and annotated for a specific study
- The syntactic analysis enhances the morphological analysis; it is an advantage to make the syntactic interpretation explicit

A user-built corpus

- A full-coverage corpus will have less bias than a database collected and annotated for a specific study
- The syntactic analysis enhances the morphological analysis; it is an advantage to make the syntactic interpretation explicit
- Several phenomena may be given elegant analyses by exploiting the interplay between the syntactic and morphological layers

A user-built corpus

- A full-coverage corpus will have less bias than a database collected and annotated for a specific study
- The syntactic analysis enhances the morphological analysis; it is an advantage to make the syntactic interpretation explicit
- Several phenomena may be given elegant analyses by exploiting the interplay between the syntactic and morphological layers
- Animacy: the genitive-accusative is always taken as genitive in the morphology, its status is determined by the syntax (OBJ? OBL? negated?)

Squeezing the empirical lemon

- The depressing life of the historical linguist

Squeezing the empirical lemon

- The depressing life of the historical linguist
- Many-layered annotation can be combined and give new insights

Squeezing the empirical lemon

- The depressing life of the historical linguist
- Many-layered annotation can be combined and give new insights
- Easy access to exhaustive data for high-frequency phenomena

Squeezing the empirical lemon

- The depressing life of the historical linguist
- Many-layered annotation can be combined and give new insights
- Easy access to exhaustive data for high-frequency phenomena
- How far can statistics take us?

Squeezing the empirical lemon

- The depressing life of the historical linguist
- Many-layered annotation can be combined and give new insights
- Easy access to exhaustive data for high-frequency phenomena
- How far can statistics take us?
- Every study improves the corpus: targeted corrections

The status of OCS *byti*

- Eckhoff, Janda and Nesset 2014: Grammatical profiling and constructional profiling to assess whether *byti* was one or two verbs
- Data layers: morphology, syntax, token alignments (Greek used as rough semantic tags)
- Radial category structure of the verb's semantics emerged from argument structure data
- *Byti* should most reasonably be seen as a single polysemous verb

Inflectional and derivational aspect in OCS

- Eckhoff and Haug to appear (soon!)
- Data layers: Morphology, syntax, prefix/stem/suffix tags, token alignments
- Conclusions:
  - Verb pairs and imperfect/aorist both express viewpoint aspect
  - The aorist is independent of telicity and has retained meanings that the new perfective doesn't have
  - These meanings can only be seen with atelic simplex verbs (delimitative, ingressive)
  - Evidence that aspect mismatches were a later development: imperfective aorist and perfective imperfect were not found in Marianus/Zographensis

Animacy and definiteness in OCS

- Eckhoff to appear (soon!)
- Data layers: Morphology, syntax, semantic tags (animacy), information status, anaphoric links, token alignments
- The gen-acc predominates with old and accessible objects
- Variation between gen-acc and nom-acc with new and anchored objects
- The nom-acc marks referential persistence
- The gen-acc may be preferred if the subject has low discourse prominence

More than a millennium on the same format

- How to control our data against modern Russian?

More than a millennium on the same format

- How to control our data against modern Russian?
- Converted SynTagRus to the PROIEL/TOROT format and will publish the full conversion on nestor.uit.no

More than a millennium on the same format

- How to control our data against modern Russian?
- Converted SynTagRus to the PROIEL/TOROT format and will publish the full conversion on nestor.uit.no
- Two dependency formats with different theoretical allegiances: Meaning-Text Theory vs. LFG

More than a millennium on the same format

- How to control our data against modern Russian?
- Converted SynTagRus to the PROIEL/TOROT format and will publish the full conversion on nestor.uit.no
- Two dependency formats with different theoretical allegiances: Meaning-Text Theory vs. LFG
- Interesting differences in argument structure handling (Berdičevskis and Eckhoff 2014)

More than a millennium on the same format

- How to control our data against modern Russian?
- Converted SynTagRus to the PROIEL/TOROT format and will publish the full conversion on nestor.uit.no
- Two dependency formats with different theoretical allegiances: Meaning-Text Theory vs. LFG
- Interesting differences in argument structure handling (Berdičevskis and Eckhoff 2014)
- Adding information: secondary dependencies (Berdičevskis and Eckhoff to appear (soon!))

Using the SynTagRus data

- Do perfective and imperfective verbs have different constructional profiles? Do they have different distributions across argument frames?
- It appears that they do
- We can track the development of simplex verbs: from aspectually neutral to imperfective

The history of simplex verbs: prediction

- **Fact:** the average imperfective and perfective profiles are different

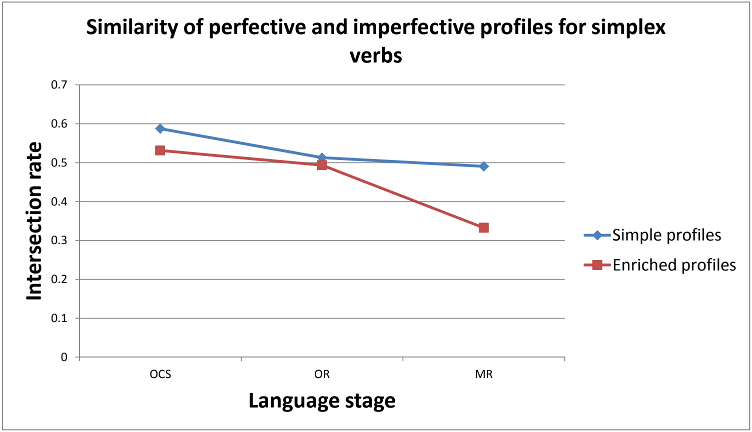
The history of simplex verbs: prediction

- **Fact:** the average imperfective and perfective profiles are different
- **Hypothesis:** for simplex verbs, the aspectual opposition is most relevant in Modern Russian, less so in Old Russian, even less in Old Church Slavonic

The history of simplex verbs: prediction

- **Fact:** the average imperfective and perfective profiles are different
- **Hypothesis:** for simplex verbs, the aspectual opposition is most relevant in Modern Russian, less so in Old Russian, even less in Old Church Slavonic
- **Prediction:** the intersection rate (measure of similarity) between the 'simplex perfective' and 'simplex imperfective' profiles will be highest for Old Church Slavonic and lowest for Modern Russian

The history of simplex verbs: results



Varangian Rus' Digital Environment: pedagogical applications

- Birds and Beasts' pedagogically oriented sister project

Varangian Rus' Digital Environment: pedagogical applications

- Birds and Beasts' pedagogically oriented sister project
- Tore Nesset (to appear): *How Russian came to be the way it is*

Varangian Rus' Digital Environment: pedagogical applications

- Birds and Beasts' pedagogically oriented sister project
- Tore Nesset (to appear): *How Russian came to be the way it is*
- Cooperation with the Higher School of Economics (Moscow)

Varangian Rus' Digital Environment: pedagogical applications

- Birds and Beasts' pedagogically oriented sister project
- Tore Nesset (to appear): *How Russian came to be the way it is*
- Cooperation with the Higher School of Economics (Moscow)
- Texts offered with morphological and syntactic analysis and philological commentary

Varangian Rus' Digital Environment: pedagogical applications

- Birds and Beasts' pedagogically oriented sister project
- Tore Nesset (to appear): *How Russian came to be the way it is*
- Cooperation with the Higher School of Economics (Moscow)
- Texts offered with morphological and syntactic analysis and philological commentary
- Dictionary resource exploiting the TOROT lemma and form inventory

Varangian Rus' Digital Environment: pedagogical applications

- Birds and Beasts' pedagogically oriented sister project
- Tore Nesset (to appear): *How Russian came to be the way it is*
- Cooperation with the Higher School of Economics (Moscow)
- Texts offered with morphological and syntactic analysis and philological commentary
- Dictionary resource exploiting the TOROT lemma and form inventory
- Expand the Old/Middle Russian part of TOROT with 100 000 more tokens

Lemmas with attested paradigms: *darъ*

	sg	du	pl
N	darъ	–	–
A	darъ	–	dary
G	daru	–	darovъ
D	daru	–	daromъ
I	daromъ, darom	–	dary
L	–	–	–
V	–	–	–

Summary

- TOROT: a treebank of OCS, Old and Middle Russian (nestor.uit.no) – and with converted data for modern Russian

Summary

- TOROT: a treebank of OCS, Old and Middle Russian (nestor.uit.no) – and with converted data for modern Russian
- Made for and by linguists

Summary

- TOROT: a treebank of OCS, Old and Middle Russian (nestor.uit.no) – and with converted data for modern Russian
- Made for and by linguists
- Belongs to a larger family of compatible treebanks for ancient languages

Summary

- TOROT: a treebank of OCS, Old and Middle Russian (nestor.uit.no) – and with converted data for modern Russian
- Made for and by linguists
- Belongs to a larger family of compatible treebanks for ancient languages
- Benefits from customised annotation application and well-established standards and guidelines

Summary

- TOROT: a treebank of OCS, Old and Middle Russian (nestor.uit.no) – and with converted data for modern Russian
- Made for and by linguists
- Belongs to a larger family of compatible treebanks for ancient languages
- Benefits from customised annotation application and well-established standards and guidelines
- Application is open-source and data are freely shared for non-commercial use

Summary

- TOROT: a treebank of OCS, Old and Middle Russian (nestor.uit.no) – and with converted data for modern Russian
- Made for and by linguists
- Belongs to a larger family of compatible treebanks for ancient languages
- Benefits from customised annotation application and well-established standards and guidelines
- Application is open-source and data are freely shared for non-commercial use
- Comprehensive annotation improves overall quality of data

Summary

- TOROT: a treebank of OCS, Old and Middle Russian (nestor.uit.no) – and with converted data for modern Russian
- Made for and by linguists
- Belongs to a larger family of compatible treebanks for ancient languages
- Benefits from customised annotation application and well-established standards and guidelines
- Application is open-source and data are freely shared for non-commercial use
- Comprehensive annotation improves overall quality of data
- This kind of data yields interesting results in long-disputed questions for OCS and Old Russian

Summary

- TOROT: a treebank of OCS, Old and Middle Russian (nestor.uit.no) – and with converted data for modern Russian
- Made for and by linguists
- Belongs to a larger family of compatible treebanks for ancient languages
- Benefits from customised annotation application and well-established standards and guidelines
- Application is open-source and data are freely shared for non-commercial use
- Comprehensive annotation improves overall quality of data
- This kind of data yields interesting results in long-disputed questions for OCS and Old Russian
- A strong, quality-controlled basis for further computational approaches to OCS, Old and Middle Russian

Summary

- TOROT: a treebank of OCS, Old and Middle Russian (nestor.uit.no) – and with converted data for modern Russian
- Made for and by linguists
- Belongs to a larger family of compatible treebanks for ancient languages
- Benefits from customised annotation application and well-established standards and guidelines
- Application is open-source and data are freely shared for non-commercial use
- Comprehensive annotation improves overall quality of data
- This kind of data yields interesting results in long-disputed questions for OCS and Old Russian
- A strong, quality-controlled basis for further computational approaches to OCS, Old and Middle Russian
- Coming: pedagogical tools and a dictionary resource