TROLLing: Open Data for Linguists

Laura A. Janda, UiT
with a lot of help from colleagues at the Department of Language and Linguistics and the University Library
How linguistics used to work...

Why is the past tense of *dive* sometimes *dove* and sometimes *dived*?
Linguistics and data

Two things happened >10 years ago

• Advent of digital corpora
  – for many languages
  – 100s of millions of words
  – balanced, annotated

• R became widely used
  – open source statistical software
  – and other tools have become available (MathLab, etc.)
Linguistics and data

Two things happened >10 years ago

- Advent of digital corpora
  - for many languages
  - 100s of millions of words
  - balanced, annotated

- R became widely used
  - open source statistical software
  - and other tools have become available (MathLab, etc.)
The view from one journal...
Over 50% since 2008
Leveling off at 75-80%?
How linguistics works today...

The usual process:
- Data is extracted from corpus or collected from experiments
- Laborious cleaning, tagging
- Statistical analysis
- Publications

BUT:
- What happens to the data after results are published?
- Can the researcher find and interpret the data later?
- What if someone else wants to use the data?
What today’s linguists need

A PLACE TO PUBLICLY ARCHIVE DATA AND CODE

A MORAL IMPERATIVE:
• Create ethical standards for sharing of data and code
• Set norms for use of statistical methods
• Learn from each other and help our community grow
• Secure and maintain scientific integrity

AN OBLIGATION:
• Funding agencies and scientific journals are increasingly requiring scholars to make their data publicly available
A PLACE FOR US:

TROLLing
The Tromsø Repository of Language and Linguistics
TROLLing

- is an international archive of linguistic data and statistical code
- is built on the Dataverse platform from Harvard University and complies with DataCite, the international standard for storing and citing research data
- is compliant with CLARIN, the EU research infrastructure for language-based resources
- assigns a permanent URL to each post
- uses metadata that ensures visibility and retrieval through international services
- is professionally managed by the University Library of Tromsø and an international steering committee
## TROLLing Dataverse

The Tromsø Repository of Language and Linguistics

**Find TROLLing at opendata.uit.no**

### Studies: 18 | Downloads: 177

<table>
<thead>
<tr>
<th>Title</th>
<th>Description</th>
<th>Last Released</th>
<th>Downloads</th>
</tr>
</thead>
<tbody>
<tr>
<td>Replication data for: Ditransitive constructions in Russian and Ukrainian</td>
<td>This paper presents new experimental data on the acquisition of structures involving ditransitive verbs in two East Slavic languages: Russian and Ukrainian. The results of an elicited production task with adults and 3...</td>
<td>Jun 18, 2014</td>
<td>39</td>
</tr>
<tr>
<td>Replication data for: Old Church Slavonic byl Part One: Grammatical Profiling Analysis</td>
<td>There is controversy over whether byl 'be' in Old Church Slavonic functioned as an imperfective verb with an unusually large number of inflected forms or as an aspectual pair of verbs, reflecting its suppletive origin from two stems (en-...</td>
<td>May 21, 2014</td>
<td>5</td>
</tr>
<tr>
<td>Spanish deverbal adjectives (1): Ivo</td>
<td>Attested adjectives in -ivo in contemporary Spanish, with information about their bases, their semantic interpretation and their morphological properties.</td>
<td>Jun 16, 2014</td>
<td>19</td>
</tr>
<tr>
<td>Metonymy in Word-Formation: Russian, Czech, and Norwegian</td>
<td>A foundational goal of cognitive linguistics is to explain linguistic phenomena in terms of general cognitive strategies rather than postulating an autonomous language module (Langacker 1987: 12-13). Metonymy is ident...</td>
<td>Jun 13, 2014</td>
<td>3</td>
</tr>
<tr>
<td>Why Russian Prefixes Aren't Empty</td>
<td>We claim that Russian verbal prefixes always express meaning, even when they are used to form the perfective partners of aspectual pairs. The prefixes in verbs like ходил/ходил-ал 'walk' and верил/верил-ал 'believe'...</td>
<td>Jun 16, 2014</td>
<td>10</td>
</tr>
<tr>
<td>Russian nu-drop verbs</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

**UIT Dataverse Network ->**

**TROLLing Dataverse**

The Tromsø Repository of Language and Linguistics

**Getting started with TROLLing**

---

**Search Studies**

**Advanced Search**

**Tips**

**Sort By:** Global ID

<table>
<thead>
<tr>
<th>Title</th>
<th>Description</th>
<th>Last Released</th>
<th>Downloads</th>
</tr>
</thead>
<tbody>
<tr>
<td>Replication data for: Ditransitive constructions in Russian and Ukrainian</td>
<td>This paper presents new experimental data on the acquisition of structures involving ditransitive verbs in two East Slavic languages: Russian and Ukrainian. The results of an elicited production task with adults and 3...</td>
<td>Jun 18, 2014</td>
<td>39</td>
</tr>
<tr>
<td>Replication data for: Old Church Slavonic byl Part One: Grammatical Profiling Analysis</td>
<td>There is controversy over whether byl 'be' in Old Church Slavonic functioned as an imperfective verb with an unusually large number of inflected forms or as an aspectual pair of verbs, reflecting its suppletive origin from two stems (en-...</td>
<td>May 21, 2014</td>
<td>5</td>
</tr>
<tr>
<td>Spanish deverbal adjectives (1): Ivo</td>
<td>Attested adjectives in -ivo in contemporary Spanish, with information about their bases, their semantic interpretation and their morphological properties.</td>
<td>Jun 16, 2014</td>
<td>19</td>
</tr>
<tr>
<td>Metonymy in Word-Formation: Russian, Czech, and Norwegian</td>
<td>A foundational goal of cognitive linguistics is to explain linguistic phenomena in terms of general cognitive strategies rather than postulating an autonomous language module (Langacker 1987: 12-13). Metonymy is ident...</td>
<td>Jun 13, 2014</td>
<td>3</td>
</tr>
<tr>
<td>Why Russian Prefixes Aren't Empty</td>
<td>We claim that Russian verbal prefixes always express meaning, even when they are used to form the perfective partners of aspectual pairs. The prefixes in verbs like ходил/ходил-ал 'walk' and верил/верил-ал 'believe'...</td>
<td>Jun 16, 2014</td>
<td>10</td>
</tr>
<tr>
<td>Russian nu-drop verbs</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Getting started with TROLLing

http://site.uit.no/trolling/getting-started/

- Promotional video
- Instructional videos
  - User guide
- TROLLing banner
Basic steps in TROLLing

• Create an account
  - Needed only for archiving -- you do not need an account to search or download data
  - This step is self-explanatory, but there is an instructional video
  - It may take a day or two for your account to be approved

• Create a study
  - TROLLing terms of use

• Search for a study
  - TROLLing terms of use
Creating a study: TROLLing terms of use

Affirm that:

• You are authorized to archive the data

• The data contains no viruses, confidential or inappropriate information

• You give UiT the permission to archive and make the data publicly available
Creating a study: The Metadata, part 1

Enter descriptive information about your study on the Cataloging Information page. These fields comprise the metadata that describe this study. Note that some fields are required [●] or recommended [★].

Fields with a light blue background [●] indicate that the field accepts HTML formatting. Fields with a light orange background [★] indicate that Date format is required (YYYY or YYYY-MM or YYYY-MM-DD). You also can add a new row [✚] if needed, such as to include multiple authors, or remove an existing row [∙].

Click Save to upload files, set permissions, and more.

Select Study Version: 1
Created: Draft study version is not created until the form is saved

Use the Data Citation Widget below to enter all the metadata needed for your citation.

Data Citation Widget

Click the links below to enter the data citation value metadata.
< Author >, < Date >, < Title >, < hdl:10037.1/10205 >, < UIT Open Research Data [Distributor] >, V1 [Version]

Publication Citation

☐ Replication Data for Original Publication (This will apply “Replication Data for:” to your title)

ID Type ID Number URL
Select

http://example-domain.edu/
What the Metadata looks like, part 1

REPLICATION DATA FOR: OLD CHURCH SLAVONIC BYTI PART ONE AND PART TWO

If you use these data, please add the following citation to your scholarly references. Why cite?


Results found in this publication can be replicated using these data.


### Creating a study: The Metadata, part 2

<table>
<thead>
<tr>
<th><strong>Data Citation Details</strong></th>
</tr>
</thead>
</table>

- **Title**: [Enter title](#)
- **Study Global ID**: [10205](#)
- **Author**: Name, [Affiliation](#)
- **Producer**: [Producer](#), [Affiliation](#), [URL](#), [Logo URL](#)
- **Production Date**: [YYYY or YYYY-MM or YYYY-MM-DD; AD or BC optional](#)
- **Grant Number**: [Grant Number](#), [Grant Number Agency](#)
- **Distributor**: [Distributor](#), [Affiliation](#), [URL](#), [Logo URL](#)
What the Metadata looks like, part 2
### Creating a study: The Metadata, part 3

**Description**

Copying and pasting from a Word document can create errors when you save this page.

**Description Date**

(YYYY or YYYY-MM or YYYY-MM-DD; AD or BC optional)

**Keyword**

Keyword: 

Vocabulary: <not applicable>

URL: 

(http://example-domain.edu/)

**Topic Classification**

<table>
<thead>
<tr>
<th>Topic Classification</th>
<th>Vocabulary</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Value</td>
<td>&lt;Field term: Choose one or...</td>
</tr>
</tbody>
</table>

URL: 

(http://example-domain.edu/)

**Topic Classification**

<table>
<thead>
<tr>
<th>Topic Classification</th>
<th>Vocabulary</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Value</td>
<td>&lt;Time depth: Choose one or...</td>
</tr>
</tbody>
</table>

URL: 

(http://example-domain.edu/)

**Topic Classification**

<table>
<thead>
<tr>
<th>Topic Classification</th>
<th>Vocabulary</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Value</td>
<td>&lt;Topic: Choose one or...</td>
</tr>
</tbody>
</table>

URL: 

(http://example-domain.edu/)
What the Metadata looks like, part 3

Abstract Part One. There is controversy over whether byti 'be' in Old Church Slavonic functioned as an imperfective verb with an unusually large number of inflected forms or as an aspeccual pair of verbs, reflecting its suppletive origin from two stems (es- and bu-). We offer an objective empirical approach to the status of this verb, using statistical analysis of 2,428 attestations of byti in comparison with 6,694 attestations of 129 other verbs. This makes it possible to accurately locate byti in the context of the verbal lexicon of Old Church Slavonic. The comparison is made via grammatical profiles, a method that examines the frequency distribution of each verb's inflected forms. This comparison is undertaken in two rounds, one assuming that byti is a single verb, and the other assuming that it is a pair of verbs. Both assumptions yield reasonable results, and although the grammatical profile analyses do not suffice to solve the controversy, they lay the groundwork for further analysis in Part Two that argues for a single-verb interpretation of byti. Data and R Scripts Part One: The Data Our analysis uses two datasets, one that presents the forms of byti as a single paradigm, verbs.csv, and one that presents it as a pair of verbs, splitverbs.csv. The R Scripts In order to represent the Church Slavonic orthography, you will need our transliteration script: translit.r. This script is sourced by the scripts for our analysis which present byti as either a single verb or a verb pair: PartOneSingleVerb.r and PartOneVerbPair.r. This script performs all of the steps for the analysis in our article and generates the plots.

Abstract Part Two: The verb byti 'be' in Old Church Slavonic appears in an unusually rich inventory of grammatical constructions that it appears in. We analyze corpus data on the distribution of constructions in order to assess the status of this verb as either a single verb or an aspeccual pair of verbs. Our study moves beyond a strict structuralist interpretation of the behavior of byti, instead recognizing the real variation and ambiguity in the data. Our findings make both theoretical and descriptive advances. The radial category structure is a central tenet of cognitive linguistics, but until now such structures have usually been posited by researchers based on their qualitative insights from data. We show that it is possible to identify both the nodes and the structure of a radial category statistically, using only linguistic data as input. We provide an enhanced description of byti that clearly distinguishes between core uses and those that are more peripheral and shows the relationships among them. While we find some evidence in support of an aspeccual pair, most evidence points instead toward a single verb. Data and R Script Part Two: The Data The dataset used in this analysis is frames.csv. The R Script The R script used in this analysis is PartTwo.r.

Keywords
- Old Church Slavonic
- grammatical constructions
- aspect

Topic Classification
- Field: Morphology
- Time-depth: diachronic
- Topic: verbs

Related Material
- Grammatical Profiles and Aspect in Old Church Slavonic
  Hanne M. Eckhoff and Laura A. Janda
  Article first published online: 15 APR 2013
  Transactions of the Philological Society
  DOI: 10.1111/1467-968X.12012

Country/Nation
- Norway

Geographic Coverage
- Eastern Europe

Kind of Data
- corpus

Data Availability

Number of Files
- 8

Terms of Use

Dataverse Network Terms of Use
- View Terms of Use [+]
What the data looks like: categories and descriptions

Use the check boxes next to the file name to download multiple files. Data files will be downloaded in their default format. You can also download all the files in a category by checking the box next to the category name. You will be prompted to save a single archive file. Study files that have restricted access will not be downloaded.

- **Data Part One**
  - **splitverbs.csv**
    - Plain Text - 5 MB - 18 downloads
    - MD5 Checksum: b169eb0766c903526cb73221b304b19d
    - Download
    - This file contains all the examples upon which the research in this study is based. In this dataset, byti is treated as two verbs. These examples are extracted from the PROIEL corpus...Continue [+]
  - **verbs.csv**
    - Plain Text - 5 MB - 13 downloads
    - MD5 Checksum: fad1d2d62fed3496eed87adc6482b7
    - Download
    - This file contains all the examples upon which the research in this study is based. In this dataset, byti is treated as a single verb. These examples are extracted from the PROIEL corpus...Continue [+]

- **Data Part Two**
  - **frames.csv**
    - Plain Text - 1 MB - 9 downloads
    - MD5 Checksum: c29509d99c36b1681cc28d7ceeb455a
    - Download
    - This dataset represents all the examples of byti extracted from the PROIEL corpus and tagged for the grammatical construction they appear in. Description of columns A=source (the man...Continue [+]

- **R Script Part Two**
  - **PartTwo.r**
    - Plain Text - 13 KB - 6 downloads
    - MD5 Checksum: c58f8d4804b31234a47a86c287a99e95
    - Download
    - This script runs the analysis for Part Two using the data in frames.csv.

- **R Scripts Part One**
  - **PartOneSingleVerb.r**
    - Plain Text - 2 KB - 4 downloads
    - MD5 Checksum: abcb5982d9b759ace60e09580689f75e
    - Download
    - This script runs an analysis of byti as a single verb using the data in verbs.csv.
  - **PartOneVerbPair.r**
    - Plain Text - 5 KB - 5 downloads
    - MD5 Checksum: 56b7ecc5db94c20e1f62d560a9ab370
    - Download
    - This script runs an analysis of byti as a pair of verbs using the data in splitverbs.csv.
  - **translit.r**
    - Plain Text - 2 KB - 4 downloads
    - MD5 Checksum: 74199a5c5d2fb91ac211c33f1919472
    - Download
    - In order to represent the Church Slavonic orthography, you will need to run this transliteration script. This script is sourced by the scripts for our analysis which present byti as...Continue [+]

- **Table Part One**
  - **VerbsEckhoff&Janda2013.pdf**
    - Adobe PDF - 126 KB - 8 downloads
    - MD5 Checksum: 9e72adbe85b9dd2b7abeda6474c68f3
    - Download
    - Figure 1 in Part One shows a plot of Old Church Slavonic verbs. Given the large number of verbs, many are illegible on this diagram. A full list of verbs with their Factor 1 values i...Continue [+]
Include a full description of all parameters

This file contains all the examples upon which the research in this study is based. In this dataset, byti is treated as a single verb. These examples are extracted from the PROIEL corpus. In this dataset, the forms of byti are represented as a single paradigm. Description of columns:

- source (the manuscript that the example comes from): values=Marianus, Supr. (Suprasliensis), Zogr. (Zographensis)
- source_division (the chapter where the example is found): values=
- verse (the verse where the example is found)
- token_id (the identification number for the example)
- sentence_id (the identification number for the sentence in which the example is found)
- verb (the form of the example)
- lemma_id (the identification number for the lemma)
- lemma (the lemma/dictionary form for the example)
- relation (the syntactic relation): values=adv (adverbial), apos (aposition), arg (unspecified argument used when more specific classification is unclear), aux (auxiliary), comp (complement clause), narg (nominal argument), obj (transitive object), obl (oblique argument), parred (parenthetical predication), part (nominal particle), pred (predicate), sub (subject), voc (vocative element), xadv (adverbial with external subject), xobj (argument with external subject)
- main_verb (is the example the main verb of the clause?): values=FALSE, TRUE
- prefix (prefix on the verb, if any)
- stem (stem of the verb)
- harmonised_stems (stems represented as single units despite any morphophonemic alternations)
- suffix (suffixes in the infinitive stem)
- reflexive (is there a reflexive clitic or not)
- mood (mood and finiteness)
- subjunctive, u(supine) tense; values=a(aorist), f(future), i(imperfect), p(present), u(past)
- voice values=act, p(passive)
- person values=1(first), 2(second), 3(third)
- number values=s(singular), u(verb_case) the case of a form; used only for participles; values=a(accusative), d(dative), g(genitive), i(instrumental), l(localive), n(nominative), v(vocative)
- semantics (semantic tag for the verb): values=acquisition, activity_interrelation, activity_physiological

Use open and persistent data formats (e.g., .csv, .txt, .pdf)
R scripts should be annotated

# This script shows how the statistics were calculated for our article
# Antonsen, Lene, Laura A. Janda, and Berit Anne Bals Baal. 2012. “Njealji davvisámi adpo
# Here we input the data on mangel from Table 2 in the article:
mangel=matrix(c(2535,1231,90,261),byrow=TRUE,ncol=2)

# Now we print out the table so that we can see what it looks like:
print("mangel+ in newspapers and literature")
print(mangel)

# Now we run a chi-square test:
test=chisq.test(mangel)
print(test)

# Now we sum all the numbers in our table:
print("This is the sum for the mangel+-table:"
print(sum(mangel))

# Now we calculate the effect size (Cramer’s V). We do this by taking the square root of
 cramer=sqrt(test$statistic/sum(mangel))
print(cramer)
print("This is the effect size")
After you have uploaded your study...

• You submit it for approval and receive an acknowledgement
• It will be approved and released by an administrator and you will receive an email
• You will be able to edit your study later if needed and resubmit
• Previous versions of your study are archived, but only the latest version shows up in initial searches
Search for a study

• All of the cataloging information (metadata) is searchable, including:
  – author
  – affiliation
  – country / nation
  – date of production and distribution
  – keywords, e.g. language
  – topic classification

• Advanced search
  – possible to include and exclude combinations of items
### TROLLing Dataverse

**Search this Dataverse**

<table>
<thead>
<tr>
<th>Study Cataloging Information</th>
<th>Operator</th>
<th>Search Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>contains</td>
<td></td>
</tr>
<tr>
<td>Producer</td>
<td>contains</td>
<td></td>
</tr>
<tr>
<td>Production Date</td>
<td>contains</td>
<td></td>
</tr>
<tr>
<td>Time Period Covered Start</td>
<td>contains</td>
<td></td>
</tr>
</tbody>
</table>

**Tabular Data Variable Information**

<table>
<thead>
<tr>
<th>Variable Information</th>
<th>Contains</th>
<th>Search Term</th>
</tr>
</thead>
</table>

*Refer to the User Guide for help with Basic Searches, Advanced Searches or Tips.*
Search results: clickable summary data on left-hand side
Users receive citation information...

If you use these data, please add the following citation to your scholarly references. Why cite?

Mykhaylyk, Roksolana; Rodina, Yulia; Anderssen, Merete, 2014, "Replication data for: Ditransitive constructions in Russian and Ukrainian", http://hdl.handle.net/10037.1/10005 UiT Open Research Data [Distributor] V5 [Version]

Results found in this publication can be replicated using these data.


Title
Replication data for: Ditransitive constructions in Russian and Ukrainian

Study Global ID
hdl:10037.1/10005

Authors
Mykhaylyk, Roksolana (UiT The Arctic University of Norway); Rodina, Yulia (UiT The Arctic University of Norway); Anderssen, Merete (UiT The Arctic University of Norway)

Producer
UiT The Arctic University of Norway

Production Date
2013
There are also terms of use that users must agree to when downloading files

Affirm that:

• You will not use the materials that you downloaded to identify individuals or organizations

• You will not download or use materials that would be prohibited by law

• You will cite the data as stipulated in the Data Citation Information in any publications or reports that you make using the data

• You understand that the distributor makes no warranties about the data
Comments

• You can also read any comments that others users have made about this data
• If you want to make a comment, you need to have an account and log in
• You can access earlier versions of a study if it has been edited by the author, but usually you will only need the latest version