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*?
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
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

Linguists got access to A LOT more 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

Linguists got access to A LOT more data
Linguists got a powerful tool to analyze data
The view from one journal...
percent quantitative articles in *Cognitive Linguistics*
1990-2012
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

WHY? Because we need to:

• 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
What today’s linguists need

A PLACE TO PUBLICLY ARCHIVE DATA AND CODE

WHY? Because we need to:

• 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

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
# Find TROLLing at opendata.uit.no

## TROLLing Dataverse

The Tromsø Repository of Language and Linguistics

### 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... Continue</td>
<td>Jun 18, 2014</td>
<td>39</td>
</tr>
<tr>
<td>Replication data for: Old Church Slavonic “beý” Part One: Grammatical Profiling Analysis</td>
<td>There is controversy over whether by is “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 (es... Continue</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>18</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... Continue</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 наносить/нанесён “write” and носить/нёсён “coo... Continue</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