The Tromsø Repository of Language and Linguistics

Scope and operation of an open repository for linguistic data

91ST ANNUAL MEETING OF THE LINGUISTIC SOCIETY OF AMERICA, JANUARY 5-8 2017, AUSTIN TX

Helene N. Andreassen, Philipp Conzett, Stein Høydalsvik, Laura Janda, Leif Longva, Tore Nesset & Odu Obiajulu

ARE YOU ABOUT TO PUBLISH YOUR RESEARCH, AND WANT TO MAKE YOUR DATA AND CODE AVAILABLE, ALONGSIDE THE PAPER? THE TROMSØ REPOSITORY OF LANGUAGE AND LINGUISTICS (TROLLING) AIMS TO MEET THE LingUISTIC COMMUNITY’S INCREASING WISH AND DEMAND FOR ARCHIVING, PUBLISHING, AND DISSEMINATION OF SCIENTIFIC RESEARCH DATA, AS WELL AS PROPER ATTRIBUTION TO THEM.

The raison d’être of TROLLing

Open science
- Transparency, accessibility, and reusability of research, using web-based tools.

Sharing of research data is increasingly encouraged/required by many funders and publishers.

As open as possible, as closed as necessary.

Where the data is not publicly available, and there are often good reasons why it is not, authors should explain why it is not made available.

Tromsø in the forefront by joining forces
- Initiative taken by the internationally oriented UiT linguistics community.

The University Library a natural partner, consisting of researchers in linguistics, open access and system development at the library.

Development guided by scientific needs and international standards, with a three-member scientific advisory board contributing to overarching strategic discussions.

- Long-term preservation of data ensured by the Department of Information Technology.

Encouraging the researchers to publish their data: measures taken
- To incite a change, consciousness-raising and support are key elements in our contact with the linguistic research community.

- Outreach is carried out by faculty and other trained linguists at workshops, scientific conferences, and in contact with journal editors.

- Support services are provided by the University Library, including curation of datasets, individual training, and online teaching material.

The whos and whats of TROLLing

- Available to linguistic world-wide for upload (which requires registration) and download.

- Archive for open data, although some data files may be locked for a certain period.

- Archive for open structural data, annotated and organized to serve as empirical basis in linguistic research.

- Corpus concordances from a electronic study of German nominalization patterns.

- New experimental data on the acquisition of structures involving ditransitive verbs in two East Slavic languages: Russian and Ukrainian.

- Data are already shared with colleagues and journal editors prior to publication via the feature Private URL.

- In-depth file descriptions required to ensure comprehensibility.

- Persistent file formats required to ensure accessibility.

- In 2017, TROLLing will apply for the Data Seal of Approval, which will declare the archive’s trustworthiness within the scientific community.

Visibility and retrieval
- Each dataset is automatically given a persistent identifier (DOI).

- Cross-referencing. The research paper contains a citation to the TROLLing post, and the research paper citation is part of the TROLLing metadata.

- TROLLing being part of a global open network, data are visible in DataCite, and will be included in other major research engines (Google Scholar, library indexes).

- TROLLing is registered in repository indexes, e.g. vrebca.org which improves the chances of potential users discovering it.

- Attribution

- Sharing/reusing data according to best practice demands clear rules and guidelines, including an international license attached to the data.

- In the case of reuse, others must refer to the data in line with good academic practice.

- In TROLLing, a dataset citation string is automatically generated, based on the registered metadata.

Areas of use and reuse
- TROLLing data are used in a wide range of publications, e.g. journal articles, edited book chapters, conference proceedings, and PhD and Master’s theses.

- TROLLing data are already reused to generate new findings, e.g. in another person’s doctoral work, where the original method has been extended and applied to new data.

- TROLLing data are further used in educational settings, to illustrate the language phenomenon under discussion.

Version control
- Datasets can be updated at any time by the contributors themselves.

- When a new version is published, the dataset citation string is automatically updated to include the new version number.

- Old versions remain accessible.

- All updates are documented in facilitate version comparisons.

Citation Metadata:
- Keyword: B Changed: C (Additional Citation Metadata: A Changed: D View Details)

Files (published): 1

Revisions: 2

This is the first published version.

TROLLing archive: opendata.uit.no/dataverse/trolling

Blog: trolling@ub.uit.no