The Tromsø Repository of Language and Linguistics

Scope and operation of an open repository for linguistic data

MØTE OM NORSK SPRÅK (MONS) 17, 22-24 NOVEMBER 2017, OS, BERGEN

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 and 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.

(From Open Access: Open Science, Language, LSA)

Tromsø in the forefront by joining forces

- Initiative taken by the internationally oriented UiT linguistics community.
- The University Library a natural partner, having a long tradition with open access archiving and publishing, and open access infrastructure.
- TROLLing developed in less than a year (March-June 2014), by a working group consisting of researchers in Linguistics, and specialists, within 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 and practiced.

Encouraging the researchers to publish their data: measures taken

- To motivate, a consciousness-raising and support are key elements in our work. TROLLing is in contact with the UiT linguistics 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.

TROLLing in numbers

- 60 published datasets
- 560 files
- 3503 downloads of entire datasets or single files

Areas of use and reuse

- TROLLing data are used in a wide range of publications: 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.

The whos and whats of TROLLing

- Available to linguistic world 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 compilation from a digital study of German nominalization patterns.
- Towed recordings were produced for the Afinj project as part of visual illustrations for students learning Norwegian Sign Language.
- … new experimental data on the acquisition of structures involving deictive verbs in two East Slavic languages, Russian and Ukrainian.

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 Datalife, and will be included in other major search engines (Google Scholar, Library indexes).
- TROLLing is registered in repository indexes, e.g. re3data.org, which improves the chances of potential users discovering the data.

Attribution

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

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 to facilitate between-version comparisons.

Structure

- Built on Dataverse, an open source platform from Harvard University.
- User-friendly interface, with metadata templates based on international standards.
- Linking to related publications, or raw or primary data stored elsewhere, possible through metadata registration.
- Data can be shared with colleagues and journal editors prior to publication via the feature Private URL.
- In-depth field descriptions required to ensure comprehensibility.
- Persistent file formats required to ensure accessibility.
- In 2018, TROLLing will apply for the CoreTrustSeal, which will warrant the archives’ trustworthiness within the scientific community.

The Tromsø Repository of Language and Linguistics

Content

- Languages: Chinese, Czech, Early New High German, English, French, German, Icelandic, Korean, Latin, Latvian, Middle Russian, Nias, North Saami, Norwegian, Old Church Slavonic, Old East Slavic, Romanian, Russian, Spanish, Ukrainian.
- Fields of study: Phonetics, phonology, morphology, syntax, semantics, lexicology, synchronic, diachronic, acquisition, language technology, sociolinguistics.
- Types of archived items: Datasets, charts, audio, video, experimental stimuli, code for linguistic or statistical analysis (e.g. R scripts).

Download a complete dataset citation template (PDF)