RAVNETRYKK

UiT goes open: Et festlig skrift til Stein Høydalsvik

The Tromsø Recommendations for Citation of Research Data in Linguistics

Collaboration illustrated

Helene N. Andreassen

Abstract

The Tromsø Recommendations for Citation of Research Data in Linguistics were published in 2019, with a twofold objective: To provide a guide on how to cite research data in linguistics according to good practices, and to contribute to making linguistics a more transparent science. This paper presents the rationale behind the recommendations as well as the development process. The goal is to demonstrate how collaboration, engagement, and different types of expertise are crucial factors to progress towards a culture of sharing of knowledge.

Introduction[1]

The Tromsø Recommendations for Citation of Research Data in Linguistics is a 15-page long document, primarily made up of templates, definitions, and annotated examples. In short: It's a document with a lot of details. Too many details? Not if the objective is to have something judged relevant and useful by all empirical linguists. It should meet the needs of the neurolinguist, the acquisitionist, the variationist, and the language documentarist, to mention but a few linguistic subfields. It should also cover all types of data collected or generated within the field, such as video and audio recordings, transcriptions, glossed text, annotations, experimental data, and introspection.

The short-term goal of the Tromsø Recommendations (Andreassen et al., 2019b) is to provide linguistic researchers, academic publishers, and data repositories with a guide on how to cite research data according to good practices. The long-term goal is to contribute to making linguistics a more transparent science, with a scholarly community adhering to a culture of sharing knowledge.

---

[1] This paper is based on Andreassen et al. (2019a). Thanks to Andrea Berez-Kroeker, Aysa Ekanger and Philipp Conzett for comments on an earlier version of it.

https://doi.org/10.7557/15.5509

This paper tells the story of the Recommendations.

Rationale

Linguistics can be defined as a data-driven social science in which scholars use observations from language use to draw inferences about cognition and social structure. Typical primary data that underpin linguistic analyses are records of language, such as audio recordings, textual productions, judgment data, and eye tracking data, and annotations of these records, such as phonetic transcriptions, frequency counts, acceptability rates, and reaction times. Despite the crucial importance of data in linguistic research, scholars too often fail to cite them properly (Berez-Kroeker et al., 2017; Gawne et al., 2017). This reduces the transparency of the research and thereby also its reproducibility.

Concerns about the use of data in linguistics publications were mentioned already in 1994 by the editor of *Language*, Sarah G. Thomason.

> Because of the traditionally high standards of *Language* regarding linguistic data, I have tried to identify cases where I may need to pay special attention to the accuracy of data: cases where the referees found problems with the data, where the data seems to be incompletely attested, or where a spot check reveals errors. When I began my term as editor, I expected that there would be cases of this kind from time to time. I did not expect that these cases would occur frequently — so frequently, in fact, that the assumption that the data in accepted papers is reliable began to look questionable. (Thomason, 1994, p. 409)

Thomason wrote the editorial note before internet, at a time when there was no infrastructure that could link publications and data, nor any online repositories where data could be archived. As a consequence, publications functioned as the main window to the empirical evidence. Today, technological solutions are in place to allow easy access to research data, but for linguistics, it is not just about improving the transparency and reproducibility of research. It so happens that linguistic data are precious on many levels, also outside the scholarly community. They capture different world views, they capture cultures at given moments as well as their contact over time with each other, and they capture cognitive capacities and variation across language users. This being said, although we now have the tools to create vast amounts of valuable linguistic data, their full potential cannot be explored unless we archive and cite them properly.

Joining forces

The Tromsø Recommendations have their origin in a network that came into existence in 2015 through the project Developing Standards for Data Citation and Attribution for Reproducible

Research in Linguistics (n.d.), funded by the National Science Foundation (NSF). The project invited more than 40 participants from North America and Europe to three multi-day workshops – researchers, archivists, research data management (RDM) specialists, scholarly publishing specialists, institutional management, and funding agencies, with the goal to reveal challenges and possible solutions to data citation and attribution practices. The project culminated in a position statement about reproducibility in linguistics (Berez-Kroeker et al., 2018b), published open access in *Linguistics*[2], and an application to the Research Data Alliance (RDA) to endorse an interest group for linguistic data. The Linguistics Data Interest Group (LDIG) was established in 2017, co-chaired by scholars from three different continents: Helene N. Andreassen from UiT The Arctic University of Norway, PhD in phonology, curator of the Tromsø Repository of Language and Linguistics (TROLLing, n.d.), and responsible for the institutional RDM training programme, Andrea Berez-Kroeker from the University of Hawai'i at Mānoa, associate professor of linguistics and specialist in language documentation and linguistic data management, and Lauren Gawne from La Trobe University, postdoctoral researcher in linguistics with an interest in critical approaches to methodology and RDM.[3]

As stated in the group charter, the overarching objective of LDIG is to "contribute to a positive culture of linguistic data management and transparency in ways that are in keeping with what is happening in the larger digital data management community", with focus on three main topics (taken from Linguistics Data Interest Group, 2017):

- Development and adoption of common principles and guidelines for data citation and attribution by professional organizations, academic publishers, and repositories for language and linguistics.
- Education and outreach efforts to make linguists more aware of the principles of reproducible research and the value of data creation methodology, curation, management, sharing, citation and attribution.
- Greater attribution of linguistic data set preparation within the linguistics profession.

Ever since the beginning, the intention of LDIG has been to function as a scholar-led, community-based project which draws on different members' expertise, experience, and local networks. This way, LDIG hopes to discover the challenges among linguists and meet their needs more efficiently, and also to evoke engagement and a sense of

[2] On 6 December 2018, the position statement was the most downloaded article of the journal (Andrea Berez-Kroeker, p.c.).

[3] Fun fact: In addition to discussions about linguistic data, the LDIG co-chairs learned a lot about collaboration across very different time zones.

commitment among the community members, which today add up to more than hundred people coming from different subfields of linguistics. Another intention has been to work within the frame of RDM specialist communities, to make sure that all outputs from LDIG are up-to-date and in line with good practices. Two examples in this regard are the overarching Research Data Alliance, with its more than 10 000 members, complex thematic community structure, and biannual plenary meetings, and the TROLLing repository, built on the open-source Dataverse platform, with its local operating group consisting of specialists in linguistics, open access, and system development.

Development of the recommendations

It is common practice among teachers to try to evoke engagement among students by teaching them the *why*s before the *how*s. If students understand the purpose of a task, beneficial to themselves and/or to society in general, it may be easier to tackle any challenging, time-consuming, or boring operation required to succeed. Professionals who plan to handle something new are not necessarily very different and may also need internal or external motivation to become engaged in the learning task. If we focus on citation of research data, an obvious external motivation comes from scientific publishers who increasingly require data underpinning research publications be available to the readers.[4] And how do we inform readers about available data? We archive them and cite them in our publication. Internal motivations undoubtedly vary, but everything suggests that more and more scholars become aware of the importance of research transparency and want to do things right.

There are already many documents on the web authored by competent, trendsetting organizations working on open science, available to scholars who want to learn about RDM. One oft-cited example is the FORCE11 Joint Declaration of Data Citation Principles (2014), a set of principles that "cover purpose, function, and attributes of citations". With this landscape as a starting point, the first task of LDIG was to create a document that could speed up the learning process in the linguistic community, an inspirational document on data citation that would speak to all scholars irrespective of their level of RDM skills and competencies. After several asynchronous meetings in the LDIG community, where all members were invited to answer questions and comment on draft versions, the Austin Principles of Data Citation in Linguistics were published in 2018. Based on the content and structure of the FORCE11 Principles, the Austin Principles (Berez-Kroeker et al., 2018a) were formulated with the goal to raise awareness among

[4] In some cases, data may not be shared because of ethical, legal, commercial, or security reasons. In many of these cases, some metadata can still be shared and as such demonstrate the existence of the data.

linguists and encourage them to make informed decisions regarding the accessibility and transparency of their research data. Information about the principles was disseminated rather widely, in local networks as well as on LINGUIST List (n.d.), and people were invited to endorse the principles on a dedicated website. Presently, more than 100 individuals, as well as 10 organizations, have officially stated that they endorse the Austin Principles and that they "support the idea that the data on which linguistic analyses are based are of fundamental importance to the field, and should be treated as such" (Berez-Kroeker et al., 2018a).

In order to help interested scholars work in line with the Austin Principles, the second task of LDIG was to create a document that could serve as a practical guide to citation of linguistic data. One could rightly ask why linguistics would need a separate guide, but as mentioned by the much respected DataCite (n.d.) on their webpage, next to their recommended citation format, different disciplines may come with different challenges. In late 2017, LDIG mounted a working group dedicated to the development of a citation guide for linguistics publications. In addition to the LDIG co-chairs Andreassen and Berez-Kroeker, the group consisted of Philipp Conzett from UiT The Arctic University of Norway, curator of TROLLing with a background in Nordic linguistics, and active in several European RDM projects, and Koenraad De Smedt from the University of Bergen, professor of computational linguistics and coordinator of CLARINO (n.d.).

In a first phase, all LDIG community members were invited to an asynchronous meeting to reflect on metadata and citation practices in linguistics (see Andreassen et al., 2018). Simultaneously, the working group started gathering information about existing citation initiatives within and outside the discipline, in order to identify which citation templates to build on. The group also collected information about metadata and citation practices in repositories for linguistic data indexed in the repository registries re3data (n.d.) and OLAC (n.d.) – a very useful task as it revealed a handful of important challenges for potential reusers of archived data. For instance, the metadata didn't always clearly specify who to cite as authors of the data. Also, in many cases, only one date was entered in the metadata, which because of lack of description of the metadata field, could be interpreted as either the recording date, the deposit date, the publication date, or the last updated date.

In a second phase, three more members joined the LDIG working group, adding more perspectives, competencies, and manpower to the citation project: Lauren Collister from the University of Pittsburgh, director of scholarly publishing with a PhD in sociolinguistics, and specialist in open access and copyright, Christopher Cox from Carleton University, assistant professor of linguistics and much involved in community-based language work, and Bradley McDonnell from the University of Hawai'i at Mānoa, assistant professor of linguistics and specialist in language

documentation. The group continued the work on the citation guide with four audiences in mind: i) academic publishers, who could add or adopt the document into their author guidelines, ii) data repositories, who could check and if needed adjust their metadata templates so as to make archived data properly citable, iii) researchers using data in their work, who could refer to the citation guide in case the author guidelines didn't specify how to cite data, and iv) researchers planning to collect and archive data, who could use the document to determine which metadata to prepare in order to make their data properly citable.

There was quite some discussion in the working group about the level of detail needed for the different audiences, and during the winter 2019, two drafts saw the light, one condensed with only key elements, and one lengthy with more examples and explanations. The group convened (with two members participating via Skype) in Philadelphia in April 2019, on the occasion of the 13th RDA plenary meeting.[5] Not only were people happy to meet and chat in person, but they finally had the chance to sit together and focus. At this point, if some of the readers of the present paper are still unsure whether linguists really need a discipline-specific citation guide: The working group spent three hours – 3 hours – discussing challenges related to the Author and Date fields in the citation template. For each element of the template, no stone was to be left unturned and potentially cause problems for scholars in the future.

Between April and November 2019, the condensed version of the citation guide was sent out for comments twice in the LDIG community, as always with lengthy and fruitful feedback in return. Using the collaborative Google Drive platform, people had access to the same document and could discuss via the Comments function. The citation guide was also sent out for comments to a list of selected linguistic data experts, journal editors, and leaders in the field, who were in a position to encourage adoption or endorsement of the document for their organization or journal after its publication.

In November 2019, a few members of the group convened (one via Skype) in Tromsø on the occasion of the 14th Munin Conference on Scholarly Publishing. All comments from the final feedback round were discussed and incorporated, and after one day of intense work, the first version was ready. Named after the city where they had been finished, the Tromsø Recommendations for Citation of Research Data in Linguistics were shipped off to the Research Data Alliance for endorsement and publication as an official RDA Supporting Output.

As for the lengthy version of the citation guide mentioned previously, this ended up serving as input to Conzett and De Smedt (to appear),

[5] A big thanks to the Linguistic Data Consortium, who generously offered to host the LDIG meeting in their offices.

a chapter on data citation to be published by MIT Press Open as part of a handbook on linguistic data management.

Lessons learned

Ever since the creation of LDIG, core members of the community have continuously worked to raise awareness among linguists, via informal discussions, emails, short presentations at conferences, RDM teaching sessions, training workshops, and summer schools. We do not know the short-term or long-term effects of our efforts, but at least, here are some lessons learned:

- Like researchers in many other disciplines, linguists experience barriers to data citation, such as the lack of awareness, training, standards, and incentives.
- It is important to involve people from different parts of the scholarly community, in order to identify practices and challenges, to get feedback on ongoing work, and eventually to implement good practices in the research and publication process.
- The world is a busy place, which makes it challenging to fully engage people in different sectors. For many researchers, moving from good intentions to practice takes time. For many academic publishers, other aspects of the publishing process are considered more pressing. For many repositories, good practices for data citation are not contained in the metadata and documentation guidelines.
- Continuous outreach seems to move things (slowly) forward, but concrete outputs, such as the Austin Principles and the Tromsø Recommendations, are key. Also, outreach must happen in the right context, with enough time for presentation and Q&A. Finally, getting the right people on board, decision-makers or trend-setters in the community, is very useful for planning ahead.

Next step

Today, the world has focus on dealing with the coronavirus disease and rapid sharing of knowledge across research institutions world-wide is essential. When the world at some point returns to a more normal state, we may hope that junior as well as senior researchers have gained increased awareness of the importance of research transparency and openness. The Tromsø Recommendations will perhaps not directly contribute to save lives, but they may nevertheless function as an important tool for linguists who wish to carry out transparent and open research.

LDIG therefore now enters a new phase, where education of young researchers and outreach to different sectors of the linguistic community are put in the center of attention. We will simultaneously continue to work on specific topics within RDM, such as metadata and archiving, in order to ensure that the Austin Principles, the

Tromsø Recommendations, and our tips and advice in general are continuously up-to-date and in line with good practices.

Concluding remarks

This paper has told the story of the Tromsø Recommendations for Citation of Research Data in Linguistics, a product that has strongly benefited from the engagement, experiences, skills and competencies in the LDIG community and its associated networks. I hope this paper may encourage practitioners in other fields to initiate similar advancements, if possible within the frame of RDA. I also hope it may inspire decision-makers and publishers to actively collaborate with and support scholar-led initiatives working toward better research practices.

Acknowledgements

Many thanks to the participants in the Data Citation and Attribution project, the Data Science for All of Linguistics project, members of the RDA LDIG, and attendees at our previous workshops, courses and presentations for fruitful discussion.

This material is based upon work supported by the National Science Foundation under Grants No. 1447886 and 1745349. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Andreassen, H. N., Conzett, P., De Smedt, K., Berez-Kroeker, A. & Gawne, L. (2018). *Data citation and metadata standards in linguistics.* Paper presented at the LDIG working session during the RDA 11th Plenary Meeting, 21–23 March 2018, Berlin, Germany. Retrieved from https://hdl.handle.net/10037/16556

Andreassen, H. N., Berez-Kroeker, A., Collister, L., Conzett, P., Cox, C., De Smedt, K., Gawne, L. & McDonnell, B. (2019a). *Data citation in linguistics publications: A scholar-led, community-based initiative.* Paper presented at the 14th Munin Conference on Scholarly Publication, 27–28 November 2019, Tromsø, Norway. https://doi.org/10.7557/5.4876

Andreassen, H. N., Berez-Kroeker, A. L., Collister, L., Conzett, P., Cox, C., De Smedt, K., McDonnell, B. & Research Data Alliance Linguistics Data Interest Group. (2019b). *Tromsø Recommendations for Citation of Research Data in Linguistics*. https://doi.org/10.15497/rda00040

Berez-Kroeker, A. L., Gawne, L., Kelly, B. F. & Heston, T. (2017). *A survey of current reproducibility practices in linguistics journals, 2003–2012*. Retrieved from https://sites.google.com/a/hawaii.edu/data-citation/survey

Berez-Kroeker, A. L., Andreassen, H. N., Gawne, L., Holton, G. Kung, S. S., Pulsifer, P. Collister, L. B., The Data Citation and

Attribution in Linguistics Group & the Linguistics Data Interest Group. (2018a). *The Austin Principles of Data Citation in Linguistics, version 1.0.* Retrieved from http://site.uit.no/linguisticsdatacitation/austinprinciples

Berez-Kroeker, A. L., Gawne, L., Kung, S. S., Kelly, B. F., Heston, T., Holton, G., Pulsifer, P., Beaver, D. I., Chelliah, S., Dubinsky, S., Meier, R. P., Thieberger, N., Rice, K. & Woodbury, A. C. (2018b). Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics 56(1)*, 1–18. https://doi.org/10.1515/ling-2017-0032

CLARINO: Common Language Resources and Technology Infrastructure Norway. (n.d.). Retrieved 18.05.2020 from https://clarin.w.uib.no/

Conzett, P. & De Smedt, K. (to appear). Guidance for citing research data. In A. Berez-Kroeker, B. McDonnell, E. Coller & L. Collister (Eds.), *The Open Handbook of Linguistic Data Management.* MIT Press Open.

Data Citation Synthesis Group, Martone, M. (Ed.) (2014). *Joint Declaration of Data Citation*. San Diego, CA: FORCE11. https://doi.org/10.25490/a97f-egyk

DataCite. (n.d.). *DataCite – Cite your data.* Retrieved 27.04.2020 from https://datacite.org/cite-your-data.html

Developing Standards for Data Citation and Attribution for Reproducible Research in Linguistics. (n.d.). Retrieved 27.04.2020 from https://sites.google.com/a/hawaii.edu/data-citation/

Gawne, L., Kelly, B. F., Berez-Kroeker, A. L. & Heston, T (2017). Putting practice into words: The state of data and methods transparency in grammatical descriptions. *Language Documentation & Conservation 11*, 157–189. Retrieved from http://hdl.handle.net/10125/24731

Linguistics Data Interest Group. (2017). *Linguistics Data Interest Group charter statement*. Retrieved from https://www.rd-alliance.org/groups/linguistics-data-ig

LINGUIST List. (n.d.). Retrieved 19.05.2020 from https://linguistlist.org/

OLAC: Open Language Archive Community. (n.d.). Retrieved 27.04.2020 from http://www.language-archives.org/

Re3data: Registry of Research Data Repositories. (n.d.). Retrieved 27.04.2020 from https://www.re3data.org/

Thomason, S. G. (1994). The Editor's Department. *Language 70(2)*, 409–413. Retrieved from https://www.jstor.org/stable/415877

TROLLing: The Tromsø Repository of Language and Linguistics. (n.d.). Retrieved 18.05.2020 from https://trolling.uit.no/