Linguistics Data Interest Group (LDIG)

The Tromsø recommendations for citation of research data in linguistics

Draft 3

26 November 2019

Helene N. Andreassen, Andrea L Berez-Kroeker, Lauren Collister, Philipp Conzett, Christopher Cox, Koenraad De Smedt, Bradley McDonnell,
and the Research Data Alliance Linguistic Data Interest Group

-- document commenting closed --

This is the third draft of the recommendations, which reflects edits made after feedback from the LDIG community and other linguistic data experts in late 2019. We invite everybody receiving the link to this document to read and comment.

Table of contents

Executive Summary

Language datasets are often not cited, or cited imprecisely, because of confusion surrounding the proper methods for citing them. For the use of researchers and scholars in the field working with datasets, we propose the following components of data citation for referencing language data, both in the bibliography and in the text of linguistics publications. As each journal may have its own stylistic conventions, we do not address specific formats or citation styles, but rather elements of citations; however, for journals or repositories seeking to update their data citation guidance, we hope this document will be helpful. Furthermore, these recommendations are intended to be only guidelines, as we cannot account for every possibility here. This guidance is based on the Austin Principles, the FORCE11 and Research Data Alliance Joint Declaration of Data Citation Principles, and the Reproducible Research in Linguistics position statement.

The template for a **minimal bibliographic reference** (i.e. in the bibliography section of a piece of academic writing) to a dataset resource is:
Author, Date, Title, Publisher, Locator.

The template for an **expanded bibliographic reference** to a dataset resource, including *conditional elements* (i.e. required in certain cases depending on resource characteristics) is:
Author, *Other Attribution (Roles),* **Date, Title, Publisher, Locator,** *Version, Date accessed, Tag.*

In-text (or in-line) citations must point to a bibliographic reference in the bibliography section of the published work. The template for a **minimal in-text citation** is:
Author, Date

The template for an **expanded in-text citation** including additional potential information is:
Author, Date, *Locator, Subset, Other Attribution (Roles)*

Please note: Definitions of the elements contained in the bibliographic reference and the in-text citation can be found in the Glossary. A longer version of the recommendations, explaining concepts, highlighting challenges and providing examples can be found in:

> Conzett, Philipp & Koenraad De Smedt. (in preparation). Guidance for citing research data. In Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller & Lauren B. Collister (eds.). To appear. *Open Handbook of Linguistic Data Management*. Cambridge, MA: MIT Press Open.

0. Introduction

0.1 Intended audience

The intended audience for these recommendations consists of i) academic publishers, ii) resource providers (e.g. repositories, archives), iii) researchers citing data, and iv) researchers making data management plans, developing and depositing data, and preparing metadata.

Academic publishers will have the opportunity to add these recommendations to their author guidelines for citation. Resource providers will learn which metadata and citation elements are crucial in order for data to be properly citable. Researchers using data in their publication will know how to cite these data in case publisher guidelines are underspecified. Researchers depositing data in a repository will know more about what metadata they should provide in order to publish their data so as to make them citable.

0.2 General recommendations

Publishers should provide guidelines for formatting details and should aim to treat data publications in a similar way to other publications as regards those aspects which they have in common. For instance, the list of contributors should be abbreviated the same way in in-text citations (with the same limit on number of names) as for other publications. The order of elements, use of initials for first names, and other formatting are publisher-dependent. The way Persistent Identifiers (PIDs) and other elements are written may also be publisher-dependent.

Resource providers might require citation of metadata elements not listed here or listed only as conditional. For instance, some resource providers include an indicator of data fixity in the citation information, e.g. Universal Numerical Fingerprint (UNF). When a recommended citation is given by a resource provider, some adjustment of formatting may be necessary to conform to stylesheets or publisher requirements; however, we advise to make every effort to include all of the same information in the citation as included in the resource provider's recommended citation.

Some resource providers might require citation of a written publication related to the resource. In this case, the written publication should be cited in addition to citing the resource itself.

This document is **not** aimed at promoting a best practice data **publication** model, only a best practice **citation** model given that data are published or made available in some way.

1. References in the bibliography

This section describes how to create full references to dataset resources for inclusion in the bibliography (or references) section of a piece of academic writing. Section 1.1 discusses whether one should cite a full dataset at the highest level or organization, or a component of the dataset. Section 1.2 provides templates for creating a reference to data in a bibliography: a minimal template containing required elements, and an expanded template also containing conditionally-required elements. Examples of references, with commentary, are given in Section 1.3.

The recommendations are kept as analogous as possible with recommendations for citation of other, more traditional, types of publications. For elements of the citation that are specific to data, the recommendations below are based upon: Data Citation Synthesis Group, Martone Maryann (ed.). 2014. *Joint Declaration of Data Citation Principles.* San Diego CA: FORCE11. https://doi.org/10.25490/a97f-egyk.

1.1 What to cite in the bibliography

Sometimes it is desirable to provide a reference to an entire resource that may be comprised of numerous components (e.g. files or folders), while at other times it is desirable to provide separate citations to the individual components that were used. The choice of what to cite in the bibliography depends both on the structure of the resource as well as whether it is overly cumbersome to cite numerous individual components. Our recommendation is that if different components of a resource have different **Authors**, provide separate entries for each of them, when feasible, in order to credit those authors properly. Otherwise, include only one reference for the entire resource, or the highest level used.

1.2 Templates for references

In this section we present two templates for citation of dataset resources in the bibliography. All elements of the templates are defined in the Glossary. Elements coded in **green and bold** are considered minimal (i.e. required), while elements coded in *purple and italics* are considered to be conditional. Conditional elements are those that may be included based on either the characteristics of the resource (e.g. references to versioned datasets should include the version number), or on subfield-specific traditions (e.g. in language documentation, it is common to acknowledge the contributions of language consultants by name).

The template for a **minimal reference to a dataset resource in the bibliography** section of a piece of academic writing is:
Author, Date, Title, Publisher, Locator.

The template for an **expanded bibliographic reference** to a dataset resource, including *conditional elements* is:

Author, *Other Attribution (Roles),* **Date, Title, Publisher, Locator**, *Version, Date accessed, Tag.*

1.3 Examples of references

The examples below are taken from a variety of sources, and are meant to illustrate various scenarios for citing data in the bibliography. For expository reasons, all examples are formatted here using the Glossa stylesheet. Please note that the stylesheet or the repository citation requirements may influence the order of elements in the citation, resulting in deviation from the suggested template above.

Citing a full dataset:

Example 1: A straightforward example.
This example shows a fairly straightforward citation of a dataset for a bibliography. The **Author** element is appended by an optional specific role (here, Collector); it also has a **Date** (2005), a **Title** (*Ma'anyan narratives*), two **Locators** (the repository-internal identifier AA4 and a DOI), and a **Publisher** (PARADISEC).

Adelaar, Alexander (Collector). 2005. *Ma'anyan narratives* (AA4). PARADISEC. https://doi.org/10.4225/72/56E979455A05E.

*Example 2: Citing a resource with an **Author** that is not a person.*
This example shows three references to resources for which the **Author** is not a person. In all three of these, the **Author** element is the organization responsible for developing the resource and making it available (see the Glossary for more information on selecting **Authors**). Note that in the second citation below, INESS is both the **Author** and the **Publisher**.

BNC Consortium. 2007. *British National Corpus, version 3 (BNC XML Edition)*. Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium. http://www.natcorp.ox.ac.uk/. (Accessed 2019-06-22).

Infrastructure for the Exploration of Syntax and Semantics (INESS). 2016. NorGram Newspaper text (30 documents from the years 2006 - 2009) in Norwegian Bokmål from the Norwegian Newspaper Corpus. In *NorGramBank Corpus*. INESS Portal. https://hdl.handle.net/11495/DB24-E30D-55EA-1. (Accessed 2019-04-01).

Princeton University. 2011. *WordNet 3.1*. Princeton University. https://wordnet.princeton.edu/. (Accessed 2019-07-11).

*Example 3: Citing resources with special **Dates**.*
This example shows citations that have special **Dates**. The Hauk dataset was published over a period of three years. The Prignitz dataset does not have a publication or deposit date, so the collection date is therefore used and marked as such.

Hauk, Bryn. 2016–2018. *Batsbi (Tsova-Tush)*. Kaipuleohone University of Hawai'i Digital Language Archive. http://hdl.handle.net/10125/42581. (Accessed 2019-06-21).

Prignitz, Gisèle. 2007 (collection date). *Enquête Burkina Faso.* Projet PFC. https://public.projet-pfc.net/. (Accessed 2019-06-22).

Example 4: Different kinds of **Locators***.*
In the Prignitz dataset from Example 3, the URL to the landing page of the main collection is used as the **Locator**. The Adelaar dataset from Example 1 above has both a PID (in this case, a DOI) *and* a repository-internal identifier (AA4); both are used to aid locating the resource. For the Mæhlum dataset below, which is published on a physical CD audio, media is specified as a **Locator**. The Ferrara and Ringsø dataset is unpublished, which is indicated here in the media specification **Locator**.

Mæhlum, Brit. 1998. *Dialektprøver fra Målselv og Bardu*. Målselv mållag. CD audio.

Ferrara, Lindsay & Torill Ringsø. 2017–2018 (collection date). *Depicting perspective in Norwegian Sign Language.* Norwegian University of Science and Technology. Unpublished video recordings and annotation files.

Example 5: Citing resources with and without Version and Date Accessed.
In the Kamoen et al. citation below, a version number is provided in the repository, so it is added to the citation.

Kamoen, Naomi, Bregje Holleman, Pim Mak, Ted Sanders & Huub Van den Bergh. 2017. *Why are negative questions difficult to answer? On the processing of linguistic contrasts in surveys*. DataverseNL. https://hdl.handle.net/10411/20857. V5.0.

However, not all resources will list a version number. The Hauk dataset in Example 3 has no version number, but the two-year period of publication indicates that the collection may have changed somewhat during that period; in cases like this it is good practice to add the date the resource was accessed to the citation. Similarly, the INESS dataset in Example 2 is a dynamic dataset, meaning it changes frequently, so the date accessed should be added to the citation.

Citing a component of a dataset:

Example 6: Variations of the element ***Author****.*
This example shows several citations to components of a dataset with variations of the element **Author**, which may not always refer to a single person in the traditional sense of the term "author." The Hauk et al. item and the Krauss et al. item are listed with several persons and their roles on the respective landing pages. With no reference recommended by the resource provider, we include all persons in the **Author** field, and also their roles. In these specific cases,

Hauk and Krauss most closely fit the definition of **Author** (i.e., the person most responsible for developing the resource; see the Glossary), so we use their names first. The UCLA Phonetics Lab Archive component does not have any named responsible person in the metadata, so we therefore put the archive in the **Author** field.

Hauk, Bryn (Researcher, Depositor), Omar P'ap'ashvili (Speaker) & Rezo Orbetishvili (Consultant). 2018. BH2-074. In *Batsbi (Tsova-Tush).* Kaipuleohone University of Hawaii Digital Language Archive. http://hdl.handle.net/10125/58935.

Krauss, Michael E. (Interviewer), Jeff Leer (Interviewer) & Anna Nelson Harry (Speaker). 1975. Interview with Anna Nelson Harry (ANLC0082). In *Krauss Eyak Recordings*. Alaska Native Language Archive. https://www.uaf.edu/anla/.

UCLA Phonetics Lab Archive. 2007. gle_word-list_1975_01. In *Gaelic, Irish.* http://archive.phonetics.ucla.edu/.

Example 7: Citing components with and without Date accessed.
The ISWOC component cited below contains dynamic data, so it requires a *Date accessed*.

Information Structure and Word Order Change in Germanic and Romance Languages (ISWOC). 2016. *West-Saxon Gospels.* INESS Portal. http://hdl.handle.net/11495/DB24-D542-3616-6. (Accessed 2019-06-22).

The reference to the entire Hauk dataset in Example 3 above included a *Date accessed* because the entire dataset was published over a two-year period. If we can assume that a single component from the full dataset is stable, then we do not need to provide a *Date accessed* when citing this component, as in Example 6 above. When in doubt about the stability of a component, include the *Date accessed.*

Example 8: Different kinds of **Locators**.
This example shows two citations with different kinds of locators. The Duke University component cited below is analog (paper) and the reference points to a specific archival box in a library for retrieval of the component; a URL for the library can also aid retrieval. The Andrade Santos item has a DOI pointing directly to the item cited, and needs no further specification.

Duke University Committee on African Studies. 1973-1976. Africa Sketches, 1973-1976 (Box 1). In *Committee on African Studies records, 1965-1976.* David M. Rubenstein Rare Book & Manuscript Library. Correspondence and sample illustrations. https://library.duke.edu/rubenstein/.

Andrade Santos, Cássio. 2018. Example_Stanza_pata.WAV. In *Singing and speech in Brazilian Portuguese: stressed, pre-stressed and post-stressed vowels*. DataverseNO. https://doi.org/10.18710/3PKATY/GQAJRB. V1.

2. In-text citations

This section describes how to create in-text (or in-line) citations in the body of a piece of academic writing (e.g. numbered examples in a linguistics article). Section 2.1 provides templates for creating in-text citations: a minimal template containing required elements, and an expanded template also containing conditionally-required elements. Examples of in-text citations, with commentary, are given in Section 2.2.

2.1 Templates for in-text citations

In-text (or in-line) citations must point to a bibliographic reference at the end of the published work. Thus, an in-text citation to data can minimally include elements required by the stylesheet. There may be field-specific conventions that apply which may differ significantly from these instructions.

The recommendations are kept as analogous as possible with recommendations for in-text citation of other, more traditional, types of publications (e.g. the Author:Year format). If your publication outlet uses a numerical citation style, use footnotes or endnotes to provide the additional information (depending on granularity). Footnotes are also recommended for long PIDs when it is necessary to refer to specific resource items, folders, sections etc.

In the absence of explicit instructions, the minimal template is:
Author, Date

Additional specificity may be indicated in in-text citations using conditional information, e.g.:

- *Locator,* like a PID pointing directly to an individual item in a resource, or URL or item name if PID is unavailable.
- *Subset*, like an individual file or files within a larger dataset, or timestamps or line numbers indicating parts of a file.
- *Other Attribution (Roles)*, like the name of the person who uttered or signed the example cited).

The expanded template including conditional information:
Author, Date, *Locator, Subset, Other Attribution (Roles)*

2.2 Examples of in-text citations

The examples in this section are meant to illustrate various types of conditional information that may be necessary for in-text citations. Again, the nature of the resource will dictate what information is necessary.

Example 9: Citation of a recording in an archived collection of recordings.
The first example presents in-text citations to the collection of materials by Hauk shown in Example 3 above.

A minimal in-text citation would appear as:

> (Hauk 2018)

It may be necessary or desirable to include more granular information. When the bibliographic reference is to the whole collection, it may be desirable to create the in-text citation by referencing a portion of the collection using a *Locator*. The *Locator* could be a component of the collection or a PID. An in-text citation to a particular item (in this case, a folder containing an audio file and a text transcription file in the collection) could be either of the following; the first refers to a particular item by its title, and the second refers to an item by its PID:

> (Hauk 2018: BH2-076)
> (Hauk 2018: http://hdl.handle.net/10125/58937)

It may be useful to refer to a particular *Subset* by its timestamp. Unless otherwise specified, we recommend adding a time range based on ISO-8601 time codes [hh]:[mm]:[ss] format.

> (Hauk 2018: BH2-076, 00:00:01–00:00:03)

If relevant, one may add *Other Attribution* to include a particular person and a *Role* in the in-text citation.

> (Hauk 2018: BH2-081, 00:00:01–00:00:03, Rezo Orbetishvili (Speaker))

Example 10: Citation of corpus available through an online interface.
The second example presents in-text citations to the Corpus of Regional African American Language (CORAAL), which provides an online interface to a corpus of recordings and aligned transcripts.

The CORAAL website indicates the following reference for the bibliography:

> Kendall, Tyler & Charlie Farrington. 2018. *The Corpus of Regional African American Language*. Version 2018.10.06. Eugene, OR: The Online Resources for African American Language Project. http://oraal.uoregon.edu/coraal.

A minimal in-text citation for CORAAL would be

> (Kendall & Farrington 2018)

More granularity may be desired. For example, to reference a single recording, the title of the recording can be appended as a *Locator*:

> (Kendall & Farrington 2018: DCA_se1_ag1_m_04_1)

Further granularity is also possible, for example, to a *Subset* like a timestamp or line numbers:

> (Kendall & Farrington 2018: DCA_se1_ag1_m_04_1, 00:00:21.1564-00:00:32.5222)
> (Kendall & Farrington 2018: DCA_se1_ag1_m_04_1, lines 18-22)

Example 11: Citation of a dynamic map created through an online database query:
In this example, we are creating a citation for a dynamic map of Balto-Slavic languages that is created by querying the Glottolog database. See Figure 1, with an in-text citation in the Figure caption.

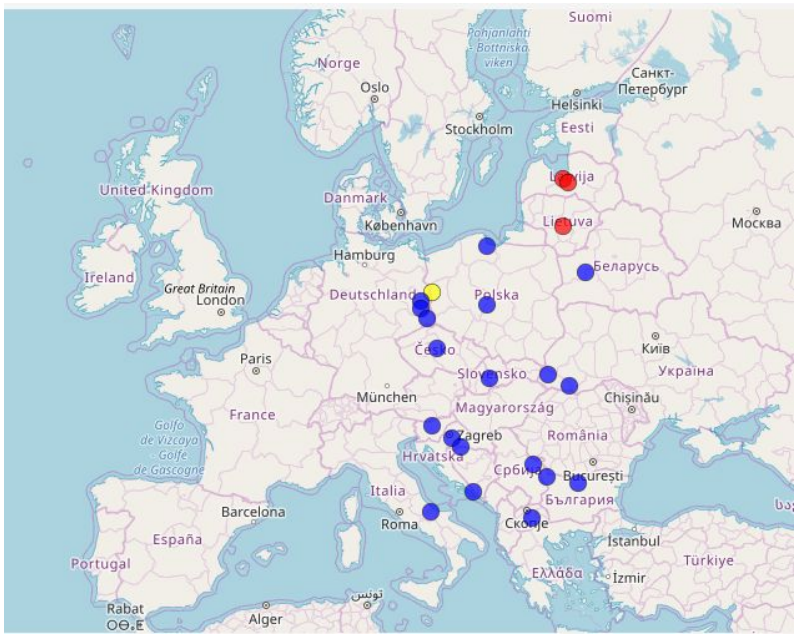

Figure 1: Map of Balto-Slavic languages (Hammarström, Forkel & Haspelmath 2019: balt1263)

Glottolog uses its repository-internal identifier balt1263 to identify the Balto-Slavic family; the identifier can serve as a *Locator*. In this case, the in-text reference points to the following citation in the bibliography, containing the URL where the repository-internal identifier found in the in-text citation could be searched to return the full map and data:

> Hammarström, Harald, Robert Forkel & Martin Haspelmath. 2019. *Glottolog 3.4.* Jena: Max Planck Institute for the Science of Human History. http://glottolog.org. (Accessed 2019-06-04). https://doi.org/10.5281/zenodo.596479.

Example 12: Citation of instances of grammaticality judgments, using line number for granularity.
This example presents an in-text citation to the following bibliographic reference, which is a series of spreadsheets of grammaticality judgments of Uzbek polar questions:

> Gribanova, Vera (Collector). 2016. *Combinatorics of the Uzbek verbal complex in polar questions 2012-2016*. http://purl.stanford.edu/bq499mh5981.

In order to properly cite a particular instance from the spreadsheet, the template **Author**, **Date**, *Locator*, *Subset* is used:

(1) *Chiroyli-mas-miz-mi?*
pretty-NEG-1PL-Q[1]
Are we not pretty?
(Gribanova 2016: -mi-inversion-dataset-2019.csv, No. 27)

Example 13: Citation of a typological survey with URL only.
This example presents an in-text citation to the following bibliographic reference, which is an online typological database with a URL for a *Locator*.

> Seifart, Frank. 2013. *AfBo: A world-wide survey of affix borrowing*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://afbo.info. (Accessed 2019-06-04).

The in-text citation refers to a particular page from the AfBo website:

> Indonesian has other influences from Dutch, for example ten Dutch derivational prefixes and three abstract-noun forming suffixes attested with non-Dutch stems (Seifart 2013, https://afbo.info/pairs/19).

3. Glossary

- **Author**: By “Author” is meant by default one or more entities (persons or organisations) responsible for developing the resource. Specific **Roles** will vary with the details of the resource and the terminology of the resource provider, but might include “Project Leader”, “Investigator”, “Researcher”, “Data Collector”, “Depositor,” "Project Contact," “Consultant,” etc. By default, only the main responsible **Authors** are listed without mentioning their roles. Other entities might be specified or required by a resource provider, publisher guidelines, or subfield norms, and should be marked by their specific

[1] The glossing has been adapted to the Leipzig Glossing Rules.

roles using parentheses, e.g. *John Smith (Data Collector).* When roles may be needed in addition to Author, we have listed these as **Other Attribution** in our templates.

For more information on contributor roles, see the following:

- https://www.casrai.org/credit.html
- Definitions in the DataCite Metadata Schema Documentation for the Publication and Citation of Research Data, Version 4.3: https://schema.datacite.org/meta/kernel-4.3/doc/DataCite-MetadataKernel_v4.3.pdf
- OLAC Role vocabulary: http://www.language-archives.org/REC/role.html

- **Date**: Date of publication (default); if not available, i.e. there is no formal publication process or it has not been completed, use the deposit date (i.e., initial date of availability, also for dynamic resources), and mark it as such using parentheses, e.g. *2018 (deposit date)*. If there is no deposit date, then use collection date (i.e. when data collection was completed or period of collection), and mark it as such, e.g. *2012-2016 (collection date).*
    - Whether **Date** is specified as the year only or a more precise date depends on publisher guidelines.
    - When the data is in the process of being collected and is still growing, the **Date** may be a range (with start and end dates) relating to the version that was used.

- **Date Accessed**: Required if the resource is dynamic, i.e. the data will change over time (e.g. treebanks that can be reparsed), or if for any other reason it is uncertain whether the cited version of the resource is stable and persistent.

- **Locator**: A Persistent Identifier (PID), sometimes also referred to as a Persistent Globally Unique Identifier (GUID) or Uniform Resource Identifier (URI), to the landing page of the resource accessed (at the collection, folder, file and/or item level as relevant, corresponding to what **Title** refers to). A PID may be a Digital Object Identifier (DOI), Handle (hdl), ARK, or other format. If there is no such identifier, mention the URL to the resource provider (repository or archive) together with the internal identifier for the resource (e.g. deposit ID), and, if only a part is referred to by the **Title**, the identifier to that part (e.g. folder or file). If no online locator exists, the **Locator** can specify the media instead (e.g. CD audio, CD-ROM text file) or analog (e.g. books, archival card files).

- **Other Attribution**: see **Author** above.

- **Publisher**: Entity responsible for providing access to the resource. In most cases this will be the name of the resource provider, e.g. the repository or archive. If possible this should be the original source, not a harvester of metadata or copier of the data.

- **Subset**: Used in in-text citations to refer to a specific portion of a resource cited in the reference list. May be a component or set of components (e.g. files). May also be a timestamp or line/row number or other indicator of granularity.

- **Tag**: A tag, e.g. “[dataset]” or “[code]” may be added to distinguish datasets from other types of publications (articles etc.).

- **Title**: Title of the resource (i.e. of the whole dataset as published). If only a well-defined part of a resource is referred to throughout the text, then that part (e.g. section, file, item, etc.) may be added to the title. If different parts of the resource are referred to throughout the text, it is preferable to specify the relevant part in each in-text citation.

- **Version**: Version number (if available and not already mentioned in the **Title**), time stamp (e.g. for nightly builds), Git commit ID, or similar. The default is that there is only one version and the resource is assumed to be stable. An alternative value is “dynamic” meaning that the resource may change without explicit versioning or time stamps; in that case, **Date accessed** is also required.

4. Contributors

In addition to input from the RDA Linguistic Data Interest Group, the following persons have contributed to the development of this document:

Helene N. Andreassen, UiT The Arctic University of Norway
Andrea L. Berez-Kroeker, University of Hawai‘i at Mānoa
Chris Cieri, Linguistic Data Consortium
Lauren Collister, University of Pittsburgh
Philipp Conzett, UiT The Arctic University of Norway
Stefano Coretta, University of Manchester
Christopher Cox, Carleton University
Koenraad De Smedt, University of Bergen
Lindsay Ferrara, Norwegian University of Science and Technology
Robert Forkel, Max Planck Institute for the Science of Human History
Susan Smythe Kung, University of Texas at Austin
Alon Lischinsky, Oxford Brookes University
Bradley McDonnell, University of Hawai‘i at Mānoa
Sebastian Nordhoff, Language Science Press
Hugh J. Paterson III, SIL International & University of North Dakota
Hiram Ring, University of Zürich
Nick Thieberger, University of Melbourne
Margaret E. Winters, Wayne State University