11 Guidance for Citing Linguistic Data

Philipp Conzett and Koenraad De Smedt

1 Introduction

Linguistic data, in their many forms, are a valuable asset in research and education on language. From the predigital age, the earliest data to reach us are written records carved in stone, wooden sticks, or clay tablets, or penned on papyrus, parchment, and such. Early field linguists recorded samples obtained from informants and other sources in notebooks and card files. Speech was recorded on analog devices such as wax cylinders, phonograph records, and magnetic tape. Consultation of such materials as cited in studies was usually cumbersome, but their citation was often relatively straightforward.

In the early digital age, materials were shipped on digital tape reels or CD-ROM, and citation consisted of references to physical media. Nowadays, most digital materials are made available online. This has clear implications for the practice of citation. Furthermore, the use of digital data in linguistics has greatly expanded in volume and variety. Primary data in the form of large digital corpora of text, audio, and video have become widely available and are often annotated at one or more linguistic levels. Some other types of digital data (in the wide sense of the term) relevant for research on language are lexicons, term banks, word nets, computational grammars, translation memories, survey results, quantitative data from experiments, and so on. Locating specific data that were used in studies would amount to looking for a needle in a haystack were it not for proper citation. Unfortunately, citation practices haven't fully kept pace with new kinds of digital data and their distribution.

In this chapter, we sometimes use the more general term *resource* when referring to different types of digital research products, including, for instance, language models and analyzers (e.g., grammars, parsers), annotation tools, statistical code associated with certain data sets, and other digital assets. Often, we mention *data* for simplicity but most guidelines for data also hold for other resources. A *data set* is a set of data items that is distributed as a whole, but often we use *data* and *data set* interchangeably.

The guidance given in this chapter is primarily targeted at authors of linguistic publications, while a secondary audience consists of academic publishers and resource providers such as repositories and archives.

2 Why and when to cite linguistic data?

In a recent position paper, a group of linguists have argued that *reproducibility* (or *replicability*)—despite its key role in verification and accountability of research—is currently underrepresented in the field of linguistics (Berez-Kroeker, Gawne, et al. 2018). They have also articulated their expectations for how linguists should manage, cite, and maintained their data for long-term access.

Fortunately, there is increasing awareness of the importance of properly citing research data in scholarly outreach. The rationale for data citation can be summarized as follows: "Data citation improves discovery, credit, and attribution of data" (Borgman 2016). A recent statement dealing with data citation that has become prominent is the Joint Declaration of Data Citation Principles put forward by the FORCE11 Data Citation Synthesis Group (Martone 2014). This declaration summarizes the recommendations of earlier studies and has been endorsed by many scholarly organizations, funders, and publishers (Cousijn et al. 2018:2). Asserting the importance of robust and accessible data as the foundation of reproducible scholarship (cf. Gawne & Styles, chapter 2, this volume), the Joint Declaration of Data Citation Principles offers a set of guiding principles on how to refer to data in scholarly communication.

More recently, the *Austin Principles of Data Citation in Linguistics* provide an interpretation of the Joint Declaration of Data Citation Principles that places it in the context of linguistic data specifically (Berez-Kroeker, Andreassen, et al. 2018). The first three of the Austin principles cover the purpose and function of proper data citation under the headings "Importance," "Credit and attribution," and "Evidence." We quote these principles (in italics) and comment on their relevance for our guidelines.

1. Importance

Linguistic data form not only a record of scholarship, but also of cultural heritage, societal evolution, and human potential. Because of this, the data on which linguistic analyses are based are of fundamental importance to the field and should be treated as such. Linguistic data should be citable and cited, and these citations should be accorded the same importance as citations of other, more recognizable products of linguistic research like publications.

For this reason, we argue and suggest as one of the main recommendations in this guidance that rules for data citation should not without reason differ from common rules for citation of other research outputs, such as publications. Our guidelines will thus whenever reasonable be in line with existing guidelines for citation of publications, and in case our guidelines fall short, we recommend the author to stick to citation rules for publication or if possible adapt them to the field of research data.

2. Credit and attribution

In linguistics, citations should facilitate readers retrieving information about who contributed to the data, and how they contributed, when it is appropriate to do so.

Providing proper attribution information is a way to credit contributors and support citation metrics and thus serves as an incentive for data sharing. Good citation standards and practices feed directly into current initiatives such as the Make Data Count project, which addresses "the significant social as well as technical barriers to widespread incorporation of data-level metrics in the research data management ecosystem."[1]

3. Evidence

Linguists should cite the data upon which scholarly claims are based. In order for data to be citable, it should be stored in an accessible location, preferably a data archive or other trusted repository. Authors should ensure that data collection and processing methods are transparent, either through links to metadata or a direct statement in the text, to make clear the relationship between the data and the scholarly claims based on it.

Citing data naturally presupposes that the cited data is findable, and preferably also accessible, interoperable, and reusable, as formulated in the FAIR principles (Wilkinson et al. 2016), which are Findable, Accessible, Interoperable, and Reusable. It is first of all findability that merits our attention in this chapter. To allow review or replication of research outputs (Berez-Kroeker, Gawne, et al. 2018) or to extend previous research, it must be made known where the data can be obtained and under which conditions it can be reused. In practice this means that it is necessary to find and refer to documentation of the data, ideally in the form of structured *metadata* associated with the data set. This information should also explain how the data is encoded and must be interpreted, and in some cases, which tools are compatible with the data. For a more comprehensive overview of different motivations for data citation, see the review provided by Silvello (2018:8–11).

The Linguistics Data Interest Group in the Research Data Alliance[2] has taken the Austin principles to heart and elaborated them through meetings and online group discussions. A summary of this work has recently been made available as "The Tromsø Recommendations for Citation of Research Data in Linguistics" (Andreassen et al. 2019). The guidelines in this chapter are consistent with these recommendations.

3 General recommendations

3.1 Which rules to follow?

Despite the importance of data citation, standards and best practice recommendations on how to cite research data are still in their infancy. We propose a set of principles and guidelines for linguistic data citation. Based on the desideratum that data citation should not unnecessarily differ from the citation of scholarly text, we suggest the following order of adherence to possible guidelines from different sources:

1. Style sheets and guidelines by the publisher of the text in which the data is cited.
2. Guidelines and suggestions by the provider of the cited data, to the extent that the publisher's guidelines for data citation are unclear or incomplete.

3. Our recommendations for best practice, which we elaborate herein, to the extent that the preceding items are missing, incomplete, or contradictory, and can be refined, clarified or extended.

Ideally, a data set or other resource has *metadata* associated with it, which is structured information about the data. A metadata record typically has fields for data type, format, location, provenance, size, license, and other descriptors. Clearly, citation of data should provide some of the same kinds of information, so it is highly desirable that a citation of data is directly based on the metadata record of the data set or is at least consistent with it.

3.2 Data availability statement

In a study resulting in a given publication, the author(s) may have used or consulted different types of data (e.g., tabular data, sound recordings, and statistical code). Irrespective of the type of data, data may be associated with a publication in three different ways, according to JATS4R[3] (Bos et al. 2018):

1. *Generated data* are data that are included or referenced and that were generated and analyzed for the study. Example: Survey data are collected and analyzed for the study.
2. *Analyzed data* are referenced data that were analyzed for the study but that were not generated for the study. Example: Data are selected and extracted from an existing corpus and analyzed for the study.
3. *Non-analyzed data* are data that were neither generated nor analyzed for the study. Example: Existing data on the topic or related topic(s) were not analyzed for the study but are somehow related and may be relevant to researchers in the field.

Reference to non-analyzed data is generally discouraged by publishers, but may be appropriate in case authors want to acknowledge that similar work has been done by other researchers, although for methodological or other reasons their data have not been analyzed (Bos et al. 2018).

For any referenced data, that is, generated, analyzed, and non-analyzed data, authors should provide full and structured citations according to the recommendations described in this chapter. For generated and analyzed data, some recommend that authors should also provide a so-called *data availability statement* (some publishers use the term *data accessibility statement*) (Bos et al. 2018). A data availability statement provides information about where data supporting the results presented in a publication can be found, including, where applicable, unique identifiers referring to the location where these data are made available (Cousijn et al. 2018:5). Some publishers provide templates for data availability statements. The following examples are partly adapted from the author guidelines of Hindawi, one of the world's largest publishers of peer-reviewed, fully Open Access journals:[4]

> Data Availability Statement: The [DATA TYPE] data used to support the findings of this study are available via the [NAME] repository ([DOI or OTHER PERSISTENT IDENTIFIER]).

Note that a data availability statement should also be provided when the data used to support your findings for any legitimate reason(s) cannot be made available:

> Data Availability Statement: The [DATA TYPE] data used to support the findings of this study have not been made available because [REASON].

or in case you have not used any data to support your study:

> Data Availability Statement: No data were used to support this study.

In all cases, data citation or referencing consists of the same two basic elements as used in citation to publications, namely (1) *bibliographic references*, usually collected at the end of the document, but sometimes placed in footnotes, and (2) *in-text citation*, at the place in the text where the data are mentioned. The exact format of in-text citations (e.g., numbered, author-year, or otherwise) and the placing and formatting of bibliographic references are not discussed here as such elements of style are usually part of the publisher's guidelines for authors and typesetting practices.

For advice on where to archive and publish your own generated and/or analyzed research data, see Andreassen (chapter 7, this volume).

4 Bibliographic references

This section describes how to create full references to linguistic data for inclusion in the references (bibliography) section of a publication or in footnotes. For

some examples, see section 7, and for a systematic list of examples, we also suggest you consult the "Tromsø Recommendations for Citation of Research Data in Linguistics" (Andreassen et al. 2019) alongside this chapter.

4.1 Template for bibliographic reference

There is some debate about what elements are necessary to make up a complete data reference. Although most accounts depart from recommendations such as the Joint Declaration of Data Citation Principles (Martone 2014), the suggested or recommended elements vary somewhat from recommendation to recommendation (Ball & Duke 2015; Silvello 2018). Adapting these recommendations to the realm of linguistics, the Tromsø recommendations suggest the following two templates for citation of data sets in the bibliography section of a piece of academic writing.

1. The template for a minimal bibliographic reference to a data set has the following elements;[5] this template is also consistent with emerging recommendations by some publishing houses.[6]

   Author, Date, Title, Publisher, Locator.

2. The template for an expanded bibliographic reference to a data set including conditional elements (i.e., required in certain cases depending on resource characteristics) is as follows:

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

Elements rendered in bold are part of the minimal template, in other words, they are always required, while elements rendered in italics are considered to be conditional. Conditional elements are elements whose presence is conditioned by either the characteristics of the resource (e.g., references to versioned data sets 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). Note that we do not assume any order in which the elements of these templates may occur. Their order and formatting may vary, depending on the bibliography style that is part of a publisher's style guide.

In this section, we briefly define and discuss the different elements in some detail. Some elements may come in different types. For instance, the **Author** field may include the name of a principal investigator, but also the name of a data collector; the **Date** may be specified by date of publication or date of deposit, and so on. In cases where such clarification can be beneficial, we propose that elements be more specifically typed, for which we will propose suitable defaults. When using an element type that is not default, the element type should be specified in parentheses, for instance: *2018 (deposit date)*; *John Smith (data collector)*.

If the metadata for the data you are to refer to do not provide information about any or some of the elements listed in the templates and you are not able to get hold of the information from the owner or responsible body for the data source, you should not make up the information. Instead, you should state in the reference the lack of information. For instance, if no date at all is available you should add "n.d." or similar in the **Date** slot of the reference, analogous to usual practice for other publications.

Author. By default, **Author** is one or more entities (persons or organizations) responsible for having developed the resource and deposited it. Specific roles may be indicated, as also practiced in other contexts, for example, with *relator* attributes (Hornik, Murdoch, & Zeileis 2012). 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," "Language User," "Consultant," "Project," and "Contact." By default, only the main responsible authors are listed without mentioning their roles. Other entities might be specified (and even required) by resource provider or publisher guidelines and/or the research community at stake and should be marked by their specific roles. Note that if the bibliographic reference specifies multiple roles, then all roles must be included, for example, "Name (Researcher, Depositor)." Whenever possible, use role names that are in line with standard or recommended vocabularies, such as the OLAC (Open Language Archives Community) Role Vocabulary (Johnson 2006) or the controlled list of contributors in the DataCite Metadata Schema (DataCite Metadata Working Group 2019:31–35). Roles other than the default **Author** are listed as *Other Attribution* in the templates. Finally, it should be pointed out that the **Author** field in a reference is not the only place for providing attribution. It is good practice to fully recognize all contributors in the metadata record of a resource. The metadata record is also the place where contributor identifiers such as

an ORCID (Open Researcher and Contributor ID) or international standard name identifier—if applicable—should be added (Ball & Duke 2015:6), or at least contact information about the author.

Other Attribution. See **Author**. *Other Attribution* is like **Author** but must mention roles.

Date. By default, **Date** is the date of publication. If the publication date is not available (i.e., there is no formal publication process), use the deposit date, that is, the date when the resource was transferred to and registered in the facility of the resource provider. If no deposit date is registered, then use the collection or production date, that is, the date when the data were collected or produced; preferably the date indicating when the collection or production was completed; alternatively, the period of collection or production. Whether the **Date** field is specified as a year or as a more precise date depends on the style guide by the publisher and the information provided by the cited resource.

Title. The **Title** is the name of the resource, at the level that comprises all citations in the study; for instance, if several parts of a data set are mentioned in the text, it is recommended to list the data set as a whole in the references, while in-text citations refer to the relevant parts. This is analogous to having a single mention to a multivolume book work in the bibliography, rather than listing the several volumes separately. If, on the other hand, the text cites a single subset or item of the data, the **Title** field may name that directly, followed by "In . . ." or similar, to refer to the full resource of which the data is a part. The level of granularity of citation will be discussed in more detail in section 4.2.

Publisher. The **Publisher** is the entity responsible for providing access to the resource. In most cases this will be the name of the resource provider (e.g., data repository or organization). If possible, this should be the original source, not a harvester of metadata or copier of the data.

Locator. The **Locator** is a digital identifier pointing to the landing page of the resource if available online. This identifier should preferably be at the level (e.g., subset, item) corresponding to what the **Title** refers to. The **Locator** of a digital object should preferably be a persistent identifier (PID) if the publisher provides one.[7] DataCite, the provider of digital object identifiers (DOIs) for data sets, recommends specifying the PID as a fully expanded resolvable persistent uniform resource locator, so instead of, for example, doi.org/10.18710/X5ZFXZ, use https://doi.org/10.18710/X5ZFXZ (DataCite DOI Display Guidelines, n.d.). If there is no such identifier, include the resource provider's internal identifier for the resource (e.g., record number, deposit identification number), in addition to the URL of the landing page. If no PID for the resource exists, include the URL, but make sure that it is available to readers: "If the URL requires a login or is session specific, meaning it will not resolve for readers, provide the URL of the database or archive home page or login page instead of the URL for the work" (American Psychological Association 2020:299, section 9.34). If no online locator exists, indicate the media type; this applies to both digital media (e.g., CD audio, CD-ROM text file) and analog media (e.g., book, archival file).

Version. The *Version* is an identifier, normally a number that is increased whenever data and/or metadata of a resource are changed, but it could also be a time stamp (e.g., for nightly builds), a Git commit identification number, or similar. The default is that there is only one version and the resource is assumed to be stable. An alternative value for *Version* is "dynamic" meaning that the resource may change without explicit versioning or time stamping; in that case, *Date accessed* is also required.

Date accessed. *Date accessed* is a date. If the resource is dynamic or it is uncertain whether the resource is stable and persistent (e.g., monitor corpora that grow in size or treebanks that are reparsed), an access date must be added to the reference.

If for some reason it is necessary or requested to explicitly distinguish data references from other references by indicating their type, authors may add an indicator (tag) at the end of the reference, often in parentheses or brackets, such as (data set), [code], and so forth. (Cousijn et al. 2018:6).

Publishers and/or resource providers might require or recommend additional elements not listed in our templates or listed only as conditional elements. For instance, some data distributors include an indicator of data fixity in the citation information, such as a Universal Numerical Fingerprint (Altman & King 2007). Finally, some resource providers might require citation of a written publication describing the data. In this case, the written publication should be cited in addition to citing the data themselves.

4.2 Granularity

A cited resource may consist of multiple parts. Thus, the question arises at what level of granularity a citation should be given. The choice of granularity level is not straightforward:

> A dataset may form part of a collection and be made up of several files, each containing several tables, each containing many data points. There are also more abstract subsets that can be used, such as features and parameters. At the other end of the scale, it is not always obvious what would constitute an intellectual whole: it can be argued, for example, that investigations should be the primary units of citation rather than individual datasets. (Ball & Duke 2015:7)

As a pragmatic solution, Ball and Duke suggest citing data at the level of granularity that the resource provider has chosen for assigning *Locators* ("identifiers") and that, where *Locators* are provided at several levels of granularity, references should be given at the finest-grained level that meets the need of the citation (7).

The following template may serve as a model for a fine-grained reference to a part of an assembled resource (including conditional and optional elements). As usual, the order of elements may vary.

> *Author of part, Date, Title of part,* In: *Author of assembled resource, Title of assembled resource, Publisher of assembled resource, Locator of part, Version of part, Date accessed of part. [resource type of part]*

The following fine-grained reference was generated by the Tromsø Repository of Language and Linguistics (TROLLing) for a file included in a data set published in the repository.

> Arkhangelskiy, Timofey, 2019, "01_rnc_borrowed_lemmata.txt", In: *Replication data for: Verbal borrowability and turnover rates*, https://doi.org/10.18710/JFNESU/LETGNY, DataverseNO, V1 [file]

Another example refers to a recording in an archive. In this example, as in the previous one, the *Locator* points directly to the landing page for the item, not to the main page of the archive.

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

As a main rule, you should choose fine-grained citation if the *Author* of the resource part is not identical with the *Author* of the assembled resource. In that case, the *Author* of the assembled resource should be mentioned, if there is one. This is also common practice in citation of scholarly publications, where reference is made to a specific paper or chapter in an anthology with contributions from different authors. If, on the other hand, there are citations to multiple parts of a resource, especially when these have the same *Author* as the assembled resource, one may wish to avoid a long list of separate references for the parts, as in the following made-up example:

> Smith, John, 2018a, "Data set 1",. . . . [dataset]
>
> Smith, John, 2018b, "file_01.txt", *Data set 1*,. . . . [file]
>
> Smith, John, 2018c, "file_02.txt", *Data set 1*,. . . . [file]
>
> . . .

Instead, the highest common denominator of all the parts should be listed in the references, while each in-text citation refers to a specific part by including a fine-grained element. How to add a more specific *Locator* in the in-text citation is explained in the next section.

5 In-text citation

In-text (or in-line) citations are meant to direct the reader to a bibliographic reference at the end of the published work, and, if relevant, indicate which part of the cited data set is referred to. For a style sheet based on the author-date format, for example, APA (American Psychological Association 2020), the minimal template is as follows.

Template for minimal in-text citation of data:

Author, Date

As mentioned in the section on bibliographic references, we recommend that data references are made at the finest-grained level. If the document refers to a single part of an assembled data set, the in-text citation can simply refer to that item in the references. If citing a part or individual item of a data set listed in the reference, you should provide details on which part of the data you wish to cite, for example, by adding a *Locator*, which can be the name or PID of a collection, file, item

(Ball & Duke 2015:7). In this case, the template is as follows.

Template for in-text citation of data, including *Locator*:

Author, Date, Locator

In certain cases, it may be useful to refer to a particular *Subset* of the data, such as a time span or a range of line numbers. If the subset is a time span, we recommend using the International Organization for Standardization's ISO 8601 time codes in the [hh]:[mm]:[ss] format (ISO 8601 2019).

Template for in-text citation of data, including *Locator* and *Subset*:

Author, Date, Locator, Subset

If the cited subset has relevant contributors that differ from the main authors, these contributors and their roles may be mentioned under *Other Attribution*.

Template for in-text citation of data, including *Locator*, *Subset*, and *Other Attribution (Roles)*:

Author (Role), Date, Locator, Subset, Other Attribution (Roles)

If your publication channel uses a number format rather than author-year format for in-text citation, use footnotes or endnotes to indicate the granularity of the citations, if needed. Footnotes (rather than a reference in the main text) are also recommended for long URLs or PIDs when it is necessary to refer to specific resource items, sections, or other parts of a resource.

Actual examples of in-text citations will be given in section 7.

6 Citing unpublished data

Sometimes one may want to cite research data that are not yet publicly available. An increasing number of publishers encourage or even require authors to provide access to the data underlying manuscripts when they are submitted for review. At that point in time, the data may still be under embargo, in a repository submission process, or for other reasons not publicly available. The rule of thumb for citing unpublished data is to provide as much information as possible, and at least the *Author* and the *Title* of resource (Ball & Duke 2015:7). In the data availability statement, one should explain the full details of the status of the resource, such as whether it is deposited, embargoed, restricted, or openly available (Ball & Duke 2015:7).

When a paper presenting new data is submitted for peer review, authors must make sure that any information given about the cited data, either through citation in the text or in the referenced data themselves, complies with any anonymity requirements of the publisher.

Once your manuscript has been accepted for publication, make sure to revisit references to unpublished data in your manuscript and bring them up to date before the final version of your article or book is published (Ball & Duke 2015:7). Analogously, once your article or book is published you should consider including a reference to the publication in the metadata record of the cited data.

7 Examples

Whereas FORCE11 has formulated very general principles for data citation, these have been adapted to a linguistics context and rationale in the above-mentioned Austin principles, but even these are fairly general and do not provide detailed guidelines for various types of language data in possible citation contexts. Drawing on the Tromsø recommendations (Andreassen et al. 2019), we will provide an overview of some data types and examples of current practice for citing these. Where appropriate, we will comment on the examples and provide supplementary information.

Note that the examples represent different citation formatting styles. As mentioned at the outset of this chapter, you should always make sure your citation follows the citation style required or recommended by the publisher.

For more examples, see the "Tromsø Recommendations for Citation of Research Data in Linguistics" (Andreassen et al. 2019). The examples presented there as well as the examples in this chapter represent a variety of good options for formatting citations.

7.1 Corpora and other collections of language materials

Example 1 illustrates how an item that is part of a language archive can be referred to. It also illustrates how the roles of contributors may be specified:

Example 1:

Hauk, Bryn (Researcher, Depositor), P'ap'ashvili, Omar (Language User), Orbetishvili, Rezo

(Consultant). 2018. Batsbi (Tsova-Tush) Collection, Item BH2–074. Kaipuleohone University of Hawaii Digital Language Archive, http://hdl.handle.net/10125/58935. Accessed on 2019-03-10.

Furthermore, the *Date accessed* is provided here because the resource provider does not provide any version information. In case there are different citations of the archive, the fine-grained element "Item BH2–074" may be omitted from the bibliographic reference. It should then be included in the in-text citation; you might also include a time code to refer to a specific excerpt, and one or more non-author contributors, if these are particularly relevant. This gives some possible variants of in-text citation for this example:

- (Hauk 2018, BH2–074)
- (Hauk 2018, BH2–074, 00:02:33–00:02:47)
- (Hauk 2018, BH2–074, 00:02:33–00:02:47, Rezo Orbetishvili (Consultant))

Example 2 has a website but no PID:

Example 2:

BNC Consortium. The British National Corpus, version 3 (BNC XML Edition). 2007. Oxford: Bodleian Libraries, University of Oxford. http://www.natcorp.ox.ac.uk/.

The following is an example of a resource on CD-ROM. The LDC catalog number is also given.

Example 3:

Liberman, Mark, et al. Emotional Prosody Speech and Transcripts LDC2002S28. CD-ROM. Philadelphia: Linguistic Data Consortium, 2002.

Examples 3 and 4, as suggested by the LDC, lack direct *Locators*. They require that the reader finds the LDC and looks up the resource number in its catalog.

Example 4:

Huang, Shudong, David Graff and George Doddington. Multiple-Translation Chinese Corpus LDC2002T01. Web download file. Philadelphia: Linguistic Data Consortium, 2002.

In quite a few cases, as illustrated in the following examples, the author of the corpus is a project team, in which case the name of the project may be given. PIDs of the *Handle* type are specified as fully resolvable URLs.

Example 5:

ISWOC, 2016. ISWOC West-Saxon Gospels. Created by Information Structure and Word Order Change in Germanic and Romance Languages (Project). Distributed by the INESS Portal. http://hdl.handle.net/11495/DB24-D542-3616-6.

Example 6:

INESS, 2016. NorGram Newspaper text (30 documents from the years 2006–2009) in Norwegian Bokmål from the Norwegian Newspaper Corpus. Created by: Infrastructure for the Exploration of Syntax and Semantics. Distributed by the INESS Portal: http://hdl.handle.net/11495/DB24-E30D-55EA-1. Dynamic data, accessed April 1, 2019.

An in-text citation of a sentence from the corpus could look like the following, including the footnote.

(INESS, 2019, Sentence #3147[8])

The PID in the footnote points directly to a specific sentence in a specific treebank archived in INESS (Norwegian Infrastructure for the Exploration of Syntax and Semantics) infrastructure. This is convenient for the reader. Such a PID can be generated by clicking on a button next to any sentence stored in INESS. The PID is a stable reference to the sentence but not necessarily to the annotation. If the treebank is a dynamic parsebank that may be reparsed, as in this example, the *Date accessed* is also required.

7.2 Databases and application data

The term *database* is used for many different kinds of resources. Examples are lexical and terminological databases, for instance, word nets, which represent semantic relations between lexical concepts. The following reference in the Modern Language Association's style is recommended by the publisher of WordNet, a lexical database for the English language, for reference to their online version:

Example 7:

Princeton University "About WordNet." WordNet. Princeton University. 2010.

This reference is somewhat underspecified, as it lacks information about the version, whereas several versions exist. Also, we recommend providing the *Locator* as a full URL, which in this example would be https://wordnet

.princeton.edu/, which does not point to the resource itself but to an information page with further instructions on how to access the resource. A more fully specified reference to a version of WordNet is given in example 8, in which the *Date* is the publication date of the specified *Version*.

Example 8:

> Princeton University, 2006. Wordnet (version 3.0). Distributed by Princeton University, https://wordnet.princeton.edu/.

Application data are data sets or databases that are meant to be used in a particular application, for instance, in a system for machine translation, a finite state transducer, a parser. Example 9 points to a parallel corpus provided in part as a translation memory in Tile Map XML format for use in machine translation systems; the *Handle* resolves to a landing page with human-readable metadata.

Example 9:

> Parra Escartín, Carla, 2012, Parallel Corpus of documents from the Technical Regulations Information System for German-Spanish (v0.3; TMX and TEI formats), Common Language Resources and Technology Infrastructure Norway (CLARINO) Bergen Repository, http://hdl.handle.net/11509/79.

7.3 Replication data sets or packages

In the common course of a research study, inputs from resources such as the ones mentioned herein are often further processed, refined, and analyzed. In recent years, it has become increasingly common—and required by funders and publishers—to make not only the raw data, but also processed data available to support the findings in a study. Processed data may consist of different resource types, such as spreadsheets, audio or video recordings and transcriptions, statistical code, software code, field notes, which are assembled into a data set that is made available to peers and the greater public.

Many replication data packages have been made for quantitative and qualitative data resulting from experiments, surveys, and similar methods. Such data are not language data in themselves, but measurements, judgments, attitudes, and so forth.

Example 10 shows a ready-made reference as generated and recommended by TROLLing:

Example 10a:

> Ji, Yinglin, 2018, "Cognitive representation of spontaneous motion in a second language", https://doi.org/10.18710/N8KO4O, DataverseNO, V1.

We recognize several of the reference elements from our minimal template:

- *Author:* Ji, Yinglin; Date: 2018;
- *Title:* Cognitive representation of spontaneous motion in a second language;
- *Publisher:* DataverseNO (TROLLing is a special collection within the DataverseNO repository);
- *Locator:* https://doi.org/10.18710/N8KO4O (a DOI in the form of a fully expanded URL);
- *Version:* V1.

Furthermore, depending on the specified citation style, the tag "[dataset]" may be added to the reference to distinguish the referenced data from other types of materials.

Example 10b:

> Ji, Yinglin, 2018, "Cognitive representation of spontaneous motion in a second language," https://doi.org/10.18710/N8KO4O, DataverseNO, V1. [dataset]

In example 10b, the reference points to a whole data set. TROLLing also assigns DOIs at file level, and thus you might want to refer to a particular file within a data set. Resuming our discussion of example 10a, based on the ready-made reference provided by TROLLing, the file "03_preference_data.txt" might be referred to as follows:

Example 11:

> Ji, Yinglin, 2018, "03_preference_data.txt," *Cognitive representation of spontaneous motion in a second language*, https://doi.org/10.18710/N8KO4O/H9PDCL, DataverseNO, V1. [file]

If, in example 11, you choose to include a reference only at data set level, the in-text reference to the file has to be more fine-grained and should preferably include either the file name or the file DOI:

- (Yinglin 2018, 03_preference_data.txt)
- (Yinglin 2018, https://doi.org/10.18710/N8KO4O/H9PDCL)

To round off this section, we would like to illustrate how the TROLLing data set would have been cited prior to publication. When submitting her article for submission, the author could have used the following (non-anonymized) reference:

Example 12:

> Ji, Yinglin, 2018, "Cognitive representation of spontaneous motion in a second language," https://doi.org/10.18710/N8KO4O, DataverseNO, DRAFT VERSION. [data set]

Once the journal article was accepted, the author would have published the data set, and replaced "DRAFT VERSION" with "V1" in the reference, before submitting the final version of the article to the journal publisher.

8 Reference management tools

Nowadays, several reference management tools are available that lessen the burden of keeping track of references and their formatting. Thus, it is paramount that reference managers have capabilities to handle at least the minimal items that we suggest. Unfortunately, widely used systems such as Zotero[9] and BibDesk[10] (based on BibTeX[11]) currently offer fragmentary support for describing research data, although changes are underway.

BibTeX does not have a dedicated type for research data among its standard reference types, but it has *url, webpage, electronic* (alias *online*), and *misc*. Our proposed elements *Author*, *Date, Publisher*, and *Title* can be mapped to BibTeX fields *Author, Date (or Year), Organization*, and *Title*, respectively. A role added for *Other Attribution* would be parsed as part of the name. For *Locator*, one can use *Url, Doi*, or *Howpublished* depending on the type. *Version* could go in *Howpublished*, as *Note*, or as part of the *Title*. *Date accessed* can be mapped to *URLdate* or *Lastchecked*.

The following is a BibTeX record that was automatically generated by the CLARINO Bergen Repository.[12] It has most of the important information, but it is unfortunate that the publisher is in the *Note* field and that *Copyright* is not a supported field for the *misc* type.

```
@misc{11509/73,
title={The Norwegian-Spanish Parallel Corpus},
author={Hareide, Lidun},
url={http://hdl.handle.net/11509/73},
note={Common Language Resources and Technology
Infrastructure Norway ({CLARINO}) Bergen Repository},
copyright={{CLARIN}\_{ACA}},
year={2013} }
```

BibLaTeX,[13] a newer alternative to BibTeX, has standard entry types for *software* and *dataset* since 2019. Whereas *software* is aliased to *misc*, the *dataset* entry type has its own definition, which includes several optional fields, including both *Publisher* and *Organization*. A possible modification and extension of the preceding example using some supported fields is the following:

```
@dataset{11509/73,
title={The Norwegian-Spanish Parallel Corpus},
author={Hareide, Lidun},
url={http://hdl.handle.net/11509/73},
location={Bergen, Norway},
publisher={Common Language Resources and Technology
Infrastructure Norway ({CLARINO}) Bergen Repository},
version={1},
note={[dataset]},
year={2013},
urldate={2020-01-17}}
```

The Research Information Systems (RIS) format, which can be imported in some bibliographic management systems, has ADVS (audiovisual material), DATA (data file), DBASE (online database), SOUND (sound recording), and so on, but no specific type for research data in general, so as to specify archives, replication packages, corpora, and such.

Zotero is a powerful bibliographic management system that also supports shared references in so-called group libraries. Currently, Zotero does not provide a dedicated citation item type for research data. However, this feature is planned to be included in the next major version upgrade.[14] As a transitional solution, Zotero recommends using the citation item type *Document* for data sets. Depending on how you create a reference in Zotero (e.g., manually vs. automatically using a web browser extension or a similar tool) you may have to edit the reference record in Zotero. In particular, you should pay attention to the following three fields, here illustrated with a data set from TROLLing. Note that the three pieces in the *Extra* field should be entered on separate lines. Do not include version if it is 1.

```
Item Type: Document
Publisher: DataverseNO
Extra: type: dataset
   version: 2
   DOI: 10.18710/BFFMPH
```

This solution works fairly well for some citation styles. Using the citation style for the seventh edition of the APA manual (American Psychological Association 2020), Zotero creates the following reference that includes the information from the *Extra* field:

> Holliday, J. J., Turnbull, R., & Eychenne, J. (2016). *K-SPAN (Korean Surface Phones and Neighborhoods)* (Version 2) [Data set]. DataverseNO. https://doi.org/10.18710/TWM79F

This reference contains all elements included in our recommended template for minimal bibliographic reference. However, fields are still lacking for other elements that might be recommended by resource providers, for example, an indicator of data fixity such as Universal Numerical Fingerprint, which is provided by the repository hosting the preceding data set.

If you want to refer to a *Subset* in your in-text citation using Zotero, you have to map this to *Part*, which is an alternative to *Page*. You may either refer to *Part* by file name or PID. For the preceding example, the in-text citation in APA style may look like this:

> (Holliday et al., 2016, pt. kspan_base.tab)

Currently, there is no convenient way to create fine-grained bibliographic references to research data using Zotero, but given the commitment and attention support services for research data management are receiving at present, this functionality is expected to come in a not-so-distant future.

9 Advice on metadata for repositories and other resource providers

Informative citation is dependent on metadata, that is, structured data describing the properties of the data. Metadata are normally entered when data are deposited in a repository. Obviously, such metadata should contain the elements necessary for citation. Metadata also serve other purposes, such as cataloging to support faceted search. For the purpose of citation, metadata should conform to the following.

- At minimum, the metadata should include the elements in the minimal templates recommended herein.
- Metadata should preferably be structured according to a standard format (e.g., component MetaData Infrastructure, RIS) so that information from it can be extracted by programs.
- Metadata should be available freely, without cost or restrictions, even if the data themselves have restrictions.
- The metadata should allow persistent reference to the data set, and to the metadata themselves, to avoid link rot. This implies that PIDs should be assigned to the data and to the metadata. Several identifiers have been proposed as standards. In the field of linguistics, the International Standard Language Resource Number[15] has been proposed (Mapelli et al. 2016). More inherently persistent identifiers are based on the Handle system,[16] which implements the identifier/resolution protocol (IRP) to enable persistence of the identifiers even if the data moves to different locations. The DOI system[17] is also based on IRP, but has additional specifications, such as those for minimal metadata.
- Data repositories and other resource providers should provide metadata in machine-readable as well as human-readable form and should preferably also generate ready-made citations, both in formats for export to reference managers and in textual format.

In a somewhat more distant future, it is expected that both data and metadata will be more inherently cloud-based, encapsulated as FAIR Digital Objects (De Smedt, Koureas, & Wittenburg 2019). This would introduce new ways for researchers and repositories to interact with data. Actionable digital objects will communicate with automated processes requesting access and data management and should be able to provide their own citations in a context-dependent way.

10 Advice for publishers

Apart from researchers and resource providers, the realization of the recommendations outlined in this chapter depends highly on the support from academic publishers. We thus strongly encourage publishers to adopt our recommendations in their work to make research more transparent and reusable. In particular, we recommend publishers to provide necessary guidance and resources for data citation to the different stakeholders in the ecosystem of scholarly communication. We are aware of the fact that several academic publishing houses and scholarly associations already have guidelines or are in the process of establishing those.

We recommend that publishers have a data policy that preferably requires authors to make data accessible

at least for reviewers at the time of manuscript submission and to make data openly available at latest at the time of publication of the article or book. Publishers may also advise authors on where and how they should deposit their data, preferably based on information available in overviews such as the Registry of Research Data Repositories.[18] Author guidelines should include guidance on data availability statements and style sheets should have clear instructions for citation and bibliographies that cover different types of linguistic resources.

Links to cited data in published texts should be written in full. In electronic publications, these links should preferably be clickable.

We also recommend that publishers make the full metadata of published data openly and freely available to researchers, catalogs, and other resource providers. This is important, for example, for services that provide overviews about citations of data sets in article and book publications. Usually this can be done by providing metadata to Crossref or similar providers.[19]

In general, publishers should follow standards and best practice recommendations for electronic publishing of scholarly results, as promoted by, for example, Crossref and JATS4R.

Acknowledgments

The authors of this chapter wish to acknowledge all who through discussion or written comments have provided ideas for or feedback on the preparation of this chapter. In particular we wish to thank the Linguistic Data Interest Group of the Research Data Alliance for their work on the development of these citation guidelines, and the Linguistic Data Consortium for providing meeting space for one of our many face-to-face discussions.

Notes

1. https://makedatacount.org/.

2. https://www.rd-alliance.org/groups/linguistics-data-ig.

3. JATS4R (JATS for Reuse; https://jats4r.org/) is a working group devoted to optimizing the reusability of scholarly content by developing best-practice recommendations for tagging content in JATS (Journal Article Tag Suite) XML.

4. See https://www.hindawi.com/research.data/.

5. In reference management tools, the term "field" is often used instead of "element."

6. E.g., Cambridge University Press, https://www.cambridge.org/core/services/authors/open-data/data-citation.

7. On the Internet, a PID has the form of a PURL, which curates redirection by means of a resolver. This scheme attempts to solve the problem of transitory locators in location-based schemes such as HTTP. Example types of persistent identifiers are the Hangle the Digital Object Identifier (DOI), and the Archival Resource Key (ARK) There are also different types of globally unique identifiers that do not involve automatic curation by a resolver, such as the International Standard Language Resource Number (ISLRN).

8. http://hdl.handle.net/11495/D8B8-3970-851A-3@lfg234387, accessed February 2, 2020. Access to this data may require login.

9. See https://www.zotero.org/.

10. https://bibdesk.sourceforge.io/.

11. See http://www.bibtex.org/.

12. A repository based on CLARIN DSpace, http://clarino.uib.no/.

13. https://www.ctan.org/pkg/biblatex.

14. https://www.zotero.org/support/dev/translators/datasets.

15. http://www.islrn.org/.

16. https://www.dona.net/handle-system.

17. https://www.doi.org/.

18. https://doi.org/10.17616/R3D, accessed November, 19, 2020.

19. See https://www.crossref.org/.

References

Altman, Micah, and Gary King. 2007. A proposed standard for the scholarly citation of quantitative data. *D-Lib Magazine* 13 (3/4). https://doi.org/10.1045/march2007-altman.

American Psychological Association. 2020. *Publication Manual of the American Psychological Association,* 7th ed. Washington, DC: American Psychological Association. https://doi.org/10.1037/0000165-000.

Andreassen, Helene Nordgård, Andrea L. Berez-Kroeker, Lauren Collister, Philipp Conzett, Christopher Cox, Koenraad De Smedt, Bradley McDonnell, and the Research Data Alliance Linguistic Data Interest Group. 2019. Tromsø recommendations for citation of research data in linguistics, *Research Data Alliance*. https://doi.org/10.15497/RDA00040.

Ball, A., and M. Duke. 2015. How to cite datasets and link to publications. DCC how-to guides. Digital Curation Centre. http://www.dcc.ac.uk/resources/how-guides.

Berez-Kroeker, Andrea L., Helene N. Andreassen, Lauren Gawne, Gary Holton, Susan Smythe Kung, Peter Pulsifer, Lauren B. Collister, the Data Citation and Attribution in Linguistics Group, and the Linguistics Data Interest Group. 2018. *The Austin Principles of Data Citation in Linguistics*. https://site.uit.no/linguisticsdatacitation/austinprinciples/.

Berez-Kroeker, A. L., L. Gawne, S. S. Kung, B. F. Kelly, T. Heston, G. Holton, P. Pulsifer, et al. 2018. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics* 56 (1): 1–18. https://doi.org/10/gft4g7.

Borgman, C. L. 2016. Data, data citation, and bibliometrics. https://escholarship.org/uc/item/98r688tr.

Bos, Ton, Paul Donohoe, Melissa Harrison, Christina Von Raesfeld, and Kelly McDougall. 2018. Data availability statements, version 1.1. JATS4R (JATS for Reuse). https://jats4r.org/data-availability-statements.

Cousijn, Helena, Amye Kenall, Emma Ganley, Melissa Harrison, David Kernohan, Thomas Lemberger, Fiona Murphy, et al. 2018. A data citation roadmap for scientific publishers. *Scientific Data* 5 (November): 180259. https://doi.org/10.1038/sdata.2018.259.

DataCite DOI Display Guidelines. n.d. https://support.datacite.org/docs/datacite-doi-display-guidelines. Accessed December 15, 2019.

DataCite Metadata Working Group. 2019. DataCite metadata schema documentation for the publication and citation of research data, version 4.2. https://schema.datacite.org/meta/kernel-4.2/index.html.

De Smedt, Koenraad, Dimitris Koureas, and Peter Wittenburg. 2019. An analysis of scientific practice towards FAIR Digital Objects. EUDAT. http://doi.org/10.23728/b2share.e14269d07ce84027a7f79ee06b994ef9.

Hornik, Kurt, Duncan Murdoch, and Achim Zeileis. 2012. Who did what? The roles of R package authors and how to refer to them. *R Journal* 4 (1): 64–69. https://doi.org/10.32614/RJ-2012-009.

ISO 8601. 2019. *Wikipedia*. https://en.wikipedia.org/w/index.php?title=ISO_8601&oldid=895977129.

Johnson, Heidi. 2006. OLAC role vocabulary. http://www.language-archives.org/REC/role.html.

Mapelli, Valérie, Vladimir Popescu, Lin Liu, and Khalid Choukri. 2016. Language resource citation: The ISLRN dissemination and further developments. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016), 1610–1613. Portorož, Slovenia: European Language Resources Association (ELRA). https://www.aclweb.org/anthology/L16-1254.

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

Silvello, Gianmaria. 2018. Theory and practice of data citation. *Journal of the Association for Information Science and Technology* 69 (1): 6–20. https://doi.org/10/gcqfk2.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data* 3 (March): 160018. https://doi.org/10.1038/sdata.2016.18.

This is a section of doi:10.7551/mitpress/12200.001.0001

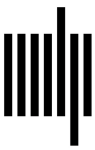
The Open Handbook of Linguistic Data Management

Edited by: Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, Lauren B. Collister

Citation:

The Open Handbook of Linguistic Data Management
Edited by: Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, Lauren B. Collister
DOI: 10.7551/mitpress/12200.001.0001
ISBN (electronic): 9780262366076
Publisher: The MIT Press
Published: 2022

The open access edition of this book was made possible by generous funding and support from the authors



The MIT Press

This book was set in Stone Serif and Stone Sans by Westchester Publishing Services.

Library of Congress Cataloging-in-Publication Data

Names: Berez-Kroeker, Andrea L., editor. | McDonnell, Bradley James, editor. | Koller, Eve, editor. | Collister, Lauren B., editor.
Title: The open handbook of linguistic data management / edited by Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller and Lauren B. Collister.
Description: Cambridge, Massachusetts : The MIT Press, [2021] | Series: Open handbooks in linguistics series | Includes bibliographical references and index.
Identifiers: LCCN 2020044363 | ISBN 9780262045261 (hardcover)
Subjects: LCSH: Computational linguistics. | Natural language processing (Computer science) | Data mining.
Classification: LCC P98 .O64 2021 | DDC 410.285—dc23
LC record available at https://lccn.loc.gov/2020044363