This is a portion of the eBook at doi:10.7551/mitpress/12200.001.0001

46 Managing Experimental Data in a Study of Syntax

Matthew Wagers

1 Introduction: Reproducibility and data management in experimental syntax

This use case considers some data management workflows that can help in studies of experimental syntax and sentence processing. For concreteness, I will use the example of data generated by an acceptability judgment experiment—a familiar sort of workhorse study in *experimental syntax* (Bard, Robertson, & Sorace 1996; Cowart 1997; Schütze 1996; Sprouse 2007). Acceptability judgments play an important role in many domains of linguistics, such as psycholinguistics and the study of sentence processing (Dillon et al. 2018), semantics, pragmatics, as well as any other linguistic domains where classes of stimuli can be discriminated by "goodness" in some sense. In an acceptability judgment experiment, participants are presented with a list of experimentally controlled sentences and respond to their acceptability, naturalness, felicity, or some related quality. The dependent variable is most commonly a binary choice (yes/no, acceptable/unacceptable) or a response along a labeled or numeric scale. Sometimes experimenters record decision-related variables such as judgment time/response time and confidence.

Acceptability judgment experiments may seem, at first blush, to have simple designs and to generate relatively uncomplicated experimental data. And yet they pose some challenging data management questions, ones that recur across a number of types of language experiments. Here I will discuss the choices that have routinely faced myself and my collaborators, and the decisions we've made—or sometimes, the ones we would aspire to make in a better world. I divide my discussion into three phases: *design*, *execution*, and *analysis*. I'd particularly like to highlight the idea that designing an experiment is itself a data-generating activity. Berez-Kroeker et al. (2018) have urged a culture of reproducible research in linguistics, whose goal is "scientific accountability" and whose means is transparency. Transparency is achieved by access not only to *source data* (observations of some linguistic behavior or function) but also to the *methods of data collection*. In a language experiment, those methods are themselves the consequences of myriad decisions about such practical issues as what lexical items to use, what presentation parameters to select, how to word the instructions, and so forth. While research reports do routinely provide that information, under a "Materials and Methods" section, it is often only in a fragmentary or ambiguous way. Having more comprehensive access to the process by which the experiment was constructed can often be just as important to evaluating a claim as being able to inspect the source data yourself. Therefore, part of managing and sharing the data of an experimental study is managing and sharing the documents and scripts that precede the actual execution of the study.

To make this point, we must consider how an acceptability experiment is designed. Suppose a researcher is interested in whether sentences with longer syntactic dependencies are less acceptable than sentences with shorter ones (they are, usually) and whether a discourse-linked *wh*-phrase dependent can counteract that length effect (it can, sometimes).[1] The researcher would design an experiment using sentences that systematically explore a logical space of dependency lengths and *wh*-phrase types. In (1) I've provided a set of sentences that span four points in that space. Observe that each token in this set exemplifies a single type of sentence, or *condition*, in the experiment: (a) long distance + bare *wh*-phrase; (b) long + D-linked *wh*-phrase; (c) short distance + bare; and (d) short + D-linked.

(1) a. Who does Nora suspect the judge must appoint?
 b. Which guardian does Nora suspect the judge must appoint?

c. Who suspects the judge must appoint someone?
d. Which guardian suspects the judge must appoint someone?

In a typical acceptability judgment experiment, a participant would see just one sentence token from (1a–d). At the same time, they would also need to see sentence tokens from all the possible conditions.[2] Therefore, there must be other groups of related sentence tokens, called *item sets* or just *items*. In a typical experiment, the number of item sets a researcher uses is a multiple of the number of conditions. For example, if there are four conditions, the researcher may create sixteen to forty item sets, depending on the sensitivity of the experiment and the expected effect size. Example (2) is another item set with the same four conditions as are in (1).

(2) a. Who was Tom worried that the patient couldn't recognize?
b. Which relatives was Tom worried that the patient couldn't recognize?
c. Who was worried that the patient couldn't recognize anyone?
d. Which relatives were worried that the patient couldn't recognize anyone?

While the items (1) and (2) instantiate the same abstract template, they necessarily vary in lots of particulars. The range and structure of that variation is determined by the researchers, who have to make decisions about what differences are allowable. They are guided by considering how differences in individual words or sentences could interact with the contrasts that the conditions are set up to measure, and whether they could help, or hurt, in obscuring from the participant the design of the experiment. For example, just comparing (1) versus (2), we can ask: is it OK that the embedded predicate is negated in (2) but not in (1)? Or that the embedding predicate is verbal in (1) but adjectival in (2)? Should the grammatical number of the *wh*-phrase be allowed to differ, as it does in (1) and (2), or should it be the same across all items?

Documenting not only the answers, but also the rationale, to such questions of item design is thus part of the data generated by the experiment. In the next sections, we consider ways to organize and preserve this data for the many other questions that arise regarding presentation of materials, selection of participants, plans of analysis, and other situations. Table 46.1 lists some of the types of data that will routinely be generated in an experimental syntax study and that will be expanded later in this chapter. Along its rows are the phases of data generation. Along its columns, I've divided the data types into *process* versus *product* data. By *product* I mean the data we need to evaluate a researcher's claims. And by *process* I refer to the information we need to exactly reproduce a researcher's study. This distinction is only heuristic, but it helps, I think, to give a sense of the complexity of even a "simple" acceptability judgment experiment.

The strongest possible view is that both the process and the product data from all phases of a study should be archived, disseminated, and cited. This view is not, I believe, deeply at odds with community standards that have emerged in psycholinguistics and experimental syntax in recent years. For some time now it has been standard to make at least the item sets public (some journals require it). And it has become increasingly common for reviewers to also request a version of the data themselves and perhaps the analysis scripts. Barriers to a more total archiving practice in this realm of experimental syntax are less social[3] and more practical—it takes

Table 46.1
Data-generating phases, and data types, in an experimental syntax study

Phase	Processes	Products
Design	Generation rubric/guidelines Lexical sets Norms, corpus queries Counterbalancing dimensions	Item sets Text Audio/video files Images
Execution	Scripts or other program data Presentation parameter settings	Experimental observations Demographic surveys Debriefings
Analysis	"Pipeline" scripts (cleaning, formatting, and shaping data) Scripts to explore, summarize, and model data	Descriptive and inferential statistics, both summative and item/participant specific Data charts or visualizations

time and forethought to collate and annotate the data. What privacy issues that exist are slight and can be solved straightforwardly by anonymizing the data sets. (Incidentally, these reasons make a basic acceptability judgment experiment a great arena for a beginning investigator to explore, play around with, or polish their familiarity with the resources and tools of data management.)

2 Data related to the design of a study

2.1 Item sets

Item sets consist of particular lexicalizations of abstract syntactic templates, as illustrated in (1) and (2). The ultimate product of the design phase of an experiment consists of a list of all item sets to be used. Practically this often exists in two forms: a *human-readable file*, which the experimenters edit themselves, and a *machine-ready file* formatted for the experiment presentation software. Preserving the human-readable file is arguably the most helpful for future researchers, who may wish to use different software, as a matter of preference, or in case the original software has become deprecated or otherwise difficult to deploy. The human-readable data should be stored in a delimited text file format[4] so that it can be directly transformed into the format required by the presentation and acquisition software; our lab uses CSV files, with commented header fields. While some currently used types of presentation software use input files with minimal formatting, such Linger (Rohde 2003), others can attain greater flexibility with more extensible, and complex, formats. Ibex (Drummond 2013) uses JSON, an open standard and language-independent file format that is organized by attribute-value pairs and that supports array data. Figure 46.1 illustrates both formats for two sentence tokens. The JSON data structure can be more richly annotated, and here that richness serves to preserve structured information about how response options were presented.

While item sets themselves are product data, there is a broad array of process data that describe the history of the item set. Most common are *norms*—data about the particular lexical items and grammatical constructions used in the study, values such as frequency or familiarity. For the tokens illustrated in figure 46.1, my collaborators and I collected information about the transitivity of the verb *praise*: how often that verb appears with and without a direct object in various corpora. While such norms could mainly serve to filter items during the design process, they might also be useful to future investigators who wish to conduct further analysis on the data. They should be preserved, therefore, to increase the transparency of the study. The norm itself, as a value, can be added in the materials file as an extra field. But it is also necessary to include descriptions of how the norms were collected—such data as corpus queries or even survey data. Indeed, norms are routinely derived from small experiments that precede the main one (see, e.g., Gahl, Roland, & Jurafsky 2004).

```
Itm,Cnd,Sentence
4,a,Who do you assume that the controversial politician would praise?
4,c,Which author do you assume that the controversial politician would praise?
```

```
[["Aa",4], DS, {s: "Who do you assume that the controversial politician would
praise?"},Question,{q: "Please indicate your confidence",as: ["Very
confident","Somewhat confident","Not confident"],randomOrder:
false,presentHorizontally: false}],
[["Ac",4], DS, {s: "Which author do you assume that the controversial
politician would praise?"},Question,{q: "Please indicate your confidence",as:
["Very confident","Somewhat confident","Not confident"],randomOrder:
false,presentHorizontally: false}]
```

Figure 46.1
Two sentence tokens in human-readable and machine-ready formats. In the upper panel, two sentence tokens from item set 4 are given. In the lower panel, the same tokens are embedded in the JSON array used by Ibex.

Another kind of process data in an acceptability judgment experiment comes from counterbalancing. *Counterbalancing* is the process by which features of the materials that are incidental or orthogonal to the design of the experiment are "balanced" against one another in a particular distribution—usually a uniform distribution, but they can also be matched to some relevant theoretical or empirical distribution (e.g., natural occurrence rate in a corpus). For example, item sets (1) and (2) include sentences with a proper name in matrix subject position, but the item set in figure 46.1 uses the local pronoun *you*. That such variation is allowed is the result of an experimenter decision to use several different subject types in the experiment, each in an equal number of item sets. For the study I've been describing, there were at least seven dimensions of counterbalancing and those dimensions were grounded in particular justifications, judgments, or prejudices–which another experimenter might not have made or held. Ideally, therefore, those justifications would be included and preserved as part of the data archive. Such data are no longer the discrete single values or expressions easily coded in a spreadsheet. They are qualitative: arguments, citations, and persuasive prose. As such, the format for preservation will be more open-ended. I prefer plain text README files structured in Markdown.

2.2 Presentation parameters

Presentation parameters are critical to the reproducibility of an experimental syntax study. For example, were sentences presented for free inspection for an unlimited amount of time on a screen? Or, were they presented in auto-paced text blocks? If they were auto-paced, how long did each chunk remain on screen? Or, were the items printed as a list in a survey? How were they randomized? How were the response options framed, and how long could the participant deliberate in making their selection?

An obvious way to preserve these data would be to archive the scripts or program files themselves used in presenting the stimuli and acquiring the data. This is advisable, but it may not be sufficient if the programs become impractical or impossible to use in the future, or simply because they may not be interpretable by everyone who would like to access the information. This concern is genuine, especially if proprietary software is used. A researcher may additionally want to consider a writing a simple "parameters" file, or folding this information redundantly into the README file.

In this context I'd like to mention pre-registration as another tool that can help to tie together the preservation of both intangible design process factors (like why certain counterbalancing procedures were adopted) and other more straightforward ones (like a list of presentation parameters). *Pre-registrations* are standardized and immutable design-and-analysis-plan documents that are registered and deposited on a server before an experiment is run. The stated reason for having such documents is to provide a public basis for distinguishing between confirmatory and exploratory analyses of a data set; as such, they can be quite bare-bones—just brief answers to a few questionnaire prompts. However, they can also be granular documents that encode more of the researchers' intentions about how the experiment will be executed. In this guise, I've found them to be useful "drafts" of the process data in an experiment (as long as any departures from the pre-registration are recorded later). Some tools, such as AsPredicted (https://aspredicted.org), are purposely very simple and user-friendly; others are more powerful or customizable, such as the facilities offered by the Open Science Framework (OSF).

3 Data collected from individual participants

3.1 Target-dependent measures

An acceptability judgment experiment generates trial-by-trial observations about the judgments participants make and, often, how long they take to make them. These direct observations are paired with the information about exact item presentation order. These observations are usually saved by the data acquisition software into files corresponding to individual participants, or to whole studies. As with presentation parameters, it is advisable to archive this "raw" data, in tandem with a reshaped version of it.

How those data are shaped depends on the analysis goals of the experiment. In a language experiment, a "long" tabular format is often the most conducive to an analysis sensitive to subject and item random effects: in this format, each row of a spreadsheet consists of a single observation or the data produced by one trial. The columns correspond both to metadata that index the trial—for example, participant, item, condition, ordinal position in the experiment/block—and to the target-dependent measures—for example, binomial judgment, or rating, judgment time, confidence. Additionally, some software may compute derived measures automatically, such as correctness of the response.

The reshaped version should be preserved in an open standard format. In most experimental syntax and behavioral psycholinguistics experiments, a plain text CSV file is the most suitable.

3.2 Observational and survey data

Information about the participants themselves are typically collected via a survey and includes such details as the participant's age and language background. Other kinds of questions are possible, though less routinely relevant in experimental syntax—information such as gender, handedness, use of corrective lenses or hearing aids, and so on. There is generally no cause to retain information that is directly personally identifiable, such as a name or ID number. These may serve a temporary administrative function, such as to grant course credit, or to pay someone on a web platform, but the information necessary to serve that function should be kept separate from the target data and destroyed as soon as feasible. Special care should be taken that any (de-identified) biographical details relevant to the analysis—and thus in need of being preserved—be stored separately from personally identifiable information.

A final kind of observational data collected in an experimental syntax study comes from the debriefing. The debriefing is the portion of the experiment in which the researcher discusses the goals or topics of the experiment and potentially reveals any misdirection (which is, frankly, uncommon in experimental syntax). In a structured debriefing, participants can be encouraged to discuss their perception of the experiment, what they believed it to get at, or whether/how they developed a response strategy. The researcher may seek targeted feedback about particular experimental items (e.g., sentences, constructions). In my own experience, the debriefing is most routinely viewed as a pro forma requirement, one ensuring that experimenters discharge their ethical obligations to the participant. But I think it would behoove us all to embrace it as another data-gathering opportunity (see, e.g., Blanck et al. 1992). The details that participants volunteer about their response strategies may be especially germane when extra-syntactic extra-grammatical factors are already theoretically implicated.[5] As such, preserving them can provide valuable clues to other investigators, particularly if future attempts at replication do not succeed. When my collaborators and I have been careful enough to collect detailed debriefing notes, we simply record them in a narrative-style text file. However, this is not a particularly consistent standard, so there are open questions, and opportunities, about how to better manage this data, and how to incorporate it with the more quantitative observations that are easily encoded in the tabular format.

4 Data related to the analysis of a study

The only reliable way to precisely and unambiguously replicate the analysis that underwrites particular research conclusions is to have access to the exact scripts that the original researcher used. Published research reports provide some of the products of this analysis, such as tables of descriptive statistics, data charts, or regression tables. However, behind those "surface" representations of the data are many decisions the experimenter made about transforming the raw data—questions of scaling, of participant exclusions, of outlier policies, among others. These decisions are sometimes reported in an article, but there is still often vagueness or ambiguity about what exactly occurred. This is much like the issues discussed earlier with materials design: for maximum transparency, and total replicability, the process data must be preserved as well.

In experimental syntax and psycholinguistics, the R language/environment is the de facto standard for statistical analysis (R Core Team 2019). It is free software distributed under the GNU General Public License. At a minimum, therefore, researchers should archive the analysis scripts, commented amply and clearly, that link specific chunks of code to parts of the published text (figures, tables, results paragraphs). My collaborators and I use R Markdown (Allaire et al. 2018; Xie, Allaire, & Grolemund 2018), a literate programming tool that allows documentation in a plain text markup language to be "woven" together with R code chunks. Archiving these R Markdown files with the target data themselves is a nearly complete solution for replicability.

5 Data format and dissemination

For experimental syntax, and many behavioral psycholinguistic studies, file formats can usually be kept relatively simple. In all cases, open formats are preferred (see Mattern, chapter 5, this volume, and Han, chapter 6, this volume). For the lightly reshaped cumulative data files, as well as for the item files or norms, CSV-formatted text is usually sufficient. And for supporting process materials, such as explanations of norming decisions or debriefing records, plain text is still preferable for its light footprint and searchability. Here my collaborators and I construct Markdown-formatted README files.

In more recent research, when we have been especially conscientious of reproducibility, my collaborators and I have begun using the version-control system Git (https://git-scm.com; see Chacon & Straub 2014) during the creation and revision of materials, writing or customization of experiment presentation scripts, and in the analysis phase. Version control is not strictly necessary for reproducibility, but it helps the researcher to track changes and alternative versions, in other words, to create a record of the process of design and analysis itself. (And the Ibex software package currently can integrate directly with Git repositories, an attractive feature for managing the transition from design to execution).

For dissemination I have taken advantage of two resources. The first is an institutional tool, the California Digital Library (https://www.cdlib.org), owned and therefore supported by the University of California. The second is the OSF (https://osf.io). OSF is a powerful platform because it brings together pre-registration, data storage, and pre-print/post-print hosting. Static snapshots of a project can be registered and associated with a digital object identifier (DOI) as well. The power of the DOI in "setting down a marker" for tracking and identifying a contribution before publication has served as a real impetus to making myself and my collaborators more open to publicly sharing data, even during review (an option that may vary as a function of journal requirements and whether authors are willing to unmask themselves during review). These snapshot versions of the project can be archived redundantly in another repository as well. In the references, I have pointed the reader to two sample OSF repositories (Dillon, Wagers, & Andrews 2018; Wagers & Chung 2018). See Andreassen (chapter 7, this volume) for greater discussion of these matters.

6 Taking stock

In this use case, I have moved through three phases of an acceptability judgment study: design, execution, and analysis. In each phase, a considerable amount of data is generated that relates not only to the proximal inputs and outputs of the experiment—that is, the items or the observations—but also to records of the researchers' decision making. Greater transparency is achieved if access is preserved to both classes of data, which are potentially more informative when interpreted jointly. It is probably impossible to keep an exhaustive and strictly veridical record of the process. But if researchers are tuned to the data-generation events that occur across the life cycle of an experiment—and create a data management plan of even modest scope (see Kung, chapter 8, this volume)—it should be possible to produce data archives and documentation in experimental syntax that more closely approach reproducibility.

Notes

1. Specifically, this design tests the effect of *discourse-linking* or *D-linking;* see Goodall (2015) for background. For purposes of illustration, I have simplified and adapted the design of an unpublished actual experiment conducted with my collaborator Brian Dillon (University of Massachusetts, Amherst), whose permission I gratefully acknowledge in describing it here. The item sets may be accessed here: https://github.com/mattwagers/killmonger.

2. Most studies in experimental syntax are *within subjects* and *within items*. What this means is that each participant (subject) is exposed to all conditions, ranging across items, and that each item set is associated with data from all conditions, ranging across subjects. There are definite advantages to a *within subjects, within items* design. For example, it can moderate strategic responding, and it can lead to improved statistical power. But it may not be appropriate for all questions or adaptable to all linguistic expressions.

3. Undoubtedly reticence does exist when it comes to sharing all details of the analysis, and I trace some of that to insecurity or uncertainty about doing a good job. Standards of statistical analysis have changed rapidly in just the past decade; a comparison of Baayen's (2004) influential critique and recent "state of the art" reviews by Vasishth and Nicenboim (2016) and Nicenboim and Vasishth (2016) give a sense of that. Linguists, and the academic units they belong to, are all grappling with how best to obtain or implement the training necessary to achieve stronger standards.

4. Here we focus on experiments that use text. Audiovisual presentation may be better suited to the experimental question at hand or to the population participating in a study. Nonetheless, an items file, as a record of the structure of the experiment, will still be a plain text file with fields: those fields could directly contain the sentences being tested or file names for the digital versions of that material.

5. As an example—here from "experimental semantics"—Anand et al. (2011) investigated the influence of monotonicity on the exclusive interpretation of plural nominals. In one of their experiments, some trials paired images to be verified with quantificational sentences whose domains were null; for example, a sentence such as "Each castle surrounded by a moat is gray" paired with a picture in which no castle was surrounded by moat. They discovered, only via debriefing, that there were some participants who accommodated these null quantificational domains, while others consistently rejected them as presupposition violations. When the data were split based on the debriefing report, presupposition-violation responders were found to be much more likely to interpret plurals as exclusive.

References

Allaire, J. J., Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, and Winston Chang. 2018. *RMarkdown: Dynamic Documents for R*. R package version 1.9. https://CRAN.R-project.org/package=rmarkdown.

Anand, Pranav, Caroline Andrews, Donka Farkas, and Matthew Wagers. 2011. The exclusive interpretation of plural nominals in quantificational environments. In *Proceedings of the 21st Semantics and Linguistic Theory Conference*, ed. Neil Ashton, Anton Chereches, and David Lutz, 176–196. Linguistic Society of America. https://doi.org/10.3765/salt.v21i0.2617.

Baayen, R. Harald. 2004. Statistics in psycholinguistics: A critique of some current gold standards. *Mental Lexicon Working Papers* 1 (1): 1–47.

Bard, Ellen Gurman, Dan Robertson, and Antonella Sorace. 1996. Magnitude estimation of linguistic acceptability. *Language* 72 (1): 32–68.

Berez-Kroeker, Andrea L., Lauren Gawne, Susan Smythe Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, et al. 2018. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics* 56 (1): 1–18.

Blanck, Peter David, Alan S. Bellack, Ralph L. Rosnow, Mary Jane Rotheram-Borus, and Nina R. Schooler. 1992. Scientific rewards and conflicts of ethical choices in human subjects research. *American Psychologist* 47 (7): 959–965.

Chacon, Scott, and Ben Straub. 2014. *Pro Git*. Berkeley, CA: Apress.

Cowart, Wayne. 1997. *Experimental Syntax: Applying Objective Methods to Sentence Judgments*. Thousand Oaks, CA: Sage.

Dillon, Brian, Caroline Andrews, Caren M. Rotello, and Matthew Wagers. 2018. A new argument for co-active parses during language comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 45 (7): 1271–1286. doi:10.1037/xlm0000649.

Dillon, Brian, Matthew Wagers, and Caroline Andrews. 2018. *A New Argument for Co-Active Parses during Language Comprehension*. OSF. August 25. doi:10.17605/OSF.IO/SD3HU.

Drummond, Alex. 2013. *Ibex*. Version 0.3.6. https://code.google.com/archive/p/webspr/. Accessed April 1, 2019.

Gahl, Susanne, Douglas Roland, and Daniel Jurafsky. 2004. Verb subcategorization frequencies: American English corpus data, methodological studies, and cross-corpus comparisons. *Behavior Research Methods, Instruments and Computers* 36 (3): 432–443.

Goodall, Grant. 2015. The D-linking effect on extraction from islands and non-islands. *Frontiers in Psychology* 5:1493. doi:10.3389/fpsyg.2014.01493.

Nicenboim, Bruno, and Shravan Vasishth. 2016. Statistical methods for linguistic research: Foundational ideas—Part II. *Language and Linguistics Compass* 10 (11): 591–613.

R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. https://www.R-project.org/.

Rohde, Doug. 2003. *Linger*. Version 2.88. Available from https://web.archive.org/web/20191220181934/http://tedlab.mit.edu/~dr/Linger/.

Schütze, Carson T. 1996. *The Empirical Base of Linguistics*. Chicago: University of Chicago Press.

Sprouse, Jon. 2007. A program for experimental syntax. PhD dissertation, University of Maryland.

Vasishth, Shravan, and Bruno Nicenboim. 2016. Statistical methods for linguistic research: Foundational ideas–Part I. *Language and Linguistics Compass* 10 (8): 349–369.

Wagers, Matthew, and Sandra Chung. 2018. *Chamorro Relative Clause Processing*. OSF. May 31. doi:10.17605/OSF.IO/B8ZWQ.

Xie, Yihui, J. J. Allaire, and Garrett Grolemund. 2018. *R Markdown: The Definitive Guide*. Boca Raton, FL: CRC Press.

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