Journal of Open Humanities Data

Language Dataset Reuse in Methodological Research on Down-Sampling Corpus Data

DISCUSSION PAPER

LUKAS SÖNNING

ubiquity press

ABSTRACT

This paper looks at opportunities and challenges of language dataset reuse in methodological research, using the development of down-sampling designs for corpus-linguistic work as a case study. This research is concerned with the construction and evaluation of strategies for down-sizing data that are too large given the resource constraints confronting a study. Work in this line of inquiry is heavily dependent on real datasets to field-test and compare techniques, which makes it a prime example of how open data can serve as a catalyst for methodological progress. In the reuse case discussed here, a methodological study made use of an informally published dataset. The original data were revised and modified, and the derived dataset archived in the Tromsø Repository of Language and Linguistics (TROLLing). The current paper draws attention to a number of questions that may arise when repurposing language data, including issues related to licensing and the question of how to determine authorship credit. It also demonstrates how depositors and reusers of data can profit from a careful review process, which is ideally assisted by trained data curators. The case study points to TROLLing as an exemplary repository that offers quality control and substantive feedback on a host of aspects including dataset documentation as well as legal and research-ethical matters.

CORRESPONDING AUTHOR:
Lukas Sönning
University of Bamberg, Germany
lukas.soenning@uni-bamberg.de

KEYWORDS:
methodological research; down-sampling; corpus linguistics; data curation; licensing; TROLLing

TO CITE THIS ARTICLE:
Sönning, L. (2026). Language Dataset Reuse in Methodological Research on Down-Sampling Corpus Data. *Journal of Open Humanities Data*, 12: 78, pp. 1–10. DOI: https://doi.org/10.5334/johd.531

DOI: 10.5334/johd.531

(1) INTRODUCTION

In the discourse surrounding open science practices, the importance of data sharing is usually linked to research transparency and reproducibility. Apart from these primary functions, open data also play an essential role for methodological research. This discussion paper focuses on the role of publicly accessible datasets for research on the improvement of down-sampling practices in corpus linguistics, an empirical branch of the language sciences that draws on large, digitized collections of naturally produced written and spoken texts. I reflect on some of the experiences made in the context of this project and demonstrate that dataset reuse holds potential that goes far beyond the provision of case studies for methods papers. Instead, an open data culture can challenge and ideally elevate methodological work to offer more circumspect, robust and practicable insights to the community.

The case study discussed in the current paper illustrates some issues that may arise when repurposing datasets and points to preventive measures that can smoothen language data reuse. A case in point is the choice of a peer-reviewed and carefully curated repository, to ensure quality control and standardized documentation. This form of 'proper' data publication not only pays attention to the FAIR[1] principles to facilitate the future reuse of data, but trained staff also offer assistance with legal aspects surrounding data sharing.

The paper is structured as follows. Section 2 discusses the relevance of open data for methodological research and provides some background on down-sampling in corpus-based work. Section 3 introduces the Tromsø Repository of Language and Linguistics (TROLLing), an exemplary data archive that will feature in the specific reuse case discussed in the current paper. Section 4 sketches the key parameters of a derived dataset that has emerged from this research program and which has been published and documented thoroughly on TROLLing. Section 5 outlines the steps taken in this particular reuse case and Section 6 describes some of the experiences made. Implications and recommendations for good practices in language data archiving and reuse are then discussed in Section 7.

(2) CONTEXT AND MOTIVATION

The current section sets the scene by considering the relevance of open data for methodological work (Section 2.1). Following this, I provide some background on the illustrative context, the development of down-sampling designs for corpus research (Section 2.2).

(2.1) METHODOLOGICAL RESEARCH AND OPEN DATA

One way in which methodological work benefits from shared data is through the availability of demonstration material for showcasing certain problems or a particular procedure. The current section argues that open data hold out further opportunities. Apart from enriching methods teaching, they have the potential to substantiate and transform methodological research.

Once a field develops interest in a new method, researchers not only profit from seeing illustrative applications but they usually require some practical training. To flatten the learning curve, materials ideally rely on a dataset to which a particular target audience, say, the participants of a hands-on workshop, can relate. In the absence of relevant open data, the audience must accept the instructor's choice of applications. Open data are therefore a valuable resource that can make methodological training and teaching more effective.

When studying a new or existing method, ideas and usage scenarios often work well in the abstract, or on paper. Efforts to transfer know-how from other disciplines, for instance, can invite premature conclusions as to the utility of a particular tool or technique for language data. For these reasons, it is often helpful – if not essential – to confront a method with actual data from the target domain(s) and observe whether it works as intended. This appears to apply in particular to observational disciplines, where the distributional characteristics of data may pose distinct and unanticipated challenges. A case in point are corpus data, which form the basis of the methodological case study discussed further below. The second added benefit of open data, then, is that they provide a testing ground for methodological ideas.

1 FAIR stands for findable, accessible, interoperable and resuable.

It is not unusual for methods papers to rely on data drawn from the author's own academic repertoire. A world in which methodologists primarily apply their ideas to research problems they are acquainted with is undesirable for at least two reasons. First, it is easy to see how certain cognitive biases may interfere and cloud our judgment. If the method in question yields results that contradict our earlier published findings, this may negatively affect the impartiality brought to the evaluation task and could create a methodological "file drawer problem", i.e. a form of publication bias (see Rosenthal, 1979) where the result of a study determines whether it enters the scientific record or is "filed away". Further, and perhaps more importantly, if methodologists do not look beyond their domain of linguistic expertise, the general utility of a method may remain unclear. The research community, however, needs reassurance as to the actual applicability and effectiveness of new techniques. Importantly, then, open data allow us to consider what may be referred to as the ecological validity of methods.

In summary, the relevance of data sharing for methodological research is not limited to the provision of material for case studies. Rather, an open data culture fosters more effective learning environments and has the potential to discipline and transform the way we work on methodological problems. The research program introduced in the next section illustrates some of these added benefits.

(2.2) ILLUSTRATIVE RESEARCH PROGRAM: DOWN-SAMPLING CORPUS DATA

The case study discussed in the current paper is drawn from ongoing work on down-sampling methodology. Section 2.2.1 describes the general motivation driving this line of research and Section 2.2.2 discusses the essential role played by open data.

(2.2.1) Down-sampling: Background

Corpus queries often return a list of hits (i.e. occurrences of the target structure) that is too large given the purpose and resources of a particular study. This is usually the case if the analysis involves some form of manual disambiguation or coding, where each occurrence must be considered and categorized individually. Researchers are then forced to restrict their attention to a subset of the data. The term down-sampling refers to the down-sizing of a pool of observations to a manageable subset.

Methodological work on down-sampling is concerned with the question of how to reduce the pool of observations in light of the stated goals of the study (see Sönning and Krug, 2022; Sönning, 2024, 2026a, 2026b). The standard approach, which is implemented in many corpus analysis tools, is to draw a random sample. The objective of down-sampling research is to devise more effective schemes, which are informed by (i) the goals of the analysis and (ii) the structure and distributional features of the data. The essential question is this: If I were able to analyse (say) only 2,000 tokens from the entire corpus, how should I go about selecting these 2,000 cases from the larger pool?

Methodological work on down-sampling can build on insights from disciplines such as sampling theory (e.g. Lohr, 2022), the design of experiments (e.g. Mead et al., 2012), and case-control research (Breslow, 1996). It involves two main elements. The first is to establish an inventory of design features, i.e. different (optional) components of the selection scheme (see Sönning, 2026b, p. 4). This search is informed by the related statistical literature and one's understanding of the general nature of corpus data. Design features can be combined in various ways to yield a concrete down-sampling plan, i.e. a selection procedure that is applied to the pool of observations.

The second element is the evaluation of down-sampling designs on actual data. This could involve the comparison of a custom scheme with the default approach (i.e. a simple random sample). For this, we need an existing dataset that hasn't been down-sampled yet. Suppose this dataset includes 10,000 observations. We then conduct a thought experiment and ask: Assuming we had only been able to analyse 2,000 observations (i.e. 20% of the full dataset), which down-sampling plan would have been preferable?

An evaluation study typically starts with an analysis of the full dataset, to have a point of reference. Then we ask which down-sampling plan gets us closer to this benchmark. This

evaluation is based on a simulation study that applies the competing schemes repeatedly, say 500 times, and then determines which strategy yields results that are, on average, closer to those from the benchmark analysis. This decision is based on several performance criteria, which are outlined in Sönning (2026b, p. 15). A simulation approach therefore allows us to see which design is likely to fare better with respect to a particular set of data.

Experience has shown that it is essential to observe the behaviour of down-sampling designs on actual corpus data. As mentioned above, certain things that appear clear in principle aren't so in practice. To give a concrete example, Sönning (2024, p. 525) considers it a "logical necessity" that an analysis of the full dataset should return estimates with higher statistical precision (i.e. narrower confidence intervals) than a down-sampled dataset. However, exceptions to this have been observed (Sönning, 2024, p. 525; Sönning, 2026b, p. 23), which may indicate that certain features of the data and/or method are not yet fully understood.

(2.2.2) Relevance of open data

To monitor how different down-sampling designs perform in the real world, a concrete dataset is required. Progress in this area of corpus-linguistic methodology is therefore very much dependent on the availability of open data. This is for two reasons.

First, the creation of datasets for this purpose is laborious – after all, to be of direct relevance, the full dataset must be sufficiently large. Its creation should also ideally involve some sort of manual work, i.e. a bottleneck that puts a ceiling on the manageable number of observations. In the early stages of down-sampling research, we followed this approach and created data, as it were, from scratch (Sönning & Krug, 2022; see Sönning & Krug (2021) for the associated dataset). This, however, unduly delays methodological progress.

Further, dataset reuse has allowed us to extend the range of research problems for the evaluation of down-sampling designs to probe the broader validity of insights. While Sönning and Krug (2022) were concerned with a sociolinguistic study of a discourse marker in Present-day British English, Sönning (2024) considered a diachronic investigation of an inflectional category in Early Modern English, and Sönning (2026b) turned to a socio-phonetic study on voice onset time (VOT) in British English, an acoustic correlate of the voicing contrast in onset stop consonants. Broadening the spectrum of testbeds not only enhances the evidence base for methodological arguments but also helps popularize techniques in different, potentially unrelated, fields of study that nevertheless confront the same empirical challenges.

This extended focus was only possible due to the availability of two datasets. Data for the socio-phonetic VOT study are from Sonderegger et al. (2017) and associated with the textbook by Sonderegger (2023); they have been published on GitHub under a CC BY 4.0 license (Sonderegger 2022). Similarly, the diachronic data on verb inflection were compiled and used by Jenset and McGillivray (2017) for their monograph and likewise published on GitHub, under an MIT license (Jenset, 2018). The latter data reuse case, which is discussed in more detail further below, led to the creation of a derived dataset, which was published in the TROLLing repository. The next section provides some background on this language data archive.

(3) LANGUAGE DATA ARCHIVE: TROLLing

The Tromsø Repository of Language and Linguistics[2] (TROLLing) is a domain-specific archive for linguistic data that operates within the *Dataverse* infrastructure and is hosted at UiT – The Arctic University of Norway (see, e.g., Andreassen, 2022). One of its central strengths, from a reuse and reproducibility perspective, lies in the structured curation workflow that accompanies data publication. Rather than functioning as a passive storage space, TROLLing actively supports researchers in preparing datasets that are well documented, reusable, and compliant with the FAIR principles.

Publishing a dataset with TROLLing involves the creation of a dedicated repository entry that combines data files with rich metadata. A key component of this package is the 'ReadMe' file, which is compiled using a standardized template developed for linguistic datasets (Conzett & Dijkstra Haugstvedt, 2024). This document serves as the main point of orientation for

2 https://dataverse.no/dataverse/trolling.

prospective users. It typically contains a high-level description of the dataset, information about data collection and processing methods, and detailed codebooks explaining the structure and contents of each data table. To support this process, TROLLing provides extensive online documentation and practical guidance for data depositors.

Once a dataset has been submitted, it enters a formal curation phase. At this stage, a trained data curator reviews the submission and provides feedback in the form of a curation report (see Supplement 2 for an example). This report identifies aspects of the dataset and its documentation that may benefit from clarification, correction, or expansion. The review addresses technical considerations relevant to FAIR compliance (such as file formats), the clarity, consistency, and transparency of the ReadMe documentation, as well as legal and ethical issues related to data sharing. Because TROLLing is specifically oriented toward the language sciences, curators are also able to comment on discipline-specific aspects of data and method description. After successful completion of this process, the dataset is assigned a clear reuse license and a persistent DOI, enabling stable citation and long-term accessibility.

Further below in Section 6, I will reflect on the particular reuse case introduced shortly. There, it will become clear that TROLLing not only provided essential support along the way, but its standardized deposit schemes ensure that many of the issues plaguing conscientious dataset reuse can be straightened through appropriate archiving of an original set of data.

(4) DESCRIPTION OF THE DATASET

The derived dataset that forms the outcome of the reuse case discussed in this paper is adapted from Jenset and McGillivray (2017). It contains tabular files recording observations of the alternation between two forms, *-(e)th* and *-(e)s*, to mark the inflectional category 'third-person singular present' on verbs in Early Modern English (*maketh/makes*, *forgiveth/forgives*). The original data, which were discussed in Jenset and McGillivray (2017, p. 190–206), are drawn from the "Penn-Helsinki Parsed Corpus of Early Modern English" (PPCEME; Kroch et al., 2004) and cover the period from 1500 to 1700. A total of 13,757 third-person singular tokens (excluding the verb BE) were annotated by these authors for a number of variables. As described in more detail in Section 5, the original dataset was reduced to a subset of 11,645 observations, and the coding of variables was in some parts revised, completed, or modified. The following listing gives the key parameters of this dataset; for more detailed documentation, see Sönning (2023). The resulting derived dataset served as a basis for the evaluation of down-sampling designs in Sönning (2024), where different strategies for drawing a subset of 2,000 tokens (out of the complete pool of 11,645 observations) were evaluated.

REPOSITORY LOCATION

https://doi.org/10.18710/5KCE4U

REPOSITORY NAME

The Tromsø Repository of Language and Linguistics (TROLLing)

OBJECT NAME

data_jenset_mcgillivray_downsampling.tsv

FORMAT NAMES AND VERSIONS

UTF-8-encoded, tab-delimited data table

CREATION DATES

Start date: 2022-11-15; end date: 2023-06-15

DATASET CREATORS

Gard B. Jenset (dataset creator, Springer Nature); Lukas Sönning (creator of derived dataset, University of Bamberg)

LANGUAGE

English

LICENSE

MIT License and CC BY 4.0 (see discussion in Section 6)

PUBLICATION DATE

2023-10-24

(5) REUSE PROCESS AND METHODOLOGY

For the purposes of the down-sampling study, I undertook a series of data selection, recoding, and restructuring steps. These were implemented during the data preparation phase of the reuse process and are fully reproducible via an R script that forms part of the dataset publication described in Section 4. In what follows, I outline the main stages of this preparation, focusing on the decisions that were necessary to adapt the existing dataset to the specific goals of the secondary analysis (i.e. the methodological study). Further details may be found in the ReadMe file associated with the derived dataset.

The preparation process comprised four broad components: filtering observations with incomplete information, revising the annotation of one particular variable, collapsing selected categories on certain variables, and applying a small number of additional adjustments to ensure consistency across the down-samples drawn during the simulation study. Each of these steps is discussed in turn.

First, I restricted the dataset to observations that contained complete information on the variables 'Author' and 'Verb', which are central to the methodological study; tokens lacking an author identifier or a verb lemma were excluded. In addition, I removed observations associated with author IDs that represent collective entities rather than individual writers, such as translator groups documented in the corpus metadata. After these exclusions, the resulting dataset comprised 11,645 tokens.

Second, the variable 'Phonological context' required revision. It captures properties of the segment preceding the suffix in question, but the annotation provided in the original dataset was incomplete. As part of the reuse process, I therefore reannotated this variable in its entirety. This involved consulting the "Irvine Phonotactic Online Dictionary" (IPhOD; Vaden et al. 2009) and carrying out manual checks and classifications where necessary. The full procedure, including decision criteria and corrections, is documented in the data preparation script.

Third, to facilitate statistical modeling and interpretation, I simplified the structure of two categorical predictors. 'Phonological context' was reduced to a binary distinction between sibilant-final stems and all other phonological environments. Similarly, the variable 'Genre' was collapsed into two broad categories reflecting relative formality. Text types such as biblical writing, sermons, legal texts, and scientific or educational works were grouped as more formal, whereas letters, diaries, travelogues, drama, and fiction were classified as more informal. Importantly, the original, more fine-grained genre labels were retained in the derived dataset for transparency and potential alternative uses.

Finally, I implemented a small number of additional adjustments aimed at aligning the structure of the dataset with the design of the simulation study. In particular, 'Genre' and 'Year' were treated as strictly between-author variables. To enforce this, I removed observations for a single author whose contributions spanned multiple genres. In addition, for a small set of authors associated with multiple publication years, I replaced individual year values with the author-specific mean year of publication. These modifications ensured that each author was associated with a single text category and a single temporal reference point. The original publication years remain available in the derived data tables, allowing future users to revisit or revise these decisions if needed.

(6) OUTCOMES AND EXPERIENCE

To take a step back for a moment, two fortunate circumstances led to the use of Jenset and McGillivray's (2017) data for my work on down-sampling. The focus on the diachrony of third-person verb inflection is due to Ole Schützler, who at the time worked on this alternation for a talk at the 2023 *International Conference on Historical Linguistics* (ICHL26).[3] As we discussed down-sampling strategies for an extension of his work to Early English Books Online (EEBO), a massive collection of Early Modern English texts (about 30,000 texts and 800 million words), I remembered reading about a related case study and dataset in Jenset and McGillivray's (2017) textbook. The associated data provided a perfect context for the further development of down-sampling designs. Thus, in our earlier work (Sönning & Krug 2022), we dealt with a setting where observations were clustered by speaker. Most speakers contributed multiple tokens to the data, and we explored the consequences of this for down-sampling practice. Jenset and McGillivray's (2017) case study included two clustering variables, 'Author' and 'Verb'. The next step was to examine and compare down-sampling methods in this more complex data setting.

Since the derived dataset on which Sönning (2024) is based differs from the original in several ways, it made sense to preserve it in some form. In the interest of transparency, this would also allow others to reproduce the results of my study. I chose the TROLLing archive, which – as we will see shortly – turned out to be immensely helpful for the publication of this set of data. I contacted Gard Jenset and Barbara McGillivray to inform them about this reuse case and to ask for permission to publish the derived dataset. For the interested reader, the supplementary materials include a copy of this email.

There were two issues that I needed to discuss with the authors. First, I was unsure about how to formulate the citation for this derived dataset. It felt inappropriate to publish it under my name, so I raised this point in the email, suggesting two formats: one with myself as the third author, and one where the dataset creators and the original publication are mentioned in the title. In his swift and friendly reply, Gard Jenset wrote that, since this is a derived dataset, he would consider the latter version the most appropriate one. The dataset therefore ended up being published as Sönning (2023) with the following title: *Background data (adapted from Jenset & McGillivray 2017) for: Down-sampling from hierarchically structured corpus data*. To me, this seemed like an acceptable solution. In general, however, guidelines for this kind of setting, where a reused and modified dataset is published, currently appear to be lacking. It would therefore seem essential to reach out to the original authors and decide together how to proceed.

The second issue concerned licensing and terms of reuse, a topic that I would assume many researchers do not deal with on a regular basis. Since the publication of the derived dataset eventually needed a clear statement on these matters, the issue naturally came up during the review and curation process with TROLLing (the supplementary materials include a copy of the curation report, where this issue is discussed). It turned out that the licensing situation for the derived dataset was rather intricate, as it involved the terms of reuse formulated for the source corpus (PPCEME) and an MIT license issued for the original dataset by Jenset (2018). The assistance provided by Huw Haugland-Grange, the data curator handling my case, was therefore invaluable. After consultation with his colleagues at TROLLing, he recommended and formulated a set of custom terms for the derived dataset. It was decided to use the standard CC0 code for all components except for the tabular file "data_jenset_mcgillivray_downsampling.tsv", which represents the derived work. As for the corpus excerpts from PPCEME that formed part of the dataset, a Fair-Use-or-Fair-Dealing assessment suggested that, since they contained a limited amount of the original work, they may be shared under copyright exemptions (see, e.g., Collister 2022, p. 1211–1222). The file was therefore published under joint terms: the MIT license, which respected the terms and conditions applying to the original dataset, and the CC BY 4.0 license, which represented my contributions to the derived version. Future reuse therefore needs to comply with both sets of regulations.

To return to the original aims of down-sampling research, the reuse case at hand allowed me to expand the scope of evaluation efforts to a morphological alternation phenomenon in

3 https://www.uni-heidelberg.de/md/slav/forschung/tagungen/ichl26/ichl26_paper_35.pdf, last accessed June 3, 2026.

Early Modern English, a research setting beyond my reach. Looking ahead, this opportunity could provide a glimpse into the future: As the inventory of design features takes form and evaluation procedures become established, down-sampling designs can be evaluated at scale: on various target constructions, at diverse levels of linguistic description, across historical periods, varieties, and languages. Future studies confronting the need to down-size their data may then be able to profit from a solid knowledge base informed by the actual performance of designs on a broad variety of corpus-linguistic research problems.

(7) RECOMMENDATIONS AND GOOD PRACTICES

I would like to conclude this discussion paper with two thoughts on language data sharing and reuse. The first is an appeal to linguists who have decided to make their data available to the community: please consider publishing your data in a proper archive. As the case study discussed above showed, a repository such as TROLLing holds benefits for depositors, who receive thorough feedback on data and documentation and advice on issues related to copyright. The greatest beneficiaries, however, are future reusers, who are provided with a well-documented, FAIR-aligned data record that clearly states the terms of reuse and is straightforward to cite (on dataset citation, see Conzett & De Smedt (2022)). Where possible, use should preferably be made of a permissive license, as this facilitates downstream decisions such as those encountered in the case study discussed here.

The second thought is directed at dataset reusers, whose responsibilities arguably go beyond giving proper credit to the original authors. The process of engaging thoroughly with an existing dataset usually exposes some (often minor) things in the original files (or their documentation) that may need clarification or correction. In the current case study, for instance, I ended up recoding one of the variables in the data. It would then be desirable to integrate sensible changes and modifications to data or their documentation into an updated version of the original dataset, provided the authors grant their permission. This form of continuing curation work contributes to the sustainability and long-term utility of data, with future reusers profiting from the work of their predecessors. If the dataset is published in a proper archive, the revised repository entry then receives a new version number, with previous versions being preserved as part of the dataset's history.

To conclude, the specific language data reuse case described in the present paper has highlighted the relevance of open data for methodological work. In this way, an open data culture can contribute to methodological innovation and consolidation, to the greater benefit of the entire community. Apart from this, the case study has thrown into sharp relief the added value of carefully curated (language) data repositories, whose work and expertise pays dividends to data depositors, dataset reusers, and the open-science culture in linguistics.

SUPPLEMENTARY FILES

- **Supplement 1.** *Reuse inquiry.* A copy of the email I sent to Gard Jenset and Barbara McGillivray to ask for permission to reuse their data and publish the derived dataset. The inquiry also kindly asks for feedback on licensing and citation format. DOI: https://doi.org/10.5334/johd.531.s1
- **Supplement 2.** *TROLLing curation report.* A copy of the curation report that I received as part of the dataset review process. Among other things, it demonstrates TROLLing's attention to detail and the active assistance provided with licensing issues. DOI: https://doi.org/10.5334/johd.531.s2

ACKNOWLEDGEMENTS

I would like to thank the TROLLing team for their careful review and curation work, and, for the dataset referenced in the current discussion paper, Huw Haugland-Grange in particular, whose assistance with questions concerning licensing and data reuse was invaluable. Further, thanks are due to Gard Jenset and Barbara McGillivray, for their very friendly and generous reactions to my inquiries as to the reuse of their data.

The publication of this article was supported by the Open Access publication fund of the University of Bamberg.

FUNDING STATEMENT

Financial support from the Deutsche Forschungsgemeinschaft (DFG) is gratefully acknowledged [grant number 548274092].

COMPETING INTERESTS

The author has no competing interests to declare.

AUTHOR AFFILIATIONS

Lukas Sönning orcid.org/0000-0002-2705-395X
University of Bamberg, Germany

REFERENCES

Andreassen, H. N. (2022). Archiving research data. In A. L. Berez-Kroeker, B. McDonnell, E. Koller, & L. B. Collister (Eds.), *The open handbook of linguistic data management* (pp. 89–100). MIT Press. https://doi.org/10.7551/mitpress/12200.003.0011

Breslow, N. E. (1996). Statistics in epidemiology: The case-control study. *Journal of the American Statistical Association*, *91*(433), 14–28. https://doi.org/10.1080/01621459.1996.10476660

Collister, L. B. (2022). Copyright and sharing linguistic data. In A. L. Berez-Kroeker, B. McDonnell, E. Koller, & L. B. Collister (Eds.), *The open handbook of linguistic data management* (pp. 117–128). MIT Press. https://doi.org/10.7551/mitpress/12200.003.0013

Conzett, P., & **De Smedt, K.** (2022). Guidance for citing linguistic data. In A. L. Berez-Kroeker, B. McDonnell, E. Koller, & L. B. Collister (Eds.), *The open handbook of linguistic data management* (pp. 143–155). MIT Press. https://doi.org/10.7551/mitpress/12200.003.0015

Conzett, P., & **Dijkstra Haugstvedt, N.** (2024). *DataverseNO README file template – General (Version 2.4)* [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.10849096

Jenset, G. B. (2018). eme_3sg_data_15112015a.txt [Data file]. GitHub. Retrieved June 3, 2026, from https://github.com/gjenset/quanthistbook/commits/master/eme_v3sng_study/eme_3sg_data_15112015a.txt

Jenset, G. B., & **McGillivray, B.** (2017). *Quantitative historical linguistics: A corpus framework*. Oxford University Press. https://doi.org/10.1093/oso/9780198718178.001.0001

Kroch, A., Santorini, B., & **Delfs, L.** (2004). *The Penn-Helsinki parsed corpus of early modern English (PPCEME)* [Dataset]. University of Pennsylvania, Department of Linguistics. Retrieved June 3, 2026, from https://www.ling.upenn.edu/ppche/ppche-release-2016/PPCEME-RELEASE-3

Lohr, S. L. (2022). *Sampling: Design and analysis* (2nd ed.). CRC Press. https://doi.org/10.1201/9780429298899

Mead, R., Gilmour, S. G., & **Mead, A.** (2012). *Statistical principles for the design of experiments*. Cambridge University Press. https://doi.org/10.1017/CBO9781139020879

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Sonderegger, M. (2022). vot_rmld.csv [Data file]. GitHub. Retrieved June 3, 2026, from https://github.com/msonderegger/rmld-v1.1/commits/main/data/vot_rmld.csv

Sonderegger, M. (2023). *Regression modeling for linguistic data*. MIT Press.

Sonderegger, M., Bane, M., & **Graff, P.** (2017). The medium-term dynamics of accents on reality television. *Language, 93*(3), 598–640. https://doi.org/10.1353/lan.2017.0054

Sönning, L. (2023). *Background data (adapted from Jenset & McGillivray, 2017) for: Down-sampling from hierarchically structured corpus data (Version 1)* [Dataset]. DataverseNO. https://doi.org/10.18710/5KCE4U

Sönning, L. (2024). Down-sampling from hierarchically structured corpus data. *International Journal of Corpus Linguistics, 29*(4), 507–533. https://doi.org/10.1075/ijcl.23079.son

Sönning, L. (2026a). Case-control down-sampling in corpus research. *Corpus Linguistics and Linguistic Theory*. https://doi.org/10.1515/cllt-2025-0074

Sönning, L. (2026b). Down-sampling strategies in corpus phonology. In P. Meer & U. Gut (Eds.), *English corpus phonetics and phonology: Current approaches and future directions* [Preprint]. OSF Preprints. https://doi.org/10.31219/osf.io/5hduq_v2

Sönning, L., & **Krug, M.** (2021). *Actually in contemporary British speech: Data from the spoken BNC corpora (Version 1)* [Dataset]. DataverseNO. https://doi.org/10.18710/A3SATC

Sönning, L., & **Krug, M.** (2022). Comparing study designs and down-sampling strategies in corpus analysis: The importance of speaker metadata in the BNCs of 1994 and 2014. In O. Schützler & J. Schlüter (Eds.), *Data and methods in corpus linguistics: Comparative approaches* (pp. 127–159). Cambridge University Press. https://doi.org/10.1017/9781108589314.006

Vaden, K. I., Halpin, H. R., & **Hickok, G. S.** (2009). *Irvine phonotactic online dictionary (Version 2.0)* [Data set]. https://www.iphod.com

TO CITE THIS ARTICLE:
Sönning, L. (2026). Language Dataset Reuse in Methodological Research on Down-Sampling Corpus Data. *Journal of Open Humanities Data*, 12: 78, pp. 1–10. DOI: https://doi.org/10.5334/johd.531

Submitted: 27 February 2026
Accepted: 27 May 2026
Published: 18 June 2026

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.