Can Computational Meta-Documentary Linguistics Provide for Accountability and Offer an Alternative to "Reproducibility" in Linguistics?

Tobias Weber
Institut für Finnougristik, LMU Munich, Germany
http://kraasna.wordpress.com
weber.tobias@campus.lmu.de

Abstract

As an answer to the need for accountability in linguistics, computational methodology and big data approaches offer an interesting perspective to the field of meta-documentary linguistics. The focus of this paper lies on the scientific process of citing published data and the insights this gives to the workings of a discipline. The proposed methodology shall aid to bring out the narratives of linguistic research within the literature. This can be seen as an alternative, philological approach to documentary linguistics.

2012 ACM Subject Classification Applied computing; Applied computing → Anthropology; Applied computing → Publishing

Keywords and phrases Language Documentation, meta-documentary Linguistics, Citation, Methodology, Digital Humanities, Philology, Intertextuality

Digital Object Identifier 10.4230/OASIcs.LDK.2019.26

Category Extended Abstract

1 Introduction

In this position paper, I will propose a methodology which draws from approaches in computational linguistics to help with the goals of meta-documentary linguistics [2]. In a broad definition, this field investigates all processes around a documentary project, including value-adding steps after the recording of data, and tracking metadata. I propose including a new layer to meta-documentations consisting of the continuous tracking of citations and instances where the outputs are used beyond the publication of results - an extended, proactive meta-documentation. The aim of the methodology proposed here is to enable this approach and, subsequently, increase transparency in linguistic research by utilising the recent advances of big data and applying them to a specific element of linguistic publications deriving from a documentary project: the examples.

1.1 Background

Accountability has been at the centre of linguistic research, as in every scientific discipline which draws heavily from primary data. Thus, both "explicit concern for accountability" and "focus on primary data" are essential features for documentary linguistics [9]. Subsequently, accountability in language documentation is facilitated by providing a sufficient amount of metadata to accompany the set of primary data, including information on the technical side of the file or recording, the content, its producers, and the circumstances of its creation (e.g. place and time of recording, type of elicitation, use of stimuli) [13]. It should furthermore highlight all value-added procedures by members of the scientific community, such as transcribing and annotating [14]. However, it is questionable whether any amount of metadata will ever be able to fully convey the narrative behind the the documentation process, with all its

2nd Conference on Language, Data and Knowledge (LDK 2019).
Editors: Maria Eskevich, Gerard de Melo, Christian Fäth, John P. McCrae, Paul Buitelaar, Christian Chiarcos, Bettina Klimek, and Milan Dojchinovski; Article No. 26; pp. 26:1–26:8

OpenAccess Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

details and peculiarities, no matter how well the corpus of language data is "mediated" [10]. To solve this issue, a possible step is to supply a meta-documentation, a narrative about the documentary work and its outputs based on the metadata [24] [2]. At the same time, this opens up the chance to examine the workings of linguistics from the perspective of the History of Scholarship, and interpret the narratives with an anthropological view to highlight the backgrounds and the human factors within such an endeavour [6]. This, in turn, would create the necessary transparency of the documentation project which provides the basis for accountability of all works drawing from its outputs.

As introduced above, meta-documentary linguistics can provide insights to the narratives behind a documentation project and its outputs. Those are usually stored in archives and can be used by researchers for their linguistic work [9]. The most common way of using archived materials is by citing them as examples within a piece of writing. However, citation objectifies language and turns the example sentences into the object of language description - and artefact of the linguistic process. The example is detached from its original context within a story, an archive, or the entire documentation project. It becomes a new instance of this example, a different "version" or "generation", or, using the idea of Basalla's genealogy of artefacts [3], a new artefact mediated by the new technological means with a different interpretative backgrounds.

The aspect of new technological means in the mediation of the linguistic example deserves some attention, as current publishing practices, even electronic ones, are still tied to the paper publication format or a digital replication thereof (e.g. a PDF file), while new formats of publication are only slowly adopted [21]. This also means that any metadata attached to the original set of primary data, for example in an XML file [1], might be deleted due to restrictions in the medium or overwritten by new metadata of the citation, i.e. contain traces from the new technology [19]. However, the focus of attention here is the different interpretative background, as data which is frequently cited will accrue a variety of contexts and descriptive/theoretical frameworks in which it is used over the course of time. It presents us with the issue of intertextuality.

Frequent use of the same example might lead to changes in the representation of the data, interpretations derived from them, and features attributed to them. This is the point where linguists assume the agency to manipulate the data to fit into their theoretical frameworks according to their own convictions and beliefs about a particular language, grammatical feature, or the nature of language itself. While the term "manipulate" sounds more severe than the action it describes, changing the grammatical tag on a morpheme, editing morpheme boundaries, or adding a new translation (all value-adding processes [2]) can lead to entirely different interpretations of the language and its grammar. And those processes of adding value are routinely executed by linguists as part of their profession and without any mischievous intent. However, this can distort the provenance of a particular interpretation or presentation to an extent where the author themselves cannot determine where an interpretation came from; the scientific community engages with the data and with other members within the community, the moment of generating new ideas and thoughts. Yet, this also means that the trajectory of thought cannot easily be traced and that important information on the narrative (or rather: for the meta-documentation) becomes obscured.

1.2 Motivation

In early 2018, a position paper by 14 linguists and 41 undersigned members of the scientific community directed broad attention to this issue, calling for "reproducibility" in linguistic research as a means to create "verification and accountability" [4]. The authors consider

"proper" data citation and attribution as the basis of scientific reproducibility, stressing the central role of primary data within the field. And while it cannot be argued that correct and transparent handling of data is the tenet of our discipline and should be fostered and enhanced, the notion of "reproducibility" is firmly based in empirical, quantitative sciences where data are reified and purely objective (this kind of data "speak for themselves"). This hints at an understanding of linguistics where conventions and the professional conduct of the researchers as objective instances guarantee identical procedures in handling data which leads to reproducible results - and it ignores all instances of researchers assuming the agency to "break" rules, come up with innovative approaches, or add value to the data. To ensure brevity of this paper, I will not discuss the position paper in further detail but move on by pointing out the crucial element which is missing in this approach: the human factor and the intertextuality.

As initially stated, a central aspect to any set of metadata is the meta-documentation which is elaborating on the data set and providing insights to the contexts of creation of the data. This concept provides a narrative of the documentary work and creates transparency about it and, most importantly, the instances of human interaction with the data. It is, thus, an opposite idea to the objective stance of "reproducibility", where the goal is to reduce the human factor in our work. And even though both approaches are tied to the original data set and require the researcher to access and make reference to the archived originals, they cannot prevent that the researcher interacts with the data in creative ways. To ensure that information on these interactions is not lost, tracking of the citation of examples should be used, either as metadata or as a meta-documentation which proactively anticipates such interactions and can expand accordingly. In my opinion, the meta-documentary approach seems more sensible, as it does not restrict the researchers' agency in working with the data and provides a basis to document and record the human factor, and the subjectivity, of linguistic research instead of imposing a rigorous objective stance.

An alternative, humanistic approach is chosen by Frank Seidel [17]. He describes documentary linguistics as a "philology" which pays attention to the contexts of data and allows for variation instead of rigour. However, the philological approach can not only be applied to primary data but should be extended to also cover all instances of data citation in the literature. It emphasises the role of the researchers and all decision makers within the publication process (e.g. editors of journals, archivists, the scientific community) and provides information relevant for tracing the trajectory of thought and commenting on variation.

Focusing on the citation of language examples, there are many instances where variation can be observed. A great survey of the "defective documentation" of Norwegian in published linguistics literature can be found in the work by Jan Engh [7]. In this survey, Engh examines mistakes in data citation for Norwegian, which, in comparison to many endangered languages, provides two literary standards with national status and various resources and reference materials available. As such, it illustrates well how the researcher's hand can influence the outcomes of citing a language example, and provides a strong counter-argument to the calls for more objective research. A second interesting point to be found in Engh's work is that, in some instances, an error was inherited from a secondary source. I would therefore like to conclude that tracking citations of language examples is necessary for a thorough meta-documentation: it should cover all instances of citation, their contexts, and their relationships to other instances. Other examples for different languages could easily be named.

In my own work [23], I have worked on legacy materials of the extinct South Estonian Kraasna dialect, building a meta-documentation and tracing the use of examples through the literature. This produced a list of differences between original transcriptions and published materials, as well as between secondary sources. While, in the case of Kraasna, the bulk of data to be handled was considerably small, frequently cited languages and documentation projects pose a different obstacle for any researcher trying to trace examples and restore a missing meta-documentation manually. However, using a mixed methods approach [11], could save a researcher from tedious work while providing a solid basis for the preparation of a meta-documentation, and furthermore enabling research into scientific practices which are invisible to the human eye but clear to the computer.

Ultimately, I would like to argue that focusing on language examples cannot only help to handle the variation in representation but also share insights on the interpretative contexts for their citation. As (false) versions can be inherited from the secondary literature, interpretations can also be inherited or at least shaped by previous analyses on an example. I would even argue that, in an instance where a previous interpretation is completely negated, there is still an inherited element - the acceptance of an example as representative for a language, of an author within the scientific community, or a piece of writing within the canon of scientific works on a topic. The same holds true for all instances where an example is identical in its version but not in its citation context. The only way to ensure that the original metadata of the "artefact" are not altered is to completely ignore and exclude the example from scientific procedures.

2 Proposed Methodology

In the previous section, I presented the ideas behind meta-documentary linguistics and how they can help linguistic research to become more transparent and accountable. I, subsequently, propose the application of computational methods to the tracking of language examples within linguistic literature and the resulting computational meta-documentary linguistic research as a combination of humanistic/philological strands of documentary linguistics and computational linguistics. Although neither the computational tools and methods in question, nor the use of automatic citation trackers are novel, the application to language examples as parts of publications is nowadays easier than before. A first approach to compiling language examples can be seen in ODIN [12], however this project appears to be resting since 2010.

2.1 Goal

The goal of the application would be to search and establish relationships (a "genealogy") of linguistic examples within (a defined set of) linguistic publications. This could help to show relations between works even if they are not indicated in the publication, and would also provide a continuous trajectory of the citations [8]. Such a tool could be useful for archivists, documenters, the scientific community, as well as publishing houses. It would provide a valuable addition to the metadata of a file, a self-updating meta-documentation beyond the initial compilation, a "database-like" platform for linguistic examples, and might integrate well with already existing anti-plagiarism software [20]. For the collectors of the data, this tool might help to check the trajectory of the own work and its distribution. Furthermore, it might uncover biases in the citation of a particular language or particular example, e.g. by a certain schools, frameworks, or researchers.

2.2 Workflow

In order to achieve these goals, the application needs to

- search literature and identify linguistic examples;
- read data from the literature and store it;
- compare variants and classify them accordingly;
- compare with cited or indicated sources / the references of the source (optional); and
- collate the results in a presentational format.

Searching

The task of searching literature might be abridged by providing a preexisting database of publications, supplying relevant URLs, or a set of PDF files [25]. However, it is also possible to think of a crawler, which could search for relevant sources online. This function needs to detect linguistic data from plain text, OCRed text, and would, ideally, be able to apply OCR techniques to images. Luckily, linguistic examples are usually presented in a conventionalised format [5] which singles out examples to a separate block in the text, generally assigning a number and providing information on metadata (e.g. language name, source, date, reference). In addition, from a plain text or a suitable OCRed version of an image, the application might receive information about the character set used, which can be an indicator in the case of a phonetic transcription like the International Phonetic Alphabet. Within running text, examples might be italicised and followed by an analysis or translation.

Reading

The step of reading and storing the data requires a high amount of storage space, or a smart search algorithm which allows to reduce the stack for comparison tasks. For this step, methods used in big data applications or anti-plagiarism software could be used. The stored examples would need to be normalised to the extent that they are comparable but not additionally altered by the software; the originally intended rendering must be preserved, or at least stored with the normalised version. It is obvious that this function needs to support Unicode with all phonetic characters and diacritics.

Comparing

The comparison task would utilise common functions of pattern matching or fuzzy string searching [15] to establish the differences between versions and group them accordingly. Should the application support the optional functionality of including metadata on the potential source found within the publication, or other relevant information like the name of the author(s) or the year of publishing, these pieces of information would have to be collected in the first step and similarly matched in the comparing step [8].

Collating

The final step would be to collate the information gathered about each group of examples and arranging them by using a hierarchy of filters, with time and type of version being the primary criteria. Drawing from philology, such a textual genealogy (in Lachmannian tradition) can be translated into a path graph with the root node being the archived original transcription and each version being attributed to a parent node. Presumably, the results of this process will contain some unclear relations which would need to be resolved; the solution to this issue, however, requires theoretical considerations to be made about the required certainty for establishing a relation, which is not primarily a technical concern.

2.3 Format of the Output

As indicated above, the result of the procedure would be a list of clustered/grouped versions, or a path graph using time as an ordering principle. The presentation of the results would depend on the nature of the query, whether a particular version is searched within the database, all versions of a single original shall be displayed, or a bulk query is made about a particular project, author, or language. Potential presentational formats range from knowledge graphs in the style of the Web of Science [22], simple tree graphs, or lists and other plain text formats (e.g. CSV). However, the functionality of the application must not depend on the format of the output, which is rather a concern for the design of an interface or front-end to this tool and, therefore, highly dependent on the integration into other digital infrastructures.

3 Potential Obstacles

Before concluding this discussion, I would like to highlight some potential issues and pitfalls of the proposed methodology. Firstly, for collecting examples from the literature, access and usage rights would have to be negotiated with the copyright holders or the publishing houses. Although there are several open access journals in the field of linguistics [16], most high-profile journals and publications are still requiring the acquisition of access rights for their articles behind a paywall. The same holds true for various archives, where access rights are limited. It might be possible to agree a collaboration with publishing houses or a library to gain access for the crawler, however, the only way forward for transparency in published research seems to be open science.

Secondly, the type of the crawled file can influence the results greatly. No matter how good OCR systems have become, there are still issues with older prints, uncommon fonts, or with text donning an array of different, yet optically very similar, diacritics (like a breve and a hačekˇ). This becomes even more difficult as citations in the running text have to be recognised and not mistaken for normal text, which can make the delimiting of the example very difficult.

Thirdly, there are various types of transcription used, which also have to be attributed to the correct original source. For example, languages using a special writing system different from the Latin alphabet (e.g. various Asian languages, languages in Russia) might be transliterated to a Latin-based version; the solution would have to be a built-in transcription tool or optional integration for such a tool from external sources which use standardised transliteration (like ISO 9 for Cyrillic). Additionally, there are subfields of linguistics which are actively using their own, traditional transcription systems like the UPA for Uralic linguistics [18]. And even though IPA is already advancing in those fields, the proposed tool should be able to handle earlier sources equally well.

Fourthly, the sheer amount of data which needs to be crawled and the required computing power will mean that this application cannot be run every time a query is made by the user, or rather that there needs to be a good balance between results which are stored in a database file (for example in a CSV format) and queries requiring a new searching process. As both, storage space and computing power would be excessively used by the task, a feasible and efficient solution for the storage and presentation of the results would have to be found.

Lastly, the question remains whether queries are sent as bulk or for specific example sentences. Should the user have the chance to enter a sentence to be compared to the database, or would there only be access to a preexisting set of examples (e.g. stored with the meta-data for a particular file containing primary data). While it would be desirable

to allow user interaction with the tool and provide a powerful application for public use, it appears that such a methodology would have to be used sparingly until the fourth point can be sufficiently resolved.

4 Outlook

In this paper, I discussed how the application of rather common methods from computational linguistics could help to make linguistic research and the use of primary data within the discipline more transparent. The proposed methodology focuses on the linguistic examples within the published research and could provide important insights to the workings of this field. This methodology should be seen as an example of the increasing use of technology within the humanities, in particular for language documentation and the theorisation thereof – a potential field for future enquiry which might be understood as computational meta-documentary linguistics.

At the moment there are no concrete plans to create an application following the outlined workflow. Yet, this project might yield interesting results about the use of primary data in linguistics, which is an explicit concern for researchers in the field and currently widely discussed among scholars. Whether or not it is possible to conduct the survey in exactly the suggested way will depend on solutions to the obstacles outlined in the previous section, especially with regards to computing power, storage, and access rights. However, as technology advances at a high pace, it will likely be possible to handle, process, and store the necessary amount of data with relative ease in the future - an obstacle which should be overcome in the next decade. In order to obtain access and usage rights from the copyright holders, there could be possible solutions to agree on a strategic partnership with publishing houses or major libraries, to acquire research grants and support by influential scientific interest groups, and to support open access and open science movements. Although this issue is not as easily resolved as others, it can be hoped that necessary agreements will be made and that a successful pilot may convince an increasing amount of rightsholders.

A subsequent idea for increasing transparency in linguistic research regarding the citation of primary data could draw from the results of the application outlined here: If there is a data set on the relations between different versions of a particular example, this could be encoded in a standardised identifier format for linguistic examples, with a full genealogy accessible in a database connected to the identifier. In other words, version control for example sentences.

Finally, I encourage all computational linguists and data scientists to consider how their knowledge and skills might be applied within their own discipline, as a tool to aid metascience from a humanistic point of view.

References

1 Peter K. Austin. Data and language documentation. In Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel, editors, *Essentials of Language Documentation*, pages 87–112. Mouton de Gruyter, Berlin, New York, 2006.

2 Peter K. Austin. Language documentation and meta-documentation. In Mari Jones and Sarah Ogilvie, editors, *Keeping Languages Alive. Documentation, Pedagogy, and Revitalisation*, pages 3–15. Cambridge University Press, 2013.

3 George Basalla. *The Evolution of Technology*. Cambridge University Press, 1988.

4 Andrea L. Berez-Kroeker, Lauren Gawne, Susan Smythe Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, David I. Beaver, Shobhana Chelliah, Stanley Dubinsky, Richard P. Meier, Nick Thieberger, Keren Rice, and Anthony C. Woodbury. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics*, 56(1):1–18, 2018.

5 Max Planck Institute for Evolutionary Anthropology Department of Linguistics. Leipzig Glossing Rules. Conventions for interlinear morpheme-by-morpheme glosses. URL: https://www.eva.mpg.de/lingua/resources/glossing-rules.php.

6 Lise M. Dobrin and Josh Berson. Speakers and language documentation. In Peter K. Austin and Julia Sallabank, editors, *The Cambridge Handbook of Endangered Languages*, pages 187–211. Cambridge University Press, 2011.

7 Jan Engh. *Norwegian examples in international linguistics literature. An inventory of defective documentation.* Universitetsbiblioteket i Oslo, Oslo, 2006.

8 Bela Gipp. *Citation-based Plagiarism Detection.* Springer Vieweg, Wiesbaden, 2014.

9 Nikolaus P. Himmelmann. Language documentation: What is it and what is it good for? In Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel, editors, *Essentials of Language Documentation*, pages 1–30. Mouton de Gruyter, Berlin, New York, 2006.

10 Gary Holton. Mediating language documentation. In David Nathan and Peter K. Austin, editors, *Language Documentation and Description 12: Special Issue on Language Documentation and Archiving*, pages 37–52. SOAS, London, 2014.

11 R. Burke Johnson and Anthony J. Onwuegbuzie. Mixed Methods Research: A Research Paradigm Whose Time Has Come. *Educational Researcher*, 33(7):14–26, 2004.

12 William Lewis. ODIN - The Online Database of Interlinear Text. URL: http://odin.linguistlist.org/.

13 David Nathan. Archives 2.0 for endangered languages: From disk space to MySpace. *International Journal of Humanities and Arts Computing*, 4(1-2):111–124, 2010.

14 David Nathan and Peter K. Austin. Receonceiving metadata: language documentation through thick and thin. *Language Documentation and Description*, 2:179–187, 2004.

15 Gonzalo Navarro, Ricardo Baeza-Yates, Erkki Sutinen, and Jorma Tarhio. Indexing Methods for Approximate String Matching. *IEEE Data Engineering Bulletin*, 24:19–27, 2001.

16 OpenEdition. URL: https://www.openedition.org/.

17 Frank Seidel. Documentary linguistics: A language philology of the 21st century. *Language Documentation and Description*, 13:23–63, 2016.

18 Emil Nestor Setälä. Über die transskription der finnisch-ugrischen sprachen. *Finnisch-ugrische Forschungen*, 1:15–52, 1901.

19 Edward Shils. *Tradition.* The University of Chicago Press, 1981.

20 Benno Stein, Sven Meyer zu Eissen, and Martin Potthast. Strategies for retrieving plagiarized documents. In Wessel Kraaij and Arjen P. de Vries, editors, *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 825–826. ACM, 2007.

21 Nicholas Thieberger. Steps toward a grammar embedded in data. In Patricia Epps and Alexandre Arkhipov, editors, *New Challenges in Typology: Transcending the Borders and Refining the Distinctions*, pages 389–407. Mouton de Gruyter, Berlin, New York, 2009.

22 Web of Science. URL: http://wokinfo.com/.

23 Tobias Weber. Kraasna - A page on the Estonian Kraasna maarahvas and its dialect. URL: https://kraasna.wordpress.com/.

24 Anthony C. Woodbury. Archives and audiences: Toward making endangered language documentations people can read, use, understand, and admire. In David Nathan and Peter K. Austin, editors, *Language Documentation and Description 12: Special Issue on Language Documentation and Archiving*, pages 37–52. SOAS, London, 2014.

25 Jian Wu, Pradeep Teregowda, Juan Pablo Fernández Ramírez, Prasenjit Mitra, Shuyi Zheng, and C. Lee Giles. The Evolution of a Crawling Strategy for an Academic Document Search Engine: Whitelists and Blacklists. In *Proceedings of the 4th Annual ACM Web Science Conference*, WebSci '12, pages 340–343, New York, NY, USA, 2012. ACM. doi:10.1145/2380718.2380762.