15

Language Documentation & Conservation Special Publication No. 15
Reflections on Language Documentation 20 Years after Himmelmann 1998
ed. by Bradley McDonnell, Andrea L. Berez-Kroeker & Gary Holton, pp. 151–156
http://nflrc.hawaii.edu/ldc/
http://hdl.handle.net/10125/24816

Reflections on descriptive and documentary adequacy

Sonja Riesberg
Universität zu Köln
CoEDL Australian National University

One of Himmelmann's primary goals in his 1998 paper was to argue for a strict division of documentation and description. Language documentation has since successfully developed to become a discipline in its own right. Nevertheless, the question concerning the interrelation of description (and thus analysis) and documentation remains a matter of controversy. This paper reflects on descriptive and documentary adequacy, focusing on two major issues. First, it addresses the question of how much analysis should enter into an adequate documentation of a language and, second, it discusses the role of language documentation and primary data in the replicability of linguistic analyses.

I started my linguistic career as a PhD student in a documentation project funded within the Volkswagen Foundation's DoBeS program. The departments I have been affiliated with since have had a strong focus on documentary linguistics and the great majority of my colleagues and friends there are documentary linguists. Thus, to me, reading Himmelmann 1998 feels almost outdated; a statement of the obvious. In the 20 years since its appearance, the field has changed dramatically. A huge number of documentation projects have been funded, and a significant part of the linguistic community naturally considers language documentation to be one of the many linguistic disciplines. Himmelmann 1998 quite obviously had an immense impact. The fact that it is still shockingly relevant was something that I only slowly started to realize when I emerged from my PhD bubble and came to see a greater variety in how people 'do linguistics'. Reflecting on the topics as they were originally raised in the paper, how they have been received and how they have developed over the years is thus probably not as redundant as it might seem at first sight.[1]

[1] I am grateful to Birgit Hellwig, Stefan Schnell, and Sarah Verlage for having reflected with me on the one or other thought addressed in this short paper. Thanks also to two anonymous reviewers and the editors of this volume, especially Bradley McDonnell, for valuable comments.

ISBN: 978-0-9973295-3-7

The delimitation of documentation and description is one of the core issues discussed in Himmelmann 1998, and one of the explicit goals of the paper was that "the collection and presentation of primary data [should] receive the theoretical and practical attention they deserve" (Himmelmann 1998: 164). Interestingly, the demand for a strict separation of documentary and descriptive activities has, at times, caused rather emotional reactions in the linguistic community, and still does today. Partly, I believe, this is due to a misunderstanding and misconception of the paper. Some of the quotes one finds about Himmelmann 1998 suggest that the paper was not read in the first place, or maybe at best skimmed, in order to take quotes out of context and thereby miss the point entirely. Anyone who insists that Himmelmann advocates data collection without analysis has obviously not read (or understood) the paper. Anyone who claims that for Himmelmann language documentation is (nothing but) "a radically expanded text collection" has started reading, but quite obviously stopped halfway through. One very recent example is the introductory chapter to the *Oxford Handbook of Endangered Languages.* In an overview that is to answer the question "what is language documentation," Campbell & Rehg list a few quotes, such as "a language documentation is a lasting, multipurpose record of a language" (Woodbury 2011: 159), the Hans Rausing Endangered Languages Project website, and the above-mentioned Himmelmann quote about the "expanded text collection" (Himmelmann 1998: 165). They then conclude that "with statements such as these, it is little wonder that some linguists [...] have misinterpreted this approach to mean: Documentary linguistics is all about technology and (digital) archiving. Documentary linguistics is just concerned with (mindlessly) collecting heaps of data without any concern for analysis and structure. Documentary linguistics is actually opposed to analysis" (Campbell & Rehg 2018: 11). I disagree. I don't believe it is the statements themselves that trigger this misinterpretation, because typically they are elaborated on in more detail in the original sources, which rarely leave room for misunderstandings. It is the way some authors choose these statements and list them in isolation, that—intentionally or not—promote a skewed picture.

In any case, it seems that an interpretation of Himmelmann has developed and continues to be spread that attributes to Himmelmann the idea that language documentation does not (or even must not) include analysis. This might be one reason for certain developments in the field, which in my opinion are rather unfortunate. One such idea is that documentation corpora should contain no elicited material at all; for instance, recordings of elicitation sessions where linguists and native speakers work on lexical, phonological/phonetic, or grammatical problems. There might, of course, be various other reasons for this, such as following the wishes of the speech community to include only 'nice' and 'clean' data in the collection, or the embarrassment felt by the researcher upon listening to themselves struggling to come to grips with complex grammatical issues, or trying to determine the exact semantics of a culturally specific lexical item. But I think the reluctance to archive elicitation data and make them available to the wider public partly also stems from the widespread idea that a language documentation is to include 'natural' data only, and that elicitation should not be part of it. To my thinking, this is an interpretation of Himmelmann's demand to separate descriptive and documentary activities which clearly overshoots his goal. In this sense, I agree with Berge's (2010) notions of adequacy in documentation, in that a documentation should include "basic phonology, morphology, syntactic constructions with context, lexicon, a full range of textual genres, registers, and dialects, and data from diverse situations and speakers" (58). It is the large variety of data, including natural data as well as elicited materials, that makes a good documentary corpus such a valuable resource. Yet, another unfortunate trend can be observed when looking at the practices of some funding agencies that support documentation projects. Here, one sometimes gets the impression that, for the agencies, the important aspect of 'value for money' is only evaluated in terms of hours of (natural language) recordings plus transcriptions that are deposited in the archives. But this is not, of course, what we conceive of as a comprehensive and adequate language documentation.[2]

These reactions and developments are all the more surprising considering that Himmelmann 1998 explicitly addresses the close interrelation of language documentation and language description at different points, and clearly states that his approach "does not imply that it is possible to make a 'pure' documentation without any descriptive analysis" (Himmelmann 1998: 165). The point Himmelmann made was that primary data should not be collected solely for the purpose of description (and then kept in some drawer (until the linguist's death) to be disposed of (after the linguist's death)). Instead, description was seen as part of the documentation. That is, description is considered a necessary part of documentation, but not its purpose. In Himmelmann's view "a good and comprehensive documentation will include all the information that may be found in a good and comprehensive descriptive grammar" (170). The opposite is not the case: The most detailed and carefully researched grammar can never cover what a language documentation covers, simply because it is a written medium that cannot compete with the multi-media resources that make up a language documentation. Exactly when description should therefore be postponed to save resources for good quality documentation, as suggested in a later paper by Himmelmann (2006: 24), is a matter of controversy (see, e.g., Evans 2008: 346).

The second reason why the paper has caused emotional reactions is because it challenges the way in which we as a discipline have been working and how we have treated primary data for decades. There are two general and substantial points to this issue. The first concerns the question of what kind of database we, as a field, want to ground our knowledge in. The second concerns the question of replicability. As Himmelmann states, "a language description aims at the record of A LANGUAGE, with 'language' being understood as a system of abstract elements, constructions, and rules that constitute the invariant underlying structure of the utterances observable in a speech community. A language documentation, on the other hand, aims at the record of THE LINGUISTIC PRACTICES AND TRADITIONS OF A SPEECH COMMUNITY" (Himmelmann 1998: 166). And, as mentioned above, the assumption is that a good documentation will include all the information that is also included in a good description. The important thing to note is that if I base my description on a comprehensive documentation, I will have to account for all the variation I find in this compilation of primary data. This is the ideal scenario, in which I will account for—or at least describe—all (morpho-)syntactic structures that occur in my corpus, all the inter- (and intra-)speaker variation, all the inter- (and intra-)genre variation, etc. If, however, I approach a speech community and their language with the sole purpose of grammar writing, it is likely that much of the above will not make it into my description, because my data base is in all likelihood a very different one. In the worst case it will consist of elicitations of phenomena I decided to investigate beforehand (e.g.,

[2] See also Himmelmann (2012) for a more detailed argumentation against the misconception that language documentation is equivalent to "(mindlessly) collecting heaps of data" and opposed to analysis (187), and, e.g., Caballero (2017) for a very nice report on how diverse a documentation corpus can be and how different data types in a corpus can be utilized by different stakeholders.

verbal and pronominal paradigms, clause chaining, relativization, etc.), elicited with one or two speakers only, and possibly illustrated with one or two carefully chosen examples from one or two narratives I collected, perhaps specifically for this reason. Obviously, the two scenarios depicted constitute the two endpoints of a continuum and most linguists will find themselves doing something in between. For instance, the 'traditional' field linguist writing a grammar will collect more than just one or two narratives and, of course, will also treat phenomena that go beyond such questions as one finds, e.g., in precast questionnaires. On the other hand, a single person working on a language documentation will not (in one lifetime, and even less so within the usual funding period) be able to describe everything there is in her collection. But the question of which approach will result in a more adequate record of the language investigated is easy to answer, and I am convinced that as a field we should strive towards the ideal, albeit in the knowledge that it represents more than a lifetime's work. Yet, it is still the case that language documentations count less than the written word. Many documentary linguists, especially early career researchers, who will often have spent years of their career compiling language documentation corpora, will probably have had the painful experience that an appointment committee has asked the question why 'so little' has been published. This is not to say that the community is unaware of this problem. It is a widely discussed deficit, and there have been serious attempts to enhance the status of documentation corpora, to review them and thus make them utilizable for application processes. We are just not quite there yet.

Turning to the question of replicability, this involves making primary data accessible. It feels like there should not be much to say about this. If we believe that science can only be taken seriously if quality can be scrutinized, primary data should to be made available (if possible), and they should be made available in a format that allows for verification or falsification of the claims made. Needless to say, this is not achieved by simply uploading an audio or video file to a digital archive, or handing over a collection of cassettes or external hard drives to physical one.[3] The falsifiability of descriptive statements was not one of the primary concerns in Himmelmann 1998 (see Himmelmann 2006: 15, where this issue is discussed in a little more detail), though it is mentioned in passing: "it is simply a feature of scientific enterprise to make one's primary data accessible for scrutiny" (p.165). Compared to other disciplines, (typological) linguistics has a lot of catching up to do. In psychology, for example, journals strongly encourage their authors to "deposit research data in a relevant data repository and cite and link to this dataset in their articles". This is considered the default, and if, for some reason, it is not possible to make data available, the author has to "make a statement explaining why research data cannot be shared".[4] In other fields data sharing is not only *encouraged* but even *required.* This is still pretty much unheard of for linguistic journals, though people have started discussing the topic of replicability and there is a serious demand for data sharing also in linguistics (see, e.g., the paper by Gawne & Berez-Kroeker, this volume). The utilization of documentation is probably one of the major foci of attention in the current discourse on documentary linguistics. Yet, until today, descriptive statements — especially about small, underdescribed languages — are hardly ever falsifiable. Extended appendices with

[3] This, of course, also holds if we want our data to be useful to researchers from other disciplines and to members of the respective speech communities.

[4] These two quotes originate from the authors' information for the journal *Cognitive Development*, cf. `https://www.journals.elsevier.com/cognitive-development`. See also `https://www.elsevier.com/authors/author-services/research-data/data-guidelines`.

interlinearized and translated texts that are added at the end of a grammar give some opportunity for falsifiability. In short research papers with limited space this becomes more difficult, especially for topics that are not well captured through transcription, not to mention studies on tonal phenomena and prosody. Language documentation corpora are one possibility, and probably the most adequate and comprehensive one, to address this issue. This is, on the one hand, because modern language archives usually (try to) guarantee long-term preservation, and on the other hand, because language documentations offer the amount and variety of data necessary to make falsifiability possible in the first place. Surprisingly, many linguists who have spent a lot of time and energy compiling these corpora (including the immensely time-consuming transcriptions, annotations, and commentaries) do not make use of these invaluable resources and often do not even reference them in their publications, even if they form the basis of their research. Others (and, I'm afraid, I partly include myself here) will make a general statement, hidden in a footnote, along the lines of "all data used in this paper is available online under xxx," but then do not make the effort to link examples to the original source. A nice example of how this can be done successfully is Seifart et al.'s recent paper on Bora drummed speech (Seifart et al. 2018), which links to the whole Bora language documentation corpus (Seifart et al. 2009), but also to single, relevant sessions (i.e., video and audio recording plus transcriptions and translations in an ELAN file) that illustrate the statements made in the paper.

Obviously, a lot has been achieved in the field of language description and documentation since Himmelmann 1998 was published (so much in fact, that it took me until about four to five years into my academic career to realize that the practices sketched in this paper are not to be taken for granted). It is just as obvious, though, that there is room for improvement. This pertains to the role of language documentations in helping linguistics to catch up with other sciences in terms of replicability of research results. But it also pertains to Himmelmann's vision quoted at the beginning of these reflections that documentary activities should receive the attention they deserve. They do not, as long as a paper in any mid-range linguistic journal on a researcher's CV still means more to potential employers and funding agencies than an annotated multi-media corpus with careful commentary of an endangered language. As mentioned above, both these issues are currently avidly discussed in the documentary community. What seems to receive less attention is the rather paradoxical development that due to the very success of establishing language documentation as a discipline, it has lost its descriptive component, which was always supposed to be—and I believe should be—an important and substantial part of it.

References

Berge, Anna. 2010. Adequacy in documentation. In Lenore A. Grenoble & N. Louanna Furbee (eds.), *Language Documentation: Practices and values*, 51–66. Amsterdam: John Benjamins.

Campbell, Lyle & Kenneth R. Rehg. 2018. Introduction: Endangered languages. In Kenneth L. Rehg & Lyle Campbell (eds.), *The Oxford handbook of endangered languages*, 1-20. Oxford University Press.

Evans, Nicholas. 2008. Review of Essentials of language documentation. *Language Documentation & Conservation* 2(2). 340–350.

Gawne, Lauren & Andrea L. Berez-Kroeker. 2018. Reflections on reproducible research. In Bradley McDonnell, Andrea L. Berez-Kroeker & Gary Holton (eds.), *Reflections on language documentation on the 20 year anniversary of Himmelmann 1998.* http://hdl.handle.net/10125/24805

Himmelmann, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36, 161–195.

Himmelmann, Nikolaus P. 2006. Language documentation: What is it and what is it good for? In Jost Gippert, Nikolaus P. Himmelmann & Ulrike Mosel (eds), *Essentials of language documentation*, 1–30. Berlin and New York: Mouton de Gruyter.

Himmelmann, Nikolaus P. 2012. Linguistic data types and the interface between language documentation and description. *Language Documentation & Conservation* 6. 187–207.

Seifart, Frank, Julien Meyer, Sven Grawunder & Laure Dentel. 2018. Reducing language to rhythm: Amazonian Bora drummed language exploits speech rhythm for long-distance communication.*Open Science* 5(4). 170354. https://doi.org/10.1098/rsos.170354

Seifart Frank, Doris Fagua, Jürg Gasché & Juan Alvaro Echeverri (eds). 2009. *A multimedia documentation of the languages of the people of the center. Online publication of transcribed and translated Bora, Ocaina, Nonuya, Resígaro, and Witoto audio and video recordings with linguistic and ethnographic annotations and descriptions.* Nijmegen, The Netherlands: The Language Archive. https://hdl.handle.net/1839/00-0000-0000-001C-7D64-2@view

Woodbury, Anthony C. 2011. Language Documentation. In Peter K. Austin & Julia Sallabank (eds.), *The Cambridge Handbook of Endangered Languages*, 159–186. Cambridge: Cambridge University Press.

Sonja Riesberg
sonja.riesberg@uni-koeln.de