1

Language Documentation & Conservation Special Publication No. 3 (August 2012):
Potentials of Language Documentation: Methods, Analyses, and Utilization,
ed. by Frank Seifart, Geoffrey Haig, Nikolaus P. Himmelmann,
Dagmar Jung, Anna Margetts, and Paul Trilsbeek, pp. 1–6
http://nflrc.hawaii.edu/ldc/sp03/
http://hdl.handle.net/10125/4510

The threefold potential of language documentation

Frank Seifart

Max Planck Institute for Evolutionary Anthropology, Leipzig

1. Introduction. In the past 10 or so years, intensive documentation activities, i.e. compilations of large, multimedia corpora of spoken endangered languages have contributed to the documentation of important linguistic and cultural aspects of dozens of languages. As laid out in Himmelmann (1998), language documentations include as their central components a collection of spoken texts from a variety of genres, recorded on video and/or audio, with time-aligned annotations consisting of transcription, translation, and also, for some data, morphological segmentation and glossing. Text collections are often complemented by elicited data, e.g. word lists, and structural descriptions such as a grammar sketch. All data are provided with metadata which serve as cataloguing devices for their accessibility in online archives. These newly available language documentation data have enormous potential in three respects

1. Given that modern language documentations are cast in a sufficiently standardized, well-structured electronic format, **computational methods** can efficiently enhance the annotations and improve the analyses of language documentation data in many ways. The combination of state-of-the-art computational methods and electronic language documentation corpora has the potential to significantly impact the way linguistic data in general is handled and analyzed.
2. The additional data available through language documentation constitute a much richer empirical basis for **analyses** in various subfields of linguistics as well as in related disciplines such as anthropology, allowing for a better understanding of the range of diversity found in human languages. The richer nature of the data may change how comparative analyses are pursued which may potentially lead to other results.
3. The multimedia documentation of linguistic and cultural practices has the potential for multiple ways of **utilization**, not only for interdisciplinary research but also for language maintenance, and it has the potential to raise awareness of language diversity and endangerment. Since documentations are in an electronic format, they can be made accessible in novel online formats.

2. New perspectives on language documentation. In order to critically discuss and make more explicit the threefold potentials of language documentation, a workshop was held in Leipzig in November 2011. The contributions to this volume are based on the presentations at this workshop. The group of contributors includes not only "language documentation practitioners" but also, crucially, "outside perspective providers", especially potential users of documentations. Among these latter are (i) experts on corpus linguistics and other computational methods; (ii) researchers from linguistics and related scientific fields who have experience with analyzing language documentation data from endangered languages or have an interest in using such data for their analyses; and (iii) experts from fields in which language documentation data are applied for language maintenance and for data curation and online presentation. The perspectives on language documentation they provide in the present volume open up, firstly, new directions for interactions of language documentation with computational methodologies; secondly, they demonstrate the potentials of novel interdisciplinary research using language documentation data; and thirdly, they discuss the various ways how language documentation data can be used for practical applications and other purposes.

Taken together, these contributions make abundantly clear that modern language documentation is not a self-serving activity guided only by abstract principles such as the preservation of cultural heritage. On the contrary, language documentation is a vibrant field with multiple connections to sophisticated computational methods, interdisciplinary research venues, and modern types of utilization.

3. Overview of the volume.

3.1. Part one: Methods.

The contributions to the first part of this volume address the following central question: How do computational methods developed for large corpora of well-known languages apply to the relatively small language documentation corpora of less well-known languages?

In the first contribution to this part of the volume, **Sebastian Drude** discusses "Prospects for e-grammars and endangered languages corpora". He describes new ways of constructing exclusively digital grammars and how these benefit from links to digital language documentation corpora.

Anke Lüdeling's contribution "A corpus linguistics perspective on language documentation, data, and the challenge of small corpora" discusses the role of variation in corpus-linguistic research and argues that in order to bring about the full potential of corpus data from language documentation for such research, these need a flexible corpus structure and explicit metadata.

Another important methodological issue in connection with language documentation corpora is distinguishing different object languages in multilingual corpora, reflecting the multilingual reality of many small and endangered language communities. **Jost Gippert** deals with these issues in his contribution "Language assignment in DoBeS and similar corpora of endangered languages".

Oliver Schreer and **Daniel Schneider**'s contribution "Supporting language research with generic automatic audio/video analysis" discusses new methods for the automatic recognition of linguistic or gestural patterns in the audio or video signal. These methods help, for instance, to segment data into utterances, recognize speakers, and identify and classify gestures, making the annotation of language documentation data and their preparation for analyses much more efficient.

Amit Kirschenbaum, **Peter Wittenburg**, and **Gerhard Heyer** provide another perspective from computer sciences on quantitative methods for language documentation corpora. They discuss "Unsupervised morphological analysis of small corpora", i.e. statistical processing and learning methods for automatic text analyses and morphological parsing and annotation of small corpora that are transcribed and translated, but have not been annotated further.

There are a number of further, new and interesting methods for quantitative computational analyses of textual data from language documentations. **Sabine Stoll** and **Balthasar Bickel**'s contribution to Part two shows how the time-aligment of these data can be used for linguistic analyses. Finally, methods for typological comparison based on parallel texts (e.g. Cysouw & Wälchli 2007, Wälchli & Cysouw forthcoming) could be applied to language documentation data by treating the transcription and translation as parallel texts.

3.2. Part two: Analyses. The recently established archives containing language documentations consist, by definition, of data from endangered, and hence generally small and often geographically isolated, language communities. These databases thus counteract the often-bemoaned bias in linguistic typology and other disciplines towards "large" languages, i.e. those that are embodied in a standardized written form, promulgated through a formal education system, and used for decontextualized communication purposes in industrialized societies. Against this backdrop, the central question discussed in the contributions to this part is: What impact has language documentation had on analyses and theorizing in linguistics and related disciplines so far and how can it make greater impact?

Peter Trudgill in his contribution "On the sociolinguistic typology of linguistic complexity loss" argues that language documentation data from small languages is absolutely essential for our understanding of language in general since these languages display distinct typological features which are not represented in the few well-studied "large" languages – these being the exceptional ones from a global perspective.

Marianne Gullberg's paper "Bilingual multimodality in language documentation data" discusses two other aspects of language documentation data. Firstly, they typically reflect the multilingual reality of humans throughout most of their history by including code-switching and other language contact phenomena. Secondly, they document the multimodal reality of human language through video recordings. Language documentation data can thus inform theoretical and empirical studies of linguistics, bilingualism, and multimodality in entirely new ways.

Two papers deal with the role of language documentation data in the study of the typology of referential hierarchies (as an example of a classical typological topic): **Jane Simpson** discusses "Information structure, variation and the Referential Hierarchy" in the light of data from languages which are now undergoing rapid change, and she illustrates how a richer documentation could have contributed to a better understanding of these data. **Stefan Schnell** in his contribution "Data from language documentations in research on referential hierarchies" focuses on the importance of textual data – as opposed to structural descriptions or elicited data – for research on topics such as the referential hierarchy.

Sabine Stoll and **Balthasar Bickel**, in “How to measure frequency? Different ways of counting ergatives in Chintang (Tibeto-Burman, Nepal) and their implications”, discuss the importance of relating the distribution of items in corpora to to the length of the time windows within which speakers and hearers use the language. This study thus makes use of the time-alignment that is typical of modern language documentation data.

From the perspective of historical and contact linguistics, **Marian Klamer** takes us on “Tours of the past through the present of eastern Indonesia”, showing how data newly available through language documentation can shed light on human prehistory and migration patterns.

3.3. PART THREE: UTILIZATION. Data from endangered languages are not only valuable for linguists but also present a repository of cultural and linguistic knowledge that can be used in various ways, including for language maintenance efforts. The central question discussed in this part is: How can language documentation data be stored, represented, and made accessible in order to be utilized in a broader context?

Gary Holton's paper “Language Archives: They're not just for linguists anymore” describes instructive examples of such utilization, namely how material from the Alaska Native Language Archive has been used for studies in ethnoastronomy and for language revitalization.

The infrastructure necessary for allowing the utilization of language documentation data over space and time is discussed by **Nick Thieberger** in “Using language documentation data in a broader context”. This includes issues such as data formats, metadata, and online cataloguing systems.

Online presentation is the major gateway for a broader utilization of language documentation. **Gabriele Schwiertz** describes a proposal for an online interface for utilization by various user groups, including the speech community in “Online presentation and accessibility of endangered languages data: The general portal to the DoBeS-archive”. On the other hand, **Hans-Jörg Bibiko** gives an introduction to “Visualization and online presentation of linguistic data”, i.e. new computational methods for creating maps, online dictionaries, and other materials that facilitate the utilization of language documentation data.

Julia Sallabank's paper “From language documentation to language planning: not necessarily a direct route” discusses the potentials and pitfalls of using language documentation for language planning, showing that language practices as observed and documented by linguists may not match how community members perceive their own linguistic behavior – or how they would prefer their language practices to be seen.

Ulrike Mosel, in “Creating educational materials in language documentation projects – creating innovative resources for linguistic research”, describes how the production of educational materials can be integrated into a language documentation project when native speakers edit the transcriptions of spontaneously spoken texts and thus create an innovative resource for the comparison of spoken and written language.

4. FURTHER POTENTIALS. Each paper of this volume makes a contribution to clarifying the potentials of language documentation. A number of additional aspects emerged from the comparative discussion of the presentations of these papers at the workshop in

Leipzig. Since they are not necessarily explicitly addressed in the individual papers, they are briefly mentioned in the following:

- There is an enormous potential in the combination of new computational methods with language documentation data due to their standardized electronic format. It appears that the possibilities of computational techniques to process, analyze, annotate, and visualize linguistic data are virtually endless (see the contributions to Part one, but also the contributions by Stoll & Bickel, Thieberger, and Bibiko). Successful approaches to putting these possibilities to use have shown two things: First, the real challenges are often conceptual, not technical. Therefore, further progress requires close collaboration between computational scientists and linguists, which is often difficult because the two fields differ in their general methodological approach. Secondly, there is often not one single ideal computational solution for a linguistic problem. Therefore such collaboration may benefit from mixed systems involving the modularization of various techniques, and from the implementation of interactive learning (see Wittenburg et al. in press).
- The study of the multimodality of language, especially of gestures, in language documentation data is a particularly promising area for future research. On the one hand, the type of language documentations now available – i.e. including video recordings of speech events with time-aligned annotation – constitute an enormous resource for the study of the cross-linguistic variability of gestures and other multimodal aspects of language, which has, so far, been largely ignored in studies on these topics (as argued by Gullberg, this volume). On the other hand, there are now methods that make further annotation and analysis of this data far more efficient than it was only a few years ago (see Schreer & Schneider, this volume).
- It appears that utilization of language documentation data by non-linguists has been happening more at archives with a focus on particular regions (see Holton, this volume) than at archives with a world-wide scope (see Schwiertz, this volume). In order for archives with a world-wide scope to be perceived as repositories for information on the region and thus attract more potential users, a possible solution is provided by regional archives which mirror data from a central archive (see http://www.mpi.nl/DOBES/regional_archives; Seifart et al. 2008)
- Language documentation activities still receive very little academic recognition in the sense that they do not count much for track records of linguists, certainly much less than, e.g. journal articles. In response to this, as one outcome of the workshop, the future issues of the journal Language Documentation & Conservation (http://nflrc.hawaii.edu/ldc) will include a special section for the review of online language documentations.

REFERENCES

Cysouw, Michael & Bernhard Wälchli. 2007. Parallel texts: Using translational equivalents in linguistic typology. *STUF - Language Typology and Universals* 60(2). 95–99.

Himmelmann, Nikolaus P. 1998. Documentary and Descriptive Linguistics. *Linguistics* 36(1). 161–195.

Seifart, Frank, Sebastian Drude, Bruna Franchetto, Jürg Gasché, Lucía Golluscio & Elizabeth Manrique. 2008. Language Documentation and Archives in South America. *Language Documentation & Conservation* 2(1). 130–140. http://scholarspace.manoa.hawaii.edu/bitstream/handle/10125/1775/seifartsmall.pdf;jsessionid=C9B53B7B701739117643CB4522ECCD2C?sequence=12 (29 May, 2012).

Wälchli, Bernhard & Michael Cysouw. forthcoming. Lexical typology through similarity semantics: Toward a semantic map of motion verbs. In Linguistics.

Wittenburg, Peter, Przemyslaw Lenkiewicz, Eric Auer, Binyam Gebrekidan Gebre, Anna Lenkiewicz & Sebastian Drude. in press. AV Processing in eHumanities – a paradigm shift. In Proceedings of the Digital Humanities Conference 2012, 16–22 July 2012, Hamburg.

Frank Seifart
frank_seifart@eva.mpg.de