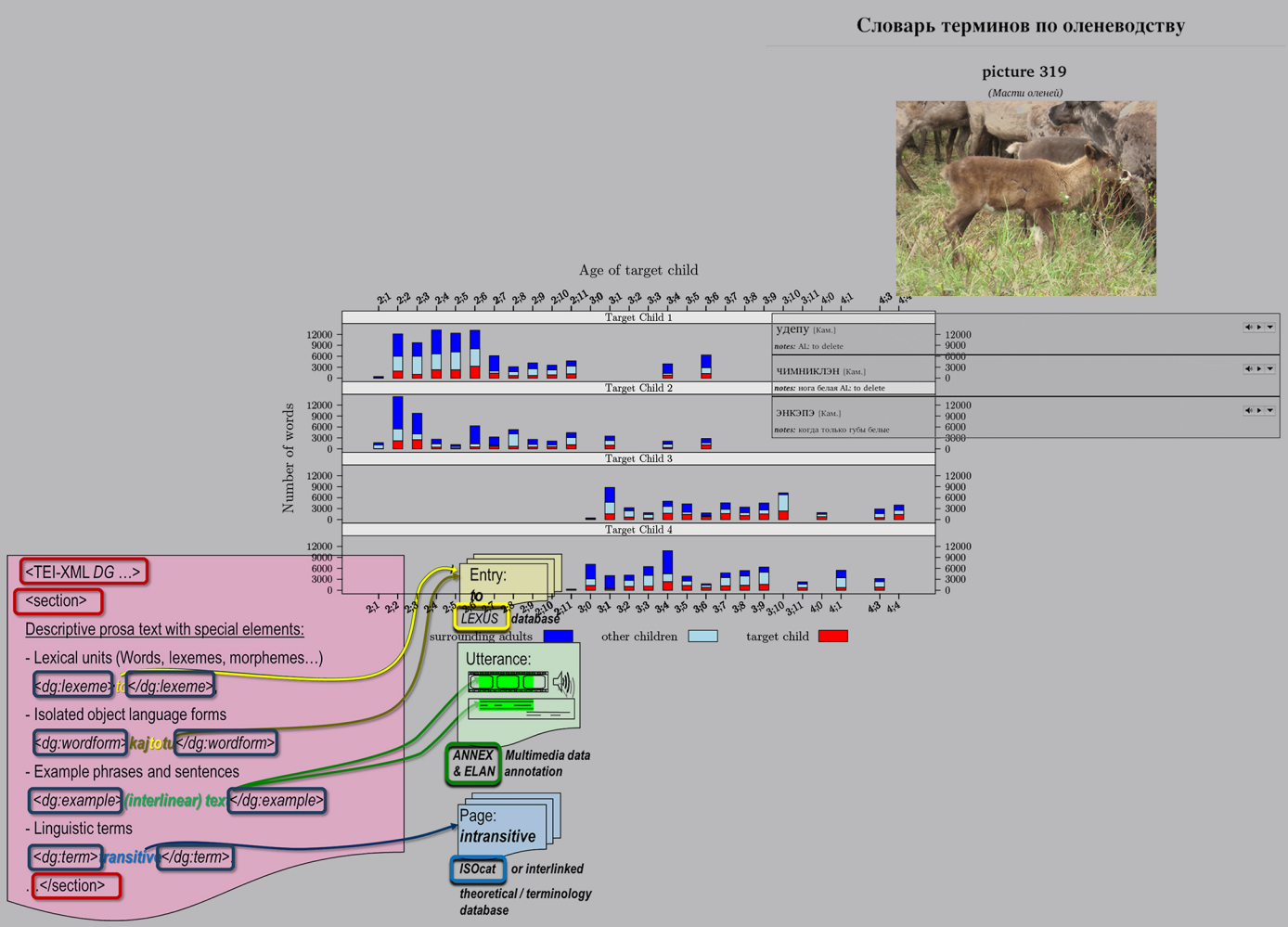
Language Documentation & Conservation Special Publication No. 3

Potentials of Language Documentation: Methods, Analyses, and Utilization



edited by

Frank Seifart, Geoffrey Haig, Nikolaus P. Himmelmann, Dagmar Jung, Anna Margetts, and Paul Trilsbeek

Potentials of Language Documentation: Methods, Analyses, and Utilization

edited by

Frank Seifart, Geoffrey Haig,
Nikolaus P. Himmelmann,
Dagmar Jung, Anna Margetts,
and Paul Trilsbeek

Language Documentation & Conservation Special Publication No. 3

PUBLISHED AS A SPECIAL PUBLICATION OF LANGUAGE DOCUMENTATION & CONSERVATION

LANGUAGE DOCUMENTATION & CONSERVATION
Department of Linguistics, UHM
Moore Hall 569
1890 East-West Road
Honolulu, Hawai'i 96822
USA

http://nflrc.hawaii.edu/ldc

UNIVERSITY OF HAWAI'I PRESS
2840 Kolowalu Street
Honolulu, Hawai'i
96822-1888
USA

Cover design by Sylvio Tüpke and Frank Seifart,
using figures from the contributions by Sebastian Drude, Sabine Stoll & Balthasar Bickel,
and Hans-Jörg Bibiko to this volume.

Library of Congress Cataloging in Publication data
ISBN 978-0-9856211-0-0

http://hdl.handle.net/10125/4540

Contents

Contributors

HANS-JÖRG BIBIKO has been working since 2004 at the Department of Linguistics of the Max Planck Institute for Evolutionary Anthropology in Leipzig/Germany. He is a computer scientist focusing on database infrastructure, software development for 'computer-aided' linguistics, digital cartography, and knowledge visualization.

BALTHASAR BICKEL got his graduate training at the Max Planck Institute for Psycholinguistics in Nijmegen in the early 1990s. After postdoctoral research in Berkeley, Zurich, and Mainz, he took over a chair in linguistic typology at the University of Leipzig in 2001. Since 2011 he has been professor of general linguistics at the University of Zurich. Bickel's core interest are the regional and universal factors shaping the worldwide distribution of linguistic diversity over time. For this, he applies methods ranging from the statistical analysis of typological databases and corpora to ethnolinguistic fieldwork and experimental methods. A special focus area is the Himalayas, where Bickel has been engaged in interdisciplinary projects on endangered languages and developing corpora of them.

SEBASTIAN DRUDE is the Scientific Coordinator of The Language Archive (TLA) at the Max Planck Institute for Psycholinguistics in Nijmegen. He is a documentary/anthropological linguist interested in language technology and infrastructure. Since 1998, he has conducted fieldwork among the Awetí indigenous group in Central Brazil, participating in the DoBeS (Documentation of Endangered Languages) research program from 2000 on. From 2008 he was a Dilthey fellow at the University of Frankfurt, before going to the MPI Nijmegen and joining the leading group of TLA in November 2011. This group hosts the central DoBeS language archive and develops tools and infrastructure for linguistics and the digital humanities.

JOST GIPPERT studied Comparative Linguistics, Indology, Japanese, and Chinese at the German universities of Marburg and Berlin (FU). Ph.D. (Dr.phil.) in 1977 with a study on the infinitive syntax of Indo-European languages. Various positions as a lecturer and university assistant in Indo-European linguistics in Berlin, Vienna, and Salzburg (1977–1990). Worked as a computational linguist in the field of Oriental languages at the University of Bamberg from 1990 to 1993. Since 1994, Chair of Comparative Linguistics at the University of Frankfurt/Main. Developer of various computer programs handling Eastern languages and non-Roman scripts. Founder and leader of the TITUS project aiming at a complete collection and Internet edition of texts in ancient Indo-European (and neighboring) languages (since 1987). Director of several projects on language documentation in the Caucasus and elsewhere (since 1999). Initiator and spokesperson of the LOEWE priority program "Digital Humanities" of Hesse, Germany (since 2010).

MARIANNE GULLBERG is professor of linguistics and director of the Humanities Lab at Lund University. Her research targets adult second language acquisition and bilingualism and the production and comprehension of gestures. She has worked on Swedish, Dutch, French, English, Turkish, Japanese, and Papiamento. She led a research group on multilingual and multimodal language processing at the Max Planck Institute for Psycholinguistics, the Netherlands, in 2002–2009, applying linguistic, psycholinguistic, and neurocognitive perspectives. She also co-founded the Nijmegen Gesture Centre in 2003. Marianne has published extensively on bilingualism and gestures and also edited a series of special issues and volumes on these topics. She is editor of three international journals in the field (Language Learning, Gesture, Language, Interaction, Acquisition) and has been vice-president of the European Association of Second Language Research.

GERHARD HEYER is chair of Automatic Language Processing at the Computer Science Department of the University of Leipzig. His field of research is focussed on automatic semantic processing of natural language text with applications in the area of information retrieval and search as well as knowledge management.

GARY HOLTON is a documentary linguist whose work focuses on the Papuan outlier languages of eastern Indonesia and the Athabaskan languages of Alaska. He is currently Professor of Linguistics at the University of Alaska Fairbanks and Director of the Alaska Native Language Archive.

AMIT KIRSCHENBAUM is a Ph.D. candidate in computer science at the University of Leipzig. His research interests include natural language processing, machine learning and psycholinguistics.

MARIAN KLAMER teaches at Leiden University and has done primary fieldwork on a dozen Austronesian and Papuan languages in Indonesia over the last two decades. Her research centres on language description and documentation, typology, and historical and contact-induced language change. Her publications include *A grammar of Kambera* (1998), *A grammar of Teiwa* (2010), *A short grammar of Alorese* (2011), and over 50 articles on a variety of topics. Klamer has coordinated numerous research projects on languages of Indonesia, including the NWO-VIDI project 'Linguistic variation in Eastern Indonesia' (2002–2007) and the EuroBABEL project 'Alor Pantar languages: Origins and theoretical impact' (2009–2012), funded by the European Science Foundation.

ANKE LÜDELING is a professor for corpus linguistics and morphology at the Institute for German Language and Linguistics at Humboldt University (Berlin). She is interested in linguistic variation and has built, annotated, and analyzed corpora of non-standard varieties of German such as learner language and historical varieties.

ULRIKE MOSEL is professor emerita of General Linguistics at the University of Kiel. After receiving her Ph.D. in Semitic languages at the University of Munich (1974), she started researching South Pacific languages and became an expert in collaborative fieldwork. Her books include *Tolai Syntax* (1984), *Samoan Reference Grammar* (1992, with Even

Hovdhaugen), and *Say it in Samoan* (1997, with Ainslie So'o). Currently, she is working on the documentation of the Teop language of Bougainville, Papua New Guinea. Together with Christian Lehmann, Hans-Jürgen Sasse, and Jan Wirrer she initiated the DoBeS language documentation programme funded by the Volkswagen Foundation.

JULIA SALLABANK is Senior Lecturer in Language Support and Revitalisation in the Endangered Languages Academic Programme at the School of Oriental and African Studies, London. She gained her doctorate at Lancaster University in 2007 and was previously commissioning editor for applied linguistics and language teaching methodology at Oxford University Press. Her main research interests are sociolinguistics/sociology of language; endangered language documentation and revitalisation; language policy and planning; and Channel Islands French. Her most recent major publication is the *Cambridge Handbook of Endangered Languages* (with Peter Austin), Cambridge University Press, 2011. She is currently preparing a book entitled *Attitudes to Endangered Languages: Identities and Policies* for publication with Cambridge University Press in 2013.

DANIEL SCHNEIDER received his diploma degree in computer science in 2006 from RWTH Aachen University, and his Ph.D. in computer science from Bonn University in the field of large-scale spoken document retrieval in 2012. He currently heads the speech technology group at Fraunhofer IAIS, St. Augustin. His current research focus is on speech recognition and speech search in large digital archives. He is actively involved as a researcher and project manager in several national and European research initiatives.

STEFAN SCHNELL is an Australian Research Council Postdoctoral Fellow at the Centre for Research on Language Diversity at La Trobe University, Melbourne. He has been working on the documentation of Vera'a, an Oceanic language with 350 speakers in North Vanuatu, on the Vures and Vera'a DoBeS project since 2007. In 2010, he received his Ph.D. from Kiel University with a thesis on referentiality and animacy effects in Vera'a morphosyntax. Together with Geoffrey Haig, he has been developing GRAID ("Grammatical Relations and Animacy in Discourse"), an annotation system for the morphosyntactic markup of texts. His current ARC project focusses on quantitative investigations of referentialty and animacy in texts from various endangered languages using the GRAID system.

OLIVER SCHREER is working as scientific project manager of the "Immersive Media & 3D Video" Group in the Image Processing Department of the Fraunhofer Institute for Telecommunications HHI. He received his Dr.-Ing. degree in electrical engineering from the Technical University of Berlin in 1999. Since August 1998, he has been with the IP department, where he is engaged in research for 3D analysis, novel view synthesis, vision-based HCI, real-time 3D video conferencing systems, and immersive media applications. Since autumn 2006, he has been Associated Professor at the Faculty of Electrical Engineering and Computer Science, Technical University Berlin. He is participating continuously in European research projects acting as workpackage leader or coordinator, as in the FP6 project RUSHES. He published more than 80 papers in conference proceedings and journals and he is editor of a book on "3D Videocommunication" published in 2005 by Wiley & Sons, UK.

GABRIELE SCHWIERTZ is an Assistant Professor at the Linguistics Department at the University of Cologne. She has worked on the documentation of Beaver (Athabaskan) within the DoBeS program. Her main research interests are intonation and prosody in lesser-studied languages, multimodal interaction, and issues in language documentation, in particular questions relating to the utilization of LD data.

FRANK SEIFART is currently Senior Researcher in the Department of Linguistics at the Max Planck Institute for Evolutionary Anthropology in Leipzig. He also coordinates a DoBeS project on the relative frequencies of nouns, verbs, and pronouns in nine language-documentation corpora. His main research interests are in linguistic typology, language contact, and language documentation theory and practice. He has done fieldwork on various indigenous languages in the North West Amazon, in particular Bora-Miraña and Resígaro, which he documented in the context of a DoBeS project.

JANE SIMPSON studies Australian Indigenous languages. She has worked on the documentation and maintenance of Warumungu and Warlpiri and on reconstructing languages of the Thura-Yura family from old sources. She is currently working on a longitudinal child language project in several northern and central Australian communities. She is a professor in the School of Language Studies at the Australian National University, Canberra.

SABINE STOLL got her Ph.D. from the University of California, Berkeley, in 2001. After postdoctoral research at the Max Planck Institute for Evolutionary Anthropology in Leipzig, she was appointed head of a new psycholinguistics unit at the University of Zurich in 2011. Stoll's research agenda centers around the question of how children can cope with the incredible variation exhibited across languages. A special focus in this is on the interplay of innate biological factors (such as the capacity for pattern recognition and imitation) with idiosyncratic and culturally determined factors (such as for instance type and quantity of input). Stoll takes a radically empirical approach to these questions, based first and foremost on the quantitative analysis of large corpora that record how children learn diverse languages.

NICK THIEBERGER is an ARC QEII Fellow at the University of Melbourne. He recorded Paakantyi (NSW) speakers in the early 1980s and then worked with Warnman speakers (Western Australia) when he was setting up the Wangka Maya language centre in Port Hedland. He built the Aboriginal Studies Electronic Data Archive (ASEDA) at AIATSIS in the early 1990s and then was at the Vanuatu Cultural Centre from 1994–1997. He wrote a grammar of South Efate, a language from central Vanuatu, which was the first grammar to cite a digital corpus of recordings in all example sentences and texts. In 2003 he helped establish the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC). He taught in the Department of Linguistics at the University of Hawai'i (2008–2010). He is a co-director of the Resource Network for Linguistic Diversity (RNLD) and the editor of the journal Language Documentation & Conservation. He is developing methods for the creation of reusable data from fieldwork on previously unrecorded languages and training researchers in those methods.

PETER TRUDGILL has been Professor of Linguistics at the Universities of Reading and Essex in England and Professor of English Linguistics at the Universities of Lausanne and Fribourg in Switzerland. He is currently Professor of Sociolinguistics at Agder University in Norway, and Honorary Professor of Sociolinguistics at the University of East Anglia in England, as well as Emeritus at Fribourg University. He is a Fellow of the British Academy and has honorary doctorates from the University of Uppsala, Sweden; the University of East Anglia; and La Trobe University, Australia. He has carried out research on dialects of English, Norwegian, Greek, Albanian, and Spanish and has published many books on sociolinguistics and dialectology, including *Dialects in Contact* (1986); and *New-dialect formation: the inevitability of colonial Englishes* (2004). His latest book, with Oxford University Press, is called *Sociolinguistic typology: social determinants of linguistic structure*.

PETER WITTENBURG was technical director at the Max Planck Institute for Psycholinguistics for many years and recently became the head of the newly founded unit "The Language Archive". This unit, funded by the Max Planck Society, the Berlin-Brandenburg Academy of Science and the Dutch Academy of Sciences is devoted to storing and preserving language resources and to developing technology to access and use these resources. The archive contains language data from various fields and in particular data about endangered languages mainly collected in the DoBeS project. The set of technologies being developed ranges from archiving and metadata tools, annotation, and lexicon tools to semi-automatic annotation tools based on statistical methods. Wittenburg is coordinating a team of about 20 highly specialized data scientists and software developers.

Acknowledgements

The Volkswagen Foundation generously financed the workshop that made this volume possible and also the proofreading and typesetting of this publication. Additionally, Volkswagen Foundation representative Vera Szöllösi-Brenig contributed to the discussion at the workshop. The Department of Linguistics, directed by Bernard Comrie, of the Max Planck Institute for Evolutionary Anthropology in Leipzig hosted the workshop. Claudia Schmidt and a team of student assistants took care of the local organization.

Each of the panels of the workshop was convened and coordinated by panel coordinators, most of whom also organized the double-blind peer review process for the respective parts of this volume. For panel one, these were Frank Seifart, Peter Wittenburg, and Daan Broeder; for panel two, Geoffrey Haig, Nikolaus P. Himmelmann, and Anna Margetts; and for panel three, Dagmar Jung and Paul Trilsbeek.

Nick Thieberger encouraged us to submit this volume to the series of Special Publications of Language Documentation & Conservation, Birgit Jänen typeset this volume in LaTeX 2_ε, and Brent Reed proofread its contributions.

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

2

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. 7–16
http://nflrc.hawaii.edu/ldc/sp03/
http://hdl.handle.net/10125/4511

Prospects for e-grammars and endangered languages corpora

Sebastian Drude

Max Planck Institute for Psycholinguistics, Nijmegen

This contribution explores the potentials of combining corpora of language use data with language description in *e-grammars* (or *digital grammars*). We present three directions of ongoing research and discuss the advantages of combining these and similar approaches, arguing that the technological possibilities have barely begun to be explored.

1. INTRODUCTION: GRAMMARS AND LANGUAGE DOCUMENTATION. Grammars, in the sense of comprehensive descriptions of the structural properties of a language, have been at the core of linguistics since its very beginning. All studies that aim to understand the patterns and limits of the variability of the human speech faculty need thorough accounts of more languages than just the few which are more well known (Zaefferer 1998). Grammars are the principal component of the so-called *Boasian triad* (i.e.: grammar, dictionary, texts; cf. Grinevald 2001) that is the customary result of linguistic fieldwork. As such, they are a well-established genre of scientific texts, a genre which recently has gained attention on its own (Ameka et al. 2006, Lehmann 2004a,b, Payne & Weber 2007).

Grammars consist mostly of prose text organized in a hierarchy of sequential chapters and sections. Certain special elements, however, distinguish grammars from other scientific books, in particular *exemplars* (in the sense of Good 2004) such as words, phrases, sentences of the language studied. These exemplars usually come with a translation and some additional analysis; in recent grammars this is often in the form of basic glossings (an interlinearized rendering of the morphs or words of the object language, as standardized by the Leipzig Glossing Rules, Comrie et al. 2008). Other typical elements found almost exclusively in grammars and other linguistic texts are, for instance, paradigms (and similar tables), and, depending on the linguistic theory which underlies the description, formal rules, or structure graphs such as trees indicating the constituent structure of sentences.

Although sometimes idealized as "theory-neutral", all descriptions of languages necessarily rely on general linguistic theories. These provide, in particular, the technical terms which are applied in the description. Often, the writer of a grammar cannot take it for granted that the underlying theoretical concepts are known to the readers. Therefore it is characteristic of many grammars to contain interspersed paragraphs explaining the underlying theory fragments and terms before they are applied to the language described.

A collection of (analysed and translated) texts, another component of the *Boasian triad* mentioned above, is now complemented or even superseded by the outcome of language documentations (in the modern sense as established by Himmelmann 1998). In this sense, documentations consist crucially of multi-purpose digital corpora of data of naturalistic language use, that is, annotated primary multi-media data. One of the uses of documentations is, of course, the study and analysis of the structure of the language, the results of which are traditionally presented in descriptive grammars (see above).

Language documentation, in this sense, crucially depends on new digital technologies (the resulting corpora with their recordings, annotations, and metadata are digital); for this reason among others linguistics has been a pacesetter in the developing field of *digital humanities*. The question now is how grammars can benefit from these new technologies and from the digital corpora that are being compiled. The following sections discuss three recent approaches to this question.[1]

2. HYPERTEXT GRAMMARS. One obvious way to enhance language descriptions with digital technology is to make use of hyperlinks, i.e. the interlinking of different locations within or across documents and other resources, giving rise to what can be called *hypertext grammars*. Already in many paper-based grammars, cross-references abound due to the truly systemic integrated character of the structure of any language. Hyperlinks are an excellent means of making these relations explicit and easy to access.

Perhaps the most needed links, however, are those linking points in the grammar text where a certain phenomenon is described with examples of this phenomenon in the corpus. As has recently been stated many times (e.g., Bird & Simons 2003: 563, Himmelmann 2006: 6), being able to illustrate statements in a grammar (or, say, a typological study) with recordings of language use would make linguistics much more accountable, providing a much more solid basis for the empirical claims and generalizations.

Other links can enhance a grammatical description. For instance, for occurrences of individual (forms of) words or morphemes discussed in the text, one would like to be directed immediately to a corresponding entry in a lexical database (or electronic dictionary). Also, details of the underlying theoretical framework could, at last, be presented apart from the description that applies these concepts, but static links or intelligent search mechanisms between both resources could provide the needed contextualization. Figure 1 (cf. Drude forthcoming) demonstrates these key features and links of a hypertext grammar.

As explained in what follows, the boxes in this figure mark elements that presuppose certain assumptions for a concrete implementation, applying certain solutions for some of the technical challenges of conceiving and implementing hypertext grammars, especially solutions which are being developed at The Language Archive (at the Max-Planck-Institute for Psycholinguistics in Nijmegen). In particular, the external resources, corresponding to the other components of the Boasian triad (and to theoretical work explaining the underlying theories and applied terms) could be instantiated by existing LEXUS, ANNEX, and ISOcat tools belonging to Language Archiving Technology (LAT, developed at the MPIPL).

[1] These and several other recent developments and projects were presented at the symposium on electronic grammaticography, organized by Sebastian Nordhoff, as part of the 2nd Conference on Language Documentation and Conservation 2011 in Hawai'i.

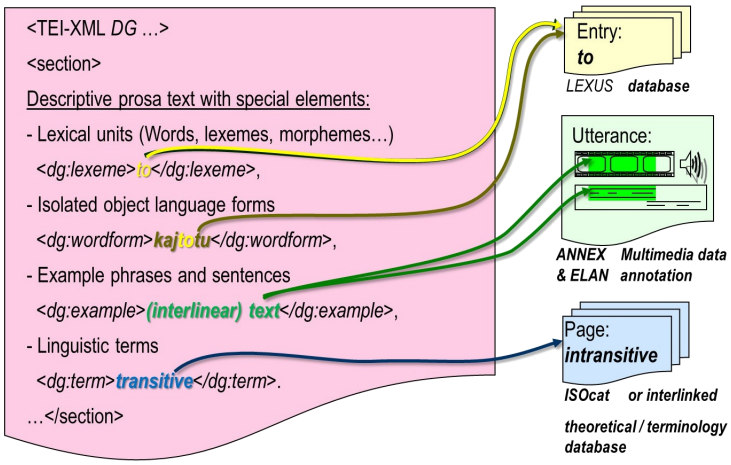


FIGURE 1: Elements and relations in a page in a hypertext grammar

For instance, an online lexical database can be developed or at least rendered with LEXUS.[2] The text corpus (containing examples) can be a collection of ELAN[3] annotation files together with the underlying primary data. If these are provided by a LAT repository,[4] the corresponding examples can be visualized with the ANNEX online service.[5] The technical terms in a description can be linked to their definition in the ISOcat registry[6] (for more comprehensive explanations of theoretical concepts, a theory-specific wiki or something similar could be employed, which still can make use of ISOcat as a point of reference).

Crucially, all elements of the external resources need some kind of persistent identifier that guarantees that the links remain stable over time even if physical locations and infrastructure change. LAT provides Handle Digital Object Identifiers.[7]

As indicated by the boxes on the left-hand side, the text body itself would probably best be encoded in some XML-based format, where the XML-tags allow specification of the hyperlinks from examples or other particular elements in the text. The norms proposed by the Text Encoding Initiative (TEI, cf. TEI Consortium 2009) could provide the basis for such a format, but they probably would have to be extended in order to serve the special needs of a digital grammar.

There are other aspects of hypertext grammars that have been explored in particular by Nordhoff (2007, 2008), for instance the conceptualization of a grammar as a living document such as a wiki. Also, others are working on certain aspects of related technology which could be integrated into a comprehensive system; an example is the EOPAS system

[2] Cf. http://tla.mpi.nl/tools/tla-tools/lexus.

[3] Cf. http://tla.mpi.nl/tools/tla-tools/elan.

[4] Cf. http://tla.mpi.nl/tools/tla-tools.

[5] Cf. http://tla.mpi.nl/tools/tla-tools/annex.

[6] Cf. http://tla.mpi.nl/tools/tla-tools/isocat.

[7] Cf. http://www.doi.org/about_the_doi.html.

being developed by Thieberger and others (Schroeter & Thieberger 2006), which provides easy direct access to spoken language examples.

3. TREEBANKS AND THE GRAMMAR MATRIX PROJECT. Hypertext grammars are conceptually close to paper grammars, although the style of writing and aspects of the content will be affected by the digital medium. But there are also *implemented grammars*, the digital equivalent to grammars as developed by computational linguistics, in particular grammar engineering. Rather than texts directed to a human reader, implemented grammars are procedural representations capable of, for instance, parsing and analyzing written word forms and sentences of a language.

As such, implemented grammars are not just tools for automatic annotation (parsing, glossing, structure assignment, etc.) but also aim at a technical representation of what we understand of the language structure – what the implemented grammar is not able to analyze points at possible gaps in our understanding of the language structure on the respective level, or at least gaps in the technical representation of our understanding. At the same time, the technical formal representation of rules/features allows for a much more precise comparison between languages, for instance for typological or historical-reconstructive research, and the same holds for the much more standardized automatically generated annotations generated by such systems.

Most implemented grammars apply only to one each of a small set of better-studied languages. For the study of linguistic diversity and less well-known languages, generic parser engines are needed that can understand different separately developed sets of rules, tailored to different languages. Kirschenbaum et al. (this volume) present research on machine learning for morphological analysis, both supervised and unsupervised. This research shares the perspective of developing systems for the automatic annotation of text in corpora, but is *per se* not rule based, i.e. it does not presuppose a technical representation of the system of the language.

One particularly promising project is the Linguistic Grammars Online (LinGO) Grammar Matrix project being developed by E. Bender and colleagues (Bender et al. 2010). In this project, a generic program applies language-specific rules to sentences and proposes syntactic trees (according to the HPSG theoretical model), which can be included in a *treebank*, a database of such trees (in recent years, one of the most often used resources for major languages).

To be more specific, the LinGO project does not speak of rules but rather of *signs*, an HPSG term that covers not only lexical units but also grammatical classes or word order patterns, each assigned to a semantic interpretation. So far, traditional descriptions have been translated into such *signs* manually by linguists working in cooperation with information scientists. For each sentence in a text, the parsing mechanism then offers a number of possible trees compatible with the signs known to the system, which is able to learn and remember the tree chosen by a linguist.

In the *Grammar Matrix* (GM), a more recent track of research by Bender and her colleagues (2010), the signs are (semi-)automatically derived from a typological profile of the language, which is elicited from the linguist in the form of a questionnaire. This is a major advance since less technical knowledge (or engineering work of a technical specialist) is needed.

Figure 2 (Bender et al. 2010: 29) represents an overview of the GM project with the elicitation of typological information (left), which enters the creation (right) of a customized grammar (i.e. a steadily improving language specific parsing automatism) together with a language independent core grammar and the analyses for previously parsed sentences.

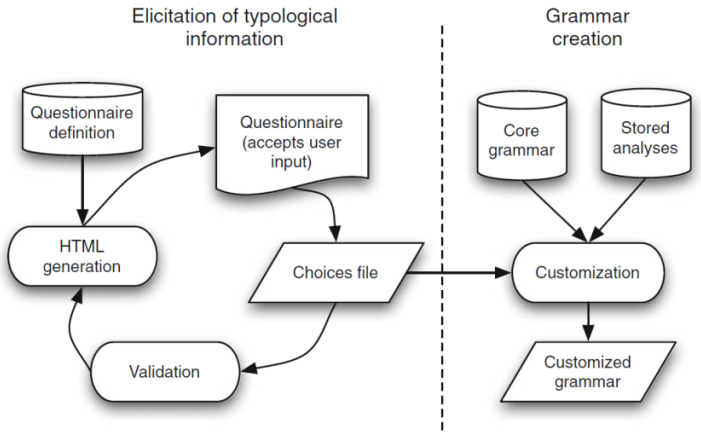


FIGURE 2: Grammar matrix components and workflow

The following screenshot from the LinGO English Resource Grammar (ERG)[8] shows the result of parsing an English sentence with the first two possible tree structures (they are ordered by probability according to what can be deduced from earlier parses of similar sentences). The formalism is, of course, highly theory-dependent, but similar systems could be conceived for other sufficiently formal explicit linguistic frameworks.

In the future, Bender and colleagues plan to develop a system that automatically derives rules (*signs*) from a sufficiently large set of texts with interlinear glosses (although these only contain morphological information and only incidentally syntactic features which are needed for constructing syntactic trees).

4. INTEROPERABLE GRAMMARS AND LITERATE PROGRAMMING. A drawback of most current parsing mechanisms is that the technical rules are difficult to build and work only with the one generic parsing mechanism they have been developed for. If a new parser is developed, the rules all have to be coded manually again. They are also not easily understood by humans and are not explicitly linked to a traditional description, although individual rules usually correspond to specific parts of a paper grammar or dictionary.

A current project developed by Maxwell and others (Maxwell & David 2008) promises to overcome these shortcomings. They build *Interoperable Grammars* (IGs), which should be comprehensible to both humans and machines. The idea is to apply the concept of *literate programming* (Knuth 1984) to grammar writing, arranging the prose descriptions around technical parts, which translate the described structural properties (rules) into a format that

[8] http://erg.delph-in.net/ (1 November, 2011).

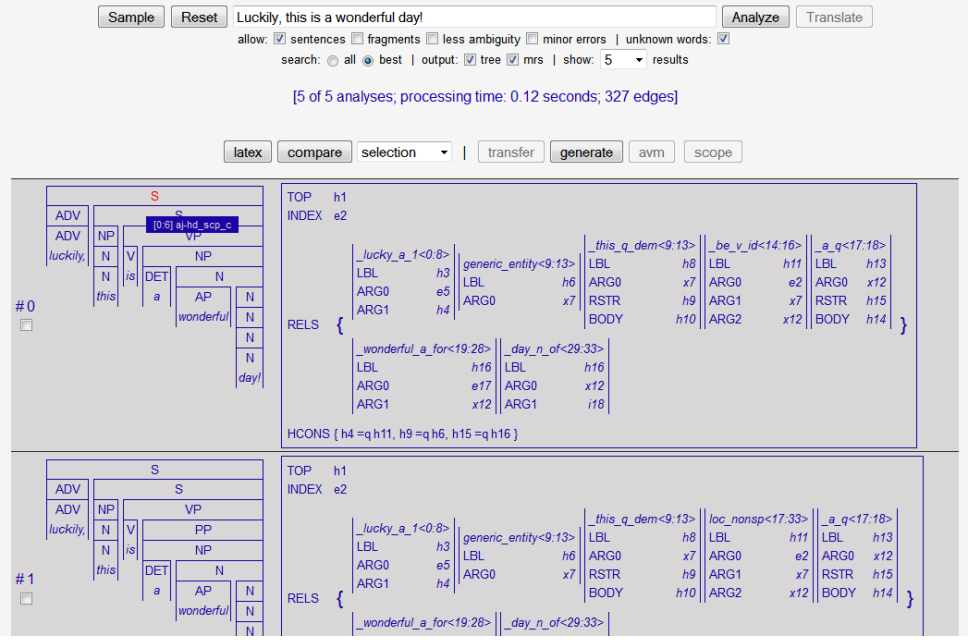


FIGURE 3: Automatically generated tree from the LinGO English Resource Grammar

can be understood by a software system. The text is the explanation, as it were, of the program code.

Crucially, the XML-based format for the rules is independent of any individual parser but allows for automatic transformation into formats usable by the respective parsers. In this manner, the code remains compatible with the evolving and frequently superseded parsers. A new parser will only need a corresponding pre-processor for rendering the rules in its specific format.

The parser applied by the IGs so far deals only with morphology (different from those in GM), and so the rules currently cover mostly morphemes and word-internal phonological variation. The outcome of applying a tailored parser to a text is a morphologically analysed and tagged text (basically, an interlinearized text).

The following diagram (Figure 4, from Maxwell forthcoming) shows the derivations of the different files and components of the IG system, where the publishable (text) grammar and the formal grammar are derived from one and the same XML document. The formal grammar is converted into a parser engine specific format and is combined with lexical information. The resulting specific parser instance can be applied to texts in the target language.

The following XML-snippet (Figure 5, from Maxwell & David 2008) exemplifies the generic technical representation of grammatical features, in this case a certain first-person-future-indicative affix with two allomorphs. (The use of UNICODE provides a means to deal with the non-western script without difficulties.)

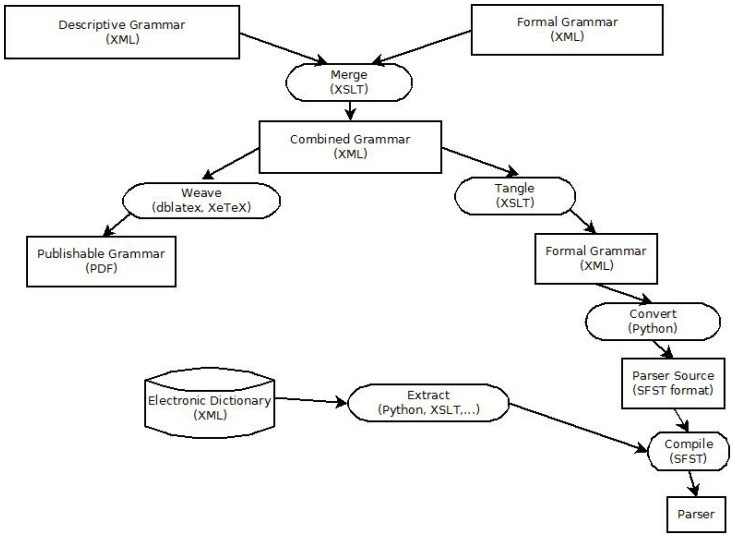


FIGURE 4: Components and workflow in the Interoperable Grammars project

```
<Mo:InflectionalAffix gloss="-1Fut" id="af1Fut">
  <!--The two "allomorphs" are really allographs-->
  <Mo:Allomorph form="(ৗ)ব">
    <!--Spelled 'bo'; usually (not always) after a C-stem -->
  </Mo:Allomorph>
  <Mo:Allomorph form="ব">
    <!--Spelled 'b'; usually (not always) after a vowel stem -->
  </Mo:Allomorph>
  <Mo:inflectionFeatures>
    <Fs:f name="Tense"><Fs:symbol value="Future"/></Fs:f>
    <Fs:f name="Mood"><Fs:symbol value="Indicative"/></Fs:f>
    <Fs:f name="Person"><Fs:symbol value="1"/></Fs:f>
  </Mo:inflectionFeatures>
</Mo:InflectionalAffix>
```

FIGURE 5: XML-snippet showing the representation of a grammatical feature

5. CONCLUSION: PROSPECTS FOR DIGITAL GRAMMARS AND LANGUAGE DOCUMENTATION. There are several other projects on implemented grammars which cannot be discussed any further here, for instance *Grammix* (Müller 2007), *ParGram* (Butt et al. 2002), *Meta-Grammar* (Kinyon et al. 2006), *OpenCCG* (Baldridge et al. 2007), or *KPML* (Bateman et al. 2005). Some of these may contribute features, conceptions, or functionalities which can be picked up by other future more integrated projects and systems. Thus, there is clearly a need for more interaction between field linguists and computational linguists. There is a huge potential in integrating the emerging systems, which are in several aspects complementary to one another, and in others resemble one another.

For instance, most computational approaches to grammar(s) aim at parsing sentences of a corpus; but the IG system is currently restricted to morphology, and the GM framework focuses on syntactic structure. The architecture of both implemented grammars relies on a generic parsing engine configured with customized rules and trained with data from individual languages. This approach is clearly most suitable for describing the plenitude of understudied languages. Advantages include gaining richer textual data and spotting gaps in analysis and description.

However, one point which continues to be problematic for all parsers, so far, is the often elliptical and in many other ways non-standard character of natural spontaneous spoken language. Syntactic parsers can usually only cope with complete grammatical sentences, or at least much better so. It remains to be explored and solved as to how parsers can be adjusted to deal with variability as typical for spoken language. (Some attempts have been made for more well-known languages, e.g. Nivre & Grönqvist 2001 for Swedish.)

Hypertext Grammars may seem uninteresting and even boring in comparison to implemented grammars since they do not create any new content or annotation. Still, linguistics needs an easy way to interlink scholarly linguistic work with corpora, lexica, and terminological / theoretical texts. Each of the pioneering works by C. Lehmann, S. Nordhoff, and N. Thieberger and others cover some aspects, but none of them covers all or even a major part of the necessary or desirable aspects which were detailed above in Section 2, and they do not address the integration with implemented grammars exemplified in Sections 3 and 4.

Overall, in a mid or long term perspective it would be ideal to combine all three (and possibly still other) approaches into one integrated framework which allows the linguist to create descriptions that satisfy two important requirements. 1) They match the high standards of traditional grammatical descriptions, including an overall pedagogically informed exposition. 2) They allow the reader to a) access the underlying primary data as well as other resources, and also b) to check the validity of derived technical or formal rules against a corpus of sentences, producing richer annotation. Automated processes would also result in more standardized annotations, which would be more suitable for cross-linguistic comparisons and typological work.

An ideal future framework would be modular; various parsers based on different theoretical frameworks could be chosen for different purposes (like the algorithms presented by Kirschenbaum et al. in this volume, or the IG morphological parser, or the HPSG/based trees created by the GM framework). Core aspects of an integrated authoring and reading environment are the hypertext interlinking with external resources and perhaps the IG *literate programming* conception. Additional possible modules not discussed here include statistical and supervised and unsupervised machine learning methods. The possibilities of combining such different approaches have not yet been explored, even initially.

With such a combination, we would gain: a) more comprehensive, empirically sound and accountable grammatical descriptions, b) more comparable and richly annotated corpora, and as a result c) a deeper understanding of language variation and ultimately even d) a ground-laying conceptual and technical framework for understanding language structure and the meaning of linguistic constructions, a necessary condition for machine-translation, as it were, the holy grail of computational linguistics.

REFERENCES

Ameka, Felix K., Alan Dench & Nicholas Evans (eds.). 2006. *Catching Language: The Standing Challenge of Grammar Writing* Trends in Linguistics. Studies and Monographs 167. Berlin: De Gruyter.

Baldridge, Jason, Sudipta Chatterjee, Alexis Palmer & Ben Wing. 2007. DotCCG and VisCCG: Wiki and programming paradigms for improved grammar engineering with OpenCCG. In Tracy Holloway King & Emily M. Bender (eds.), *Proceedings of the GEAF 2007 Workshop Stanford, CA. CSLI* CSLI Studies in Computational Linguistics ONLINE, 5–25. Stanford: CSLI.

Bateman, John A, Ivana Kruijff-Korbayová & Geert-Jan Kruijff. 2005. Multilingual resource sharing across both related and unrelated languages: An implemented, open-source framework for practical natural language generation. *Research on Language and Computation* 3(2–3). 191–219.

Bender, Emily M., Scott Drellishak, Antske Fokkens, Laurie Poulson & Safiyyah Saleem. 2010. Grammar customization. *Research on Language and Computation* 8(1). 23–72.

Bird, Steven & Gary Simons. 2003. Seven dimensions of portability for language documentation and description. *Language* 79(3). 557–582. http://www.sil.org/~simonsg/preprint/Seven%20dimensions.pdf (5 April, 2012).

Butt, Miriam, Helge Dyvik, Tracy Holloway King, Hiroshi Masuichi & Christian Rohrer. 2002. The parallel grammar project. In John Carroll, Nelleke Oostdijk & Richard Sutcliffe (eds.), *Proceedings of the Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics*, 1–7. http://www.aclweb.org/anthology/W/W02/W02-1503.pdf (5 April, 2012).

Comrie, Bernard, Martin Haspelmath & Balthasar Bickel. 2008. *The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses*. Leipzig: Max Planck Institute for Evolutionary Anthropology & Department of Linguistics of the University of Leipzig. Online at http://www.eva.mpg.de/lingua/pdf/LGR08.02.05.pdf (5 April, 2012).

Drude, Sebastian. forthcoming. Digital grammars: Integrating the Wiki/CMS approach with Language Archiving Technology and TEI. In Nordhoff, Sebastian (ed.), Electronic Grammaticography. University of Hawai'i Press. [Talk given at the grammaticography colloquium at the 2.ICLDC, Hawai'i, Februrary 2011].

Good, Jeff. 2004. The descriptive grammar as a (meta)database. In *Proceedings of the E-MELD Workshop 2004: Linguistic Databases and Best Practice, July 15–18, 2004, Detroit, Michigan*, .

Grinevald, Colette. 2001. Encounters at the brink: Linguistic fieldwork among speakers of endangered languages. In Osamu Sakiyama (ed.), *Lectures on Endangered Languages 2*, 285–313. Osaka: ELPR Publications.

Himmelmann, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36(1). 161–195. Full version online at http://www.hrelp.org/events/workshops/eldp2005/reading/himmelmann.pdf (5 April, 2012).

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: Mouton de Gruyter.

Kinyon, Alexandra, Owen Rambow, Tatjana Scheffler, SinWon Yoon & Aravind K Joshi. 2006. The metagrammar goes multilingual: A cross-linguistic look at the V2-phenomenon. In *Proceedings of the 8th International Workshop on Tree Adjoining Grammar and Related Formalismss*, 17–24. Sydney: Association for Computational Linguistics.

Kirschenbaum, Amit, Peter Wittenburg & Gerhard Heyer. this volume. Unsupervised morphological analysis of small corpora: First experiments with Kilivila.

Knuth, Donald E. 1984. Literate programming. *The Computer Journal* 27(2). 97–111.

Lehmann, Christian. 2004a. Documentation of grammar. In Osamu Sakiyama, Fubito Endo, Honore Watanabe & Fumiko Sasama (eds.), *Lectures on endangered languages: 4. From Kyoto Conference 2001* Endangered Languages of the Pacific Rim Publication Series, C-004, 61–74. Osaka: Osaka Gakuin University.

Lehmann, Christian. 2004b. Funktionale Grammatikographie. In Waldfried Premper (ed.), *Dimensionen und Kontinua: Beiträge zu Hansjakob Seilers Universalienforschung* Diversitas Linguarum 4, 147–165. Bochum: Brockmeyer.

Maxwell, Michael. forthcoming. Electronic grammars: Taking advantage of the possibilities. In Nordhoff, Sebastian (ed.), Electronic Grammaticography. University of Hawai'i Press. [Talk given at the grammaticography colloquium at the 2.ICLDC, Hawai'i, Februrary 2011].

Maxwell, Michael & Anne David. 2008. Interoperable grammars. In *Proceedings of the First International Conference on Global Interoperability for Language Resources (ICGL 2008)*, Hong Kong. http://hdl.handle.net/1903/11611 (5 April, 2012).

Müller, Stefan. 2007. The Grammix CD-ROM: A software collection for developing typed feature structure grammars. In Tracy Holloway King & Emily M Bender (eds.), *Proceedings of the GEAF 2007 Workshop Stanford, CA. CSLI* CSLI Studies in Computational Linguistics ONLINE, Stanford: CSLI.

Nivre, Joakim & Leif Grönqvist. 2001. Tagging a corpus of spoken Swedish. *International Journal of Corpus Linguistics* 6(1). 47–78.

Nordhoff, Sebastian. 2007. The Grammar Authoring System GALOES. Talk given at the Workshop on "Wikifying research". MPI Leipzig. http://www.eva.mpg.de/lingua/staff/nordhoff/pdf/The%20grammar%20authoring%20system%20GALOES.pdf (5 April, 2012).

Nordhoff, Sebastian. 2008. Electronic reference grammars for typology: Challenges and solutions. *Language Documentation and Conservation* 2(2). 296–324. http://hdl.handle.net/10125/4352 (5 April, 2012).

Payne, Thomas E & Davis J Weber (eds.). 2007. *Perspectives on Grammar Writing*. Amsterdam: Benjamins.

Schroeter, Ronald & Nicholas Thieberger. 2006. EOPAS, the EthnoER online representation of interlinear text. In *Sustainable Data from Digital Fieldwork. Proceedings of the conference held at the University of Sydney, 4–6 December 2006*, Sydney: Sydney University Press. http://hdl.handle.net/2123/1297 (14 December, 2011).

TEI Consortium. 2009. TEI P5: Guidelines for electronic text encoding and interchange, 1st edn. http://www.tei-c.org/Guidelines/P5/.

Zaefferer, Dietmar. 1998. *Deskriptive Grammatik und allgemeiner Sprachvergleich* Linguistische Arbeiten 383. Tübingen: Niemeyer.

Sebastian Drude
Sebastian.Drude@mpi.nl

3

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. 17–24
http://nflrc.hawaii.edu/ldc/sp03/
http://hdl.handle.net/10125/4512

Language-specific encoding in endangered language corpora

Jost Gippert
Goethe University of Frankfurt/Main

The paper addresses problems of corpus building and retrieval resulting from code-switching, which is a characteristic feature of endangered language recordings. The typical appearance of code-switching phenomena is first outlined on the basis of data collected in the DoBeS 'ECLinG' project, which dealt with three endangered Caucasian languages spoken in Georgia: Tsova-Tush (Batsbi), Udi, and Svan. The problem of language-specific retrieval is illustrated with examples showing the usage of the word *da* in Tsova-Tush contexts, which represents, as a homonym, either a native copula form ('it is') or the Georgian conjunction 'and'. The subsequent section discusses the annotation requirements that are necessary to automatically distinguish the languages involved in code-switching, with a focus on the emerging ISO standard 639-6. It is argued that the fine-grained distinction of varieties and subvarieties and their interrelationship – as aimed at in this standard – requires a thorough reconsideration if it is to be applied in the markup of corpus data.

1. INTRODUCTION. It is well known that recorded texts of natural speech in endangered languages abound in code-switching, mostly between the endangered vernacular and dominant languages, but also other languages involved in the bi- and multilingual settings that are typical for language endangerment. This multilingual data is crucial for all kinds of language-specific or cross-linguistic research into endangered languages, as well as for the theory of language endangerment in general (see also Gullberg, this volume) .[1] However, at present, annotation schemes such as those developed in the DoBeS framework do not admit of an easy differentiation of linguistic units pertaining to different linguistic layers, and language-specific search functions are still wanting. The present paper first illustrates the presence of multiple languages in the documentation of Caucasian languages (section 2) and then discusses ways to cope with this, considering, among other things, the advantages of the emerging ISO standard 639-6 'Language Names' (section 3).

[1] Cf. Gippert 2008: esp. 174–188, for a case study based on the three Caucasian languages Svan, Tsova-Tush and Udi. Cf. Gullberg (this volume) for a more general view on the impact of bi- and multilingualism in endangered language communities for linguistic theory.

2. GEORGIAN ELEMENTS IN TSOVA-TUSH (BATSBI) AND UDI. The effect of a missing distinction between the languages involved in bi- or multilingual settings can easily be demonstrated with the materials that were collected between 2003 and 2007 by the DoBeS 'ECLinG' project, which addressed three endangered Caucasian languages spoken in Georgia, viz. Svan, Tsova-Tush (Batsbi), and Udi. In the text recordings that were provided by the project to the DoBeS Archive[2], we can clearly see that Georgian as the dominant language of the area in question has left its traces everywhere in both monologic and dialogic speech of speakers of all generations.

In the case of Tsova-Tush, an East-Caucasian ('Nakh') language closely related to Chechen and Ingush but unrelated to (South-Caucasian) Georgian, this has brought about a peculiar homonymy, given that one of its most frequent verb forms, the copula form *da* '(it) is', is indistinguishable from the most frequent particle of Georgian, the conjunction *da* 'and'. Executing an 'annotation content search' for the word form *da* in the DoBeS Tsova-Tush materials with the TROVA tool[3] yields both Tsova-Tush and Georgian contexts from the annotated text recordings,[4] with the latter representing ca. 10%. Among them we find Georgian *da* 'and' in the following contexts:

(a) utterances not pertaining to a given narrative but addressing people present in the recording session as in *is sḳami, gadmodit, švilo,* ***da*** *axlos daǯekit, ḳaco* 'That chair, come over, boy, **and** sit down close by, man!';[5]

(b) sentences of reported Georgian speech inserted into a Tsova-Tush narrative as in the case of *ṗeṭresac avaġebine švebulebao* ***da*** *čamovedit alvanšio…* "'I will make Peter take a vacation (too), **and** let's go down to Alvani"…(he said)', introduced by Tsova-Tush *ḳoxiv var – ḳoxiv…* 'He was a Georgian, a Georgian';[6]

(c) idiomatic formulae such as *me magisi ase* ***da*** *ise* 'I …his …this **and** that way', i.e., 'I could do his mother this and that', embedded between the Tsova-Tush sentences *oḳuyxvas āl'i*^{n} *sog mē aḥ bʿivnaḥēn* 'That Kakhetian (man) said to me, "you killed

---

[2] Cf. http://corpus1.mpi.nl/ds/imdi_browser?openpath=MPI533677%23. The recordings stored in the archive consist of about 500 texts (ca. 70 hrs.) for Svan, 312 texts (ca. 37 hrs.) for Tsova-Tush, and 43 texts (ca. 6 hrs.) for Udi; all recordings are fully transcribed and provided with a Georgian and English translation, ca. 10% additionally with a multilevel grammatical analysis.

[3] Cf. http://corpus1.mpi.nl/ds/annex/search.jsp?transferuid=1&nodeid=MPI534223%23&row=29. The search yields 774 hits (27.11.2011, 20:01h) in 'Single Layer' mode set to 'exact match'. A similar result (768 hits) is achieved searching for *da* in Georgian script because the annotations were mostly provided in both Georgian and Latin scripts.

[4] As an example of Tsova-Tush *da* 'it is' we may quote the sentence *vir ma aṭṭan da, ḳaǩon daʾoḥe, davina da* 'However, the donkey is light, it is small, it is light' (from a monologue on donkey breeding, http://corpus1.mpi.nl/ds/annex/runLoader?nodeid=MPI793920%23&time=170662&duration=666&tiername=trs@AS; the sentence in question starts at 00:02:46 in the recording).

[5] In a monologic account on Tushian house-building, cf. http://corpus1.mpi.nl/ds/annex/runLoader?nodeid=MPI793880%23&time=107000&duration=500&tiername=tl@EC, sentence starting at 00:01:45.

[6] In a biographical narrative, cf. http://corpus1.mpi.nl/ds/annex/runLoader?nodeid=MPI793894%23&time=215816&duration=727&tiername=tl@EC, sentence starting at 00:03:33. Note that the interviewer admonishes the narrator to return to the Tsova-Tush language by interjecting *vēġeš…vēġeš…vēġeš…*, i.e. 'in our (language), in our (language), in our (language)' after the quoted sentence. – Note also that Tsova-Tush *var* 'he was' and Georgian *var* 'I am' form another remarkable pair of homonyms.

it"', and *as ōqpinivḥē, beḳxētlex co vas^{o}, moḥ āl'i^{n}* 'I had released him, (so) I am really surprised how he could say (so)';[7]

(d) in insertions of Georgian geographical denominations such as *zemo* ***da*** *kvemo alvani* 'Upper **and** Lower Alvani' (in the given case dependent as a quasi-genitive on *amgēgmav* 'planner', which is in turn an integrated loan from Georgian *(da)mgegmavi* 'id.'),[8] but also in

(e) Georgian phrases mixed without any obvious motivation with Tsova-Tush contexts as in the case of *isev ċava* ***da*** *mova, ċaiq̇vans* 'she (the cat) will go **and** come back (and) bring it away', linked as an apodosis to a Tsova-Tush protasis, *me qe eyl'čeḥatx me dolix o qenaʾā doʾdebēn* 'afterwards, when we tell her "come on, bring that other one away, too"'.[9]

In one instance, we even find the inner-Georgian homonym of *da* 'and', viz. the noun *da* 'sister', in a Georgian sentence embedded in a Tsova-Tush context: with *beladiant enċeras... beladiant elane ro ari, ḳutxeši rom cxovrobs moxuci kali, ai imis* ***da*** *iq̇o* 'Beladianti Entsera...(she) who is Beladianti Elane, the old woman who lives at the corner, look, her **sister** it was', the speaker replies to the question *veċer ḥan?* 'Who was in love with him?'.[10]

In the case of Udi, things are different in that we have no homonymous equivalent of Georgian *da* in this language. Nevertheless we arrive at 63 hits of *da* searching through the Udi ECLinG corpus,[11] all representing the Georgian conjunction. Different from the Tsova-Tush examples illustrated above, we here even find cases where *da* is not used in a longer Georgian phrase or chunk but isolated, in a plain Udi context, as if being a loan; cf., e.g., *zu qayzupe metġo tärämišbaki garxox* ***da*** *evaxṭe qayzupe...* 'I discovered the places of their emergence, **and** when I had discovered them...',[12] or *mya buyanq̇e q̇eiri, q̇eiri ʒüra ḳinigiux serbayan, manote bakale oxari äʿyluġo baxṭink,* ***da*** *manote äʿyluġo sṭudenṭġo baxṭink, aspiranṭur baxṭink...* 'Here we want to make other, other types of books, which

[7] In a narrative on a shepherd's life, cf. http://corpus1.mpi.nl/ds/annex/runLoader?nodeid=MPI689484%23&time=296926&duration=642&tiername=tl@EC, sentence starting at 00:04:53. The fixed Georgian formula *mainc da mainc* 'nevertheless, however' occurring several times in Tsova-Tush contexts can already be taken to be a loan.

[8] In the biographical narrative mentioned above, cf. http://corpus1.mpi.nl/ds/annex/runLoader?nodeid=MPI793894%23&time=296800&duration=600&tiername=tl@EC, phrase starting at 00:04:55.

[9] In a dialogue on cat breeding, cf. http://corpus1.mpi.nl/ds/annex/runLoader?nodeid=MPI793914%23&time=145800&duration=400&tiername=tl@CD, sentence starting at 00:02:21.

[10] In a dialogue discussing the contents of a folk song, cf. http://corpus1.mpi.nl/ds/annex/runLoader?nodeid=MPI793871%23&time=30929&duration=533&tiername=tl@EA, sentence starting at 00:00:24.

[11] With the TROVA tool; cf. http://corpus1.mpi.nl/ds/annex/runLoader?nodeid=MPI1360405%23&time=247290&duration=1458&tiername=tl@MN.

[12] In a monologue on the origin of the Udi people, cf. http://corpus1.mpi.nl/ds/annex/runLoader?jsessionid=BE2633512018044F706DBF19BD1CDDAF&nodeid=MPI1360405%23&time=247290&duration=1458&tiername=tl@MN, sentence starting at 00:04:00.

will be easy (to read) for the children, **and** which (will be) for the children, the students, for becoming aspirants…'[13]

3. DEMARCATION OF LANGUAGES. It is clear that any automatic retrieval mechanism aiming at a distinction of the Georgian elements in Tsova-Tush or Udi contexts presupposes an adequate demarcation of the languages in question. In the ECLinG project, we have, for lack of more suitable means, started by inserting curly braces to mark the beginning and end of Georgian insertions.[14] This is not an expedient method, however, as braces may easily be neglected by retrieval engines (and the TROVA search function of the DoBeS Archive does neglect them). Instead, a consistent language-specific retrieval would require the linguistic affinity to be marked for every single word form,[15] a task that can easily be achieved using a semi-automatic annotation software such as the Summer Institute of Linguistics (SIL) Toolbox[16] where the information in question can be stored in a lexicon and transferred to annotation tiers in the text files (see Figure 2 below). This, however, presupposes a thorough grammatical analysis of the texts which would require the morphology of the 'mixed-in' language to be accounted for alongside that of the 'basic' vernacular (cf. the case of Georgian *da* sister in the Tsova-Tush example above which would have to be defined as a Georgian nominative or absolutive singular). This task, too, could be fulfilled in connection with an additional lexicon-based markup, but 'complete' grammatical annotations of this type cannot always be provided in the course of a given documentation project. As a matter of fact, only ca. 10% of the ECLinG data could be prepared in this way so that the searches are mostly restricted to the sentence level, which does not allow for a markup of individual words.

A peculiar problem arises if a language-specific search is to be executed not within a given corpus (with, maybe, an idiosyncratic demarcation of languages) but across resources of different origins. In this case it is inevitable to provide the information as to the linguistic affinity of word-forms in a standardized way. As a matter of fact, unique codes denoting languages have been the object of standardization endeavors for many years,[17] and computer users have for long been acquainted with two-letter codes such as EN for English or DE for German indicating the keyboards they use or other language-relevant information.[18] Dealing with endangered vernaculars, two-letter codes of this type are of little help, however, given that it is a maximum of (26^2 =) 676 languages that can be assigned by a pair of characters, and languages such as Tsova-Tush/Batsbi, Udi, or Svan are not among those

[13] In a monologue on the foundation of an Udi school, cf. http://corpus1.mpi.nl/ds/annex/runLoader?nodeid=MPI1360403%23&time=56676&duration=473&tiername=tl@MN, sentence starting at 00:00:51. – A notable Georgian-Udi homonym occurring in the texts is *xe*, which means 'water' in Udi and 'tree' in Georgian.

[14] In a similar way, square brackets have been used to denote Russian passages. The same denotations were also used in the materials of Caucasian languages recorded in the 'SSGG' project ('The sociolinguistic situation of present-day Georgia', project funded by the Volkswagen Foundation from 2005 to 2009) which are as well stored in the Archive of the MPI Nijmegen (cf. http://corpus1.mpi.nl/ds/imdi_browser?openpath=MPI663243%23).

[15] A prototypical distinction of linguistic affinities as represented in the ECLinG and SSGG recordings has been developed for the TITUS search engine which covers the texts of the recordings, too (cf. http://titus.fkidg1.uni-frankfurt.de/database/titusinx/titusinx.htm).

[16] Cf. http://www.sil.org/computing/toolbox/ for the software in question.

[17] They are the objects of the ISO standard 639 ('language names').

[18] The two-letter codes are standardized in ISO 639-1, a sub-standard of ISO 639.

registered in the standard. To overcome this, a standard consisting of three-letter codes (ISO 639-3) was conceived a few years ago,[19] under the aegis of the SIL as a 'registration authority'.[20] Albeit this standard would theoretically comprise (26^3 =) 17,576 entries, only about 6,900 codes have been assigned so far, obviously in accordance with SIL's 'Ethnologue' and the '6,909 languages' identified in it.[21] It is true that among these, we do find codes for Tsova-Tush ('Bats', BBL), Udi (UDI), and Svan (SVA), but there is no distinction possible yet of dialectal variants such as Upper Bal, Lower Bal, Lashkh, and Lentekh in the case of Svan or Vartashen (Oghuz) and Nidzh (Nij) Udi.[22] We must further consider that the elements 'mixed in' in code-switching are not necessarily representative of a given 'standard language' but usually dialectally or sociolectally biased.[23] Therefore it is clear that a much more fine-grained reference system is needed to adequately represent the diversity we are dealing with in the contexts of endangered languages.

Such a reference system has recently been initiated, with the four-letter code inventory of ISO 639-6, which is meant to cover all human language varieties including dialects, sociolects, historical stages, and the like. Different from the former sub-standards of ISO 639, the new standard, which implies a maximum of (26^4 =) 456,976 individual assignments,[24] is not restricted to a mere list of entries but comprises information as to the mutual interdependency of entries in terms of parent-child-relations; a system that would help a lot indeed if, e.g., a given search is not to be restricted to a given variety but to be expanded to a larger scope. Unfortunately, a first analysis of the standardization work undertaken by the responsible Technical Committee of the International Standardization Organization (ISO/TC 37/SC 2) and the institution authorized for the registration of the codes[25] reveals remarkable inconsistencies in the varieties accounted for and their hierarchical arrangement. E.g., we do find 'Spoken Bats' with the code BBLS as a 'child' of Bats, i.e. Tsova-Tush (BBL), and the latter is correctly subordinated to NXAX, i.e., the 'Nakh' subfamily of (North-)East-Caucasian languages (CCNE). Similarly, we find the Tushian ('Tush') dialect of Georgian (TXSH) as a child of KATS, i.e. 'Georgian spoken', in its turn depending on KAT = 'Georgian', which is a child of GGNC = 'Georgian cluster' and a grand-child of CCNS = 'South Caucasian'. On the other hand, Georgian dialects such as Imeretian ('Imeruli', IMRI), Rachian ('Rachuli', RCLI), Gurian ('Guruli', GRLI) or sociolects such

[19] In 2007; cf. http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=39534.

[20] Cf. http://www.sil.org/iso639-3/.

[21] Cf. Lewis 2009. - How dubious the calculation of languages in 'Ethnologue' is, becomes obvious immediately if we consider that it contains 21 entries (with appertaining three-letter-codes) under 'High German' (including 2 varieties of Yiddish) plus 10 entries under 'Low Saxon', but only 2 entries under 'English' (viz. 'English' and 'Scots'). As the criteria and standards applied for counting vary between different countries, regions, or investigators, the number of 6,500 languages world-wide, consistently repeated in both scientific and popular publications since the 10th edition of 'Ethnologue' (ed. by Barbara F. Grimes) was published in 1984 (with 6,519 languages counted), is nothing but a popular myth.

[22] Cf. Gippert 2008: 162–163 and 187–188 on the importance of these dialectal varieties.

[23] Cf. ib.: 175 as to an example.

[24] Given that the four-letter codes include the existing three-letter codes, the number of possible codes must be increased by the 6,900 entries of ISO 639-3.

[25] This is the World Language Documentation Centre, Wales (cf. http://www.thewldc.org/); cf. the website in http://www.geolang.com/, which makes queries about the standard available in http://www.geolang.com/iso639-6/.

as 'Judeo-Georgian' (JGE) are direct children of GGNC (and accordingly, siblings of the 'Georgian' standard language, KAT). As a matter of fact, the arrangement of varieties of Georgian in the dependency tree (cf. Figure 1) is enigmatic, and all linguists interested in providing data for cross-corpus retrieval should try to influence this on-going standardization process before its results have been accepted. This is all the more true as the standard is also meant to encompass sociolectal and historical varieties, which renders the application of one simple tree-like structure with parent-child-relations rather problematical.

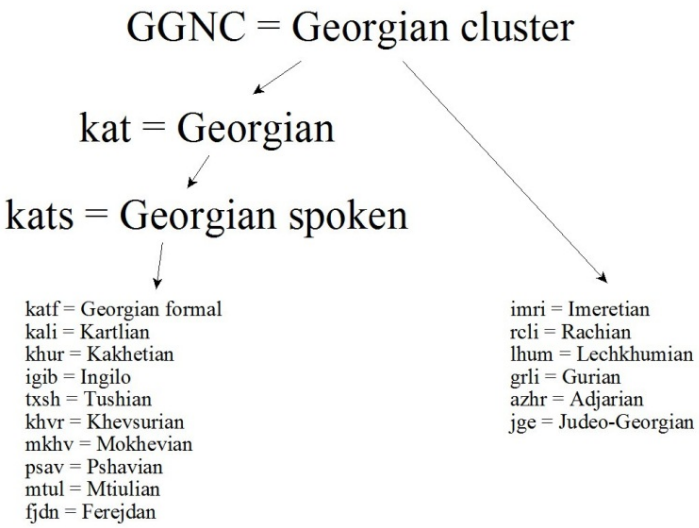


FIGURE 1: Varieties of Georgian in ISO 639–6

At the same time, we should prepare for applying fine-grained language codes in our linguistic analyses, given that they provide a means of clearly distinguishing the different layers we usually have to deal with in recordings of endangered languages. Figure 2 shows a first example of four-letter language codes applied to indicate the languages involved in the setting of Khinalug, an East-Caucasian language spoken in Azerbaijan, which has been the object of a DoBeS project since July, 2011.[26] The task of developing means to use these codes in language-specific corpus retrieval remains still to be solved. One solution might consist in assigning the language as a property of a given annotation, rather than storing the

[26] My thanks are due to Monika Rind who provided the given example from her fieldwork in Khinalug. In the language-related tiers, KJJS stands for 'Spoken Khinalug', AZJS for 'Spoken Azeri', RUSS for 'Spoken Russian', and ARA for 'Arabic'. The tier \lan indicates whether a given word (form) is part of Khinalug speech (i.e., with grammatical properties such as case endings of this language) or of code switched to Azeri (with Azeri grammar), while \src indicates the immediate source of a word (form) in question (usually Azeri, as this is the main contact language of the Khinalug speakers). In addition, \etylan indicates the etymological origin of a word (e.g., Arabic) wherever applicable. Thus, e.g., the language (\lan) of *turistin* is styled as being Azeri (AZJS) because the word bears the Azeri genitive ending, *-in*, while *turizmi* is styled as being Khinalug because it bears the Khinalug genitive ending, *-i*. The source (\scr) is Azeri in both cases, while the etymological origin (\etylan) is Russian (further derivation from French etc. notwithstanding).

information in separate markup tiers.[27] This, however, would not help for annotations on the sentence level as instances of code-switching would not be coverable in this case.

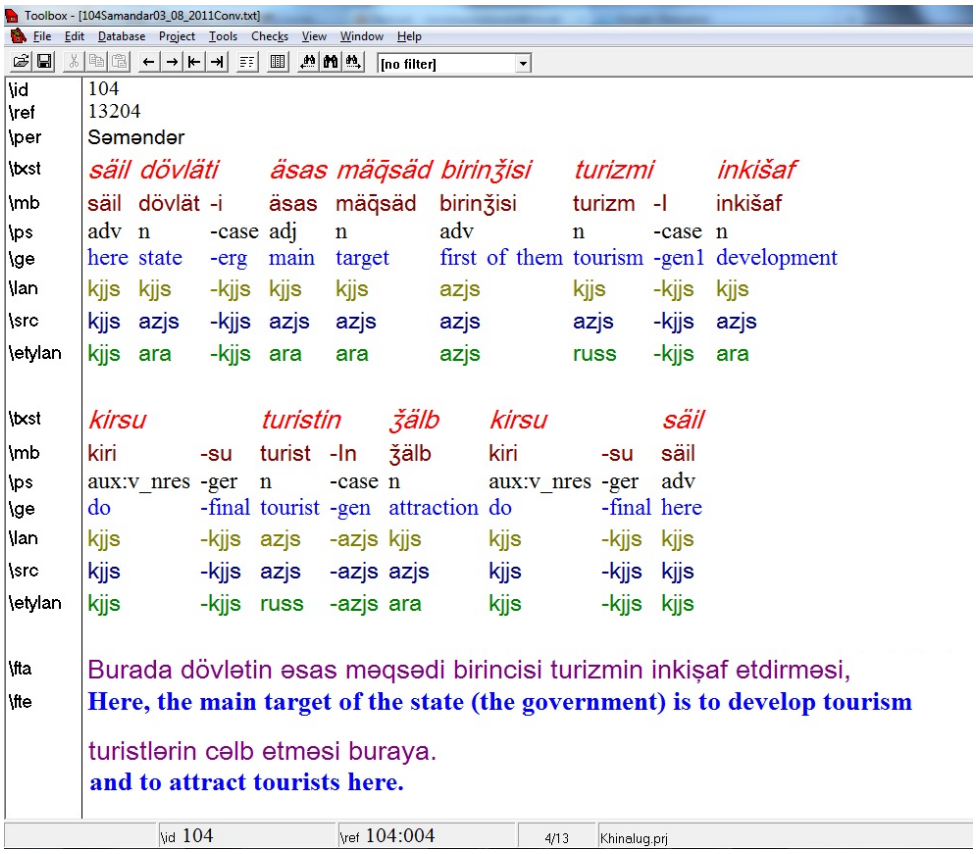


FIGURE 2: Toolbox example sentence with indication of languages

4. CONCLUSION AND OUTLOOK. The multilingual nature of many endangered languages corpora makes them especially interesting for a number of research questions. At the same time, the demarcation of material from different languages in these corpora poses severe problems for the annotation of data and for automatic retrieval mechanisms. The standardization of fine-grained language codes is one important prerequisite for coping with these problems but the ongoing endeavor towards this requires input from specialists to avoid misleading solutions. Scholars working on endangered languages are especially encouraged to bring in their expertise in this respect.

[27] It goes without saying that this would require a major addition to the functionality of the ELAN tool and the XML structure it relies upon.

REFERENCES

ECLING Corpus. ECLING Project (Endangered Caucasian Languages in Georgia). DoBeS Language Resource Archive. http://corpus1.mpi.nl/ds/imdi_browser?openpath=MPI533677%23.

Gippert, Jost. 2000–2012. TITUS Datenbank (Thesaurus Indogermanischer Text- und Sprachmaterialien). http://titus.fkidg1.uni-frankfurt.de/database/titusinx/titusinx.htm.

Gippert, Jost. 2008. Endangered Caucasian languages in Georgia: Linguistic parameters of language endangerment. In K. David Harrison, David S. Rood & Arienne Dwyer (eds.), *Lessons from Documented Endangered Languages*, 159–194. Amsterdam: Benjamins.

Grimes, Barbara F. (ed.). 1984. *Ethnologue: Languages of the World, Tenth edn*. Dallas, Tex: SIL International.

Gullberg, Marianne. this volume. Bilingual multimodality in language documentation data.

International Organization for Standardization (ISO). 2007. ISO 639–3:2007. http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=39534.

Lewis, M. Paul (ed.). 2009. *Ethnologue: Languages of the World, Sixteenth edn*. Dallas, Tex: SIL International. Online version: http://www.ethnologue.com/.

SIL International (Summer Institute of Linguistics). 2007. ISO 639–3. http://www.sil.org/iso639-3.

The World Language Documentation Center. Wales. http://www.thewldc.org/, http://www.geolang.com/, http://www.geolang.com/iso639-6/.

Toolbox. SIL International (Summer Institute of Linguistics). http://www.sil.org/computing/toolbox/.

Jost Gippert
gippert@em.uni-frankfurt.de

4

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. 25–31
http://nflrc.hawaii.edu/ldc/sp03/
http://hdl.handle.net/10125/4513

Unsupervised morphological analysis of small corpora: First experiments with Kilivila

Amit Kirschenbaum[a], Peter Wittenburg[b], and Gerhard Heyer[a]

[a]*University of Leipzig*
[b]*Max Planck Institute for Psycholinguistics, Nijmegen*

Language documentation involves linguistic analysis of the collected material, which is typically done manually. Automatic methods for language processing usually require large corpora. The method presented in this paper uses techniques from bioinformatics and contextual information to morphologically analyze raw text corpora. This paper presents initial results of the method when applied on a small Kilivila corpus.

1. INTRODUCTION. Unsupervised approaches to language processing attempt to discover the structure of language based on machine-learning techniques applied to unannotated text. Present unsupervised methods are typically developed for large corpora and are not suitable for small corpora of a few hundred thousand or even just thousands of words. However, this is the typical size of corpora in language documentation (see Klamer, this volume; Holton, this volume), where the potential of unsupervised learning methods has already been acknowledged:

> "Basic linguistic descriptions of lexicon and grammar made on the basis of transcribed recordings still form an important component of language documentation, however, and with the realization that languages are disappearing at a far faster rate than linguists can document them, it is natural to look for ways of making this process less labor-intensive." (Hammarström & Borin 2011)

In particular, unsupervised methods for morphological analysis aim to learn the internal structure of words from a raw text corpus of a given language, and this typically means segmenting words into morphemes.

Unsupervised morphological analysis can be traced back to an algorithm introduced by Harris (1955, 1967), and later improved by Hafer & Weiss (1974). The algorithm detects morpheme boundaries as a function of the number of distinct letters that follow, or precede, a letter sequence which is part of a word (Letter Successor Variety). If a peak is reached in that number, then it would probably be due to a morpheme boundary.

A different approach incorporates the Minimum Description Length (MDL) principle (Rissanen 1978), which is based on information-theoretic grounds. This principle follows

the idea that regularities in data can be used "to describe it using fewer symbols than the number of symbols needed to describe the data literally" (Grünwald 2007). It therefore seeks to minimize a cost function which is the sum of the description length[1] of the model explaining the data, and of the description length of the data, when encoded with this model. Goldsmith (2001, 2006) used MDL to construct lists of stems, suffixes, and signatures, i.e. structures that indicate which stems may appear with which suffixes. Another MDL based method is presented by Creutz & Lagus (2002). In this method, the model is a lexicon of morphs, which may either be prefixes, suffixes, or stems. The data is encoded by sequences of pointers to the lexicon. Each input word is segmented in various ways. For each way, the cost is evaluated, and the segmentation which achieved the minimum cost is selected. This method is recursive, and each new morph may be subject to further splitting.

Another type of algorithms also uses contextual features as part of the morphological segmentation process. Freitag (2005) utilizes local co-occurrence information to create clusters that correspond roughly to syntactic classes. The method then induces affix transformation rules which express relations between clusters and show possible affixation patterns. Further description of other methods devoted to unsupervised morphology can be found in a detailed survey by Hammarström & Borin (2011).

The purpose of the method presented in this paper is to help alleviate the situation in language documentation as described above by providing a way to automatically segment small corpora on the morphological level in order to facilitate lexical and morphological analysis and description.

The present method is (in principle) language independent and should work for languages with various properties (e.g. with both concatenative and non-concatenative morphology, various word orders, etc.). It employs word-distributional similarity and sequence alignment, a technique borrowed from bioinformatics in which protein or DNA sequences are compared, and similar regions among them are then identified.

For the purpose of this paper, the method was applied to a small corpus of Kilivila (Senft 1983–1997), an Austronesian language spoken by the Trobriand Islanders of Papua New Guinea.[2] The examples in the following section come from this corpus as well.

2. METHOD. The method starts with computing a word co-occurrence model to discover distributional similarities between words. The model is represented in a high-dimensional vector space, where every word in the corpus is associated with a *context co-occurrence vector*. The context vector for a word *w* is defined by words which co-occur with *w* within a sentence context.[3] To reduce noise, the context vector for *w* is represented only by significant co-occurrences of *w*. To extract the significant word co-occurrences we utilize the method described in Quasthoff & Wolff (2002). This method is based on comparing the expected number of joint occurrences of two words in a corpus, under the independence assumption, to their actual number of co-occurrences in that corpus. Examples of some

[1] The length is measured in bits.

[2] We would like to thank Prof. Gunter Senft for providing the corpus and the accompanying linguistic analysis.

[3] The term *co-occurrence*, henceforth, will refer to joint occurrence of two words within sentential context.

co-occurrences (and their morphological analyses) of an input word in Kilivila *bukuninamsisi* [2FUT-think-PL][4] from our test corpus are given in Table 1.

In the next step, similarity relations between words are computed by comparing their context vectors. The underlying rationale is based on the distributional hypothesis by Harris (1968) according to which words with similar distributional properties (i.e., contexts) tend to be semantically similar. We employ the method described in Bordag (2008) to compare context vectors and obtain distributionally similar words. Table 2 shows some of the distributionally similar words for the word *bukuninamsisi*.

mankawa	[DEM-DEM-CP.thing]
ekau	[3PRS-take]
yegulaga	[I-Emph-Emph]
sogu	[friend-my]
isiligaga	[3PRS-important]
makaukweda	[our-veranda]
bibwadi	[3FUT-be.possible]
beya	[here,there,this]
bagisi	[1FUT-see]
tabu	[taboo]
...	...

TABLE 1: Co-occurrences examples for the word *bukuninamsisi* [2FUT-think-PL]

lalilivali	[1PST-tell]
bukusisusi	[2FUT-be-PL]
lumkola	[feeling]
bukukanukwenusi	[2FUT-lie.down-PL]
biboda	[3FUT-be.good]
ibubulisi	[3PRS-work-PL]
nanomi	[mind-your]
biyapu	[3FUT-be.good.and.bad]
bukulilolasi	[2FUT-walk-PL]
evagisi	[3PRS-make-PL]
...	...

TABLE 2: Examples for distributionally similar words of the word *bukuninamsisi* [2FUT-think-PL]

The set of distributionally similar words of w is then filtered based on edit distance (Needleman & Wunsch 1970) from w. The resulting target set consists of words which are both distributionally and orthographically similar to w. Orthographically close words

[4] Abbreviations: 1, 2, 3 – 1st, 2nd, 3rd person; CP – Classificatory Particle; DEM – demonstrative; Emph – emphatic; FUT – future; PL – plural; PRS – present; PST – past

may be derivations or inflections of one stem, or words from one word class sharing a set of morpho-syntactic features (e.g., verbs in the same tense).

The words from the target set are aligned using a *multiple sequence alignment* (MSA) method. Multiple sequence alignment aims to discover functional, structural, or evolutionary relationships among a set of biosequences by searching for character patterns in these biosequences. The alignment process inserts "gap" into the sequences allowing equivalent characters from different sequences to be positioned in the same column.

We employed BioJava (Holland et al. 2008), a bioinformatics toolkit, to perform the alignment. The strategy used was, first, to align *w* and its orthographically most similar word from the target set and then to gradually align less similar words in a cumulative fashion. The result is a set of aligned words from which one or more segmentation patterns can be extracted based on character overlap in the aligned sequences. Table 3 demonstrates a part of the alignment of the word *bukuninamsisi* and its target set. '–' signs mark the "gaps" inserted in words during the alignment process.

b	*u*	*k*	*u*	–	–	–	*n*	*i*	*n*	*a*	–	*m*	*s*	*i*	*s*	*i*
b	*u*	*k*	*u*	–	–	–	*n*	–	–	*a*	–	*m*	*s*	*i*	*s*	*i*
b	*u*	*k*	*u*	–	–	–	*n*	*o*	*k*	*a*	*p*	*i*	*s*	*i*	*s*	*i*
–	–	*k*	*u*	–	–	–	*n*	*a*	*n*	*a*	–	*m*	*s*	*a*	*s*	*i*
b	*u*	*k*	*u*	–	–	–	*l*	*i*	*g*	*e*	–	*m*	*w*	*e*	*s*	*i*
b	*u*	*k*	*u*	–	–	–	*n*	*u*	*k*	*w*	*a*	*l*	–	*i*	*s*	*i*
b	*u*	*k*	*u*	–	–	–	*s*	*i*	–	–	–	–	*s*	*u*	*s*	*i*
b	*u*	*k*	*u*	–	–	–	*m*	*a*	–	–	–	–	*s*	*i*	*s*	*i*
b	*i*	*t*	*a*	–	–	–	*n*	*i*	*n*	*a*	*m*	–	–	–	*s*	*i*
b	*u*	*k*	*u*	–	–	–	*t*	*e*	*m*	*a*	*l*	–	–	*i*	*s*	*i*
...																

TABLE 3: Part of the alignment for the word *bukuninamsisi* and its target set

The patterns are scored according to their length in characters and according to the number of words in the target set which they match. We record the pattern with the highest score as a possible segmentation of the words in the target set. The pattern extracted for the set presented in Table 3 is *buku-si*, which encodes [2FUT-VERB-PL], and words that match this pattern would be segmented according to it. There is no restriction for the form of that extracted pattern, and it is indeed possible that patterns would also include e.g., infixes.

However, the word *w* can also be a member of target sets of different words {*w'*} for which different patterns might be recorded as possible segmentations. Consequently, for a given *w*, several segmentation patterns may be recorded, and each of them may appear more than once. Hence, we get a weighted set of possible patterns for morphological segmentation of *w*, and we select (in the current implementation) the pattern with the highest frequency.

3. EVALUATION. We applied the method on a small corpus of narrations in Kilivila (Senft 1983–1997). The corpus consists of ca. 13,000 words. The corpus contains

morphological annotations constructed by an expert and thus serves as our "gold standard" reference.

The evaluation method compares the segmentation decisions derived from our method for each word to the actual segmentations in the reference. Segmentation points that were marked by the method and correspond to actual morphological boundaries are called true positives (*tp*). Segmentation points which were identified by the method but do not correspond to morphological boundaries are called false positives (*fp*). Morphological boundaries which were not detected by the method are called false negatives (*fn*). Precision and recall are then calculated for each word, based on the amount of segmentation points in each of the above categories.

$$Precision_w = \frac{\#tp}{\#tp + \#fp} \qquad Recall_w = \frac{\#tp}{\#tp + \#fn}$$

Thus, the precision measures the portion of segmentation points which are correct, out of the reported segmentation points, and recall measures the portion of those correctly found segmentation points out of the actual segmentation points. Both measures reach a maximum of 1 when there are no mistakes in segmenting the word by the method. The value decreases when there are prediction errors, i.e. redundant segmentation points in the case of precision, or missed ones in the case of recall. When no segmentation point is identified correctly, the value of these measures is 0. The average precision (P) and recall (R) are then calculated based on the results for each word.

Table 4 summarizes the results. The first line is our baseline, which randomly assigns segmentation points as morpheme boundaries. The second line presents the results for applying the method on the whole corpus. The third line presents the evaluation results for using the method after setting a reliability threshold on the derived patterns.

METHOD	P	R
Random	0.22	0.44
Unsupervised	0.381	0.569
Unsupervised+thresh	0.682	0.133

TABLE 4: Evaluation results

4. FUTURE WORK. The method presented here still requires much exploration. We plan to experiment further with the ways of extracting patterns from target sets and determining the final segmentation, and we also plan to experiment with different corpus sizes. The present version of the method assumes existing sentence boundaries; however, we plan to experiment with sentence independent context windows as well.

A future version of this method should also be able to derive the morphology of the analyzed language, in the sense of supplying the user with generalizations regarding, for example, inflectional and derivational paradigms. This method is intended to be a component in

a larger framework of automatic annotation which would consist of both unsupervised and supervised algorithms.

The unsupervised module of the system would attempt to compensate the data sparseness problem by using linguistic information of various sources (morphological, parts of speech, semantic levels) and by taking advantage of the interaction between these levels. As a result, the module would produce suggestions for linguistic analyses on the three levels, which the annotator can manually correct. The supervised module would then train a model on the corrected data and would produce the final annotation of the corpus.

This system of interactive annotation is planned to be integrated into existing and widely used environments such as ELAN[5] (Wittenburg et al. 2006) or LEXUS[6] (Kemps-Snijders et al. 2006) in order to make the annotation process more efficient.

REFERENCES

Bordag, Stefan. 2008. A comparison of co-occurrence and similarity measures as simulations of context. In *Proceedings of the 9th International Conference Computational Linguistics and Intelligent Text Processing (CICLing)*, 52–63. Springer.

Creutz, Mathias & Krista Lagus. 2002. Unsupervised discovery of morphemes. In *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning*, 21–30. Association for Computational Linguistics.

Freitag, Dayne. 2005. Morphology induction from term clusters. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, 128–135. Association for Computational Linguistics.

Goldsmith, John. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics* 27(2). 153–198.

Goldsmith, John. 2006. An algorithm for the unsupervised learning of morphology. *Natural Language Engineering* 12(4). 353–371.

Grünwald, Peter D. 2007. *The Minimum Description Length Principle*. Cambridge, Mass: MIT Press.

Hafer, Margaret A & Stephen F Weiss. 1974. Word segmentation by letter successor varieties. *Information Storage and Retrieval* 10. 371–385.

Hammarström, Harald & Lars Borin. 2011. Unsupervised learning of morphology. *Computational Linguistics* 37(2). 309–350.

Harris, Zellig S. 1955. From phoneme to morpheme. *Language* 31(2). 190–222.

Harris, Zellig S. 1967. Morpheme boundaries within words: Report on a computer test. *Transformations and Discourse Analysis Papers* 73. Philadelphia: University of Pennsylvania.

Harris, Zellig S. 1968. *Mathematical Structures of Language*. New York: Wiley.

Holland, Richard C. G, Thomas A Down, Matthew R Pocock, Andreas Prlic, David Huen, Keith James, Sylvain Foisy, Andreas Dräger, Andy Yates, Michael Heuer & Mark J Schreiber. 2008. BioJava: An open-source frame-work for bioinformatics. *Bioinformatics* 24(18). 2096–2097.

Holton, Gary. this volume. Language archives: They're not just for linguists any more.

[5] http://www.lat-mpi.eu/tools/elan/

[6] http://www.lat-mpi.eu/tools/lexus/

Kemps-Snijders, Marc, Mark-Jan Nederhof & Peter Wittenburg. 2006. LEXUS, a web-based tool for manipulating lexical resources. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, 1862–1865.

Klamer, Marian. this volume. Tours of the past through the present of eastern Indonesia.

Needleman, Saul B & Christian D Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* 48(3). 443–453.

Quasthoff, Uwe & Christian Wolff. 2002. The Poisson collocation measure and its applications. In *Proceedings of the Second International Workshop on Computational Approaches to Collocations*, Vienna.

Rissanen, Jorma. 1978. Modeling by shortest data description. *Automatica* 14(5). 465–471.

Senft, Gunter. 1983–1997. Tales of the Trobriand Islanders. Transcribed, morpheme-interlinearized and glossed corpus of Kilivila (fairy-)tales. Nijmegen: Mimeo.

Wittenburg, Peter, Hennie Brugman, Albert Russel, Alex Klassmann & Han Sloetjes. 2006. ELAN: A professional framework for multimodality research. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, 1556–1559.

Amit Kirschenbaum
amit@informatik.uni-leipzig.de

Peter Wittenburg
Peter.Wittenburg@mpi.nl

Gerhard Heyer
heyer@informatik.uni-leipzig.de

5

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. 32–38
http://nflrc.hawaii.edu/ldc/sp03/
http://hdl.handle.net/10125/4514

A corpus linguistics perspective on language documentation, data, and the challenge of small corpora

Anke Lüdeling
Humboldt University of Berlin

This paper deals with issues of corpus design that might prove problematic for the study of under-resourced languages, e.g. in language documentation. It argues that it is not yet well understood which linguistic and extra-linguistic (predictor) variables cause linguistic variation (i.e. the response variable), which means that the scope of a linguistic finding cannot always be assessed. In order to deal with this problem, it is argued that we need a flexible corpus architecture with the option of adding meta-data to corpora/sub-corpora at any point in time.

1. INTRODUCTION. On an abstract technical level, there are no categorical differences between a large corpus for a well-researched language with many resources and a standardized orthography and a corpus of an endangered language or small variety without codified standards: In both cases one needs to represent a source text and annotations to it. Which levels of annotation are needed depends on the research question or purpose of the corpus; the corpus architecture needs to be flexible enough to represent annotation layers in different formats. A large number of corpora have been built in recent language documentation projects; and because many of these corpora are multi-modal and diverse, language documentation has had a significant impact on the development of corpus tools, formats, and architectures (see e.g. Auer et al. 2010, Wittenburg et al. 2010). Many of these tools, formats, and architectures are also used for other corpora – typically smaller corpora, corpora of spoken language, or corpora of non-standardized varieties, such as learner varieties, dialects, or historical language stages (see e.g. Ostler 2008 for an overview of smaller corpora, or Wittenburg 2008 for an overview of multimodal corpora) where linguistic analysis and annotation is crucial, and the methods developed for well-resourced languages or language-independent methods sometimes have fed back into language documentation projects (see e.g. Kirschenbaum et al., this volume, for a language-independent method of automatic morphological analysis). The same multi-layer standoff architecture and search tool (Annis, cf. Zeldes et al. 2009, Zipser & Romary 2010), for example, is used for a historical corpus of German (Petrova et al. 2009), deeply annotated corpora of African languages (Chiarcos et al. 2011), a modern German learner corpus (see below), or a large treebank of modern German newspaper texts.

On a more detailed level, however, there are issues of linguistic variation that can be understood only by working with many well-documented varieties of a language, and it might be that language documentation projects can learn from corpus work in the well-researched and richly-resourced languages.

2. CORPUS DESIGN. Research on linguistic variation is probably among the most interesting fields for which corpus data can be used, and I will focus on variation research here. It has been known for a long time that language production depends on many parameters – 'the same' (the variable) can be expressed in many different ways (the variants).[1] Research on variation focuses on the language-internal and language-external co-variables that influence the choice of one variant over another. In an ideal world, a corpus could be designed according to well-understood parameters that only depend on the research question at hand.[2] In addition to issues of time, money, and feasibility, this means that we would have to understand all the parameters that influence language production and could find portions of language for each necessary combination of parameters. This would mean that we select a sample (or several samples) of a language according to some design criteria and add these as meta-data to the corpus sources. I want to argue here that we are far from understanding which parameters influence variation and that we continue to discover new (not previously discussed) parameters that also influence linguistic variation. This means that we cannot work with a static set of meta-data categories.

Dedicated corpora as well as large reference corpora (such as 'national' corpora, the BNC, the ANC, DeReKo, etc.) that are not built for one specific research question include material for some of the possible combinations (such as n % of spoken informal dialogue or m % of written formal language, etc., see e.g. Hunston 2008 on corpus design). The rationale behind this is that for a given combination of parameters, each piece of material would be as good as any other piece of material. If we wanted to do research on, say, the use of adjectives in informal conversation we could take any informal conversation and generalize from that – there are countless publications on many aspects of the grammar of 'newspaper language', a given dialect, a given register, etc. that seem to (implicitly) follow this reasoning.[3]

However, linguistic research from many domains (sociolinguistics, dialectology, etc.) has discovered more and more parameters that seem to influence the choice of variants. These include among many others

- properties of the speaker: dialect, socio-economic factors, age, gender, education, etc., see e.g. Labov (1972, 2001)
- mode of communication: spoken, written, computer-mediated, etc., see e.g. the papers in Androutsopoulos & Beißwenger (2008) for CMC research or Siekmeyer (2011) for

[1] This is the view formulated in variationist sociolinguistics or in variationist models for language change (see e.g. Labov 1972, 1994, 2001 etc. or Rissanen 2008).

[2] In addition to corpora with a specific and documented design, we find opportunistic corpora and Web-based corpora where the composition is not known (such as the WaCky corpora, Baroni et al. 2009). While these might be used as 'example banks' or for NLP purposes, the lack of meta-information is problematic for more fine-grained variation research (see Lüdeling et al. 2007 for a discussion of the problems).

[3] This issue has sometimes been discussed under the header of 'representativeness'; see e.g. Biber (1993).

a corpus-based study of spoken versus written texts; mode is often discussed together with other parameters such as distance or level of formality (Koch & Oesterreicher 1994, Maas 2006)

- purpose of the communication: text type, functional variation, register, see e.g. Biber (1995, 2006).

There are several problems for corpus design here: First, for many of these co-varying variables it is unclear what the variants would be: Maas (2006), for example, distinguishes between three registers while Biber (1993, 1995, 2006) shows that many more distinctions can be made reliably. Second, without more research it is unclear which other factors play a role in language production. I want to illustrate this with an example from the German learner corpus Falko (Lüdeling et al. 2008, Reznicek et al. 2010). The corpus consists of written argumentative essays produced by advanced learners of German as a foreign language. It is well-known that the L1 has an influence on a learner's text production in the L2 (transfer, interference, see Ellis 2008). Golcher & Reznicek (2011, in preparation) show that a machine learning method that takes into account only substring frequencies is able to classify learner texts by L1 with high accuracy. In addition to using substring frequencies of the original texts, they also compare the part-of-speech annotation, again modeled as a string, and show that simply by looking at the frequencies of part-of-speech (POS) sequences, the L1 of a learner can be predicted with high accuracy. While this in itself might not be too surprising, Golcher & Reznicek use the same method to classify learner texts by their topic. All learners were asked to choose one of four topics; all other production parameters (time, setting, access to assistance, etc.) were kept equal for all texts. Figure 1 (from Golcher & Reznicek 2011: 31) shows that classification by topic is possible with an accuracy of around 80%, even by only taking into account POS sequences (thus not looking at lexical information). Automatic classification by topic is more reliable than classification by L1.

Findings such as these show that people use different grammatical means (represented by sequences of POS tags) when writing about different topics, even within the same text type (argumentative essay) written under the same circumstances.[4] This means that we might have to add 'topic' to the list of co-varying variables.[5] And if we look at further parameters, we might find further variables. The findings by Mosel (this volume), who shows that certain narrative styles are associated with specific stories in Teop, and Gullberg (this volume), who speaks about the different parameters that influence L2 acquisition, provide further cases in point.

3. CONSEQUENCES. Detailed studies that analyze variable after variable and their interaction with text production can only be done for languages with rich resources. But there

[4] The method here is a stylometric method developed in Golcher (2007). Golcher & Reznicek take care to eliminate any confounding L1 evidence that might have been introduced by the fact that learners with a given L1 tend to choose certain topics over others.

[5] Further studies are needed to find the reasons for these differences (Golcher & Reznicek, in preparation, provide some suggestions). It might, for instance, turn out to be the case that the differences are only indirectly due to the topic but point to register differences instead because some of the topics prompt the writers to argue more emotionally than others.

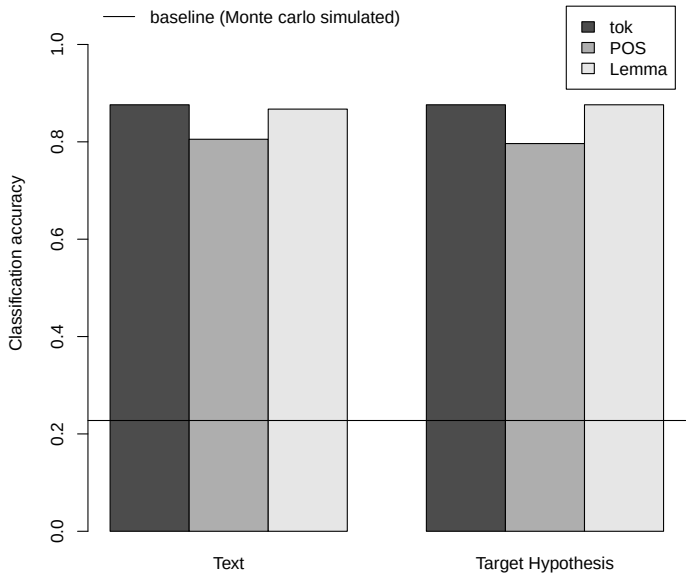


FIGURE 1: Classification of Falko texts by topic, using substring frequencies (figure taken from Golcher & Reznicek 2011: 31). The target hypothesis is a slightly modified version of the learner text where grammatical errors are corrected; results for classification are similar.

is no reason to assume that the findings from such research on richly-resourced languages will not in principle carry over to less-researched languages. Does this mean for language documentation that we only make grammatical statements about individual texts and do not generalize at all in such a context? This is clearly not the case because, as has been shown time and again. Speakers have a good awareness of appropriate variants (this is, of course, the basis for the fact that automatic classification works at all). But it is necessary that we understand which grammatical elements (on all grammatical levels, from sounds to text structure) are specific to external co-varying variables and which are not, and how variables interact with each other.[6]

On a technical level, we need powerful flexible corpus architectures where annotation layers and meta-data can be added at any point in time. Today very often a corpus can only be accessed via an interface and only the original corpus provider has the power to change it. If new linguistic analyses on this material are carried out and new co-varying variables

[6] There are many statistical models that can be used to reduce dimensions (such as principal component analysis) or can be used to look at the effect of different variables (also as random effects models). Interaction is crucial, and its effects have often been underestimated, cf. Gries, forthcoming.

are found after the corpus was first constructed, we need the option to add this information to the corpus. The technical means are there: Several multi-layer standoff architectures for corpora are being developed, and I am hopeful that the remaining technical issues (with respect to conversion frameworks and search tools, etc.) will be solved within the next few years. What seems to be just as necessary (but perhaps more difficult) is an awareness of the fact that we do not yet understand variation and cannot know all meta-data categories or annotation layers when we construct a corpus. This awareness would result in better corpus exchange and update functionalities and a fundamental work-flow change.

REFERENCES

ANC. American National Corpus. http://americannationalcorpus.org/ (11 December, 2011).

Androutsopoulos, Jannis & Michael Beißwenger. 2008. Introduction: Data and methods in computer-mediated discourse analysis. *Language@Internet* 5(2). http://www.languageatinternet.org/articles/2008/1609 (12 December, 2011).

Auer, Eric, Albert Russel, Han Sloetjes, Peter Wittenburg, Oliver Schreer, Stefano Masnieri, Daniel Schneider & Sebastian Tschöpe. 2010. ELAN as flexible annotation framework for sound and image processing detectors. In *European Language Resources Association LREC 2010: Proceedings of the 7th International Language Resources and Evaluation*, 890–893. Paris: ELRA.

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi & Eros Zanchetta. 2009. The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation* 43(3). 209–226.

Biber, Douglas. 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8(4). 243–257.

Biber, Douglas. 1995. *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge: Cambridge University Press.

Biber, Douglas. 2006. *University Language: A corpus-based study of spoken and written register*. Amsterdam: John Benjamins.

BNC. British National Corpus. http://www.natcorp.ox.ac.uk/ (11 December, 2011).

Chiarcos, Christian, Ines Fiedler, Mira Grubic, Katharina Hartmann, Julia Ritz, Anne Schwarz, Amir Zeldes & Malte Zimmermann. 2011. Information structure in African languages: Corpora and tools. *Language Resources and Evaluation; Special Issue on African Language Technology* 45(3). 361–374.

DeReKo. Deutsches Referenzkorpus. http://www.ids-mannheim.de/kl/projekte/korpora/ (11 December, 2011).

Ellis, Rod. 2008. Language transfer. In Rod Ellis (ed.), *The Study of Second Language Acquisition*, 349–403. Oxford: Oxford University Press.

Falko. Ein fehlerannotiertes Lernerkorpus des Deutschen als Fremdsprache. http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko (11 December, 2011).

Golcher, Felix. 2007. A new text statistical measure and its application to stylometry. In Matthew Davies, Paul Rayson, Susan Hunston & Pernilla Danielsson (eds.), *Proceedings of the Corpus Linguistics Conference (CL2007)*, University of Birmingham. http://amor.cms.hu-berlin.de/~golcherf/cl07.pdf (12 December, 2011).

Golcher, Felix & Marc Reznicek. 2011. Stylometry and the interplay of topic and L1 in the different annotation layers in the FALKO corpus. In Amir Zeldes & Anke Lüdeling (eds.), *QITL-4 – Proceedings of Quantitative Investigations in Theoretical Linguistics 4* QITL-4, 29–34. Berlin: Humboldt-Universität zu Berlin. http://edoc.hu-berlin.de/conferences/qitl-4/golcher-felix-29/PDF/golcher.pdf (12 December, 2011).

Golcher, Felix & Marc Reznicek. (in preparation). Stylometry and the interplay of L1 and text topic in second language texts.

Gries, Stefan Th. forthcoming. Statistische Modellierung. Zeitschrift für Germanistische Linguistik 2 (2012).

Gullberg, Marianne. this volume. Bilingual multimodality in language documentation data.

Hunston, Susan. 2008. Collection strategies and design decisions. In Anke Lüdeling & Merja Kytö (eds.), *Corpus Linguistics: An International Handbook*, vol. 1, 154–168. Berlin: Mouton de Gruyter.

Kirschenbaum, Amit, Peter Wittenburg & Gerhard Heyer. this volume. Unsupervised morphological analysis of small corpora: First experiments with Kilivila.

Koch, Peter & Wulf Oesterreicher. 1994. Schriftlichkeit und Sprache. In Hartmut Günther & Otto Ludwig (eds.), *Schrift und Schriftlichkeit / Writing and Its Use: Ein interdisziplinäres Handbuch internationaler Forschung / An Interdisciplinary Handbook of International Research*, 587–604. Berlin: Walter de Gruyter.

Labov, William. 1972. *Language in the Inner City: Studies in the Black English Vernacular*. Philadelphia: University of Pennsylvania Press.

Labov, William. 1994. *Principles of Linguistic Change. Volume I: Internal Factors*. Oxford: Blackwell.

Labov, William. 2001. *Principles of Linguistic change. Volume II: Social Factors*. Oxford: Blackwell.

Lüdeling, Anke, Seanna Doolittle, Hagen Hirschmann, Karin Schmidt & Maik Walter. 2008. Das Lernerkorpus Falko. *Deutsch als Fremdsprache* 2(2008). 67–73.

Lüdeling, Anke, Stefan Evert & Marco Baroni. 2007. Using web data for linguistic purposes. In Marianne Hundt, Nadja Nesselhauf & Carolin Biewer (eds.), *Corpus Linguistics and the Web*, 7–24. Amsterdam: Rodopi.

Maas, Utz. 2006. Der Übergang von Oralität zu Skribalität in soziolinguistischer Perspektive / The Change from Oral to Written Communication from a Sociolinguistic Perspective. In Ulrich Ammon, Norbert Dittmar, Klaus J. Mattheier & Peter Trudgill (eds.), *Sociolinguistics. An International Handbook of the Science of Language and Society / Soziolinguistik. Ein internationales Handbuch zur Wissenschaft von Sprache und Gesellschaf*, vol. 3, 2147–2170. Berlin: Walter de Gruyter.

Mosel, Ulrike. this volume. Creating educational materials in language documentation projects – creating innovative resources for linguistic research.

Ostler, Nicholas. 2008. Corpora of less studied languages. In Anke Lüdeling & Merja Kytö (eds.), *Corpus Linguistics: An International Handbook*, vol. 1, 457–483. Berlin: Walter de Gruyter.

Petrova, Svetlana, Michael Solf, Julia Ritz, Christian Chiarcos & Amir Zeldes. 2009. Building and using a richly annotated interlinear diachronic corpus: The case of Old High German Tatian. *Traitement Automatique des Langues* 50(2). 47–71.

Reznicek, Marc, Maik Walter, Karin Schmid, Anke Lüdeling, Hagen Hirschmann & Cedric Krummes. 2010. Das Falko-Handbuch: Korpusaufbau und Annotationen. Version 1.0. http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko/pdf/Falko-Handbuch_Korpusaufbau%20und%20Annotationen.pdf (12 December, 2011).

Rissanen, Matti. 2008. Corpus linguistics and historical linguistics. In Anke Lüdeling & Merja Kytö (eds.), *Corpus Linguistics: An International Handbook*, vol. 1, 53–68. Berlin: Walter de Gruyter.

Siekmeyer, Anne. 2011. *Strukturelle Kategorien des sprachlichen Ausbaus bei Jugendlichen mit Deutsch als Erstsprache und Deutsch als Zweitsprache. Eine korpuslinguistische Untersuchung des Gebrauchs komplexer Nominalphrasen als Merkmale literater Strukturen in verschiedenen Registern*: Universität Saarbrücken dissertation.

Wittenburg, Peter. 2008. Preprocessing multimodal corpora. In Anke Lüdeling & Merja Kytö (eds.), *Corpus Linguistics: An International Handbook*, 664–685. Berlin, New York: Mouton de Gruyter.

Wittenburg, Peter, Paul Trilsbeek & Przemek Lenkiewicz. 2010. Large multimedia archive for world languages. In *SSCS'10 - Proceedings of the 2010 ACM Workshop on Searching Spontaneous Conversational Speech, Co-located with ACM Multimedia 2010*, 53–56.

Zeldes, Amir, Julia Ritz, Anke Lüdeling & Christian Chiarcos. 2009. ANNIS: A search tool for multi-layer annotated corpora. In *Proceedings of Corpus Linguistics 2009, Liverpool*, http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/mitarbeiter-innen/amir/pdf/CL2009_ANNIS_pre.pdf (07 June, 2012).

Zipser, Florian & Laurent Romary. 2010. A model-oriented approach to the mapping of annotation formats using standards. In *Proceedings of the Workshop on Language Resource and Language Technology Standards, LREC 2010*, Malta. http://hal.archives-ouvertes.fr/inria-00527799/en/ (12 December, 2011).

Anke Lüdeling
anke.luedeling@rz.hu-berlin.de

6

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. 39–45
http://nflrc.hawaii.edu/ldc/sp03/
http://hdl.handle.net/10125/4515

Supporting linguistic research using generic automatic audio/video analysis

Oliver Schreer[a] and Daniel Schneider[b]

[a]*Fraunhofer Heinrich-Hertz-Institute, Berlin*
[b]*Fraunhofer Institute for Intelligent Analysis and Information Systems, Sankt Augustin*

Automatic analysis can speed up the annotation process and free up human resources, which can then be spent on theorizing instead of tedious annotation tasks. We will describe selected automatic tools that support the most time-consuming steps in annotation, such as speech and speaker segmentation, time alignment of existing transcripts, automatic scene analysis with respect to camera motion, face/person detection, and the tracking of head and hands as well as the resulting gesture analysis.

1. INTRODUCTION. Language documentation activities have produced large audio and video corpora, which can be used to investigate various topics such as the relation between spoken language and gestures or special characteristics of endangered languages. Meaningful annotations of these corpora are required as the basis for their analyses, and so the annotations ultimately contribute to the development of new theories. One of the aims of the AVATecH project (Auer et al. 2010, Tschöpel et al. 2011) is to implement algorithms that allow for automatic and semi-automatic creation of pre-annotations for such corpora, hence reducing the time needed to perform the manual annotation task. Automatic analysis of typical language documentation data (such as data from the DoBeS archive) is a difficult task due to two factors. First of all, the size of the media corpora is very significant (currently about 70 TB in the DoBeS Archive). Secondly, the recordings are highly diverse in terms of language, conditions, and situations. Effective methods for automated processing of such content are not widely available or do not exist at all.

Automatic audio/video annotation algorithms will be important for two reasons. Firstly, they lead to a dramatic decrease in the time necessary to perform simple but time-consuming pre-annotation tasks. Secondly, automation of some parts of the process can significantly increase the uniformity of the annotations created worldwide by different researchers. This would contribute to the consistency and comparability of the available language data. In this paper, we present a short overview of the tools for automatic audio/video annotation that are currently available; these have been developed within the AVATecH project. We will also present some initial results indicating the potential benefits for researchers.

2. THE AVATECH CONCEPT. The system concept of AVATecH is detailed in Figure 1. The Fraunhofer Institutes are technology providers delivering recognizers in the form of executables. These recognizers are integrated into existing annotation tools using a common recognizer interface that is based on a derivate of the CMDI (Component Metadata Infrastructure) specification, developed within the CLARIN research infrastructure project (Váradi et al. 2008, Broeder et al. 2010). The annotation tools are developed and maintained by the Max-Planck-Institute for Psycholinguistics (MPI-P). The interactive ELAN tool is an open source annotation tool with a graphical interface for annotating audiovisual content for linguistic research. ELAN is now used by many different types of researchers worldwide. The main areas of application include language documentation, sign language research, and gesture research. An additional tool, ABAX, has been created in the AVATecH project. In contrast to ELAN, it is used to perform a series of annotation tasks on multiple files. ABAX provides a CMDI-interface as well. Researchers can use either ELAN or ABAX to create enriched annotations using the recognizers developed by the Fraunhofer Institutes. Researchers provide media files to be annotated and specify the required parameter settings for the recognizer that will be applied. In addition, some recognizers accept existing annotations or additional feedback information, which can be used to optimize the performance of the recognizer. The main contribution of AVATecH to ELAN is that for the first time, automatic analysis tools for audio/video analysis are available and integrated into the system. In this way, the ELAN tool is not only a graphical user interface but also becomes a powerful automatic annotation engine.

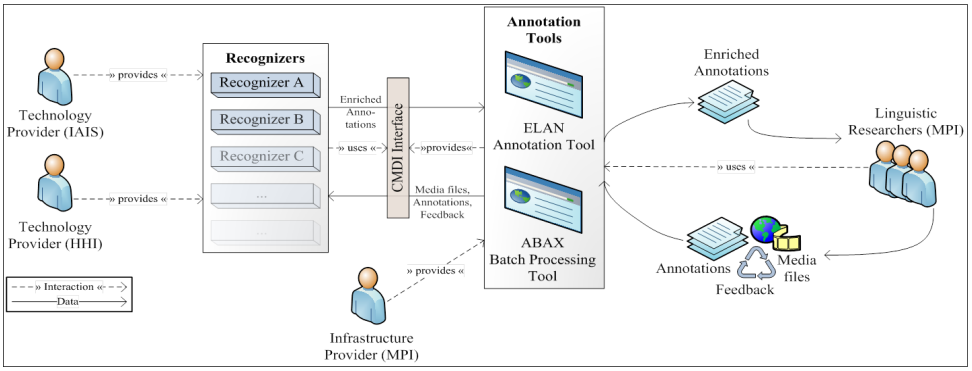


FIGURE 1: The AVATecH concept

3. AUDIO AND VIDEO ANALYSIS ALGORITHMS. All algorithms have been created with the aim of performing well on recordings with a wide variety of acoustic and light conditions as well as quite different scenarios (single vs. multiple persons). The baseline versions have also been designed to work without user interaction (except for the initial configuration by means of a few numerical parameters) to allow batch processing on multiple videos. The implementation was performed using a highly modular structure so that future automatic annotators can be easily integrated into the current framework, using as input the results provided by the previous detectors. In the following paragraphs, the recognizers which are currently available are briefly described in order to demonstrate the large

variety of analysis tools that have been developed up to now. First, we describe the audio-related detectors (from acoustic segmentation to speaker clustering), and then we introduce the video-based detectors (from shot detection to head and hands tracking).

3.1. AUDIO SEGMENTATION. For linguistic annotation, segmentation into units corresponding to utterances is of high importance but is difficult to achieve automatically without errors. This recognizer provides a fine-grained segmentation of the audio stream into homogeneous segments, e.g. between speakers or according to other significant acoustic changes such as pauses. The user can control the granularity of segmentation by tuning a corresponding feedback parameter. An important issue is the choice of the target segment granularity. Some researchers are interested in a segmentation according to who is speaking, e.g. during an interview, while others are interested in a very fine-grained utterance segmentation. The first group will perceive results from the second as being over-segmented, while the second group will consider a speaker segmentation to be under-segmented. Hence we decided to provide two baseline versions for segmentation, one focusing on speaker changes, and one optimized for fine-grained segmentation. For the utterance segmentation, we expect that more corrections will be required by the user since utterances are often not only separated acoustically but are also based on the content of the spoken words (which is not exploited by the current language-independent algorithm).

3.2. SPEECH DETECTION. This recognizer is able to label audio segments containing human speech, regardless of the language of the recording. To enhance the performance of this detector, the user can manually provide a small amount of speech and non-speech samples in order to adapt the model to the given data, which leads to a more robust detection.

3.3. SPEAKER CLUSTERING. A language-independent speaker clustering recognizer is able to find segments spoken by the same person within a given recording. The results can be used to remove utterances by the interviewer in a recording or to extract material from specific speakers from a recorded discussion. For optimization of the detection performance, we use manual user input such as the number of speakers or speaker audio samples. If the user can provide samples of a speaker, we can combine unsupervised speaker clustering with supervised speaker identification, i.e. the algorithm labels all segments where a specific known person speaks with the corresponding name.

3.4. VOWEL AND PITCH CONTOUR DETECTION. The pitch contour detector can allow researchers to graphically specify pitch contours and search for similar patterns. The detector can tag segments in audio recordings and annotate with pitch and intensity properties such as minimum, maximum, initial or final f0 frequency, or volume. The detector invokes PRAAT to calculate f0 and volume curves of the input over time. Those are then used to find characteristic segments and to annotate them.

3.5. SHOT/CUT DETECTOR AND KEY FRAMES EXTRACTOR. A shot is defined as a set of video frames that have been continuously recorded with a single camera operation and represent therefore the basic unit of a video. This recognizer is able to detect such shots and label them. All video analysis algorithms described further provide results for a given

shot and therefore rely on the results of the shot/cut detection. Sub-shots are defined as a sequence of consecutive frames showing one event, or a part thereof, taken by a single camera act in one setting with only a small change in visual content.

3.6. GLOBAL MOTION DETECTION. Accurate motion analysis allows different types of video content to be distinguished. It can be used to segment a video in order to select only those parts which are relevant for the researchers. For example, the presence of zooms and motion inside of a scene are usually the most interesting, while shots containing only panning and a low amount of internal motion are usually of little interest and can usually be discarded without further analysis. The algorithm developed performs a frame-based analysis and detects when global motion (pan, tilt, zoom in, or zoom out) occurs inside a shot. For each frame in the video, a motion vector map is computed using the Hybrid Recursive Matching (HRM) algorithm (Atzpadin et al. 2004).

3.7. SKIN COLOR ESTIMATION. A prerequisite for head and hands tracking, e.g. for gesture studies, is skin color estimation. There is no unique set of skin color parameters which can achieve good results for all recordings, and therefore typical approaches that make use of a training set to collect the parameters for skin detection on the entire dataset cannot be applied. The estimation scheme uses both the temporal information provided by the change between one frame and the next and the spatial information provided by the fact that skin color pixels tend to cluster in well defined regions. This skin color estimator does not need a training dataset but rather estimates the color ranges identifying skin color for each frame in each video.

3.8. HEAD AND HANDS TRACKING. The algorithm works first by segmenting the image in skin vs. non-skin pixels, using the information provided by the skin color estimator. The subsequent step in the detection process involves the search of seed points where the head and hands regions most likely occur. A region-growing algorithm is then applied to the seed points in order to cluster together all the skin pixels in the neighborhood. Each region is approximated by an ellipse, characterized by the position of the center, its orientation, and the length of its axes, and for tracking purposes, each of them is assigned a label (Figure 2). The tracking is performed by analyzing the change in position and orientation of the ellipses along the timeline, assigning labels based on the position of the regions in the current and previous frames.

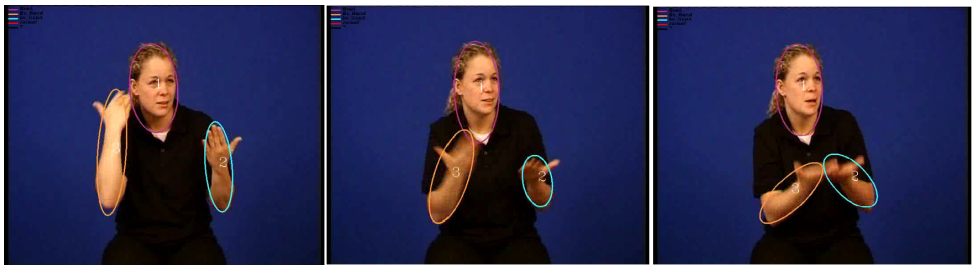


FIGURE 2: Sequence of images showing the head/hand tracking approach

4. USER INTERACTION. The expected output is very heterogeneous, and in some cases baseline recognizers can perform poorly with no additional adaptation. Furthermore, the researchers cannot accept annotation errors, such as a segment being wrongly labeled as no-speech but having speech in it (false negative). Therefore the analysis components support adaptation and feedback-loop mechanisms. The adaptation mechanism offers the researcher the opportunity to give examples of speech or video he/she is interested in, e.g. samples of a speaker for automatic speaker detection or sample segments with and without speech for the automatic detection of speech. The feedback-loop mechanism offers the user the opportunity to give feedback about the quality of the result of a first run and then perform a second process with the updated information again. For example, this could be applied for the speaker identification process: The user adapts the recognizer before running the component the first time by selecting some examples of the speaker, then runs the recognizer, and then verifies a number of segments, and the recognizer would use this response to adapt the algorithm before running the process a second time.

5. EXPERIMENTS. The development of the procedures outlined above requires intensive collaboration with researchers from the humanities in order to provide them with the most helpful and powerful tools for their annotation tasks. Therefore, a range of tests have been performed to assess how the methods can increase the effectiveness of the researchers' work. The measure of effectiveness is defined as the difference between the time necessary to create annotations for given media with and without the developed algorithms. This value cannot be easily calculated as the time necessary for annotating a time unit of media depends on factors such as 1) the purpose of the recording and contents of the media; 2) what exactly needs to be analyzed and annotated; 3) the person performing the annotation process and their expertise. Also, the level of applicability of the methods can be different for different scenarios, resulting in a different amount of help they can offer. In order to estimate the usefulness of the methods, a scenario has been created in which a researcher had to perform a number of annotation tasks aimed at different linguistic research questions. The tasks have been chosen to represent a common set of actions undertaken by a researcher annotating his/her recordings and included: 1) marking utterances of all speakers in the recording; 2) marking the size of the gesture space of a recorded person; 3) marking where speech overlaps with gestures; 4) marking specific gesturing or behavior throughout the entire recording, like nodding, raising the arms from a resting position to the level of the body, etc.; 5) marking when gesturing action happens and segmenting it into stroke, hold, and retreat. These tasks were first performed by several researchers manually, and the time necessary to carry them out was measured and averaged.

In this paper, we are describing the results of a preliminary experiment. Our aim is to gain initial insights into the potential of automatic annotation in the described scenario and use these insights to design in-depth experiments on a larger scale. For the preliminary experiment, the recognizers described above have been executed on selected exemplary recordings taken from the MPI-P archive, creating a set of initial annotations. Then, the initial automatic annotations were evaluated and manually corrected by a researcher. Figure 3 presents the time necessary to carry out three particularly time-consuming annotation tasks, namely speech utterance segmentation, gesture segmentation, and speech-gesture overlap. When the task was performed with the help of recognizers, the annotation time represents the time

required for correcting the results obtained from automatic analysis in order to make them useful for the researcher. Using the recognizers, the annotation time was greatly reduced for all three annotation tasks. The marking of utterances required some corrections of the boundaries of the annotations and also required splitting and merging some of the results of the applied algorithms. Detecting and segmenting the gestures, which are the most complex tasks, also required a significant amount of corrections. However, all test cases have proven to save a substantial amount of time required for annotation, which can be spent on other tasks.

For this preliminary experiment, recordings were not selected systematically in terms of recording conditions, length, or corresponding research scenario, but rather selected by the researchers because they were interested in the corresponding annotation task. In the next step, we plan to carry out the described experiment on a larger range of different documents and also on a larger scale such that we can analyze the effect of different recording conditions and research scenarios on the annotation speedup through automatic recognizers. High detector error rates will render our approach unusable as the number of corrections becomes intractable. Hence there is also need for investigating the minimum detector accuracy that is required for our semi-automatic approach.

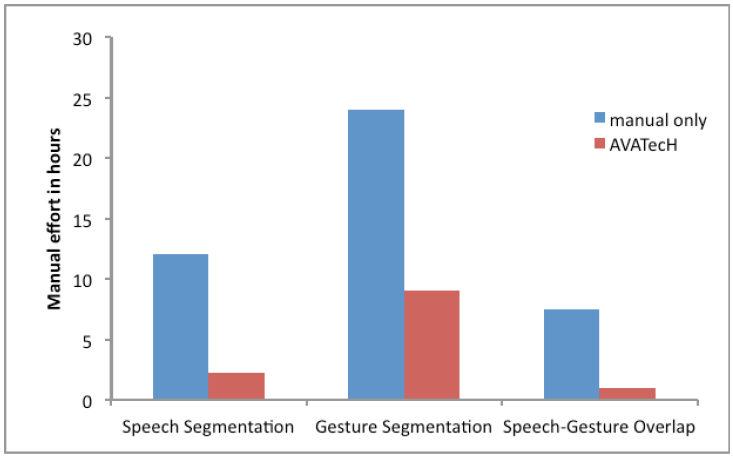


FIGURE 3: Initial evaluation results

6. CONCLUSIONS AND OUTLOOK. We have demonstrated the immense potential of automatic audio visual analysis for the processing of language documentation as a basis for linguistic research. Besides the scenarios investigated in the context of the AVATecH project, many more research domains from linguistics, and in particular gesture studies, may benefit from these techniques. An important issue for further applications of AVATecH is the availability of this technology to the international community. One challenge in this context is to find a solution as to how the technology developed by the Fraunhofer institutes can be used in the research community. Several licensing models are currently under discussion, and the common goal is to find an appropriate solution by mid 2012. It is clear that AVATecH is relevant for a large variety of research questions beyond the ones considered

in the collaborative project so far. Hence, the linguistic research community, and in particular the gesture research community, is invited to approach the AVATecH partners and to provide additional audio-visual test material and related research questions for further investigation. Only intensive collaboration between technology providers and researchers from the humanities can help to improve the methods and to adapt them to the needs and desires of the end users. We believe this collaborative work could contribute to a significant increase in the amount of annotations that can form the basis for further linguistic research.

REFERENCES

Atzpadin, Nicole, Peter Kauff & Oliver Schreer. 2004. Stereo analysis by hybrid recursive matching for real-time immersive video conferencing. *Transactions on Circuits and Systems for Video Technology, Special Issue on Immersive Telecommunications* 14(3). 321–334.

Auer, Eric, Peter Wittenburg, Han Sloetjes, Oliver Schreer, Stefano Masneri, Daniel Schneider & Sebastian Tschöpel. 2010. Automatic annotation of media field recordings. In Caroline Sporleder & Kalliopi Zervanou (eds.), *Proceedings of the ECAI 2010 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2010)*, Lisbon: University of Lisbon.

Broeder, Daan, Marc Kemps-Snijders, Dieter Van Uytvanck, Menzo Windhouwer, Peter Withers, Peter Wittenburg & Claus Zinn. 2010. A data category registry- and component-based metadata framework. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner & Daniel Tapias (eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 43–47. Valletta, Malta: European Language Resources Association (ELRA).

ELAN. Language Archiving Technology. http://www.lat-mpi.eu/tools/elan/.

PRAAT. Paul Boersma and David Weenink. Amsterdam: University of Amsterdam. http://www.fon.hum.uva.nl/praat/.

Tschöpel, Sebastian, Daniel Schneider, Rolf Bardeli, Oliver Schreer, Stefano Masneri, Peter Wittenburg, Han Sloetjes, Przemyslaw Lenkiewicz & Eric Auer. 2011. AVATecH: Audio/Video technology for humanities research. In Cristina Vertan, Milena Slavcheva, Petya Osenova & Stelios Piperidis (eds.), *Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage, Hissar, Bulgaria, 16 September 2011*, 86–89. Shoumen, Bulgaria: Incoma Ltd.

Váradi, Tamás, Peter Wittenburg, Steven Krauwer, Martin Wynne & Kimmo Koskenniemi. 2008. CLARIN: Common language resources and technology infrastructure. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis & Daniel Tapias (eds.), *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, 1244–1248. Marrakech, Morocco: European Language Resources Association (ELRA).

Oliver Schreer
oliver.schreer@hhi.fraunhofer.de

Daniel Schneider
daniel.schneider@iais.fraunhofer.de

7

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. 46–53
http://nflrc.hawaii.edu/ldc/sp03/
http://hdl.handle.net/10125/4516

Bilingual multimodality in language documentation data[1]

Marianne Gullberg
Lund University

Most people in the world speak more than one language, making bilingualism the norm rather than the exception. Furthermore, speakers generally also move their hands – they gesture – in coordination with speech and language in nontrivial ways. Bilingualism and multimodality should thus be on research agendas focused on the nature of linguistic systems and language use in context, yet they are often overlooked. Conversely, research and theorizing on bilingualism and multimodality is often based on Western-European, standardized languages, and little is known about other linguistic contexts. This paper makes the point that language documentation data has the potential to inform theoretical and empirical studies of linguistics, bilingualism and multimodality in entirely new ways, and, conversely, that documentation work would benefit from taking the bilingual and multimodal nature of its data into account.

1. INTRODUCTION. It is frequently (and probably accurately) claimed that most people in the world speak more than one language, meaning that bilingualism[2] – not monolingualism – is the norm rather than the exception (e.g., Grosjean 1982). Furthermore, when we speak, we also move our hands and arms, producing gestures in coordination with speech and language in nontrivial ways. The fact that most speakers are bilingual under some definition and that they gesture while they speak should put bilingualism and multimodality at the heart of research agendas focused on the nature of linguistic systems and language use in context. Moreover, the communities represented in documentation data of endangered languages are often bilingual, and the data at hand typically contain multimedia recordings of interactions where gestures are a natural part of language use (see Gippert, this volume). This paper therefore makes the point that language documentation data has the potential to inform theoretical and empirical studies of linguistics, bilingualism and multimodality (meaning the combined study of speech and gesture as one ensemble) in entirely new ways,

[1] I express my thanks to the organizers of the DoBeS workshop "Potentials of Language Documentation: Methods, Analyses, and Utilization" for inviting me to think about these issues. I also thank the panel coordinators, the editors, and two anonymous reviewers for helpful comments on this paper.

[2] I will use the term bilingual to refer to speakers who regularly use more than one language, regardless of the actual number of languages used, acquisition history, or proficiency level. The term therefore includes multilinguals, early simultaneous bilinguals, and speakers who have acquired their languages in adulthood ('second language speakers') but use them regularly.

and, conversely, that documentation work would benefit from taking the bilingual and multimodal nature of its data into account.

2. BILINGUALISM. Studies of bilingualism and adult second language acquisition (SLA; traditionally two separate fields) are inherently concerned with cross-linguistic and typological variation in that the properties of languages that come into contact in an individual mind are assumed to make a difference for the nature of bilingualism and language learning and use. Different linguistic domains are examined for the ways in which a bilingual's languages interact and determine the nature of the bilingual's language system. In the vast body of work in this domain, data typically come from standardized languages (European languages, Japanese, Korean, and Mandarin Chinese) studied and acquired in classroom settings. Based on such data sets, research addresses theoretical concerns. A first example of such a question is the monolingual native speaker norm. Most studies assume that the goal of bilingualism and SLA is to become like a monolingual native speaker, the golden standard against which all bilinguals and learners are measured. This is particularly obvious in the literature probing the role of maturational constraints on learning (the critical period hypothesis), asking why children seem to be 'better' language learners than adults (e.g., DeKeyser & Larson-Hall 2005, Hyltenstam & Abrahamsson 2003; for the standard view in SLA that adult learners are faster than children, who in turn may reach higher final levels of proficiency in some linguistic domains, see Krashen et al. 1982). There is a strong bias towards the standardized language in its written, academic form as opposed to the variation found in actual spoken language use. Furthermore, there is surprisingly little recognition of the fact that if most speakers in the world are bilingual, then it is unclear who the monolingual native speaker is and what the norm represents (cf. Davies 2003). Even studies that recognize that bilinguals are not two monolinguals in one body (cf. Grosjean 1998) nevertheless compare bilinguals to monolinguals, and usually find fault.

A second issue is the question of general and supposedly universal developmental trajectories in adult acquisition whereby learners, for instance, acquire grammatical morphemes in a set order (e.g. Dulay & Burt 1972, 1978) or pass through specific developmental stages of word order acquisition regardless of the patterns in the first and second language (Hyltenstam 1978, Klein & Perdue 1997, Pienemann 1998).

A third question concerns the role of the other language, known as transfer or cross-linguistic influence (Jarvis & Pavlenko 2008, Kellerman & Smith 1986, Odlin 1989). Traditionally, the first language (L1) has been assumed to 'leak' into the second language (L2) in the form of foreign accent, lexical choice, structures, etc. Similarity between languages is supposed to be helpful; differences are assumed to cause difficulties in learning (Ringbom 2007 for an overview). However, recent work highlights the fact that languages affect each other in all directions such that the second language also affects the first, making the pure monolingual speaker even less likely (e.g. Brown & Gullberg 2011, Cook 2003). The psycholinguistic literature focuses on the costs of controlling this interaction. Bilinguals are often shown to be slower to name pictures than monolinguals, leading to the assumption that it is costly to be bilingual. This assumption is based on ideas of co-activation in processing, inhibition, and executive control (e.g. Costa 2005, Meuter 2005).

Finally, research on code-switching – that is, the use of two languages in the same utterance, clause, or phrase – examines what sort of social and linguistic constraints can be

placed on mixing, asking whether mixing is orderly or random (e.g. Auer 1998, Clyne 2003, Muysken 1995, Myers-Scotton & Jake 2001, O'Shannessy 2011). Recently, researchers also ask what psycholinguistic constraints may be placed on mixing.

Two things are noteworthy. First, empirical facts and theories about bilingualism/acquisition (within the four domains outlined, but also beyond them) are based on very small samples of languages that are formally acquired, supporting the bias towards monolingual norms. Findings may therefore not reflect typical patterns across larger samples and in the untutored 'wild'. It is unclear how the standard claims would fare if language documentation data were taken into account. For example, the claims might look different if the full extent of linguistic variation in contexts where speakers are functionally bilingual with complex sociolinguistic and functional divisions of labor between languages were taken into account. Similarly, the claims might take a different form if contexts where speakers are acquiring languages for different purposes other than passing academic proficiency tests were considered. Put differently, there are substantial gaps in our knowledge about bilingualism and adult acquisition. The variation and rich bilingual contexts in language documentation data raise important challenges for bilingual/acquisition studies with implications for the validity of the theorizing around these issues.

Second and conversely, studies of language contact and documentation work rarely consider theoretical claims in SLA and bilingualism concerning acquisition, language processing, and the nature of bilingual linguistic systems in general. This means that descriptions of endangered systems as if they were monolingual also risk misrepresenting the linguistic reality under study. In this case, insights from bilingualism studies raise challenges for descriptive work, and this has implications for the theorizing around the nature of linguistic systems (cf. Evans & Levinson 2009a,b, Levinson & Evans 2010).

3. MULTIMODALITY – HEARING AND SEEING LANGUAGE. When we speak, we typically also gesture, and we deploy this expressive resource in systematic and nontrivial ways. Contemporary gesture studies indicate that gestures are closely linked to language and speech in production, comprehension, and development, and interact with cultural, social, psychological, and linguistic aspects of communication (cf. Goldin-Meadow 2003, Kendon 2004, McNeill 2005). The connections are manifold: Gestures improve speech comprehension (e.g. Rogers 1978); their presence or absence influences both the content and the fluency of speech production (e.g. Bernardis & Gentilucci 2006, Graham & Heywood 1975); and the modalities develop in parallel in childhood (e.g. Capirci et al. 2005, Hickmann et al. 2011, Iverson & Goldin-Meadow 2005), and break down together in disfluency and stuttering (Mayberry & Jaques 2000, Seyfeddinipur 2006). Speech and gestures also combine in complex ways to regulate speakerhood, turn-taking, and other aspects of conversation (Duncan 1972, Schegloff 1984).

A number of theoretical issues are being discussed in this field. One concerns the relationship between speech and non-conventionalized gestures (so-called co-speech gestures that lack standards of well-formedness) and the assumption that they form an integrated and co-orchestrated system (Clark 1996, Goldin-Meadow 2003, Kendon 2004, McNeill 2005). Specifically at stake is their common conceptual origin. Tight semantic and temporal cross-modal coordination (similar meaning expressed at the same time, e.g., a gesture tracing a

circling movement rightward as the speaker says *roll down the street*) is taken as evidence for such a joint stage, suggesting that gestures are planned and organized with speech.

A second and related question is the extent to which different linguistic communities have different gestural repertoires and whether they are determined by cultural convention or by linguistic factors. If speech and gestures express the same meaning at the same time (*roll down the street* in speech and a rightward circling gesture), and different languages express different meanings in speech (e.g., *descend la rue* 'descend the street' [path but not manner of motion] vs. *roll down the street* [manner and path of motion], then logically, different linguistic communities should have different gestural repertoires. Moreover, traditional claims about cultural differences between, for instance, gesture-rich (e.g. Italian) vs. gesture-poor (e.g. Japanese) communities are persistent but not supported by empirical facts. Studies of culture-specific meanings of conventional gestures (e.g. the ring-gesture which can mean 'zero', 'excellent', 'money' or 'orifice', depending on location) are also relatively rare. Therefore, cross-linguistic and cross-cultural issues remain relatively underexplored in gesture studies. Existing work predominantly focuses on gestural practices in European languages, Japanese and Mandarin Chinese (Kita 2009 for an overview). Very few studies have examined how speech and gestures are mobilized and orchestrated in a wider set of languages. When they have, studies have often focused on deixis and pointing (e.g. Kita & Essegbey 2001, Sherzer 1973, Wilkins 2003). Much less is known about the full repertoires of conventional and non-conventional gestures, forms and meanings, and their deployment in discourse (e.g. Barakat 1973, Brookes 2004, Newbury 2011, Núñez & Sweetser 2006).

A third related issue is how children and other learners become competent, native speakers and gesturers of a particular language in a given community. The question of whether gestures are learned (through imitation and molding) or whether their development is based on linguistic development remains a contentious topic (e.g. Bates et al. 1977, Gullberg et al. 2008). In adults and bilinguals, issues of cross-linguistic influence and general developmental trends become pertinent with questions concerning whether one repertoire leaks into another, whether bilinguals code-switch in gesture along with speech, etc. (e.g. Brown & Gullberg 2008, Gullberg 2009, Nicoladis et al. 1999). These are largely open empirical questions at this point.

Thus, as with bilingualism studies, the empirical facts and theories around speech, language, and gestures are based on very small samples of standardized languages, and typically on adult native monolingual speakers. Current empirical findings may represent subpatterns, and again, it is unclear how theoretical claims would fare if language documentation data were taken into account to explore the full extent of gestural, cultural, and linguistic variation found in contexts typically described. It seems likely that the current theoretical landscape would change considerably. The gaps in our knowledge will have consequences for the validity of the theorizing around these issues.

Conversely, although language documentation now often comprises multimodal resources such as audio and video, these are rarely exploited to address theoretical questions regarding the multimodal nature of language, acquisition, and bilingualism. More complete descriptions of language in contextual use – the natural habitat of language – should obviously take gestures into account. Gesture analysis can open new windows on the nature of the systems under description and on the interactive practices. Insights from gesture studies

raise challenges for descriptive work and this has implications for theories pertaining to the nature of linguistic systems.

4. CONCLUDING REMARKS. I have outlined a situation where bilingualism/SLA studies and studies of multimodality remain uninformed by the knowledge accumulated in language documentation work and the linguistic analyses resulting from it. I have also suggested that theoretical issues raised in the fields of bilingualism and gesture studies are not discussed in the domain of linguistic analysis or contact linguistics. This mutual ignorance is unfortunate. Clearly, vital theoretical and empirical gains could be made in all fields if researchers collaborated and considered data and frameworks on both sides. Bilingualism and multimodality studies are rife with testable claims and theories waiting to be challenged and informed by the rich resources in documentation data. Documentation work would also benefit from acknowledging the bilingual and multimodal nature of the data collected. The way forward then, to achieve the "sea change in linguistics" that Levinson & Evans (2010) discuss, lies in joint ventures and collaborative interdisciplinary work to exploit the data already available in archives and to inform new data collection. In this way, the language sciences can become truly accumulative, and a broader view of multimodal language in situated use can inform our theories. We have everything to gain and very little to lose.

REFERENCES

Auer, Peter. 1998. *Code-Switching in Conversation: Language, Interaction and Identity*. London: Routledge.

Barakat, Robert. 1973. Arabic gestures. *Journal of Popular Culture* 6(4). 749–793.

Bates, Elizabeth, Laura Benigni, Inge Bretherton, Luigia Camaioni & Virginia Volterra. 1977. From gesture to first word: On cognitive and social prerequisites. In Michael Lewis & Leonard A. Rosemblum (eds.), *Interaction, Conversation, and the Development of Language*, 247–307. New York: Wiley.

Bernardis, Paolo P. & Maurizio M. Gentilucci. 2006. Speech and gesture share the same communication system. *Neuropsychologia* 44(2). 178–190.

Brookes, Heather. 2004. A repertoire of South African quotable gestures. *Journal of Linguistic Anthropology* 14(2). 186–224.

Brown, Amanda & Marianne Gullberg. 2008. Bidirectional crosslinguistic influence in L1-L2 encoding of manner in speech and gesture: A study of Japanese speakers of English. *Studies in Second Language Acquisition* 30(2). 225–251.

Brown, Amanda & Marianne Gullberg. 2011. Bidirectional crosslinguistic influence in event conceptualization? The expression of Path among Japanese learners of English. *Bilingualism: Language and Cognition* 14(Special Issue 01). 79–94.

Capirci, Olga, Annarita Contaldo, M. Cristina Caselli & Virginia Volterra. 2005. From action to language through gesture: A longitudinal perspective. *Gesture* 5(1–2). 155–177.

Clark, Herbert H. 1996. *Using Language*. Cambridge: Cambridge University Press.

Clyne, Michael. 2003. *Dynamics of Language Contact: English and Immigrant Languages*. Cambridge: Cambridge University Press.

Cook, Vivian (ed.). 2003. *Effects of the Second Language on the First*. Clevedon: Multilingual Matters.

Costa, Albert. 2005. Lexical access in bilingual production. In Judith F. Kroll & Annette M.B. de Groot (eds.), *Handbook of Bilingualism: Psycholinguistic Approaches*, 308–325. Oxford: Oxford University Press.

Davies, Alan. 2003. *The Native Speaker: Myth and Reality*. Clevedon: Multilingual Matters.

DeKeyser, Robert & Jenifer Larson-Hall. 2005. What does the critical period really mean? In Judith F. Kroll & Annette M B. de Groot (eds.), *Handbook of Bilingualism: Psycholinguistic Approaches*, 88–108. Oxford: Oxford University Press.

Dulay, Heidi C. & Marina K. Burt. 1972. Goofing: An indicator of children's second language learning strategies. *Language Learning* 22(2). 235–252.

Dulay, Heidi C. & Marina K. Burt. 1978. Natural sequences in child second language acquisition. In Evelyn Marcussen Hatch (ed.), *Second Language Acquisition: A Book of Readings*, 347–361. Rowley, MA: Newbury House.

Duncan, Jr., Starkey. 1972. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology* 23. 283–292.

Evans, Nicholas & Stephen C. Levinson. 2009a. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences* 32(5). 429–492.

Evans, Nicholas & Stephen C. Levinson. 2009b. With diversity in mind: Freeing the language sciences from Universal Grammar [Authors' Response]. *Behavioral and Brain Sciences* 32(5). 472–492.

Gippert, Jost. this volume. Language-specific encoding in endangered language corpora.

Goldin-Meadow, Susan. 2003. *Hearing Gesture: How Our Hands Help us Think*. Cambridge, MA: Harvard University Press.

Graham, Jean Ann & Simon Heywood. 1975. The effects of elimination of hand gestures and of verbal codability on speech performance. *European Journal of Social Psychology* 5(2). 189–195.

Grosjean, François. 1982. *Life with Two Languages: An Introduction to Bilingualism*. Cambridge, MA: Harvard University Press.

Grosjean, François. 1998. Studying bilinguals: Methodological and conceptual issues. *Bilingualism: Language and Cognition* 1(2). 131–149.

Gullberg, Marianne. 2009. Reconstructing verb meaning in a second language: How English speakers of L2 Dutch talk and gesture about placement. *Annual Review of Cognitive Linguistics* 7(1). 221–244.

Gullberg, Marianne, Kees de Bot & Virginia Volterra. 2008. Gestures and some key issues in the study of language development. *Gesture* 8(2). 149–179.

Hickmann, Maya, Henriette Hendriks & Marianne Gullberg. 2011. Developmental perspectives on the expression of motion in speech and gesture: A comparison of French and English. *Language, Interaction and Acquisition* 2(1). 129–156.

Hyltenstam, Kenneth. 1978. *Progress in Immigrant Swedish Syntax: A Variability Analysis*. Lund: Lund University.

Hyltenstam, Kenneth & Niclas Abrahamsson. 2003. Maturational constraints in SLA. In Catherine J. Doughty & Michael H. Long (eds.), *The Handbook of Second Language Acquisition. Blackwell Handbooks in Linguistics*, 539–588. Malden, MA: Blackwell.

Iverson, Jana M. & Susan Goldin-Meadow. 2005. Gesture paves the way for language development. *Psychological Science* 16(5). 367–371.

Jarvis, Scott & Aneta Pavlenko. 2008. *Crosslinguistic Influence in Language and Cognition*. New York: Routledge.

Kellerman, Eric & Michael Sharwood Smith. 1986. *Crosslinguistic Influence in Second Language Acquisition*. New York: Pergamon.

Kendon, Adam. 2004. *Gesture: Visible Action as Utterance*. Cambridge: Cambridge University Press.

Kita, Sotaro. 2009. Cross-cultural variation of speech-accompanying gesture: A review. *Language and Cognitive Processes* 24(2). 145–167.

Kita, Sotaro & James Essegbey. 2001. Pointing left in Ghana: How a taboo on the use of the left hand influences gestural practice. *Gesture* 1(1). 73–95.

Klein, Wolfgang & Clive Perdue. 1997. The basic variety (or: Couldn't natural languages be much simpler?). *Second Language Research* 13(4). 301–347.

Krashen, Stephen D, Michael H Long & Robin C Scarcella. 1982. Age, rate, and eventual attainment in second language acquisition. In Stephen D. Krashen, Robin C. Scarcella & Michael H. Long (eds.), *Child-Adult Differences in Second Language Acquisition*, 161–172. New York: Newbury House.

Levinson, Stephen C. & Nicholas Evans. 2010. Time for a sea-change in linguistics: Response to comments on 'The Myth of Language Universals'. *Lingua* 120(12). 2733–2758.

Mayberry, Rachel I. & Joselynne Jaques. 2000. Gesture production during stuttered speech: Insights into the nature of gesture-speech integration. In David McNeill (ed.), *Language and Gesture*, Cambridge: Cambridge University Press.

McNeill, David. 2005. *Gesture and Thought*. Chicago: University of Chicago Press.

Meuter, Renata F. I. 2005. Language selection in bilinguals: Mechanisms and processes. In Judith F. Kroll & Annette M. B. de Groot (eds.), *Handbook of Bilingualism: Psycholinguistic Approaches*, 349–370. Oxford: Oxford University Press.

Muysken, Pieter. 1995. Code-switching and grammatical theory. In Lesley Milroy & Pieter Muysken (eds.), *One Speaker, Two Languages: Cross-Disciplinary Perspectives on Code-Switching*, 177–198. Cambridge: Cambridge University Press.

Myers-Scotton, Carol & Janice L. Jake. 2001. Explaining aspects of code-switching and their implications. In Janet L Nicol (ed.), *One Mind, Two Languages: Bilingual Language Processing*, 84–116. Malden: Blackwell.

Newbury, Kendra. 2011. Gesture and language shift on the Uruguayan-Brazilian border. In Gale A. Stam & Mika Ishino (eds.), *Integrating Gestures: The Interdisciplinary Nature of Gesture*, 231–241. Amsterdam: Benjamins.

Nicoladis, Elena, Rachel I. Mayberry & Fred Genesee. 1999. Gesture and early bilingual development. *Developmental Psychology* 35(2). 514–526.

Núñez, Rafael E. & Eve Sweetser. 2006. With the future behind them: Convergent evidence from Aymara language and gesture in the crosslinguistic comparison of spatial construals of time. *Cognitive Science* 30(3). 401–450.

Odlin, Terence. 1989. *Language Transfer: Cross-Linguistic Influence in Language Learning*. Cambridge: Cambridge University Press.

O'Shannessy, Carmel. 2011. Language contact and change in endangered languages. In Peter K. Austin & Julia Sallabank (eds.), *The Cambridge Handbook of Endangered Languages*, 77–99. Cambridge: Cambridge University Press.

Pienemann, Manfred. 1998. *Language Processing and Second Language Development. Processability theory*. Amsterdam: Benjamins.

Ringbom, Håkan. 2007. *Cross-Linguistic Similarity in Foreign Language Learning*. Clevedon: Multilingual Matters.

Rogers, William T. 1978. The contribution of kinesic illustrators toward the comprehension of verbal behavior within utterances. *Human Communication Research* 5(1). 54–62.

Schegloff, Emanuel A. 1984. On some gestures' relation to talk. In John Maxwell Atkinson & John Heritage (eds.), *Structures of Social Action: Studies in Conversation Analysis*, 266–296. Cambridge: Cambridge University Press.

Seyfeddinipur, Mandana. 2006. *Disfluency: Interrupting Speech and Gesture*, vol. 39 MPI Series in Psycholinguistics. Nijmegen: Radboud University, Nijmegen.

Sherzer, Joel. 1973. Verbal and nonverbal deixis: The pointed lip gesture among the San Blas Cuna. *Language in Society* 2(1). 117–131.

Wilkins, David. 2003. Why pointing with the index finger is not a universal (in sociocultural and semiotic terms). In Sotaro Kita (ed.), *Pointing: Where Language, Culture, and Cognition Meet*, 171–215. Mahwah, NJ: Erlbaum.

Marianne Gullberg
marianne.gullberg@ling.lu.se

8

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. 54–63
http://nflrc.hawaii.edu/ldc/sp03/
http://hdl.handle.net/10125/4517

Tours of the past through the present of eastern Indonesia

Marian Klamer
Leiden University

The past twenty years have seen a variety of data being collected from largely undocumented languages in eastern Indonesia, an area hitherto almost unknown. Such data are valuable in reconstructing the history of this area at a macro-level. In addition, as research in particular areas becomes more fine-grained, it is possible to combine linguistic data with data from oral history and ethnographic observation in order to reconstruct the migration histories of specific speaker groups. A case study of such a micro-level reconstruction is presented here.

1. INTRODUCTION. One fundamental idea in historical and contact linguistics is that similarities between two geographically close languages are not accidental but point to a shared history of their speakers. The speakers either descend from a common ancestor so that the similar features were passed down the generations; or they were once in contact with each other and features from the one language were adopted into the other. The nature and spread of similarities between languages spoken today can thus be studied to reconstruct pictures of the past of their speakers. This is particularly relevant in regions where written historical records are generally lacking, and archaeological records, if they exist, have yet to be uncovered. One such region is eastern Indonesia.

Obviously, a comparison of similarities across languages can only be done once the languages have been described. This paper therefore addresses the following question: How did language description and documentation in eastern Indonesia contribute to our knowledge of the history of people living in the area? After presenting an overview of descriptive and documentation efforts in eastern Indonesian languages (Section 2), the impact of the documentation will be illustrated at the level of the language families in the area (Section 3) and at the level of individual language communities (Section 4). Section 5 presents a summary.

2. RECENT DESCRIPTIVE AND DOCUMENTATION EFFORTS IN EASTERN INDONESIA. Linguistically and ethnically, eastern Indonesia constitutes an interface between the Austronesian and Papuan worlds. Papuans have lived in this area for more than 40,000 years, whereas Austronesians came down from Taiwan less than 6,000 years ago.

What was originally the Papuan area became largely "austronesianized" through the incoming Austronesians, who assimilated with the original populations. However, in some locations, Papuan languages continued to be spoken – in Papua itself, as well as in outlier groups located west and east of Papua. The westerly outlier groups are in Halmahera in the north and Timor Alor Pantar in the south, as indicated in Figure 1. Note that 'Papuan' is a cover term for numerous mutually unrelated non-Austronesian language groups in Papua or its vicinity.

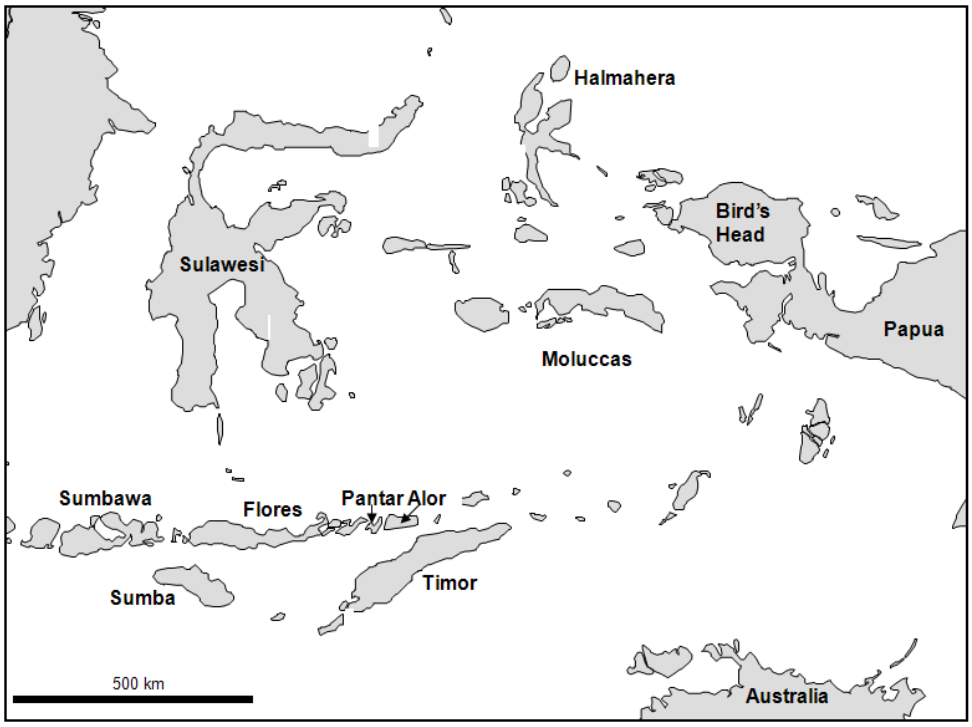


FIGURE 1: Eastern Indonesia

The number of languages spoken in the area of eastern Indonesia as discussed here may be estimated at 200–250 – a figure that is vague for lack of data (see also Hammarström & Nordhoff, forthcoming). Since the early 1990's, when eastern Indonesia was considered "perhaps the least known area in the Austronesian world today" (Tryon 1995: 12), we have gone from almost no information to a body of documentation on 25–30 languages. Most of these are grammars of individual languages that were produced by PhD students or postdoctoral researchers from institutions in Australia and the Netherlands. In addition, and with varying results, documentation projects have focused on Rongga in Flores, Waima'a and Fataluku in East Timor, languages of Central Maluku (Alang, Sou Amana Teru, Alune), and Totoli in Northern Sulawesi. Documentation is ongoing for Iha in West Papua, Mor on the Bomberai peninsula, and Wooi in Yapen Island in the Geelvink Bay, north of Papua's mainland. In sum, during the past two decades about 10% of the languages in eastern Indonesia

have been grammatically and lexically described (Klamer & Ewing 2010: 5 provide references), and a handful have been documented.

3. IMPACT AT THE LEVEL OF LANGUAGE FAMILIES. Research relating to language family groupings sets out to know whether or not two languages have a common ancestor, and if so, what their ancestor would have looked like. A linguist who encounters an undescribed language will first try to relate that language to a language family in the area, since it is easier to analyze an undescribed language when it can be related to existing knowledge of other languages. In eastern Indonesia, this means asking the question whether the language is Austronesian or belongs to one of the Papuan families.

Which parts of the new language do we consider in our attempt to group it with an existing family? Classic comparative research hypothesizes family relations on the basis of geographical closeness of languages and then proceeds to demonstrate the relatedness by comparing sets of cognate words (lexicon) and paradigms (morphology). But what if geographically close languages do not have cognate words and morphology? Can we group languages together on the basis of other types of evidence, such as shared structural features in the domains of phonology, morphology, and syntax? Examples of structural dimensions (also referred to as 'typological features') along which languages in eastern Indonesia have been grouped, include segment inventory (how many consonants and vowels do languages have; and which ones?), syllable structure, the position of affixes (prefixing or suffixing?), word order (verb-object, or object-verb?), types of grammatical alignment systems (nominative-accusative, active-stative, or split S?), and the encodings of nominal possession (Foley 2000, Klamer 2002, Himmelmann 2005, Pawley 2005, Donohue 2007).

It is generally agreed that shared structural features in isolation are too instable to signal deep time family relations between languages. It is, however, commonly agreed that if languages share a set of such features, this might point to a history of contact between their speakers. For example, when people adopt a second language (to trade, get an education, when they marry, and so on), they will – mostly unconsciously – use some structures of their first language in the new language they adopt. As a result, their second language will contain features that are similar with the first, and these features can be passed down to the language their children acquire. Such similarities between languages are then due to linguistic contact rather than to a common ancestor language.

With the amount of data on languages of eastern Indonesia increasing, it is now possible to group the languages in this region not only according to their genealogy but also according to their structural features. When the groupings overlap, it is possible to identify certain shared structural features as 'typical for Austronesian' or 'typical for Papuan' and then formulate hypotheses about the direction in which the features transferred – from Austronesian to Papuan or the other way around.

For some Austronesian languages in eastern Indonesia with structures unlike those found in western Austronesian languages, and similar to those found in Papuan languages, we believe the structures have a Papuan donor; while some Papuan languages have adopted Austronesian features (see Klamer et al. 2008, and references cited there). The study of similarity patterns across languages of different families thus enables us to reconstruct where there has been contact between Austronesian and Papuan groups, even in locations where no such contact exists today. It also allows us to formulate hypotheses about the origin of the

adopted features as Austronesian or Papuan, and the direction of transfer; this is information which is valuable in the study of the population history of the region.

However, it must be stressed that the amount of linguistic information for this immense area is still extremely limited, and the set of languages that can be compared constitute a convenience sample, not a geographically representative one. Any macro-level contact scenarios we propose must therefore remain hypothetical, and often we only have evidence for diffusion through (pre-historic) substrate contact between Austronesian and Papuan in a particular region. This can mean three things: (i) Papuan speakers shifted to an Austronesian language and introduced Papuan traits into it. This presumably happened in the languages of Halmahera such as Taba (Bowden 2001); (ii) Austronesian speakers shifted to a Papuan language and introduced Austronesian traits into it. This is what must have happened in Fataluku in East Timor (McWilliam 2007); and (iii) there has been a diffusion of features, but we do not know in which direction the transfer took place. Note that the absence of any evidence for diffusion in certain areas (e.g. Sumba) is equally interesting as it suggests a different population history.

In sum, description and documentation efforts in eastern Indonesia are having an impact at the macro-level in the sense that, with the newly available data, more detailed genealogical groupings and typological characterizations of both Austronesian and Papuan language systems can be designed. And these typological 'fingerprints' can be used to compare languages for the presence of features that might have crossed family borders. As far as such features signal contact between members of the two families, they constitute evidence that in certain zones of Indonesia, speakers of Austronesian and Papuan languages have been in contact with each other, even if we see no such contacts today, while in other areas no traces of such contacts exist.

4. IMPACT AT THE LEVEL OF LANGUAGE COMMUNITY. In addition to the larger-scale comparisons discussed in the previous section, the last ten years have seen some regions within eastern Indonesia being studied with more granularity. One such region is the Alor Pantar archipelago, just north of Timor. The Alor Pantar islands are home to some twenty closely related Papuan languages (Holton et al. 2012), see Figure 2.

Six of these languages have now been studied in depth, and we also have extensive lexical and syntactic survey data that is representative for the whole group. One language is an exceptional case: It is spoken on the coasts of Pantar and Alor and is locally referred to as *Bahasa Alor* 'Alorese'. Alorese has less than 5% lexical similarity to its Papuan neighbors, and its lexicon and morphology clearly show it belongs to the Austronesian family. The Alorese are considered 'non-indigenous' in contrast to the mountain populations of Alor and Pantar (Anonymous 1914: 75–78). The fact that they only inhabit coastal areas (see Figure 2) indeed suggests a sea-faring past.

What is the origin of this coastal Austronesian group that is surrounded by Papuan speakers? When did they arrive? To investigate this, I compared Alorese with a language assumed to be its closest genealogical relative, Lamaholot. Lamaholot is spoken in a number of dialectal varieties on the islands of Flores, Solor, Adonara, and Lembata, located some 200 km east of Pantar Island, see Figure 3.

For the comparison I used the following types of data: Survey word lists and recent grammar sketches on Lamaholot and Alorese that contain information on derivational and

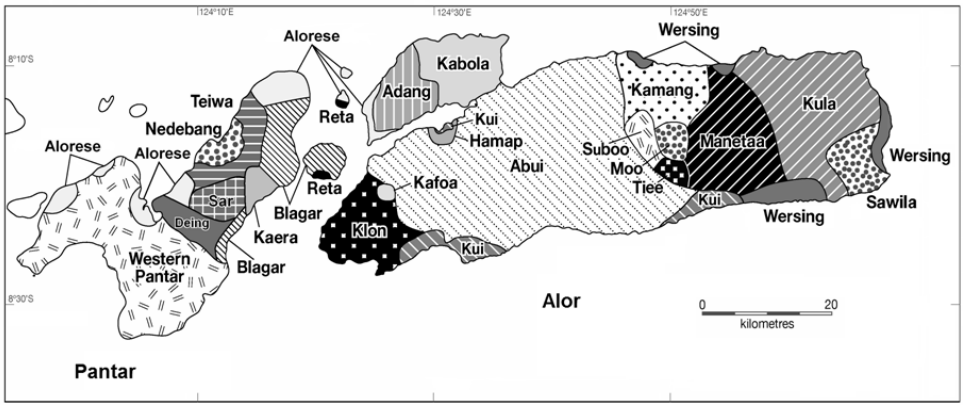


FIGURE 2: Languages on Pantar and Alor

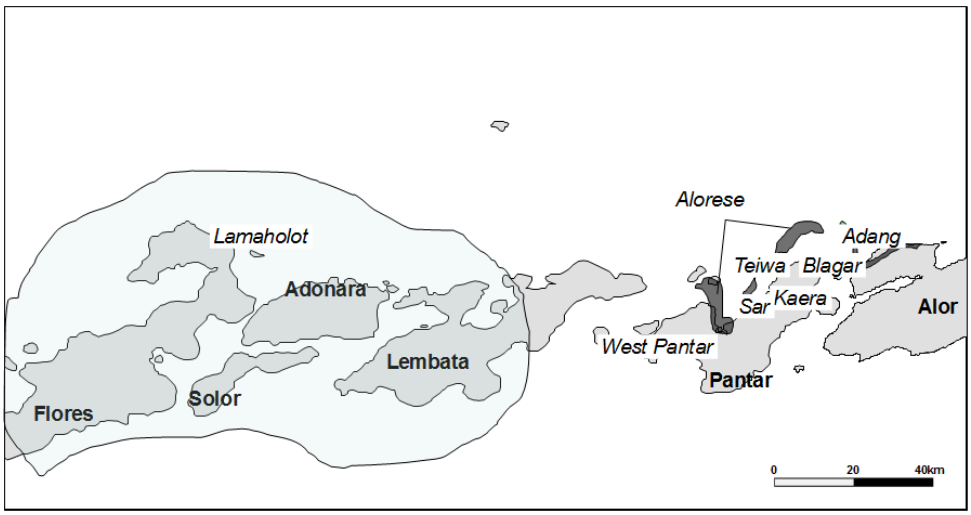


FIGURE 3: Lamaholot, Alorese, and Papuan languages surrounding Alorese

inflectional morphology and syntax (Nishiyama & Kelen 2007, Klamer 2011), published information on proto-Austronesian and proto-Malayo-Polynesian morpheme reconstructions (Blust 2009), historical and ethnographic sources (Anonymous 1914, Lemoine 1969, Barnes 1973: 86, Barnes 2001: 280, Rodemeier 2006), and personal communications about Alorese oral history (for details see Klamer 2011, Klamer, forthcoming).

I found that Lamaholot and Alorese have around 60% lexical similarity (Klamer 2011: Appendix), which suggests that they are closely related but separate languages, comparable to German and English, which are also 60% lexically similar. The most striking contrast between Lamaholot and Alorese is their morphological profile: While Lamaholot has a significant amount of derivational and inflectional morphology, Alorese is completely isolating,

as illustrated in Table 1. Comparing the Lamaholot derivational morphemes with forms reconstructed for Proto-Austronesian (PAN) or Proto-Malayo-Polynesian (PMP) (a subgroup of Austronesian) clearly shows they are Austronesian.

DERIVATIONAL MORPHOLOGY	LAMAHOLOT	ALORESE
Consonant replacement, e.g. ***p**et* 'bind' > ***m**et* 'belt' < PAN *ma 'stative'	yes	no
Prefix *bə(C)-*, e.g. *rawuk* 'hair' > ***bə**-rawuk* 'have hair' < PMP *maR- 'intransitive verb'	yes	no
Prefix *pə-*, e.g. *tua* 'palm wine' > ***pə**-tuak* 'taste like palm wine' < PMP *pa-ka- 'treat like X'; *tutu'* 'speak', ***pə**-nutu'* < PMP *paR 'deverbal noun'	yes	no
Prefix *kə-*, e.g. *pasa* 'swear' > ***kə**-pasa* 'oath' < PMP *ka- 'formative for abstract nouns'	yes	no
Infix *-ən-*, e.g. *tali* 'add' > *t-**ən**-ali* 'added thing?' < PAN *-um- 'Actor voice'	yes	no
Prefix *mən-*, e.g. *ba'at* 'heavy' > ***mən**-a'at* 'something heavy' < PAN *ma 'stative'	yes	no
INFLECTIONAL MORPHOLOGY		
Possessor suffix	yes	no
Prefix encoding subjects on vowel-initial transitive verbs	yes	a few fossilized
Suffix encoding subjects of intransitive verbs	yes	no
Agreement on adjectives	yes	no
Agreement on numerals	yes	no

TABLE 1: Derivational and inflectional morphology in Lamaholot and Alorese[1]

The contrasts in the inflectional morphology of Lamaholot and Alorese are illustrated for possessor encoding in examples (1) and the encoding of transitive subjects in examples (2).

(1) Possessor suffix present in Lamaholot (LMH), absent in Alorese (ALR)[2]

a. LMH (Nishiyama & Kelen 2007: 24)
*lango-**nən***
house-3SG
'his house'

b. ALR (Klamer 2011: 52)
ni *uma*
3SG house
'his house'

[1] Lamaholot data from Nishiyama & Kelen 2007: 48–54; PAN and PMP data from Blust 2009: 359, 363–366, 370.

(2) Subject prefix on transitive verb in Lamaholot (LMH), absent in Alorese (ALR)

a. LMH (Nishiyama & Kelen 2007: 124)

Go	***k**-a'an*	*tembok*	*pe'en*	*me'an*
1SG	**1SG**-make	wall	the	red

'I made the wall red.'

b. ALR (Kmalner n.d., AAS: 021)

Go	*lelang*	*tale*	*te*	*gato*
1SG	make	rope	that	snap

'I snapped that rope.'

The data in Table 1 illustrate that Austronesian morphology is abundant in Lamaholot and virtually absent in Alorese. Both languages are closely related – they descend from a common ancestor language, which we may refer to as 'proto-Lamaholot-Alorese'. As morphemes are more easily lost over time than they are gained, I propose that Proto-Lamaholot-Alorese had at least the amount of morphology of today's Lamaholot. When Alorese and Lamaholot split, Lamaholot kept reflexes of much of the proto-morphology, while almost all of it was lost in Alorese.

Given the geographical distance between the Lamaholot and Alorese settlements (see Figure 3), the split must have occurred when speakers emigrated from the Lamaholot homeland on Flores, Solor, and Lembata, to settle on Pantar and Alor some 200 kilometers further east. This migration occurred before or around ~1,400 AD. This date *ante quem* is based on local legends, which contain reference to a 'colony of Javanese' or *orang djawa* settled on the north coast of Pantar '5 to 600 years ago' [in 1914] (Anonymous 1914: 77). Note that in regional Malay (used by Anonymous to record the legends) the notion *orang djawa* does not necessarily denote people from Java but probably refers to immigrants from elsewhere in the archipelago.[3] In other words, the coastal settlers mentioned in the legends which are referred to as 'Javanese' may be considered 'foreigners'. This ties in with observations that no linguistic or cultural links exist between Lamaholot and Javanese, while today's Alorese and Lamaholot are very close on both counts. In sum, I propose that the *orang djawa* in these old legends were in fact Lamaholot speakers from Flores, Solor, or Lembata.

The following is additional supporting evidence. The legend referred to by Anonymous (1914) is the first of two legends also reported in Lemoine (1969), Barnes (1973: 86, 2001: 280), and Rodemeier (2006). The first legend recounts that two 'Javanese' brothers, Aki Ai and his younger brother Mojopahit, sailed to Pantar, where Aki Ai treacherously abandoned Mojopahit. Mojopahit's descendants eventually established four kingdoms on the coasts of Pantar and Alor, called Baranusa, Pandai, Munaseli, and Alor Besar. The second legend recounts that 'Javanese' immigrants who were allied to one of the Pantar kings destroyed one of the other kingdoms, after which the defeated population fled to Alor Island. It thus seems clear that four foreign kingdoms were in place around 1,300 AD in northern Pantar; that they were established by non-indigenous colonizers who came from the west; and that the same groups colonized the Alor coast (see Rodemeier 2006 for more

[2] Abbreviations: 1, 3 – 1st, 3rd person; SG – singular

[3] Cf. footnote 1 in Anonymous (1914: 89): 'on these islands, "orang djawa" is interpreted as everything ("alles") which comes from elsewhere in the archipelago'; compare Kambera (Sumba) *tau Jawa* (lit. 'Javanese') 'stranger, someone not from Sumba' (Onvlee 1984: 115).

discussion). It is then interesting to observe that today, the location names Baranusa, Alor Besar, Pandai, and Munaseli are still known, and that these are the places where speakers of Alorese live.[4] In sum, I propose that the legendary 'Javanese' who settled on Pantar some 700 years ago were in fact the ancestors of today's Alorese speakers.

Having established that Alorese is a relative newcomer on Pantar Island, and seeing that it is surrounded by Papuan neighbors, we can investigate the amount of contact that existed between the Alorese and their neighbors by examining the features they borrowed from Papuan. Data from recent research now make it possible to compare the lexicon and syntax of Alorese with five of its Papuan neighbor languages: Teiwa, Blagar, Kaera, Sar, Western Pantar, and Adang, see Figure 2.

The somewhat surprising outcome of this comparison is that very few Papuan words (less than ~5% of the basic Alorese lexicon, from 5 different Papuan languages) were borrowed into Alorese, and also very little Papuan syntax was borrowed. The local Papuan languages have had minimal influence on Alorese, which suggests that contact between Alorese and Papuan speakers cannot have been very intensive or long-term. Note that the fact that Alorese and Papuan speakers are currently in contact is irrelevant as Indonesian is the lingua franca on the islands.

The limited amount of lexical and syntactic borrowing from Papuan into Alorese does not connect straightforwardly with the observation that Alorese lost all the morphology of its ancestor language. Inflectional and derivational morphology is known to be seriously problematic for post-adolescent second language learners who have passed the 'critical threshold' (Lenneberg 1967) for language acquisition (Kusters 2003: 21, 48). In other words, when morphology gets lost, that usually involves a stage where the language is learned as a second language by post-adolescent speakers. There is, as of yet, no evidence of Papuan speaker groups shifting *en masse* to Alorese. There is however reason to believe (see Klamer 2011, forthcoming) that Alorese men married women from various neighboring Papuan clans, thus bringing adult speakers of different Papuan languages into the Alorese community. These individuals acquired a simplified Alorese variety, which became nativized and was passed down the generations. The question remains why these Papuan speakers did not introduce more of their Papuan syntax and lexicon into their Alorese, or if they did, why their children did not acquire it. Was there community pressure to speak Alorese in its lexically and syntactically 'original' form, while the morphology was lost for other reasons? Additional research on social positions and language attitudes of speakers in Alorese communities may help to shed light on this puzzle.

5. SUMMARY. The past twenty years have seen a variety of data being collected from largely undocumented languages in eastern Indonesia, a hitherto almost unknown area where Austronesian and Papuan languages meet. Such data have an impact on higher level groupings of languages into families and on the typological characterizations of Austronesian and Papuan languages. Typological 'fingerprints' of genealogical groups are used to compare individual languages for the presence of features that might have crossed the border of the family they belong to. Cross-cutting features like these indicate where speakers

[4] Cape Muna in north Pantar is still considered the location of the mythical kingdom Munaseli. The language spoken there is referred to as 'Muna' or *Kadire Senaing* 'speech we understand' (Rodemeier 2006: 49), and it is clearly an Alorese dialect.

of Austronesian and Papuan languages have been in contact with each other, even if no such contact exists today. This information is useful in the study of the macro-level linguistic ecology of the area.

In addition, certain regions within eastern Indonesia have been the focus of more fine-grained research. For example, a significant body of data on individual languages on Pantar and Alor is now available, and when this is combined with oral histories and ethnographic observations, we can reconstruct histories of specific speaker groups on a micro-level. In the case discussed here, we proposed that speakers of Alorese, the only Austronesian language spoken on Pantar and Alor, are relative newcomers to the islands. They arrived on Pantar some 700 years ago from a homeland that is different from the one recounted in their legends. Since then, Alorese has lost all of its morphology, but there are no linguistic traces of intensive contact between Alorese and Papuan speakers.

REFERENCES

Anonymous. 1914. De eilanden Alor en Pantar, Residentie Timor en Onderhoorigheden. *Tijdschrift van het Koninklijk Nederlandsch Aardrijkskundig Genootschap* 31. 70–102.

Barnes, Robert H. 1973. Two terminologies of symmetric prescriptive alliance from Pantar and Alor in Eastern Indonesia. *Sociologus. Zeitschrift für Völkerpsychologie und Soziologie* 23. 71–89.

Barnes, Robert H. 2001. Alliance and warfare in an Eastern Indonesian Principality: Kédang in the last half of the nineteenth century. *Bijdragen tot de Taal-, Land- en Volkenkunde* 157(2). 271–311.

Blust, Robert A. 2009. *The Austronesian languages*. Canberra: Pacific Linguistics.

Bowden, John. 2001. *Taba: Description of a South Halmahera language*. Canberra: Pacific Linguistics.

Donohue, Mark. 2007. Word order in Austronesian from north to south and west to east. *Linguistic Typology* 11(2). 349–391.

Foley, William A. 2000. The languages of New Guinea. *Annual Review of Anthropology* 2. 357–404.

Hammarström, Harald & Sebastian Nordhoff. forthcoming. Achievements and challenges in the description of languages of Melanesia. In Nicholas Evans and Marian Klamer (eds.), Melanesian languages on the edge of Asia. Special issue of Language Documentation and Conservation.

Himmelmann, Nikolaus P. 2005. The Austronesian languages of Asia and Madagascar: Typological characteristics. In Alexander Adelaar & Nikolaus P Himmelmann (eds.), *The Austronesian Languages of Asia and Madagascar*, 110–181. London: Routledge.

Holton, Gary, Marian Klamer, František Kratochvíl, Laura Robinson & Antoinette Schapper. 2012. The historical relation of the Papuan languages of Alor and Pantar. *Oceanic Linguistics* 51(1). 87–122.

Klamer, Marian. 2002. Typical features of Austronesian languages in Central/Eastern Indonesia. *Oceanic Linguistics* 41(2). 363–383.

Klamer, Marian. 2011. *A Short Grammar of Alorese (Austronesian)*. Munich: Lincom Europa.

Klamer, Marian. forthcoming. Papuan-Austronesian language contact: Alorese from an areal perspective. In Nicholas Evans and Marian Klamer (eds.), Melanesian languages on the edge of Asia. Special issue of Language Documentation and Conservation.

Klamer, Marian. n.d. Alorese Toolbox corpus. Leiden University.

Klamer, Marian & Michael Ewing. 2010. The languages of East Nusantara: An introduction. In Michael Ewing & Marian Klamer (eds.), *East Nusantara: Typological and Areal Analyses*, 1–24. Canberra: Pacific Linguistics.

Klamer, Marian, Ger P. Reesink & Mirjam van Staden. 2008. Eastern Indonesia as a linguistic area. In Pieter Muysken (ed.), *From Linguistic Areas to Areal Linguistics*, 95–149. Amsterdam: Benjamins.

Kusters, Wouter. 2003. *Linguistic Complexity: The Influence of Social Change on Verbal Inflection* PhD Thesis. Leiden University. Utrecht: LOT Publications.

Lemoine, Annie. 1969. Histoires de Pantar. *L'Homme: Revue Francaise d'Anthropologie* 9(4). 5–23.

Lenneberg, Eric H. 1967. *Biological Foundations of Language*. New York: Wiley.

McWilliam, Andrew. 2007. Austronesians in linguistic disguise: Fataluku cultural fusion in East Timor. *Journal of Southeast Asian Studies* 38(2). 355–375.

Nishiyama, Kunio & Herman Kelen. 2007. *A grammar of Lamaholot, Eastern Indonesia. The Morphology and Syntax of the Lewoingu Dialect*. Munich: Lincom Europa.

Onvlee, Louis. 1984. *Kamberaas (Oost-Soembaas) – Nederlands Woordenboek*. Dordrecht: Foris.

Pawley, Andrew. 2005. The chequered career of the Trans New Guinea hypothesis. In Andrew Pawley, Robert Attenborough, Jack Golson & Robin Hide (eds.), *Papuan Pasts: Studies in the Cultural, Linguistic and Biological History of the Papuan-Speaking Peoples*, 67–107. Canberra: Pacific Linguistics.

Rodemeier, Susanne. 2006. *"Tutu kadire" in Pandai, Munaseli: Erzählen und Erinnern auf der vergessenen Insel Pantar (Ostindonesien)* [Tutu Kadire – Erzählen und Erinnern lokalgeschichtlicher Mythen am Tanjung Muna in Ostindonesien. PhD-dissertation: University of Leipzig. 2004]. Berlin: Lit Verlag.

Tryon, Darrell T. (ed.). 1995. *Comparative Austronesian Dictionary: An Introduction to Austronesian Studies*. Berlin: Mouton de Gruyter.

Marian Klamer
m.a.f.klamer@hum.leidenuniv.nl

9

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. 64–72
http://nflrc.hawaii.edu/ldc/sp03/
http://hdl.handle.net/10125/4518

Data from language documentations in research on referential hierarchies

Stefan Schnell
Kiel University

This paper outlines potentials of documentary linguistics for typological research in referential hierarchies. Specifically, I will demonstrate how the analysis of original text data from the Oceanic language Vera'a enhances knowledge about referential hierarchy effects in the domains of number marking and morphosyntactic properties of objects. With this language-specific research as a background, I will outline ways in which original text data from language documentation projects can be used in cross-corpus investigations of aspects of referential hierarchies across languages.

1. INTRODUCTION. This paper[1] outlines potentials of language documentation for typological research in referential hierarchies. After a brief summary of typological grammar-based research in referential hierarchies in Section 2, I will show in Section 3 that certain patterns of number marking and object realization in the Oceanic language Vera'a emerge only through the investigation of original, culture-specific text data. Section 4 outlines how this type of corpus-based research may supplement the established typological approach.

2. TRADITIONAL RESEARCH ON REFERENTIAL HIERARCHIES IN LINGUISTIC TYPOLOGY. Traditional typological research in referential hierarchies has focused on the comparison of languages in terms of the structural variation or restrictions within a specific type of construction. Two classic examples are number marking in referential expressions (Smith-Stark 1974) and the differential realization of arguments in the clause, most notably differential case marking of objects (Bossong 1985). In both construction types, a split is observed between a positive and a negative value for the formal variable in question (presence vs. absence of number marking / case marking, respectively). Positive and negative values are associated with elements in different areas on the Referential Hierarchy (i.e. Silverstein's hierarchy; cf. Silverstein 1976), and the construction split is thus mapped onto

[1] The research reported in this paper was generously sponsored by Grant II/81 898 from the Volkswagen Foundation whom I would like to acknowledge hereby. I am grateful to two anonymous reviewers for valuable comments on an earlier version of this paper. Also, I would like to thank the general editor of this issue of LD&C, Frank Seifart, and the convenors of the Analysis panel of the Leipzig Workshop "Potentials of Language Documentation: Methods, Analyses, and Utilization" Leipzig, 3–4 November 2011, Geoffrey Haig, Nikolaus P. Himmelmann, and Anna Margetts, for further comments and suggestions. I am of course responsible for all remaining errors.

a (cluster of) functional domain(s) comprising person, referentiality (which roughly corresponds to "activation" or "accessibility" (cf. Lambrecht 1994)), and animacy. Figure 1 is a reproduction of this hierarchy rendering the well-known patterns of restricted plural marking in five different languages for different types of referential expression (cf. Croft 2003: 130/134), and it shows two types of distributions that are unattested and indeed precluded. The Referential Hierarchy thus represents a model of possible and impossible linguistic structures or languages (cf. Croft 2003). In analogy, patterns of differential object marking can be mapped onto the Referential Hierarchy; however, additional notions like number distinctions and definiteness have been shown to be relevant here.

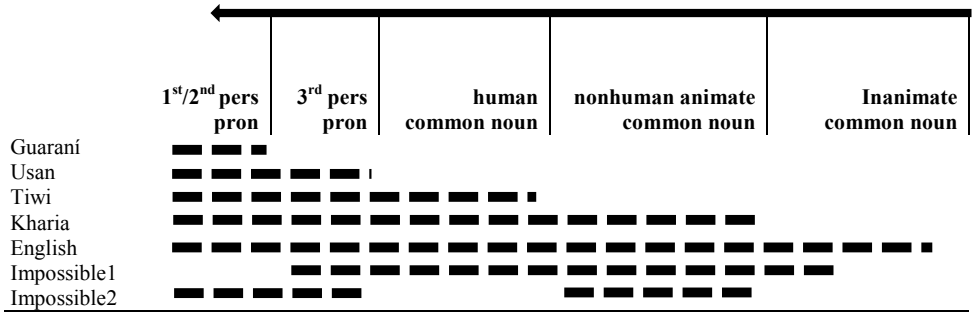


FIGURE 1: Distribution of number distinctions on *Extended Animacy Hierarchy* after Croft (2003: 134)

Crucially, the data bases for this kind of typological work comprise grammatical descriptions or specific studies dedicated to the phenomenon in questions. The latter are often based on focused elicitations of typologically relevant information. The information obtained in this way are interpretations and – to some degree – abstractions of linguistic structures.

Different types of morphosyntactic constructions in individual languages are associated with fixed, rather small feature sets in order to enhance clear distributional descriptions and generalizations. For marking of plurality, only one pair of (usually binary) formal features of a particular element may be considered, i.e. presence vs. absence of plural-marking affix. The values for this variable can then be associated with areas on the Referential Hierarchy (cf. *WALS* Feature #33A, Dryer 2011). Of course, more complex systems may be considered in this way, for instance number systems with more than two values or differential case marking in a P as well as an A function, possibly with more than two possible values (cf. Bickel & Witzlack-Makarevich 2008).

Two problems with this approach remain unresolved in this line of research and can probably only be tackled by use of corpus data: 1. The general neglect of language-internal variation; 2. Treatment of epiphenomenal associations of construction splits with the Referential Hierarchy as connected to factors of discourse structure (cf. Simpson, this volume). Text data has, however, rarely been used for this kind of research (cf. Wälchli 2006), and the purpose of the following sections is to outline how such data can supplement our understanding of animacy and referentiality facts as observable cross-linguistically.

3. INVESTIGATING REFERENTIAL HIERARCHIES IN VERA'A. The Oceanic language Vera'a was documented in a DoBeS project, and the text corpus provided in this documentation served as the main, and almost sole, database for a study in animacy and referentiality effects on the morphosyntax of the language (Schnell 2010)[2]. I will briefly summarize the findings concerning number marking of referential expressions and the differential treatment of P arguments.

3.1. NUMBER MARKING. Possible number distinctions and the means of number marking in Vera'a depend on the type of referential expression and their animacy properties. Pronouns – which most often have human referents (94.9% of pronominal S, A, and P arguments; Schnell cf. 2011b) – show an obligatory 4-way SG-PL-DU-TL/PAUC distinction. Common nouns designating kin relations show an obligatory 3-way SG-PL-DU distinction. Other human nouns designating age- and sex-defined subclasses of humans obligatorily distinguish singular vs. non-singular number and can optionally be marked for dual. With the exception of nouns designating natural forces, all other nouns optionally distinguish singular and plural.

	DISTINCTIONS	MARKING DEVICE
Pronouns	SG-PL-DU-TL/PAUC obligatory	inflection as in Table 2
Kin terms	SG-PL-DU obligatory	+ reduplication pers. DU/PL
'man', 'woman', 'child', … (+hum referent)	SG-NSG oblig; opt. DU	+ reduplication pers. DU/PL
'human being', spirits, animals, inanimate Ns	opt. SG-PL	+/-reduplication PL particle
forces ('hurricane', 'sun', 'fire')	–	–

TABLE 1: Number distinctions and means of marking with different types of nouns in Vera'a

Means of number marking also correlate with referential form and animacy: Personal pronouns are inflected for person and number as shown in Table 2. Kin terms and nouns designating age- and sex-defined subclasses of humans are reduplicated and form a personal NP with *raga* 'people (PL)' or *ruwa* 'two people (DU)' as head noun, as in (1)[3]. All other

[2] Corpus of the Vera'a language compiled by the author is available at http://corpus1.mpi.nl/ds/imdi_browser/?openpath=MPI649371%23
Another text corpus of Vera'a compiled by Alexandre François can be found at http://lacito.vjf.cnrs.fr/archivage/languages/Vera_a_en.htm

[3] Abbreviations: 1, 2, 3 – 1st, 2nd, 3rd person; A – agent-like argument of canonical transitive verb; ART – article; CS – construct suffix; DEM – demonstrative; DISC – discourse particle; DU – dual; INCL – inclusive;

nouns (except for those designating forces) are preceded by the pluralizing particle *'erē* where non-singular number is to be made explicit, as in (2). Example (2) also demonstrates the optionality of plural marking with the noun *'añsara* 'person, human being' and other common nouns (cf. Table 1).

	SINGULAR	DUAL	TRIAL / PAUCAL	PLURAL
1 INCL	–	*gidu(ō)*	*gidō'ōl*	*gidē*
1 EXCL	*no*	*kamadu(ō)*	*kamam'ōl*	*kamam*
2	*nik(ē)*	*kumru(ō)*	*kimi'ōl*	*kimi*
3	*di(ē)*	*duru(ō)*	*dir'ōl*	*dir(ē)*

TABLE 2: Vera'a free personal pronouns

(1) [1.PALA.009]

e	***ruwa***	***re-reñe***	*anē*	*duru*	*=m*	*da'ō*	*duruō*
ART	two.people	RED-woman	DEM	3DU	=TAM	care.for	3DU

'The two girls, they (DU) [the parents] looked after them [the girls].'

(2) [HHAK.002]

di	*ga*	*kurkur*	***ēn***	***'erē***	***'añsara***	*di*	*ga*	*kur*	***ēn***
3SG	TAM	RED:devour	ART	PL	person	3SG	TAM	devour	ART

'añsara	*delñe*	*=n*	*Vunu Lava*
person	around	=ART	place.name

'He ate the people, he ate the people around Vanua Lava.'

Referentiality and animacy are both relevant for the variable expression of number, with (almost always human) pronouns making all available distinctions, followed by kin terms and other human nouns. Lesser number distinctions are made with non-human and inanimate nouns. With high-ranking common nouns, more complex means of number marking are employed, and these are obligatory; number marking of lower-ranking expressions is less complex and optional. Furthermore, preliminary observations suggest that the occurrence of the optional pluralizing particle *'erē* depends on certain referential properties of the noun quantified and seems to be more likely with human nouns than with other animate or inanimate ones. Crucially, the likelihood of *'erē* occurring with different nouns under different contextual conditions could hardly be determined on the basis of elicitations or isolated examples, but instead requires quantitative investigations of text data.

INTERJ – interjection; LIG – ligature; LOC – locative; P – patient-like argument of canonical transitive verb;PAUC – paucal; PL – plural; POSS.FOOD – classifier food possession; POSS.VES – classifier vessel possession; RED – reduplication; S – single argument of canonical intransitive verb, SG – singular; TAM – tense aspect mood; TL – trial

3.2. P ARGUMENTS. The realization of P arguments is another area of Vera'a morphosyntax where referential hierarchies are relevant. A more comprehensive treatment of P realization is provided in (Schnell 2011a: 34 ff.) and Schnell (Forthcoming), and I will confine myself here to non-lexical topical P arguments. Such topical P arguments are either realized as pronouns within the verb complex ('Pro'), or left implicit ('zero'). The former choice is largely restricted to human, while the latter is preferred with non-human discourse participants, as shown in examples (3) and (4):

(3) Pronominal P argument, human referent (in 4c) [JJQ.120–123]

a. *'ō* *ko-n* *e* *$_i$Qo* *=m* *sal* *[...]*
INTERJ POSS.VES-CS ART Qo =TAM float

b. *ei* *'aluwō* *$_k$dē* *=k* *da* *mē* *$_i$diē*
INTERJ tomorrow 1PL.INCL =TAM do DAT 3SG

c. *$_k$dē* *=k* *van* *'ō* *$_i$di* *mē* *=n* *sisidiñ*
1PL.INCL =TAM go carry 3SG DAT =ART bird.catching
'Oh, Qo's canoe is floating. [...] Hey, tomorrow we will do [the following] to him: we will go with him catching birds.'

(4) Zero P argument, inanimate referent [ISAM.005–006]

a. *$_i$[dir]^A* ***=ēk*** ***bigbig*** *$_j$[ēn* *gorē* *=n* *vovoñodo]^P*
3PL =TAM RED:eat ART POSS.food-3PL =ART RED:fish

b. *$_i$[dir]^A* ***=ēk*** ***mul*** ***'ō*** ***kal$_j$Ø^P*** *lē* *=n* *vono-re*
3PL =TAM go carry up LOC =ART home-3PL

c. *$_i$[dir]^A* ***=ēk*** ***big*** *$_j$Ø^P*
3PL =TAM eat
'Then they eat their catch, the take (it) up [the shore] to their village and have (it).'

The correlation between referential form and animacy features of P arguments is, however, merely a soft constraint which is reflected in a strong tendency and may be violated to some degree. Table 3 gives the combined scores of these correlations for three narrative texts (Texts IDs: ISAM, JJQ, PALA)[4]. Hence, contrary to the general tendency, human Ps may occur as zeros (cf. (5)) and non-human Ps as pronouns (cf. (6))[5]:

(5) Zero P argument, human referent [JJQ.200]

dir *=m* *bol* *ēn* *gunu-m* *dir* *man* *row* *'ō'*
3PL =TAM steal ART spouse-2SG 3PL TAM flee carry
'They stole your wife and fled with (her).'

[4] Recorded texts with annotation in ELAN are available in the Vera'a language corpus (under narrative texts) at http://corpus1.mpi.nl/ds/imdi_browser/?openpath=MPI649371%23

[5] **Bold face font** marks those constituents that are analyzed as VC-internal constituents.

(6) Pronominal P argument, inanimate referent [JJQ.165]
dir ***=m*** ***var*** *ēn* *'ekē* *anē* *dir* ***=k*** ***var*** ***diē*** *di*
3PL =TAM stump ART place DEM 3PL =TAM stump 3SG 3SG
ne ***ōn*** ***'abilin***
TAM lie askew
'They stumped this place, and as they stumped it, it lay askew.'

	PRO	% OF PRO	ZERO	% OF ZERO	TOTALS	% OF
+hum	67	97.1%	6	9.5%	**73**	**28.9%**
% OF +HUM	94.5%		5.5%		**100.00%**	
-HUM	2	2.9%	57	90.5%	**59**	**71.1%**
% OF -HUM	3.4%		96.6%		**100.00%**	
TOTALS	**69**	**100.00%**	**63**	**100.00%**	**132**	**100.00%**

TABLE 3: Humanness and referential form of topical P arguments in Vera'a[6]

The observation that human P arguments are preferably granted pronominal realization while non-human participants are left implicit can only be verified once a sufficiently large amount of text data is investigated. Isolated examples alone, like the ones cited above, would only be suggestive but never decisive (cf. Stoll & Bickel, this volume, for a nuanced statistical treatment of variation in referentiality in Chintang). A further point worth mentioning here is the preliminary observation that this pattern looks slightly different in non-narrative texts where non-human participants with a P function appear to be more readily pronominalized. Hence, the pattern observed for narrative texts may well be an artifact of this particular text type, and 'discourse topicality' may be the real issue. Future investigation of these text data will show whether this observation is borne out.

4. CROSS-CORPUS RESEARCH IN REFERENTIAL HIERARCHIES – THE GRAID INITIATIVE. Given that investigations of original text data in individual languages may contribute enormously to our understanding of referentiality and animacy, the investigation of such text data *across* languages seems to be the most obvious and most urgent thing to do; all the more as large-scale language documentation projects around the world have produced unprecedentedly large amounts of original text data that are easily accessible for linguists.

There appear to be two main obstacles preventing linguists from directly comparing original texts across languages in order to scrutinize the effects of animacy and referentiality across languages (cf. Wälchli 2006: 1). The need for annotating corpora for the relevant features and the enormous workload involved therein (cf. Schultze-Berndt 2006: 2). The need for text data to be minimally comparable. In order to overcome the problem of comparability, researchers have used either parallel texts, i.e. translational equivalents like the *Declaration of Human Rights* or (parts of) the *Bible* (Wälchli 2009, cf. Cysouw & Wälchli

[6] Spreadsheets containing the scores cited here available at http://vc.uni-bamberg.de/moodle/course/view.php?id=9488

2007 for an overview of available parallel texts), or 'content-equivalent' texts elicited with the help of stimuli like the Pear Film (Chafe 1980) or the *Frog Story* picture book (Mayer 1994[1969]). As for parallel texts, although these have been proven a useful database in research within, for instance, lexical typology (cf. Wälchli 2009, 2006), they may not be suitable for research on referentiality and animacy. This is at least suggested by observations from *Bible* translations: As (Mosel & Hovdhaugen 1992: 10f.) show, 'Biblical' Samoan texts differ from texts of indigenous registers in that the former show an unnaturally high proportion of pronominal reference, while the latter prefer zero anaphora. *Pear Stories* have been shown to be a suitable database for research in referential density (Bickel 2003, Stoll & Bickel 2009). *Frog Stories*, on the other hand, have been shown to feature an unnaturally high referential density and do not seem to be amenable for cross-corpus research in referential hierarchies (Foley 2003).

Despite the obvious advantages of and need for controlled data, the comprehensive text corpora compiled in language documentation projects comprise data of the highest quality in terms of authenticity and cultural embeddedness. The GRAID initiative (Haig & Schnell 2011, Haig et al. 2011) touches on this potential by applying a cross-linguistically applicable and easily practicable set of glossing conventions to texts from language documentation projects. GRAID glosses register the referential form, animacy features, and grammatical function of (mainly core) arguments. Hence, once texts are coded in this way, questions like the one concerning the pronominality of P arguments can be tackle d in an immediate and detailed manner, yielding exact figures about correlations between animacy, referential form, and syntactic function. In this way, texts from different languages can be analyzed quantitatively and – at least to some extent – compared in terms of referentiality and animacy. Haig et al. (2011) demonstrate that the original text data they use for their study of pronominal reference shows a surprisingly high degree of uniformity, suggesting that the lack of control for content may actually be of lesser relevance than would be expected. Thus, while Cysouw & Wälchli (2007: 98) state that traditional typology is fruitfully supplemented by parallel-text typology, the study of original texts in linguistic typology may likewise be a worthwhile enterprise (cf. Wälchli 2006). The GRAID initiative opens up such opportunities in the area of referential hierarchy research.

REFERENCES

Bickel, Balthasar. 2003. Referential density in discourse and syntactic typology. *Language* 79(4). 708–736.

Bickel, Balthasar & Alena Witzlack-Makarevich. 2008. Referential scales and case alignment: reviewing the typological evidence. In Marc Richards & Andrej L. Malchukov (eds.), *Scales. Linguistische Arbeitsberichte. ARBEITS BERICHTE 86*, 1–37. Leipzig: Universität Leipzig.

Bossong, Georg. 1985. *Empirische Universalienforschung: differenzielle Objektmarkierung in den neuiranischen Sprachen*. Tübingen: Narr.

Chafe, Wallace L. 1980. The deployment of consciousness in the production of a narrative. In Wallace L. Chafe (ed.), *The Pear Stories: Cognitive, Cultural and Linguistic Aspects of Narrative Production*, 9–50. Norwood, NJ: Ablex.

Croft, William. 2003. *Typology and Universals. 2nd edn*. Oxford: Oxford University Press.

Cysouw, Michael & Bernhard Wälchli. 2007. Parallel texts: Using translational equivalents in linguistic typology. *Sprachtypologie und Universalienforschung – STUF* 60(2). 95–99.

Dryer, Matthew S. 2011. Coding of nominal plurality, feature 33A. In Matthew S. Dryer & Martin Haspelmath (eds.), *The World Atlas of Language Structures Online*, Munich: Max Planck Digital Library. http://wals.info/feature/33A (03 March, 2012).

Foley, William A. 2003. Genre, register and language documentation in literate and preliterate communities. In Peter K. Austin (ed.), *Language Documentation and Description 1*, 85–98. London: School of Oriental and African Studies.

Haig, Geoffrey & Stefan Schnell. 2011. Annotations using GRAID (Grammatical Relations and Animacy in Discourse). Introduction and guidelines for annotators. Version 6.0. http://vc.uni-bamberg.de/moodle/course/view.php?id=9488.

Haig, Geoffrey, Stefan Schnell & Claudia Wegener. 2011. Comparing corpora from endangered language projects: Explorations in typology with original texts. In Geoffrey Haig, Nicole Nau, Stefan Schnell & Claudia Wegener (eds.), *Documenting Endangered Languages: Achievements and Perspectives*, 55–86. Berlin: Mouton de Gruyter.

Lambrecht, Knud. 1994. *Information structure and sentence form. Topic, focus and the mental representations of discourse referents*. Cambridge: Cambridge University Press.

Mayer, Mercer. 1994[1969]. *Frog, where are you?* New York: Dial Books for Young Readers.

Mosel, Ulrike. 1982. The influence of the church missions on the development of Tolai. In Rainer Carle, Martina Heinschke, Peter Pink, Christel Rost & Karen Stadtlander (eds.), *Gava': Studies in Austronesian Languages and Cultures. Dedicated to Hans Kähler*, 155–172. Berlin: Reimer.

Mosel, Ulrike & Even Hovdhaugen. 1992. *Samoan Reference Grammar*. Oslo: Scandinavian University Press.

Schnell, Stefan. 2010. *Animacy and referentiality in Vera'a*: Kiel University dissertation.

Schnell, Stefan. 2011a. A grammar of Vera'a, an Oceanic language of North Vanuatu. PhD thesis. Kiel University.

Schnell, Stefan. 2011b. Pronominal reference in Vera'a narrative discourse. Paper presented at the International Workshop on Vanuatu Languages. 20–23 October 2011, A.N.U., Kioloa Coastal Campus.

Schnell, Stefan. Forthcoming. Referential hierarchies in three-participant constructions in Vera'a. In: Eva van Lier (ed.), Referential hierarchies in three-participant constructions. Special issue of Linguistic Discovery, in memory of Anna Siewierska.

Schultze-Berndt, Eva. 2006. Linguistic annotation. In Jost Gippert, Nikolaus P. Himmelmann & Ulrike Mosel (eds.), *Essentials of Language Documentation*, 213–251. Berlin: Mouton de Gruyter.

Silverstein, Michael. 1976. Hierarchy of features and ergativity. In R.M.W. Dixon (ed.), *Grammatical Categories in Australian Languages*, 112–171. Canberra: Australian Institute of Aboriginal Studies.

Simpson, Jane. this volume. Information structure, variation and the Referential Hierarchy.

Smith-Stark, Thomas Cedric. 1974. The plurality split. *Chicago Linguistic Society* 10. 657–661.

Stoll, Sabine & Balthasar Bickel. 2009. How deep are differences in referential density? In Jiansheng Guo, Elena Lieven, Nancy Budwig, Susan Ervin-Tripp, Keiko Nakamura & Seyda Özçaliskan (eds.), *Crosslinguistic Approaches to the Psychology of Language: Research in the Tradition of Dan Isaac Slobin*, 543–555. London: Psychology Press.

Stoll, Sabine & Balthasar Bickel. this volume. How to measure frequency? Different ways of counting ergatives in Chintang (Tibeto-Burman, Nepal) and their implications.

Wälchli, Bernhard. 2006. Descriptive typology, or, the typologist's expanded toolkit. Unpublished ms. http://ling.uni-konstanz.de/pages/home/a20_11/waelchli/waelchli-desctyp.pdf (21 March, 2012).

Wälchli, Bernhard. 2009. *Motion Events in Parallel Texts: A Study in Primary Data Typology*. Unpublished Habilitationsschrift, University of Bern.

Witzlack-Makarevich, Alena. 2011. *Typological variation in grammatical relations*. Leipzig: University of Leipzig dissertation. http://www.uni-leipzig.de/~witzlack/Witzlack2010Typological.pdf (21 March, 2012).

Stefan Schnell
s.schnell@latrobe.edu.au

10

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. 73–82
http://nflrc.hawaii.edu/ldc/sp03/
http://hdl.handle.net/10125/4519

Information structure, variation and the Referential Hierarchy

Jane Simpson
Australian National University, Canberra

Silverstein (1976)'s hierarchy of features and ergativity (Referential Hierarchy) was proposed to capture apparent systematic variation with respect to word-class (pronouns versus nouns) in the expression of the grammatical functions Subject and Object and the semantic roles Agent and Undergoer linked to these functions. An assumption of the original hierarchy was obligatoriness of marking, rather than optionality (i.e. choice of marker or its absence). Optionality is often associated with a semantic/pragmatic force additional to straight expression of grammatical function. This additional meaning may determine reanalysis and subsequent change in the morphosyntactic expression of Subject/Object/Agent/Undergoer. Along the way, apparent counter-examples to the Referential Hierarchy may be created. To understand the counter-examples, and test the descriptive adequacy of the Referential Hierarchy, better language documentation is needed.

1. INTRODUCTION[1]. In the 1970s Michael Silverstein made a proposal about the linguistic means of expressing arguments and the accessibility of arguments for other processes. This was the hierarchy of features and ergativity (later called the Referential Hierarchy) (Silverstein 1976). It engendered considerable typological work as linguists sought to test the claims cross-linguistically, both within languages which they were researching and across languages. Grammars were the main source of publicly available data on many languages, although a few authors, especially those influenced by Franz Boas's approach (cf. Voegelin & Hymes 1953), also published text collections and dictionaries. The language data was primarily available as transcribed texts without accompanying audio and had to be searched by hand rather than digitally. Thus testability of the predictions of the hierarchies was limited to what could be found by hand in the available data. Modern language documentation has expanded the number of languages that can be searched, as well as the range of language documentation (more texts and dictionaries and linked multimedia) and the ways in which language documentation can be searched (different search possibilities along with greater reliability).

[1] I thank for their comments the participants in the workshop on *Potentials of Language Documentation* (November 3–4, 2011), Max Planck Institute for Evolutionary Anthropology, Leipzig, Frank Seifart and Geoffrey Haig, and an anonymous referee.

In this paper, I illustrate the need for electronically searchable text corpora linked to sound with respect to understanding how changes in the case-marking of pronouns in the Australian language Arrernte could create a counter-example to the Referential Hierarchy.

2. BACKGROUND: THE REFERENTIAL HIERARCHY. Silverstein (1976)'s hierarchy of features and ergativity (Referential Hierarchy) was proposed to capture apparent systematic variation in the morphosyntactic expression of the grammatical functions Subject and Object and the linked semantic roles Agent and Patient with respect to word-class (pronouns versus nouns).

1/2 pron > 3 pron > proper N >	human N >	animate N > other
Agent Subject		Undergoer Object [2]

It was intended to capture two aspects of the linguistic expression of Subject and Object arguments relating to case and verb agreement respectively. On Case, Silverstein proposed that the higher an element is on the hierarchy, the less likely it is to have Ergative case if it is acting as an Agent Subject, and conversely that the lower an element is on the hierarchy, the less likely it is to have Accusative case if it is acting as an Undergoer Object. In terms of verb agreement, Silverstein proposed that the lower an element is on the hierarchy, the more likely there is to be special verb agreement such as inverse marking if it is acting as an Agent Subject and if the Undergoer Object is higher on the hierarchy. Thus it is less marked for first person to act on third person than vice versa, and the more marked situation may be expressed by use of special inversion markers or reversal of agreement marker order.

Silverstein observed that the Australian language, Arrernte was a counter-example to his proposal that first person would be less likely to be marked Ergative than second or third person. Table 1 gives the modern Arrernte pronouns, showing that nouns and first person singular follow an Ergative-Absolutive pattern, while plural pronouns and second and third person singular pronouns follow a Nominative-Accusative pattern. The odd behavior of first person singular pronouns is striking.

Since the Referential Hierarchy reflects strong statistical tendencies, a counter-example such as this needs discussion. One type of explanation lies in their historical origins – what were the pressures in the system that allowed them to emerge?

At some time there must have been a bridging context, that is, where two variants of a form arose in slightly different meaning contexts (semantic or pragmatic) to express the same idea (or part of the same idea) in different meaning-contexts. One form-variant+meaning pair followed the hierarchy in expressing grammatical relationships; the other was used in a context where the grammatical relations were downplayed (e.g. a case form was used to express a non-case grammatical concept, such as emphasis). At a later stage, the form-variant+meaning pair that expressed the grammatical relation stopped being

[2] Abbreviations and conventions: 1, 2, 3 – 1st, 2nd, 3rd person; A – Subject of transitive verb; ABS – Absolutive; ACC – Accusative; ALY – Alyawarr; ARR – Arrernte; C – underspecified consonant; CV – sequence of consonant and vowel; ERG – Ergative; IS – Subject of intransitive verb; N – noun; NOM – Nominative; NPST – non-past tense; O – Object; PST – Past Tense; PREA – pre-Arandic; PRS –present tense; PRON – pronoun; S – Subject; SG – singular; V – verb, underspecified vowel; '-' – clear morpheme boundary; '=' – clitic boundary; '.' – frozen morpheme boundary; '*' – reconstructed form.

ARRERNTE	AGENT SUBJECT	INTRANSITIVE V SUBJECT	UNDERGOER OBJECT
1 SG	*(a)the*	*ayenge*	*ayenge*
2 SG	*unte, nge*	*unte, nge*	*nge-nhe*
3 SG	*re*	*re*	*re-nhe*
Plural Pronouns	A/IS	A/IS	O*nhe*
Nouns	Ergative affix	unmarked (Absolutive)	unmarked (Absolutive)

TABLE 1: Modern Arrernte singular pronouns

used, leaving the other form-variant+meaning pair either on its own, or to expand to cover both contexts. This then created a violation of the hierarchy.

In order to see how statistically unlikely systems arise, we need to be able to observe the variation at an earlier stage of the language, or, failing that, in a genetically and typologically similar language. The main sources for such observations are grammars and text corpora.

Grammar writers, especially those writing grammars of small languages of which they are not native speakers, naturally focus on the statistically most prevalent forms, and not on variants. As well, the descriptive capability of grammar-writers increases as we learn more about the expressive capabilities of languages. Topics such as information structure are treated in much more detail in modern grammars than in grammars even of the 1980s. So, the grammarian may not have made the needed observations of variation. Therefore, it is essential to have large balanced text corpora in order to observe the variation, so as to determine what the bridging contexts are.

3. BACKGROUND: INFORMATION STRUCTURE. To understand the situation that gives rise to counter-examples to case-expression such as the Arrernte example, it is helpful to look at counter-examples to verb agreement, the other means of expression of argument relation that Silverstein linked to the hierarchy. These have been explored typologically (Bickel 2008, Filimonova 2005), and in general such counter-examples seem to be more numerous than the counter-examples to case-expression. Bickel observes that the Referential Hierarchy may work differently for case-expression and verb agreement because "these two ways of marking arguments have a fundamentally different relationship to referential distinctions".

Bickel's point is important. A hierarchy which has pronouns at one end and nouns at the other is a hierarchy of information structure, of how the Speaker organizes information about reference for the Hearer. **Pronouns** assume that the Hearer knows who is being referred to and are likely to be used for pragmatic functions such as continuing topic: *John came in and **he** sat down*. They may also be used for contrast: *It was **him** that did it, not me*, but very often languages have special means for marking these (as strong pronouns as opposed to the weak pronouns of continuing topic). **Proper names** are allegedly unique identifiers, but their use also assumes that the Hearer knows in general who is being referred to (*John* is a proper name, but in the sentence *John came in and **he** sat down*, it is assumed that the Hearer knows which *John* is being referred to). **Nouns** are more likely to be used for

new information, switch of referent for subject or object. Thus the order on the Referential Hierarchy reflects an ordering of information structure roles from those that the Speaker expects the Hearer to recognise, to those that that the Speaker expects to be new to the Hearer.

'Verb agreement' relates to this hierarchy of information structure because in many languages 'verb agreement' is not just agreement with a co-present noun; it is **the** linguistic means of expression for referents. Consider the following two examples from the Central Australian language Warlpiri:

(1) Warlpiri [V agreement only]
Wangka-mi *ka=**rna***
talk-NPST PRS=1S
'I am talking.'

(2) Warlpiri [V agreement + pronoun]
Ngaju *ka=**rna*** *wangka-mi*
1S PRS=1S talk-NPST
'**I** am talking. / It's me that's talking.'

The 'agreement marker' can express continuing topic on its own as well as agreeing with another element. In this respect, verb agreement lines up with 'weak' pronouns as opposed to 'strong pronouns'. Therefore verb agreement may actually **be** an element on the hierarchy, rather than something the choice of which expresses the hierarchy – like case. That is, the elements of the Referential Hierarchy are actually part of the information structure resources of a language.

The information structure resources of a language involve many linguistic means of expression other than part of speech and morphological properties as represented in the Referential Hierarchy. A Hearer's attention may be directed to a referent via other morphemes, e.g. clitics and demonstratives, or by word order, by prosodic means such as pauses and intonation changes, and by non-verbal means such as gesture and eye-gaze. It follows then that to understand putative counter-examples to the proposed typological universal of the Referential Hierarchy, we need to know how information structure as a whole works in the language.

Labelling information structure functions is hard to do precisely and reliably. To see what the information flow is, and to understand why something is considered old or new information or contrastive, we need continuous text, not just example sentences. To understand the linguistic means of expression of information structure, we need a corpus which covers conversational speech and dialogue talk as well as the monologues, narratives, descriptive texts, and procedural texts that make up most of the traditional text collections accompanying grammars. A corpus which contains transcriptions of text and intonation, and is linked to audio-visual files, is essential for providing empirical support for claims about information structure. Claims about the linguistic expression of hanging topics, contrastive topics, afterthoughts, and emphasis need to be backed up by reference to examples from such a corpus, and by quantifying the use of particular means of expression. Corpora produced

before the software became available for linking audio-visual files to text usually lack the fine-grained representations of intonation so necessary for understanding how information structure is expressed.

So, in sum, elements of the Referential Hierarchy (parts of speech/morphological properties) are part of the means speakers have for expressing information structure. How they operate within a language must be seen in the light of all means of structuring information in the language. For example, intonation often provides cues to information structure, and so audio-visual information is needed, not just transcripts.

4. ARRERNTE AS A COUNTER-EXAMPLE. With this as background, we return to the putative counter-example from Arrernte. When finding something that at first glance seems to run counter to a fairly well established typological universal such as the Referential Hierarchy, the first step is to establish how this counter-example fits into the patterns found in the neighboring and genetically close languages. Arrernte is a member of the Arandic language family. Examination of the neighboring, and closely genetically related, language, Alyawarr, reveals that it differs in having a three-way split for both first and second person singular (Wilkins 1989). This split is reconstructed for pre-Arandic as well (Koch 2004).

		AGENT SUBJECT	INTRANSITIVE V SUBJECT	UNDERGOER OBJECT
1 SG	ARR	*(a)the*	*aye.nge*	*aye.nge*
	ALY	*athe*	*aye-nge*	*aye-nhe*
	***PREA**	***ngathu**	***ngay(V)-nge**	***ngay(V)-hna**
2 SG	ARR	*unte, nge*	*unte, nge*	*nge-nhe*
	ALY	*ntwe*	*nge*	*nge-nhe*
	***PREA**	***nyuntu**	***nyun.-nge**	***i.nge-nhe**
3 SG	ARR	*re*	*re*	*re-nhe*
	ALY	*re*	*re*	*re-nhe*
	***PREA**	***CV.re**	***CV.re**	***CV.re-nhe**[3]

TABLE 2: Modern Arrernte and Alyawarr compared with Koch's reconstructions for an earlier stage of Arandic

In Arrernte, Alyawarr, and pre-Arandic, *-nhe* (**-nha*) represents an Accusative case-marker. In Alyawarr and pre-Arandic there is some evidence for a morpheme *-nge* on intransitive subjects, but it appears to be frozen in Arrernte.

From the point of view of the Referential Hierarchy, the pre-Arandic Ergative-Absolutive-Accusative of the Speaker-Hearer singular pronouns and the Ergative-Absolutive of the nominals would be not unexpected (and see Haig 2008 for a similar pattern in Vafsi). What is unexpected is the proposed break in the hierarchy through the third person Nominative-Accusative pattern, which retains the marked Accusative object. Since the evidence for the absence of Ergative-marking on the pre-Arandic reconstructed pronoun is basically absence of evidence for its presence, we leave this and focus on the changes in the modern systems.

[3] CV stands for a reconstruction of an indeterminate C followed by an indeterminate V.

In sum, the three-way pre-Arandic system for Speaker/Hearer singular pronouns has been retained in Alyawarr but in Arrernte has moved to two different two-way systems: a two-term Ergative-Absolutive for first person singular (creating an unmarked Object), and, for second person singular, a three-term Nominative-Accusative system with two different terms for the Nominative.

The change appears to have taken place in two stages. The collapse of the second person appears to have been relatively recent (Wilkins 1989), who cites evidence from an early recorder, R. H. Mathews:

> "There are two district [sic] forms of the first person in the singular number, namely, *ta* and *yinga*. *Ta* is always used when connected with a transitive verb... *Yinga* is employed when connected with an intransitive verb... In the second person singular there are also two forms of the pronuon [sic] – *unta* for use with transitive verbs, and *nga* with intransitive verbs." (Mathews 1907: 325)

That is, in the late nineteenth century, when Mathews carried out work on Arrernte, both first and second person singular had Ergative forms. So how did a three-way pre-Arandic system change to the modern system with the discontinuity between first person singular (ERG-ABS), other pronouns (NOM-ACC), and nouns (ERG-ABS)? There must be bridging contexts which allow the change from one meaning to another.

5. BRIDGING CONTEXT: OPTIONAL ERGATIVE MARKING. I propose that the likely bridging context for the second person change is optional Ergative marking. When the hierarchy was first proposed, little attention was paid to the possible optionality of marking for case. The situations discussed were those of obligatory marking. But it has become clear that in many Australian languages, Ergative case marking is optional (McGregor & Verstraete 2010). Verstraete shows that in Umpithamu Ergative marking is obligatory for inanimate transitive subjects, but optional for animate transitive subjects (as predicted by the Referential Hierarchy), and that contrast and answers to questions are likely to favour the use of Ergative. Thus, Ergative case-marking has several interrelated associations: indicating the grammatical function of subject, indicating the semantic roles of agent, and identifying referents that the hearer needs to be alerted to, most likely as new information.

At any time a speech community may highlight one or other of these associations. If the grammatical function of Subject is highlighted, then Ergative may be obligatory for marking the subject of transitive clauses. If the semantic role is highlighted, then there may be splits in how the subjects of transitive clauses are marked (e.g. using Absolutive for the subjects of perception verbs). If discourse function is highlighted, then subjects of transitive verbs may have Ergative marking in certain non-prominent discourse contexts, or lack it in positions that are otherwise discourse- prominent (because Ergative marking would be redundant).

An association between optional Ergative marker and discourse prominence is found in the neighboring language Warlpiri (which is related to Arrernte, but not closely). In traditional Warlpiri spoken in Yuendumu, Ergative marking is optional on first and second person singular pronouns in initial position. Elsewhere (demonstratives, nouns), Ergative marking appears to be obligatory (Mary Laughren, p.c.).

(3) Warlpiri[4]

a. ***Ngaju**=rna* *paka-rnu.*
IABS=1S hitPST

b. *Paka-rnu=rna* ***ngaju.lu-rlu***.
hit-PST=1S I-ERG

c. ***Ngaju.lu-rlu=rna*** *paka-rnu.*
I-ERG=1S hit-PST

d. *??Paka-rnu=rna* ***ngaju***.
hit-PST=1S I(ABS)
'I hit (it).'

Initial position in Warlpiri is a position where discourse prominent elements are placed (Laughren 2002, Legate 2002, Simpson 2007). This association of lack of Ergative marking with a discourse prominent position has several possible explanations. One is that the word order is changing so that Subject (whether Agent or Intransitive V Subject) is becoming fixed in initial position, and so only Subjects in non-initial position need Ergative marking to identify their function. But this needs testing to find out what the discourse contexts are that favor *Ngaju V* over *Ngajulurlu V* and *V ngajulurlu*. Unfortunately this is hard to do. The Warlpiri corpus (which was transcribed or composed between 1959 and the early 1990s, and has been the basis for the Warlpiri Dictionary) is large by Australian standards, containing at least 600,000 words. But it has certain gaps. There is limited meta-data on speakers. There is as yet almost no interlinear glossing and no linking to sound. The genres include elicitation, mythological and personal narratives, meta-linguistic discussion, and written texts, but no natural conversation.

Recent changes in Warlpiri further support the association of position and Ergative case. In Lajamanu, a Warlpiri community to the north of Yuendumu, a new mixed language, Light Warlpiri, has developed, based on the interaction of Warlpiri and Kriol (O'Shannessy 2006). This language has Ergative marking, but it has become generally more optional than in traditional Warlpiri. Using spontaneous texts, and comprehension and production tasks, O'SShannessy has shown that Ergative marking occurs 60% or so of the times when it could be expected on nouns in traditional Warlpiri, where it is categorical. Ergative marking is more likely when subjects of transitive clauses appear post-verbally (the language is moving to SVO order), and on unusual Agents (i.e. when the hierarchy is inverted, that is, when Subjects of transitive verbs are inanimate and Objects are animate).

The fact that Ergative marking appears post-verbally fits with its optional absence in traditional Warlpiri on initial first person pronouns. There is less need to mark elements in initial position as being discourse prominent because the initial position itself indicates discourse prominence. By contrast, post-verbal position is less prominent, and so a post-verbal expression which represents an Agent is more likely to receive special marking to indicate that it is expressing an Agent. Thus in Light Warlpiri, Ergative marking has shifted from being mostly a marker of grammatical function to being a marker of discourse prominence or of inverted expectations (cf. Umpithamu (Verstraete 2010), where inanimates must receive Ergative case, but animates are optionally marked with Ergative).

[4] *ngaju* 'Absolutive'; *ngajulu-rlu* 'Ergative', where *-rlu* is a common allomorph of Ergative

The end state (=current state) of Light Warlpiri is well understood from O'Shannessy's description (although she notes the difficulty of operationalizing pragmatic factors such as contrast and continuing topic for testing their relevance to the choice of Ergative case marking in comprehension and production tests). However, the transitional states and the reasons for change cannot be tested properly since we cannot easily find answers without time-consuming hand-coding of the existing corpus.

Having established that, in a neighboring language, Ergative case-marking may change to being associated with discourse prominence when this is not marked by other means, such as word order, we return to Arrernte. Wilkins writes:

> "The two forms [*unte* and *nge*], are basically equivalent in meaning, although, of the two second singular pronouns, *nge* may be considered the unstressed and less emphatic form." (Wilkins 1989: 125)

Thus, in his view, the second singular form *unte* (the old Ergative) is more associated with discourse prominence (like a strong pronoun, and like the use of Ergative in Light Warlpiri), while the second singular form *nge* (the old Absolutive) is more associated with continuing topic (like a weak pronoun).

We can posit the following sequence of changes:

- **Pre-Arandic**: Ergative only used to mark grammatical functions: three-way split in first and second person with *-nhe* as Accusative on all singular pronouns.
- **Pre-modern Arrernte** [Mparntwe Arrernte 1907]
  - 1/2 singular pronouns *athe/unte* ERG, *aye.nge/nge* ABS, (Mathews does not give the forms for Object)
- **Mparntwe Arrernte** 1989
  - 1 singular pronoun *athe* ERG, *ayenge* ABS
  - 2 singular pronoun: 2 pronouns representing NOMINATIVE: *unte* Discourse-prominent, *nge* Continuing topic
- Proposed direction:
  - Pre-Arandic → pre-Modern Arrernte
    Accusative form of 1st person is lost, possibly by conflation of the Accusative marker *-nhe*, with the augment *-nge* of the form used for Subject of Intransitive verb. This resulted in an Ergative-Absolutive system for 1st person singular (cf. Koch 2004).
  - pre-Modern Arrernte → Modern Arrernte
    Ergative form of 2nd person singular *unte* moves from marking grammatical function to marking discourse function of emphasis. The difference between *unte* (Subject of transitive verb) and *nge* (Subject of intransitive verb) is reanalyzed as a difference between *unte* as discourse prominent Subject and *nge* as continuing topic Subject, regardless of whether the verb is transitive or intransitive. This results in a

Nominative-Absolutive system for 2nd person singular parallel to that of 3rd person singular.

The explanation of the changes in second person requires finding out more about the discourse function of Ergative-marked elements. For its time, Wilkins' grammar is rich in including 12 complex interlinear texts (approximately 360 clauses) covering instruction, morality fables, mythological story, description of natural phenomena, and co-constructed written texts. But even despite the variety of genres, it is hard to find examples. The second person *nge* does not occur at all in his texts, and the second person *unte* only occurs twice in an instructional text ("If you have a cold then you inhale…") and four times in polite commands presented as reported speech in a story. This is not enough to test patterns.

6. CONCLUSION. Putative counter-examples to proposed typological universals can often be understood in part by considering paths of grammatical change. But to understand and analyze how languages change from one typological state to another, we need more than just grammars, even rich ones such as Wilkins' grammar, because they inevitably reflect the descriptive preoccupation of the time. Information structure is an area of the language whose importance in grammatical description only became widely accepted in the 1990s. Grammars written before that time frequently lack a discussion of information structure; for example, they may tell the reader how to ask a question (because that involves special pronouns or word order or clitics) – but not how to answer it (and so they lack examples of the clearest means of expressing new information).

We need, in addition, large corpora of texts linked to audio-visual recordings to observe the variation and determine the bridging contexts that lead to the emergence of counter-examples. Ideally, these would cover a wide range of genres and range of speakers, be annotated for pragmatic factors, and would be refreshed over time.

REFERENCES

Bickel, Balthasar. 2008. On the scope of the referential hierarchy in the typology of grammatical relations. In Greville G. Corbett & Michael Noonan (eds.), *Case and grammatical relations: Studies in honor of Bernard Comrie*, 191–210. Amsterdam: John Benjamins.

Filimonova, Elena. 2005. The Noun Phrase Hierarchy and relational marking: Problems and counterevidence. *Linguistic Typology* 9(1). 77–113.

Haig, Geoffrey. 2008. *Alignment Change in Iranian Languages: A Construction Grammar Approach*. Berlin: Mouton de Gruyter.

Keenan, Edward L. & Bernard Comrie. 1977. Noun Phrase Accessibility and Universal Grammar. *Linguistic Inquiry* 8(1). 63–99.

Koch, Harold. 2004. The Arandic subgroup of Australian languages. In Claire Bowern & Harold Koch (eds.), *Australian Languages: Classification and the Comparative Method*, 127–50. Amsterdam: Benjamins.

Laughren, Mary. 2002. Syntactic constraints in a 'free word order' language. In Mengistu Amberber & Peter Collins (eds.), *Language Universals and Variation*, 83–130. Westport, CT: Praeger.

Legate, Julie Anne. 2002. *Warlpiri: Theoretical implications*. Massachusetts: Massachusetts Institute of Technology dissertation.

Mathews, R. H. 1907. The Arran'da language, Central Australia. *Proceedings of the American Philosophical Society* 46(187). 322–339. www.jstor.org/stable/983471?origin=JSTOR-pdf (12 December, 2011).

McGregor, William B. & Jean-Christophe Verstraete. 2010. Optional ergative marking and its implications for linguistic theory. *Lingua* 120(7). 1607–1609.

O'Shannessy, Carmel Therese. 2006. *Language contact and children's bilingual acquisition: learning a mixed language and Warlpiri in northern Australia*. Sydney: University of Sydney dissertation. http://hdl.handle.net/2123/1303 (21 March, 2012).

Silverstein, Michael. 1976. Hierarchy of features and ergativity. In Robert M.W. Dixon (ed.), *Grammatical Categories in Australian Languages*, 112–171. Canberra: Australian Institute of Aboriginal Studies.

Simpson, Jane. 2007. Expressing pragmatic constraints on word order in Warlpiri. In Annie Zaenen, Jane Simpson, Tracy Holloway King, Jane Grimshaw, Joan Maling & Chris Manning (eds.), *Architectures, Rules, and Preferences: Variations on Themes by Joan W. Bresnan*, 403–427. Stanford CA: CSLI.

Verstraete, Jean-Christophe. 2010. Animacy and information structure in the system of ergative marking in Umpithamu. *Lingua* 120(7). 1637–1651.

Voegelin, Charles (Carl) F. & Dell H Hymes. 1953. A sample of North American Indian dictionaries with reference to acculturation. In *Proceedings of the American Philosophical Society*, vol. 97 5, 634–644. http://www.jstor.org/stable/3149275 (6 June, 2012).

Wilkins, David P. 1989. *Mparntwe Arrernte (Aranda): Studies in the structure and semantics of grammar*. Canberra: Australian National University dissertation.

Jane Simpson
jane.simpson@anu.edu.au

11

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. 83–89
http://nflrc.hawaii.edu/ldc/sp03/
http://hdl.handle.net/10125/4520

How to measure frequency? Different ways of counting ergatives in Chintang (Tibeto-Burman, Nepal) and their implications

Sabine Stoll and Balthasar Bickel
University of Zürich

The frequency of linguistic phenomena is standardly measured relative to some structurally defined unit (e.g. per 1,000 words or per clause). Drawing on a case study on the acquisition of ergativity by children in Chintang, an endangered Tibeto-Burman language of Nepal, we propose that from a psycholinguistic point of view, it is sometimes necessary to measure frequencies relative to the length of the time windows within which speakers and hearers use the language, rather than relative to structurally defined units. This approach requires that corpus design control for recording length and that transcripts be systematically linked to timestamps in the audiovisual signal.

1. INTRODUCTION. Both in historical linguistics and language acquisition research, frequency is generally assumed to be one of the most important features influencing language development (e.g. Bybee & Hopper 2001). One of the main assumptions of the usage-based approach is that distributions of patterns, i.e. frequency distributions and repetitions, play a key role in language change and language learning, underlying the gradual emergence of constructions diachronically (e.g. Hopper 1988) and developmentally (e.g. Tomasello 2003).

However, since frequency is a relational measure, any counting is meaningless unless we have a unit over which we can reasonably assume that the relevant items are tracked by speakers and hearers when processing language: we can count phenomena per linguistic unit (words, phrases, clauses etc.), per non-linguistic context and genre, per content unit (the choice of specific topics), or per time unit (in, say, minutes of speech or hours of conversation). It is unclear *a priori* what kind of unit is most useful for a given research question. Although the choice of counting unit has fundamental consequences on the results, this issue has received surprisingly little attention. The issue is particularly pressing, however, when we design and compile relatively small corpora, such as corpora of spoken and endangered languages, because the choice of counting unit predetermines the kinds of factors one needs to consider: to what extent is it important to balance or control for content types, recording time length, number of words, etc., and which of these is important for what research purpose?

In this paper we discuss some of the consequences of choosing among a variety of counting units. We exemplify these issues with a study on the role of frequency in the acquisition of ergative case in Chintang (`ISO639.3:ctn`, Tibeto-Burman/Sino-Tibetan, Eastern Nepal), based on a corpus that we compiled as part of a DoBeS project.[1] A key advantage of the corpus is that it is systematically linked to time stamps in the audiovisual recordings, and this makes it possible to consider not only counting units that are defined in terms of grammar or content but also in terms of time flow.

2. DATA. Chintang is a polysynthetic language spoken in a village in Eastern Nepal by about 6,000 people, who are all bilingual in Nepali, the *lingua franca* of Nepal (e.g. Bickel et al. 2007, 2010, Stoll et al. 2012). The language is endangered, but there is still a substantial number of children who learn the language as their first language. Our study is based on a longitudinal language acquisition corpus of 4 children learning Chintang (all from different families). Two children were aged 2 years and two children aged 3 years at the beginning of the study. The children were recorded over a period of 18 months for about 4 hours per month, while playing in their natural environment (mostly outdoors), with many different interlocutors around, both children and adults. A minimum of one and a half hours of recordings per month were used for the present study. The data were transcribed, translated, morphologically glossed, and tagged for part of speech properties (for more information see http://www.spw.uzh.ch/clrp). Figure 1 shows the amount of data available for the different children and recording sessions.

3. ERGATIVE MARKING IN CHINTANG. Ergative case in Chintang is distributed along a split system conditioned by person. The ergative marker (*-ŋa*) occurs obligatorily only with third person noun phrases. For first and second person the marker is optional, and for first person exclusive it is ungrammatical. Additional complications come from the fact that arguments are very frequently dropped in Chintang discourse (Stoll et al. 2012) and that the same case form *-ŋa* also doubles as an instrumental and an ablative marker (Bickel et al. 2010). As a result, ergative case does not seem to have a very high cue validity in Bates & MacWhinney's sense, even in third person contexts. This would make the acquisition of ergative case particularly challenging and difficult to account for.

But the question arises whether this impression of low cue validity is in fact empirically justified. In order to examine this, we need to chart the actual distributions in the speech of native adult speakers. In the following we analyze the adult speech surrounding our target children in the corpus.

4. MEASURING FREQUENCY. As noted in the introduction, the key issue in exploring frequencies is the choice of unit over which we count frequency. Usually there is more than one option. Each option leads to very different results, but more importantly, each option

[1] The data are available in the DoBeS archive, http://corpus1.mpi.nl. We use a snapshot of the corpus from October 2010, with a total size of ca. 280,000 words. Development of the corpus was made possible by a DoBeS grant (PI Balthasar Bickel) and a Dilthey fellowship to Sabine Stoll, both from the Volkswagen Foundation. Our research is embedded in the Chintang Language Research Program (http://www.spw.uzh.ch/clrp), and we are grateful to our colleagues in the program, especially Sebastian Sauppe, Taras Zakharko, and Robert Schikowski for help in preparation of the corpus for the present study. All corpus analyses were performed in R (R Development Core Team 2011) and visualized using the package `lattice` (Sarkar 2008).

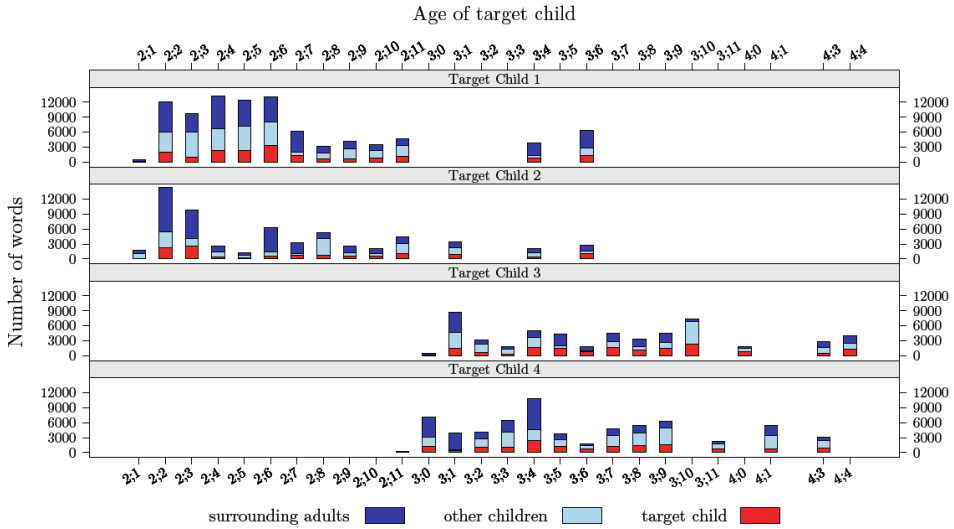


FIGURE 1: Distribution of data in number of words.

also makes strong but implicit assumptions about language processing and memory, both when learning a language for the first time in acquisition and when replacing one variant with another in language change. While this is not the place to review the psychological literature on these assumptions, we present a case study in the following that explores the general kinds of assumptions and overall results that are tied to four specific ways of measuring frequencies. We take as an example the acquisition of ergative case in Chintang.

4.1. RAW NUMBERS PER AGE IN MONTHS. A first relational option is the use of raw numbers per age, e.g. per month of age. This measure is rarely chosen because it is probably not very useful in most contexts without knowing what these numbers relate to. It obviously makes a huge difference if we find 5 instances in a corpus of 1,000 words or in a corpus of 10,000 words. Thus, the relational component is crucial for evaluating the numbers, and it should be explicitly stated. This is so in the options for counting that we consider in the following.

4.2. ERGATIVES PER WORD. Another option counts how often per word unit the ergative would occur, i.e. the proportion of words with an ergative marking. This would give us an impression of how often a child hears such a marker independently of its syntax or semantics. Results are shown in Figure 2.

If we used this measure, ergative marking would indeed appear to be exceedingly rare, never exceeding about .03%. However, to the extent that we would not want to assume that children parse language completely without any semantic or structural analysis, this measure might not be very revealing. Further, counting simply ergatives per word ignores the fact that ergatives can only occur in certain syntactic contexts: they are limited to noun phrases functioning as transitive agent ('A') arguments of transitive verbs. This brings up another relational type of counting ergatives.

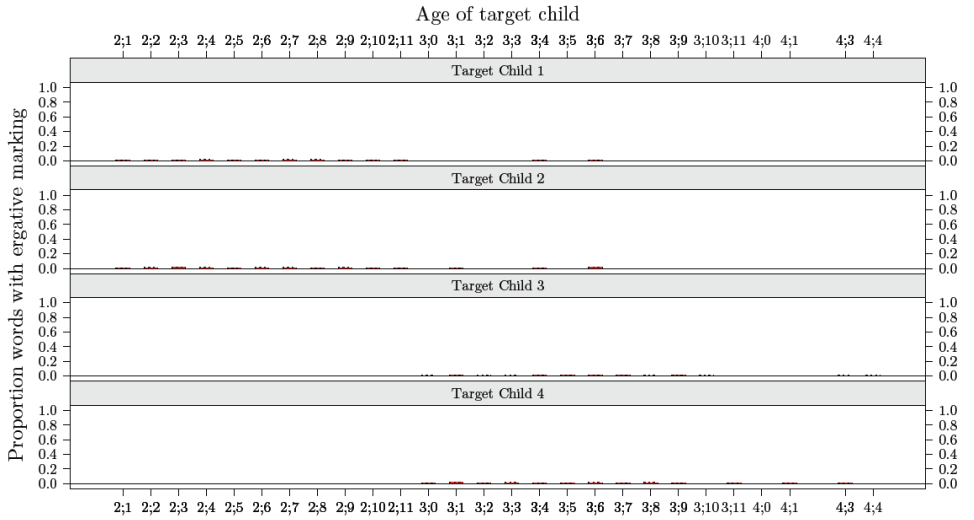


FIGURE 2: Proportion of words with ergative marking in adult speech surrounding the target children in the study

4.3. ERGATIVES PER TRANSITIVE VERB. Figure 3 illustrates the proportions of transitive verbs with an ergative marker per child and age. We exclude from this the occurrence of the same marker in an instrumental or ablative function, i.e. we limit our attention to the transitive A role.[2] On this count too, ergatives seem to be rare, although with a maximum of 11%, not as rare as when counting ergatives per word (.03%, Figure 2). At any rate, this would still be in line with the expectations derived from purely structural considerations.

However, counting ergatives per transitive verbs begs a number of questions: why should we choose all transitive verbs, rather than only verbs that are actually used with transitive syntax (cf. Note 2)? If we choose all transitive verbs, should we include A arguments across all persons, or should we limit our attention to third persons since it is only here that ergatives are compulsory? Regardless of what answer we give, it will invariably make the psychologically very strong assumption that the child has abstract knowledge over all these features of grammar (such as lexical vs. syntactic transitivity, or person categories), i.e. that the child parses the input on the basis of a fairly fine-grained distributional analysis. It is not at all clear, however, whether such an assumption is indeed warranted. Similar issues arise when considering the psychological bases on which speakers, regardless of their age, engage in language change: when new forms are innovated and especially when (as is often the case) forms are extended to new contexts, it is unclear whether and to what extent speakers make consistent distributional assumptions about the context from which the innovation starts.

[2] Transitive verbs can also be used in detransitivized constructions, where the A argument receives nominative case. For present purposes we gloss over these different uses and only consider the bare opportunity for ergative case marking which is associated with every transitive verb. For further discussion see Bickel et al. (2010), Schikowski et al. (2010).

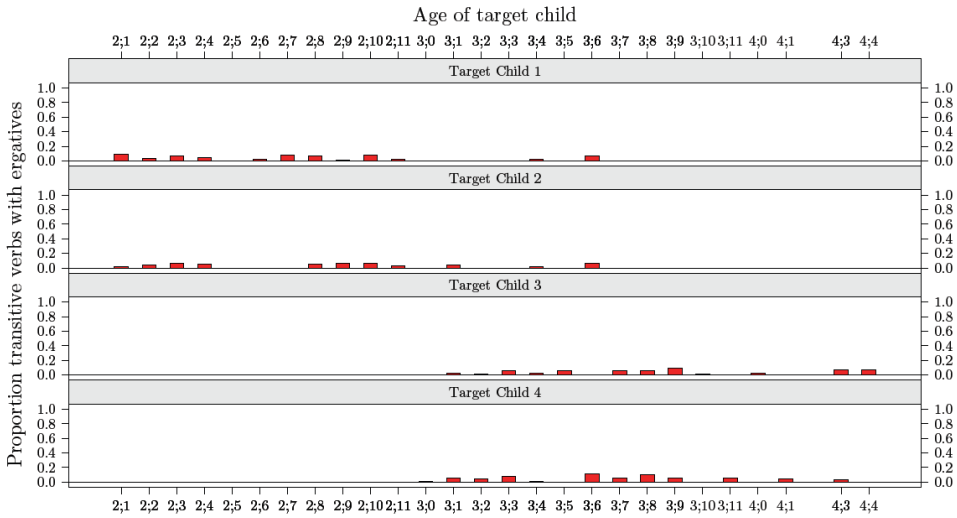


FIGURE 3: Proportion of transitive verbs combining with an A argument marked by ergative case in adult speech surrounding the target children in the study

An additional difficulty in Chintang concerns the fact that agents are often named in isolation, with the verb dropped (e.g. because it was mentioned in a previous conversational turn). These cases are excluded when counting ergatives per verb, but ergatives in isolation might provide key contexts that help children learn their use.

4.4. ERGATIVES PER TIME UNIT. Under this approach we consider the density in which ergatives are offered to children (or hearers more generally). Density of occurrence is arguably a psychologically important unit since it directly relates to well-known memory demands on processing and learning. Figure 4 shows the counts of ergatives per hour of speech. This includes all ergatives, regardless of their context.

In stark contrast to all previous frequency counts, counting ergatives per hour, i.e. in terms of the density of occurrence, suggests that the number of cases that a child hears is not so small after all. Children hear the ergative on average every two minutes (30 occurrences per hour), sometimes even every minute. To the extent that density of occurrence is psychologically relevant, this relatively high density would seem to facilitate the learning process considerably.

5. CONCLUSIONS. The present study suggests a distinction between two types of frequency measures. One measure relies on the frequency of X relative to the structural opportunity for X. This is the standard in corpus linguistics and also in usage-based theory. However, the psychological relevance of this type of frequency measure is unclear because it relies on very strong assumptions about the extent to which the 'opportunity for X' is in fact known and taken into account by hearers when learning a language or when being involved in language change.

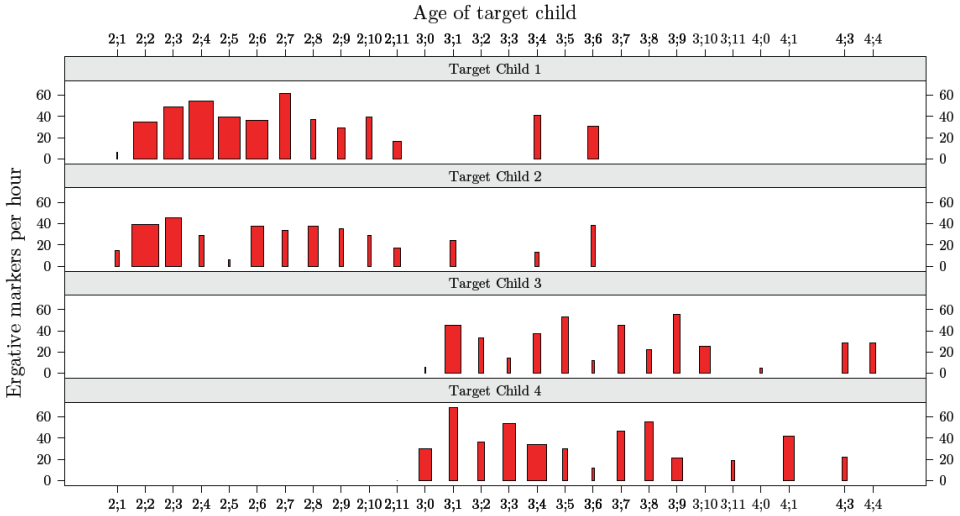


FIGURE 4: Proportion of ergatives per hour in adult speech surrounding the target children in the study. (Bar width is proportional to corpus size in number of words.)

An alternative measure relies on the frequency of X within a given time window and aims at estimating the density of occurrence of X. This measure directly relates to the demands on memory and processing that are relevant for language learners. This measure makes minimal assumptions about the level of analysis that a hearer uses, and at the same time, it gives an impression of how often a hearer is confronted with the feature in question.

For such a measure to be applicable, corpora need to control not only for genres, register, contexts, etc. (as emphasized by Lüdeling, this volume), but also for recording length. For this to be possible, transcripts need to be systematically linked to timestamps in the audiovisual signal.

REFERENCES

Bates, Elizabeth & Brian MacWhinney. 1982. Functionalist approaches to grammar. In Eric Wanner & Lila R. Gleitman (eds.), *Language acquisition: the state of the art*, 173–218. Cambridge: Cambridge University Press.

Bickel, Balthasar, Goma Banjade, Martin Gaenszle, Elena Lieven, Netra Paudyal, Ichchha P. Rai, Manoj Rai, Novel K. Rai & Sabine Stoll. 2007. Free prefix ordering in Chintang. *Language* 83. 43 – 73.

Bickel, Balthasar, Manoj Rai, Netra Paudyal, Goma Banjade, Toya Nath Bhatta, Martin Gaenszle, Elena Lieven, Iccha Purna Rai, Novel K. Rai & Sabine Stoll. 2010. The syntax of three-argument verbs in Chintang and Belhare (Southeastern Kiranti). In Andrej Malchukov, Martin Haspelmath & Bernard Comrie (eds.), *Studies in ditransitive constructions: a comparative handbook*, 382–408. Berlin: Mouton de Gruyter.

Bybee, Joan L. & Paul J. Hopper. 2001. Introduction. In Joan L. Bybee & Paul J. Hopper (eds.), *Frequency and the emergence of linguistic structure*, 1–24. Amsterdam: Benjamins.

Hopper, Paul. 1988. Emergent grammar and the a priori grammar postulate. In Deborah Tannen (ed.), *Linguistics in context*, 117–134. Norwood, NJ: Ablex.

Lüdeling, Anke. this volume. A corpus linguistics perspective on language documentation, data, and the challenge of small corpora.

R Development Core Team. 2011. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. http://r-project.org.

Sarkar, Depayan. 2008. *Lattice: Multivariate data visualization with R*. Berlin: Springer.

Schikowski, Robert, Netra P. Paudyal & Balthasar Bickel. 2010. Fluid transitivity in Chintang. Paper presented at the workshop on Valency Classes, MPI for Evolutionary Anthropology, Leipzig, 21 August 2010 [http://www.spw.uzh.ch/schikowski/work/2010-fluid-transitivity.pdf] (21 March, 2012).

Stoll, Sabine, Balthasar Bickel, Elena Lieven, Goma Banjade, Toya Nath Bhatta, Martin Gaenszle, Netra P. Paudyal, Judith Pettigrew, Ichchha P. Rai, Manoj Rai & Novel Kishore Rai. 2012. Nouns and verbs in Chintang: children's usage and surrounding adult speech. *Journal of Child Language* 39. 284 – 321.

Tomasello, Michael. 2003. *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Cambridge, Mass.: Harvard University Press.

Sabine Stoll
sabine.stoll@uzh.ch

Balthasar Bickel
balthasar.bickel@uzh.ch

12

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. 90–95
http://nflrc.hawaii.edu/ldc/sp03/
http://hdl.handle.net/10125/4521

On the sociolinguistic typology of linguistic complexity loss

Peter Trudgill
Agder University

The nature of the human language faculty is the same the world over, and has been so ever since humans became human. This paper, however, considers the possibility that, because of the influence which social structure can have on language structure, this common faculty may produce structurally different types of language under different sociolinguistic conditions. Changing sociolinguistic conditions in the modern world are likely to have the consequence that, in time, the only languages remaining in the world will be severely atypical of how languages have been throughout most of human history.

1. SOCIOLINGUISTIC TYPOLOGY. Workers involved with language documentation are only too well aware that a very large proportion of the world's languages are likely to be lost to us before very long. There is one aspect of this language loss, however, which is not often mentioned and which makes the task of documentation and description even more urgent. In this paper I discuss this particular aspect of language loss from the perspective of *sociolinguistic typology*. By sociolinguistic typology, I mean a form of linguistic typology which is sociolinguistically informed, asks sociolinguistic questions, and tries to supply sociolinguistic answers. The assumption behind sociolinguistic typology is that the nature of the human language faculty is the same the world over, and has been so ever since humans became human. But because of the influence social structure can have on language structure, I have long argued that this common faculty of the human mind may produce different types of language in different places and at different moments in human history (see, for example, Trudgill 1992).

A further assumption made in this paper is that the "equicomplexity hypothesis" has no validity (Shosted 2006). This hypothesis – that all languages are equally complex – has for many decades been rejected implicitly – or even explicitly (e.g. Trudgill 1983) – by sociolinguists, who have long accepted that language contact can produce simplification – as in the development of creoles, and creoloids such as Afrikaans, through language contact; and of koinés through dialect mixture. Obviously, languages are less complex after simplification than before. And if a language can be more or less complex at different stages of its history, then clearly some languages can be more complex than others. This point is more explicitly tackled in a number of papers in Sampson et al. (2009).

A sociolinguistic-typological approach to linguistic complexity (Trudgill 2011) then leads us to ask: Why are some languages more complex than others? Are sociolinguistic answers to this question available? And what are the socio-structural conditions necessary for the development of linguistic complexity?

To answer this question we need first to discuss what complexity might be. My brief answer (see further Trudgill 2011) is – and I acknowledge that there are a number of other possible answers – that complexification consists of factors such as: increase in irregularity, increase in morphological opacity, increase in syntagmatic redundancy, and increase in morphological categories. These are all factors which make for *L2* [second language] *difficulty* – linguistic phenomena which are difficult for post-critical-period adult learners to acquire.

2. MATURE PHENOMENA AND SOCIETIES OF INTIMATES. According to Östen Dahl, "there is a significant overlap between ... those linguistic features that are most recalcitrant in second language learning" and "mature features" (Dahl 2004: 286). Much L2 difficulty, that is, has to do with *mature phenomena*. According to Dahl, mature phenomena are linguistic features which imply a lengthy period of historical development – they "presuppose a non-trivial prehistory" (2004: 2). Dahl mentions, for example, syntactic agreement as "belonging to the later stages of maturation processes" (2004: 197). Other examples of features which take very many generations to come into being include, amongst others, the development of inflectional morphology and fusional languages generally; grammatical gender systems; and the large-scale grammatical marking of evidentiality. (On the L2 difficulty of inflectional morphology, see Dahl 2004: 286, and Meisel 1997.)

The socio-structural conditions necessary for development of complexity will thus be those which favor the growth of mature phenomena. And these conditions will then rather clearly be those which produce long – generally centuries or even millennia – periods of uninterrupted linguistic change. Note, however, that such conditions simply provide a background against which complexity *can* develop. They provide matrices "which permit but do not necessarily and inexorably produce complexity. We cannot say that complexification will inevitably occur under such conditions – languages spoken in low-contact communities will certainly be found which demonstrate no great complexity" (Trudgill 2011: 89).

The types of interruption most likely to interfere with the development of mature phenomena will be those which lead to the simplification which is brought about by certain types of adult language learning (in certain types of sociolinguistic situation Trudgill see 2011: Chapter 2), due to factors such as social instability and linguistic contact. We can deduce that social factors which favor complexity thus include *high social stability* and *low contact*, where "contact" must be specifically interpreted to mean certain forms of contact involving post-critical-period – i.e. adult and adolescent as opposed to child-language-learning.

Linguistic complexity also develops most readily – although, once again, not inevitably – in *societies of intimates* (see Trudgill 2011 for detailed argumentation; and see also Wray & Grace 2007, Wohlgemuth 2010). According to Givón & Young (2002), hunter-gatherers belong to "societies of intimates" – societies "where all generic information is shared" (Givón 1979: 297); and which contrasts with "societies of strangers", the larger and more complex human groups which began to develop around 10,000 BC and which most of us inhabit today. For nearly all of human history, we lived in societies characterized not only by

stability, and low contact, but also by *small size*, *dense social networks*, and *informational homogeneity = large amounts of shared information*. Examples of linguistic complexity which linguists have specifically associated with societies of intimates include evidentials – "complex evidential systems, in their vast majority, are confined to languages with smallish numbers of speakers, spoken in small, traditional societies" (Aikhenvald 2004) – and the remarkable generationally-marked pronoun system of Onya Darat which, as described by Tadmor (forthcoming), cannot work except in a community where everybody knows everybody else.

I therefore suggest that the major complexity-producing social factors include: small size, dense social networks, large amounts of shared information, high stability, and low contact. The relevance of these societal features stems from the fact that, linguistically, complexification at the morphological and morpho-syntactic levels arises as a result of linguistic processes such as fusion, reanalysis, and refunctionalization, plus a complex of processes leading to grammaticalization – and with an important role at many points for phonology – which are all processes requiring considerable periods of time in order to develop undisturbed and go to completion.

My sociolinguistic-typological point of view is that in large, high-contact, unstable communities with loose social networks, such lengthy periods are less likely to be available. And not only are mature phenomena less likely to develop, they are also very vulnerable to being lost through simplification if high-contact situations develop: mature phenomena "are highly prone to being filtered out in suboptimal language acquisition" (Dahl 2004: 286).

3. THE PRESENT IS NOT LIKE THE PAST. One of the fundamental bases of modern historical linguistics has been the *uniformitarian principle* (Labov 1972). Knowledge of processes that operated in the past can be inferred by observing ongoing processes in the present. This leads to the methodological principle of *using the present to explain the past*: we cannot try to explain past changes in language by resorting to explanations that would not work for modern linguistic systems.

But there is one very important respect in which the present is not like the past at all. Human language came into existence perhaps something like 100,000 years ago (Corballis 1999, Dixon 1997) or even considerably earlier (Evans 2010). The earliest date for a post-neolithic society anywhere in the world is about 5,000-6,000 years ago, in the Middle East (Langer 1987), and later, sometimes very much later, everywhere else. This means that human languages were spoken in neolithic and pre-neolithic societies for at least 95% of their history (cf. Wichmann forthcoming). It is therefore probable that widespread adult-only language contact is a mainly post-neolithic and indeed a mainly modern phenomenon, associated with the last 2,000 years, during which time the world's human population has grown enormously from about 200 million to 7,000 million. Given that the development of large, fluid societies of strangers is also a post-neolithic and indeed mainly modern phenomenon, then a sociolinguistic-typological perspective suggests that the dominant standard modern languages in the world today are likely to be seriously atypical of how languages have been for nearly all of human history.

I recall once asking an erudite generative linguist how he would handle switch-reference in his current theoretical model, and he replied that he did not know, explaining that switch-reference was "something you only get in exotic languages" which he did not know anything

about. I would argue however that in fact these "exotic languages" are precisely what we must make sure that we do know about. These languages, with their mature phenomena, are actually, especially from a diachronic perspective, not exotic at all. They are normal. This is what human languages must have been like throughout most of the tens of thousands of years of human history on this planet. It is the creoloids and koinés and creoles that have developed in the last two thousand years, and particularly in the last 500 years, that are weird and unrepresentative.

Robert Orr (1999) pointed out that Bob Dixon's (1972) ground-breaking study of the Australian language Dyirbal, which "revolutionised our view of ergativity" (Comrie 1978: 393), was a close-run thing: Dyirbal was a dead language before the next decade was out. As Schmalstieg (1980: 18) says, if it had not been for the timing of Dixon's fieldwork, "it seems quite possible that the Dyirbal population might have disappeared without a trace, and notions of ergativity would have remained unhampered by new facts".

How many other "new facts" about human language are we going to miss out on? A very large proportion of the world's isolated languages and dialects, and of languages and dialects spoken in small tightly-knit communities, may not be with us much longer. It would not be at all surprising if in a few generations' time there were no languages at all in the world with any of their typical social characteristics, and therefore, I would argue, with any of their typical linguistic characteristics. The human language faculty can be assumed to have remained unchanged for very many millennia indeed; but the sociolinguistic-typological matrices in which linguistic change occurs have changed significantly. It is therefore not totally unreasonable to suppose that, in the future, we are increasingly unlikely ever again to see the development of, say, highly inflectional fusional language varieties. And I submit that complex features such as large evidential systems, quadral number, switch-reference systems, polysynthesis, and generationally marked pronouns as described by Tadmor, are unlikely to be created afresh in the societies of strangers that predominate in modern conditions. These are mature linguistic phenomena that will no longer be given time to develop. And where they have already developed, they are also now relatively likely to be lost, because of the current increase in adult high-contact situations, because of increases in community size, and because of language death. In other words, they might eventually disappear from the world's languages altogether.

4. CONCLUSION. If therefore we want to learn more about the inherent nature of linguistic systems, we have to urgently focus most of our attention on linguistic structures and linguistic changes of the types that occur in the ever-dwindling number of low-contact, dense social network varieties of language in the modern world. It is of course highly relevant that, happily, current forms of language documentation aim to document not only particular languages, but also the sociolinguistic characteristics of the speech communities in which they are spoken. But if we want to build up an accurate picture of the nature of human languages throughout human history, we have to hurry, not only because, as workers in the field of language documentation are only too well aware, most of the world's languages are in danger, but also because *most of the languages that will be left behind will increasingly tend to be of a single, historically atypical type*.

REFERENCES

Aikhenvald, Alexandra Y. 2004. *Evidentiality*. Oxford: Oxford University Press.

Comrie, Bernard. 1978. Ergativity. In Winfred P. Lehmann (ed.), *Syntactic Typology: Studies in the Phenomenology of Language*, 329–394. Austin: University of Texas Press.

Corballis, Michael C. 1999. The gestural origins of language: Human language may have evolved from manual gestures, which survive today as a "behavioral fossil" coupled to speech. *American Scientist* 87(2). 138–146.

Dahl, Östen. 2004. *The Growth and Maintenance of Linguistic Complexity*. Amsterdam: Benjamins.

Dixon, R.M.W. 1972. *The Dyirbal Language of North Queensland*. Cambridge: Cambridge University Press.

Dixon, R.M.W. 1997. *The rise and fall of languages*. Cambridge: Cambridge University Press.

Evans, Nicholas. 2010. *Dying Words: Endangered Languages and What They Have To Tell Us*. Oxford: Wiley-Blackwell.

Givón, Talmy. 1979. *On Understanding Grammar*. New York: Academic Press.

Givón, Talmy & Phil Young. 2002. Cooperation and interpersonal manipulation in the society of intimates. In Masayoshi Shibatani (ed.), *The Grammar of Causation and Interpersonal Manipulation*, 23–56. Amsterdam: John Benjamins.

Labov, William. 1972. *Sociolinguistic Patterns*. Philadephia: University of Pennsylvania Press.

Langer, William L. 1987. *An Encyclopedia of World History*. London: Harrap / Boston: Houghton Mifflin Company.

Meisel, Jürgen M. 1997. The acquisition of the syntax of negation in French and German: Contrasting first and second language development. *Second Language Research* 13(3). 227–263.

Orr, Robert A. 1999. Evolutionary biology and historical linguistics. *Diachronica* 16(1). 123–159.

Sampson, Geoffrey, David Gil & Peter Trudgill (eds.). 2009. *Language Complexity as an Evolving Variable*. Oxford: Oxford University Press.

Schmalstieg, William. 1980. *Indo-European Linguistics: A New Synthesis*. University Park: Pennsylvania State University Press.

Shosted, Ryan. 2006. Correlating complexity: A typological approach. *Linguistic Typology* 10(1). 1–40.

Tadmor, Uri. forthcoming. The grammaticalization of generational relations in Onya Darat. In Randy LaPolla (ed.) The Shaping of Language: Relationships Between the Structures of Languages and Their Social, Cultural, Historical, and Natural Environments.

Trudgill, Peter. 1983. *On Dialect: Social and Geographical Perspectives*. Oxford: Blackwell.

Trudgill, Peter. 1992. Dialect typology and social structure. In Ernst Håkon Jahr (ed.), *Language Contact: Theoretical and Empirical Studies*, 195–212. Berlin: Mouton de Gruyter.

Trudgill, Peter. 2011. *Sociolinguistic Typology: Social Determinants of Linguistic Complexity*. Oxford: Oxford University Press.

Wichmann, Søren. forthcoming. Neolithic linguistics. In Gojko Barjamovic, Irene Elmerot, Adam Hyllested, Benedicte Nielsen, and Bjørn Okholm Skaarup (eds.). Language and Prehistory of the Indo-European peoples – A Cross-Disciplinary Perspective. Budapest: Archaeolingua.

Wohlgemuth, Jan. 2010. Language endangerment, community size and typological rarity. In Jan Wohlgemuth & Michael Cysouw (eds.), *Rethinking Universals: How Rarities Affect Linguistic Theory*, 255–277. Berlin: De Gruyter.

Wray, Alison & George W Grace. 2007. The consequences of talking to strangers: Evolutionary corollaries of socio-cultural influences on linguistic form. *Lingua* 117(3). 543–578.

Peter Trudgill
peter.trudgill@unifr.ch

13

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. 96–104
http://nflrc.hawaii.edu/ldc/sp03/
http://hdl.handle.net/10125/4522

Visualization and online presentation of linguistic data

Hans-Jörg Bibiko

Max Planck Institute for Evolutionary Anthropology, Leipzig

This contribution gives an introduction to state-of-the-art techniques for the visualization and online presentation of linguistic data and world-wide linguistic diversity, such as linguistic maps and online dictionaries, using a software environment called R. The aim is to draw linguists' attention to the possibilities offered by these techniques and to give some practical hints as to how they can be used specifically for linguistic and language documentation data.

1. R AS A TOOL FOR CREATING THE VISUALIZATION AND ONLINE PRESENTATION OF LINGUISTIC DATA.

Visualization by way of diagrams, charts, animations, maps, etc. and their online presentation is an important means of enhancing the usability of linguistic data for various scientific and educational purposes and of presenting linguistic facts to the general public. Nowadays these visualizations can be created relatively easily with R, an open-source scripting-language based software environment for doing statistics and generating high-quality graphics (see www.R-project.org) . Since R is open source, a huge amount of so-called packages, i.e. libraries of functions for a particular purpose created by the world-wide community of R users, are available that extend the functionality of R enormously. A particularly important point in this context is that with R it is possible to generate so-called vector PDF graphics; these have various advantages over pixel images and can be edited and used in publications or presentations. In the following sections, I give examples of how R can be used to visualize complex data sets and their geographic distribution (Section 2), how custom maps can be generated (Section 3), and how structured linguistic data such as wordlists can be transformed into user-friendly resources such as online thematic dictionaries (Section 4).

2. GENERATING MAPS WITH R.

2.1. DISPLAYING WORDLISTS BY USING MAPS.

An initial example of the application of R is illustrated by the geographic distribution of the word for "three" in about 3,000 languages. The data used in this example are from Eugene Chan's compilation of "Numeral Systems of the World's Languages" (http://lingweb.eva.mpg.de/numeral). A static wordlist from 3,000 languages is difficult to work with. Mapping this as online information helps

one to visualize relationships, e.g. between cognate terms or areally diffused terms. Clicking the link in the caption of Figure 1 will display a movie illustrating such a visualization. The HTML/KML file that underlies this visualization was generated by means of R, since R is also a powerful scripting language, making it unnecessary for the user to have to acquire knowledge in classic scripting languages such as Python and Perl, etc.

FIGURE 1: The geographic distribution of words for "three" in Northwest Mexico [video: http://youtu.be/4bb0y7lWsqQ, http://scholarspace.manoa.hawaii.edu/bitstream/handle/10125/4522/13-wordlist.m4v]

2.2. DISPLAYING STRUCTURAL FEATURES ON MAPS. It is interesting to study the geographic distribution not only of lexical information, as in the previous example, but also of structural linguistic information, as in the World Atlas of Language Structures (WALS) (Dryer & Haspelmath 2011). Maps similar to those of WALS can be created using R, as in Figure 2; this map shows the distribution of the WALS feature "Order of Adposition and Noun Phrase" for languages spoken in China and Mongolia.

2.3. DISPLAYING VARIOUS FEATURES AND VALUES IN PIE CHARTS. In comparative and quantitative linguistics, the assignment of only one value for a language feature is often difficult. Therefore it is sometimes more meaningful to give information on the degree to which various values are true for a given feature. R can be used to create pie charts displaying this information, adding another dimension to the representation of the

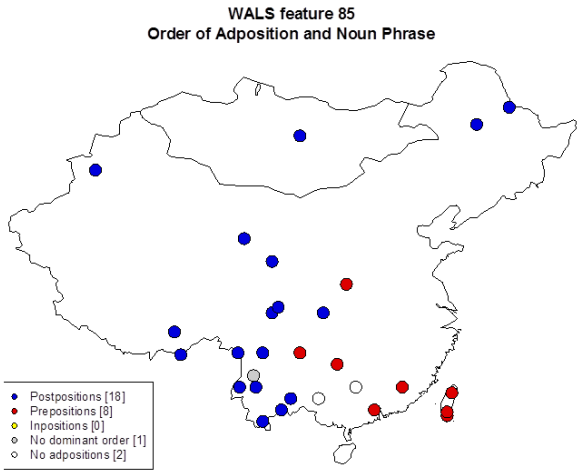


FIGURE 2: Map showing "Order of Adposition and Noun Phrase" for languages spoken in China and Mongolia

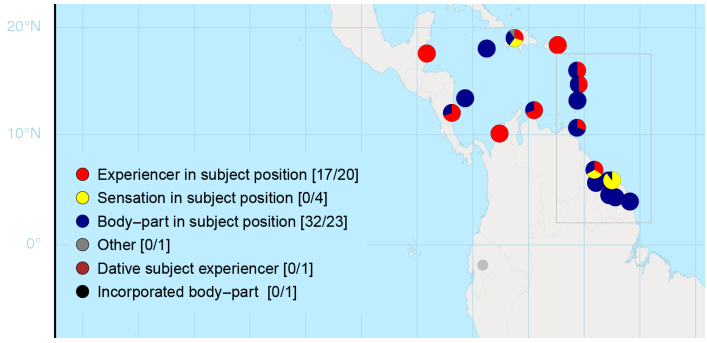


FIGURE 3: Values for "Experiencer constructions: 'headache' " from APiCS

geographic distribution of linguistic features. For example, Figure 3 shows the values for a feature called "Experiencer constructions: 'headache' " from *The Atlas of Pidgin and Creole Language Structures* (APiCS) (Michaelis et al. forthcoming) in various languages of the Caribbean area.

Figure 4, taken from "The electronic World Atlas of Varieties of English" (eWAVE) (Kortmann & Lunkenheimer 2011), shows how different features can be represented in pie charts. The color of the bottom right section of the pie charts represents the values of the feature "She/her used for inanimate referents", the top section of the feature "Generalized third person singular pronoun: object pronouns", and the bottom left section of the feature "Me instead of I in coordinate subjects". Such a representation allows one to visually inspect if and where such features correlate.

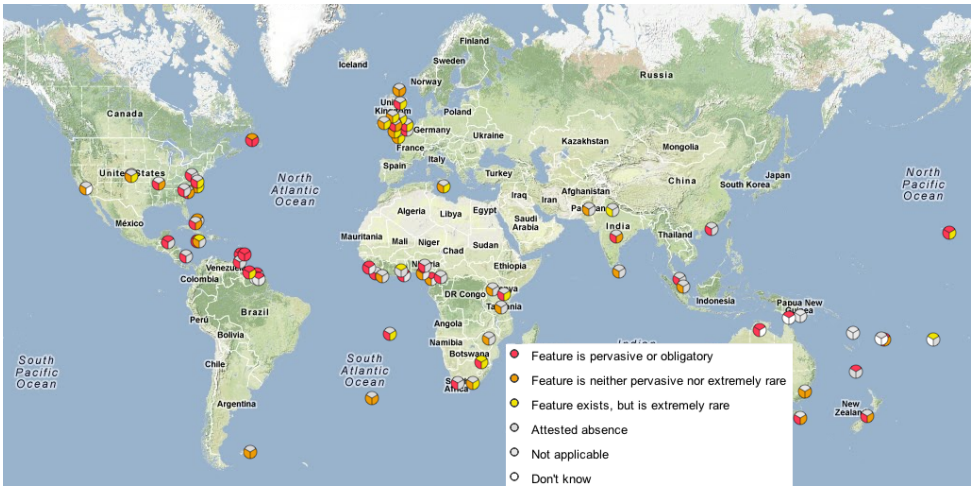


FIGURE 4: Simultaneous display of three features in eWAVE

3. INTERACTION BETWEEN R AND GEOGRAPHIC INFORMATION SYSTEMS (GIS).

3.1. USING GIS DATA TO CREATE HIGH-RESOLUTION MAPS. The contextualization of language documentation data often requires detailed custom maps of the often remote areas where the documented languages or dialects are spoken. Such maps can be created by freely available datasets stored as part of so-called geographic information systems (GIS) (see http://wiki.gis.com/wiki/index.php/Geographic_information_system). R is able to use the freely available GIS data with the packages "maptools" and "sp", as well as others. This is especially useful for generating maps of smaller scale, e.g. a smaller island.

Each custom map requires a particular resolution. For instance, a very rough resolution of the Hawaiian island O'ahu, as provided by one of the various packages developed within R itself, may be appropriate for a map of the world or a larger region (Figures 5a–5c); but the exact locations of small language or dialect communities often require a much higher resolution, and this can only be provided by GIS data sets, which can be imported into R (Figure 5d). With these data sets it is possible to zoom in to very fine levels of granularity (Figure 5e).

3.2. ADDING FURTHER INFORMATION TO MAPS. There are many data sets available containing further potentially relevant information for creating language maps, including population data, climate zones, land use, etc. For instance, topographic information, such as elevation, is also freely available and can be added to maps created in this way. For example, the GIS topographic data set "Oahu, HI 1/3 arc-second MHW DEM" (http://www.ngdc.noaa.gov/dem/squareCellGrid/download/3410) can be added as a background to the map of O'ahu (Figure 6).

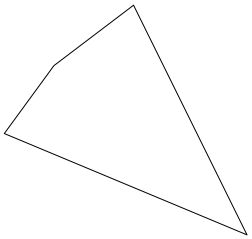


FIGURE 5a: Map created with the R package “maps” – 5 points

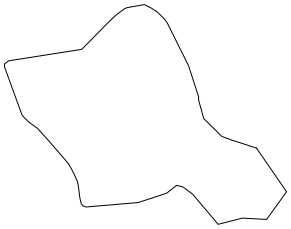


FIGURE 5b: Map created with the R package “sp” (wrld_simple data set) – 50 points

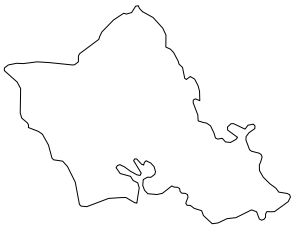


FIGURE 5c: Map created with the R package “mapdata” – 173 points

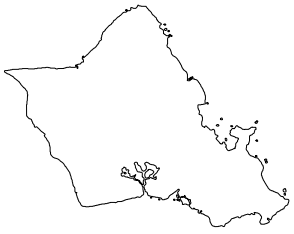


FIGURE 5d: Map created with the import of the GIS data set “gshhs” [2] – 2254 points

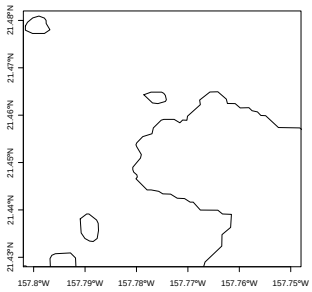


FIGURE 5e: High-resolution detail of map created with the GIS data set "gshhs"

FIGURE 6: Map of O'ahu using the GIS data set Oahu, HI 1/3 arc-second MHW DEM [http://www.ngdc.noaa.gov/dem/squareCellGrid/download/3410]

3.3. DRAWING GEOREFERENCED MAPS. Linguists often need to mark specific areas on a map, e.g. a language location, isoglosses, etc.; in other words, they need to draw a georeferenced area, which is then compatible with the maps and further geographic information discussed so far. This can be done using R's ability to import KML files, which are, for instance, used in Google maps and Google Earth. A relatively easy way to create a self-defined georeferenced polygon, like an isogloss, and include it in a map is to use the free version of Google Earth (http://www.google.com/earth). Google Earth allows one to create and to edit a georeferenced polygon, which can be saved as a KML file and imported into R. Clicking the link in the caption of Figure 7 will display a movie illustrating this workflow.

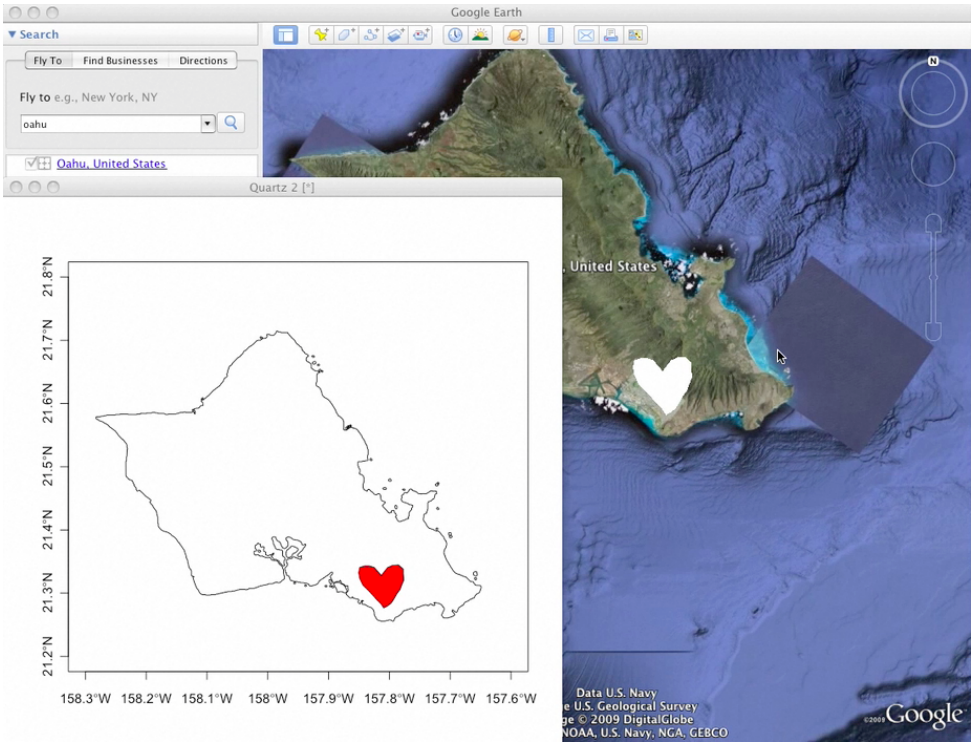


FIGURE 7: A workflow for drawing georeferenced maps [video: http://youtu.be/mMSaSaAxCp8, http://scholarspace.manoa.hawaii.edu/bitstream/handle/10125/4522/13-polygon.m4v]

4. GENERATING WEBSITES FROM STRUCTURED LINGUISTIC DATA. For the general public, as well as for some scientific purposes, it is useful to present linguistic data, such as thematic wordlists or entire lexica, in easily accessible formats, such as websites. If data such as wordlists are stored in formats such as Toolbox (see www.sil.org/computing/toolbox) and ELAN (see http://www.lat-mpi.eu/tools/elan), the creation of such HTML-based websites can be carried out relatively easily using R. As already mentioned, R is a powerful scripting language which works with data stored in many different formats including XML (for instance ELAN's eaf files) and plain text (for instance Toolbox files). In other words, instead of learning another scripting language like Python or Perl, such conversions (including the usage of regular expressions) can be done using R. Figure 8 shows an interactive HTML-based Even-Russian online dictionary of reindeer terms with pictures and linked sound files that was created with the help of R, using a Toolbox file as a source. The data for this example comes from Aralova et al. (in preparation).

5. CONCLUSION. The previous sections provided some examples of how R, an open-source software environment, can be used to process linguistic data, including data that is

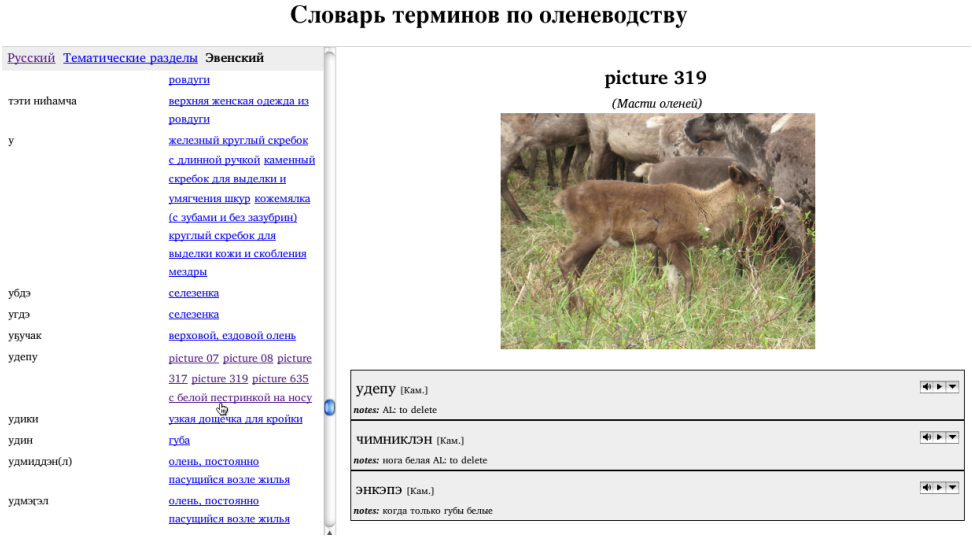


FIGURE 8: An online dictionary created from a Toolbox file using R

generated in language documentation, for various forms of visualizations and online presentations, which are appealing both for scientific research as well as for use by speech communities and the general public. While it is true that applying these techniques in R requires some familiarity with scripting languages in general, it is hoped that the examples given here will stimulate linguists to engage in these techniques themselves or else to cooperate with information scientists who are familiar with these environments.

REFERENCES

Amante, Christopher & Barry W Eakins. 2009. *ETOPO1 1 Arc-Minute Global Relief Model: Procedures, Data Sources and Analysis*. NOAA Technical Memorandum NESDIS NGDC-24. http://www.ngdc.noaa.gov/mgg/global/global.html.

Aralova, Natalia, Alexandra Lavrillier, Dejan Matić, Brigitte Pakendorf, Evgeniya Zhivotova & Luise Zippel. in preparation. Even dialectal dictionary of reindeer herding terminology. Leipzig: Max Planck Research Group on Comparative Population Linguistics.

Bivand, Roger S., Edzer J. Pebesma & Virgilio Gómez-Rubio. 2008. *Applied Spatial Data Analysis with R* Use R! New York: Springer.

Dryer, Matthew S. & Martin Haspelmath (eds.). 2011. *The World Atlas of Language Structures Online*. Munich: Max Planck Digital Library. http://wals.info/.

Kortmann, Bernd & Kerstin Lunkenheimer (eds.). 2011. *The electronic World Atlas of Varieties of English (eWAVE)*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://www.ewave-atlas.org/.

Michaelis, Susanne, Philippe Maurer, Martin Haspelmath & Magnus (eds.) Huber. forthcoming. The Atlas of Pidgin and Creole Language Structures (APiCS). http://lingweb.eva.mpg.de/apics/index.php/The_Atlas_of_Pidgin_and_Creole_Language_Structures_%28APiCS%29.

National Geophysical Data Center (NGDC). A Global Self-consistent, Hierarchical, High-resolution Shoreline Database (GSHHS). http://www.soest.hawaii.edu/wessel/gshhs/.

R Development Core Team. 2011. *A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org.

Hans-Jörg Bibiko
bibiko@eva.mpg.de

14

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. 105–110
http://nflrc.hawaii.edu/ldc/sp03/
http://hdl.handle.net/10125/4523

Language archives: They're not just for linguists any more

Gary Holton
Alaska Native Language Center, Fairbanks

While many language archives were originally conceived for the purpose of preserving linguistic data, these data have the potential to inform knowledge beyond the narrow field of linguistics. Today language archives are being used by people without formal linguistic training for purposes not necessarily envisioned by the original creators of the language documentation. The DoBeS Archive is particularly well-placed to become an important resource for cultural documentation, since many of the DoBeS projects have been interdisciplinary in nature, documenting language within its broader social and cultural context. In this paper I present a perspective from a legacy archive created well before the modern era of digital language documentation exemplified by the DoBeS program. In particular, I describe two types of non-linguistic uses which are becoming increasingly important at the Alaska Native Language Archive.

1. LESSONS FROM THE ANALOG ERA. Within the language archiving community we now face the problem of how to preserve and access a continually growing body of born-digital documentation of endangered languages. But at the Alaska Native Language Archive (ANLA, http://www.uaf.edu/anla) we have always been playing catch-up in the digital realm. For the past ten years I have been involved with ANLA in various capacities, and throughout that time the Archive has been not a dry, quiet repository but rather a dynamic part of ongoing language work in Alaska. Indeed, the Archive remains an integral part of language research and language revitalization efforts in Alaska. In many ways the Archive is less a repository and more an active research tool.

The ANLA collection includes nearly everything written in or about each of Alaska's twenty indigenous languages. This amounts to some 15,000 items, including everything from primary field notes to published grammars and dictionaries. In addition, ANLA holds significant collections in related languages outside Alaska; in particular, the coverage of Eskimo languages spoken in the Russian Far East is the most extensive in the world. Not all materials are original or unique; in many cases the Archive holds copies of documents housed elsewhere. (These copies may turn out to be the only extant copies, as happened recently with Bittle's Kiowa Apache field notes.) The aim is for comprehensive coverage. In addition to the print materials, the Archive also contains more than 5,000 audio recordings, though in contrast to the print materials no attempt has been made at comprehensive coverage.

As a legacy archive still struggling to enter the 21st century, ANLA cannot tell us much about best practices in digital preservation or about new technologies for accessing and sharing language resources. What ANLA and other legacy archives can tell us is how language archives have been and are being used. More than half a century has passed since collection began at ANLA, and in that time two important points have emerged regarding archive users and the uses they make of the archive. First, the primary users of the Archive are Native speakers and their descendants, that is, members of Alaska Native language communities. This situation is typical of established legacy archives (Austin 2011). Second, many users seek information which is not primarily linguistic in nature. Neither of these points should be taken as absolutes. The Archive continues to be used by academic researchers, and users continue to seek linguistic documentation. What is notable is that academics are not the *only* users, and linguistic information is not the *only* type of documentation sought. These observations may be relevant as we consider how to make use of language documentation being generated by current projects.

2. LANGUAGE ARCHIVES AS SOURCES OF CULTURAL DOCUMENTATION. As linguists – creators or collectors of language documentation – we tend to think first of the *linguistic* uses of archival materials. How can these materials inform our understanding of language? What does a text tell us about how serial verbs are employed? What does a recording tell us about prenasalization? Linguists tend to ask linguistic questions.

But language documentation encodes much more than just linguistic information. Language documenters are first and foremost field workers, interacting with speaker consultants whose interests lie in the documentation of many types of knowledge, be they marriage traditions or navigational techniques. A field worker documenting names for kin terms or stars is very likely to also document knowledge of marriage customs or stellar navigation, respectively, even if inadvertently. Furthermore, much language documentation has been collected without regard to the nature of non-linguistic content. A text may be recorded because it represents a particular genre or style, such as narrative or conversation. The content of that recording – i.e., what that narrative or conversation is about – is generally not constrained by the documenter. As a result language archives now present a veritable treasure trove of non-linguistic information encoded in the signal of the subject language. And users of language archives are very often interested in this type of information.

This interest can be gauged by considering some examples of recent inquiries at ANLA. These include requests for: information on ethnobotany in the Yukon Flats region; a copy of a eulogy for a 19th century missionary; a copy of Yup'ik music recordings, for use in writing a libretto; information on genealogy in the Upper Koyukuk region; and information on Russian influence in the middle Kuskokwim River region. All of these requests have a linguistic component; for example, ethnobotanical documentation includes indigenous names for plants, and the eulogy for Father Rysev was delivered in Aleut, translated into English. What distinguishes these patron requests is that they reflect an interest specifically in the *non-linguistic* content. A user sought information on ethnobotany in the Yukon Flats region not because she was interested in indigenous plant names but rather because she was interested in how plants were used. A user sought out the eulogy for Father Rysev not to study the use of Aleut language but rather because he was annotating Rysev's diaries. In this type of research it is the content rather than the linguistic code which is relevant.

Given the diversity of non-linguistic research at ANLA and constraints on space in this article I will elaborate here on just one area: ethnoastronomy. In late 2011 I began working with a user who was attempting to identify names for stars, constellations, and other atmospheric phenomena across Alaska. Before we met, this patron, Chris Cannon, had spent more than two years researching star names without much success; only two relevant published sources of information had been identified. MacDonald's (1998)s study of Inuit ethnoastronomy is one of the best works in this genre for any language; however, it is focused on the Eastern Canadian Arctic and includes only a few comparative terms in the Alaskan dialect of the Inuit language. The remaining 19 Alaska Native languages are completely outside the scope of this work. Bradley (2002) is a study of celestial navigation in Yup'ik and is easily the best modern documentation available. However, it is far from a comprehensive study and is focused on only one Alaska Native language.

It turns out that the best record of Alaska Native ethnoastronomy is to be found within the collections of ANLA. Cannon ((Unpublished ms.)) has worked tirelessly to compile this information using ANLA sources. Some language documentation contains extremely detailed information about stars. For example, Knut Bergsland's (1952) Aleut field notes include a star chart and a hand-drawn, labeled map of Aleut star names. While most linguists have not included such detail, it is still possible to compile star names in many languages. A name is recorded for the constellation *Ursa major* or some portion thereof in each of the twenty Alaska Native languages. In many cases these names are buried in obscure sources. They are not always accompanied by literal translations, but comparing the literal translations reveals some interesting patterns, as shown in Table 1.

LANGUAGE	FAMILY	LITERAL TRANSLATION
Aleut	Eskimo-Aleut	'caribou'
Yup'ik	Eskimo-Aleut	'caribou'
Siberian Yupik	Eskimo-Aleut	'caribou'
Inupiaq	Eskimo-Aleut	'caribou'
Tsimshian	Tsimshianic	'spoon'
Haida	Haida	'sea-otter-stretching-board'
Tlingit	Na-Dene	'all stone'
Ahtna	Na-Dene	'the one that moves above us'
Den'ina	Na-Dene	'one that turns over us', 'stars stretched'
Koyukon	Na-Dene	'it rotates its body', 'according to it the year is measured'
Tanacross	Na-Dene	'dipper'
Upper Tanana	Na-Dene	'I'm sitting'
Gwich'in	Na-Dene	'the seat'

TABLE 1: Literal names for constellation *Ursa major* in Alaska Native languages

Even without examining the actual linguistic form of the names, it is immediately clear from the literal translations that the Eskimo-Aleut languages share a common conceptualization of *Ursa major* as a caribou – something not found in languages of the other families. Examination of other star names reveals further insights. For example, the Inupiaq

constellation *iglupeaqtalik*, literally, 'the turf house', is formed from a combination of stars in constellations known to English speakers as Orion, Auriga, and Gemini. These facts are not purely linguistic observations; rather, they are facts about how indigenous Alaskan communities conceive of the sky. They reflect indigenous knowledge embedded in the linguistic code.

3. CREATION OF DERIVED LANGUAGE MATERIALS. All Alaska Native languages are extremely endangered, and for most, the youngest speakers are already age 70 or older. As the number of speakers continues to decline, there has been a marked increase in language revitalization efforts (cf. Gaul & Holton 2005). And these efforts increasingly turn to ANLA as a resource for developing derived or secondary language materials. For no language is this more evident than it is for Eyak. In 2008 Eyak became the first Alaska Native language to disappear in recent times. (The last speaker of Tsetsaut, an Athabaskan language formerly spoken on the Portland Canal, passed away in the first half of the twentieth century.) While all of Alaska's 19 remaining languages are severely endangered, the Eyak situation is in many ways exceptional. In most situations of language shift, it is difficult to identify a "last speaker". Rather, as knowledge of language erodes, the criteria which define a fluent speaker adapt, creating new last speakers (cf. Evans 2001). But in the case of Eyak the break-up and scattering of the community of speakers led to a large generational gap in transmission, with the last few speakers actually outliving their immediate descendants. In a very real sense, Marie Smith Jones was indeed the last Native speaker of Eyak; no partial or semi-speakers survive. Thus, for Eyak the only surviving sources of information about the language are found in the archival documentation at ANLA.

Since the passing of the last speaker there has been a surge of community interest in Eyak language. Truly an Eyak renaissance. The Eyak Language Project has created a multimedia website (sites.google.com/site/eyaklanguageproject) using information harvested from ANLA. Language workshops have been organized in Anchorage and Cordova. While some might characterize these efforts as too little, too late, that characterization is not appropriate, owing to the large body of documentation materials. Thanks to this documentary record, Eyak language revitalization is not "too late". The first known documentation of Eyak is found in a 1308-item vocabulary by Rezanov (1805). Eyak was then "re-discovered" in the 1930s through the work of de Laguna (Krauss 2006). Then, in 1961, serious linguistic documentation of the language began under the direction of Michael Krauss. At that time there were only four remaining speakers with a good command of Eyak; in other words, Eyak was already severely endangered nearly half a century ago. Krauss began documentation in earnest, eventually compiling about 3,000 pages of field notes; a massive dictionary of some 3,600 pages and 7,000 lexemes (Krauss 1970); and about ten hours of transcribed narrative texts (Krauss et al. 1970). Krauss continued his documentation work sporadically through the 1970's and 1980's, but by 1992 only one speaker remained and the corpus became essentially closed. Nonetheless, in spite of the small number of speakers and the lack of a thriving language community, in the span of less than a decade Eyak went from being almost unknown to being among the best documented of the Alaska Native languages. This invaluable documentary record forms the basis for new language initiatives today.

For Alaska Native languages other than Eyak, the situation is not so dire. Fluent speakers remain and can continue to contribute to ongoing language documentation and revitalization

programs. Nevertheless, archival documentation continues to be of great value because it documents an earlier stage of the language – a stage no longer accessible or known to speakers today. This documentation can form the basis for new language materials. For example, the Jesuit missionary Jules Jetté made elaborate and detailed records of place names in the middle Yukon River region in the early 20th century. Information from these maps can serve as the basis for new projects such as MapTeach (www.mapteach.org), which encourages geographic understanding and sense of place among descendents of Koyukon speakers, as well as new types of cultural reference materials (cf. YRDFA 2008). The Alutiiq Museum has converted Jeff Leer's grammatical notes into a multimedia website (http://www.alutiiqmuseum.org/language/learn-alutiiq.html).

4. CONCLUSION. The massive amount of documentation compiled over the past decade by various DoBeS projects and others represents an unprecedented contribution to the documentary linguistic record. As linguists begin to analyze this body of data, they will no doubt uncover new insights into the nature of human language. But the experience of legacy archives such as ANLA shows us that the DoBeS archive has a great potential to impact other areas of knowledge as well. Legacy archives can thus inform the way we think about the utilization of emerging language archives. In particular, archive users may be interested not only in linguistic applications but also in non-linguistic information encoded in the linguistic record; this highlights the importance of including information of relevance to non-linguists in the metadata. Furthermore, users may wish to repurpose linguistic data to create derived language materials for use in language revitalization efforts. Archives will better serve users to the extent that they facilitate these two types of uses.

REFERENCES

Austin, Peter. 2011. Who uses Digital Language Archives? Post to Endangered Languages and Cultures Blog. http://www.paradisec.org.au/blog/2011/04/who-uses-digital-language-archives/ (21 March, 2012).

Bergsland, Knut. 1952. Notes on Aleut wordlists regarding the natural world. Alaska Native Language Archive, Ms., ANLA Item AL950B(B170)1952.

Bradley, Claudette. 2002. Traveling with Fred George: The changing ways of Yup'ik star navigation in Akiachak, Western Alaska. In Igor Krupnik & Dyanna Jolly (eds.), *The Earth is Faster Now: Indigenous observations of arctic environmental change*, 240–265. Fairbanks: Arctic Research Consortium and Smithsonian Arctic Studies Center.

Cannon, Chris. (Unpublished ms.). Working list of indigenous star and constellation names of Alaska and their European counterparts.

Evans, Nicholas. 2001. The last speaker is dead – Long live the last speaker! In Paul Newman & Martha Ratliff (eds.), *Linguistic Fieldwork*, 250–281. Cambridge: Cambridge University Press.

Gaul, Karen K. & Gary Holton. 2005. Speaking across generations: Dena'ina language revitalization in Southcentral Alaska. *Alaska Park Science* 4(1). 26–31.

Krauss, Michael E. 1970. Eyak Dictionary. Massachusetts Institute of Technology, Ms., ANLA Item EY961K1970b.

Krauss, Michael E. 2006. A history of Eyak language documentation and study: Fredericæde Laguna in memoriam. *Arctic Anthropology* 172–217.

Krauss, Michael E., Anna Nelson Harry, Lena Saska Nacktan, Marie Smith, George Johnson & Galushia Nelson. 1970. Eyak Texts. Massachusetts Institute of Technology, Ms., ANLA Item EY961K1970a.

MacDonald, John. 1998. *The Arctic sky: Inuit astronomy, star lore, and legend*. Ottawa: Royal Ontario Museum and Nunavut Research Institute.

Rezanov, Nikolai Petrovich. 1805. Slovar' Unalashkinskago, Kad'yakskago, Kinayskago, Kolyuzhskago; Ugalykmutskago i Chugatskago Yazykov. Fond Adelunga, Academy of Sciences Leningrad, Leningrad. Unpublished ms.

Yukon River Drainage Fisheries Association (YRDFA). 2008. *Middle Koyukuk River of Alaska: At Atlas of Fishing Places and Traditional Place Names*. Anchorage, Alaska: Yukon River Drainage Fisheries Association.

Gary Holton
gary.holton@alaska.edu

15

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. 111–117
http://nflrc.hawaii.edu/ldc/sp03/
http://hdl.handle.net/10125/4524

Creating educational materials in language documentation projects – creating innovative resources for linguistic research

Ulrike Mosel
Kiel University

In its first two sections this paper briefly discusses two models of language documentation projects: the hierarchical model, in which the language documentation corpus (LDC) serves as a resource for the development of educational materials (EMs), and the integrative model, which integrates the production of EMs into the LDC and makes them a resource for linguistic research. The third and the fourth section describe how the integrative model was applied in the Teop Language Documentation Project and what kind of linguistic research topics it provides.

1. THE HIERARCHICAL MODEL OF LANGUAGE DOCUMENTATION. The hierarchical model of language documentation sees as its primary goal the compilation of a LDC (Himmelmann 2006) which can function as a resource for writing descriptive grammars, dictionaries, or educational materials, as shown in Figure 1.

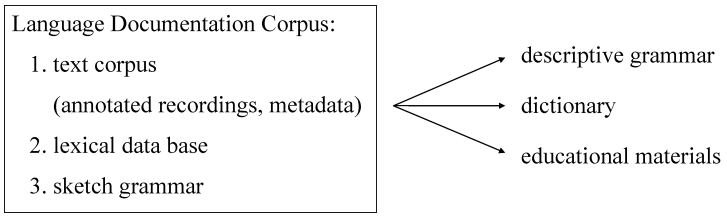


FIGURE 1: The hierarchical model of language documentation

While the exclusion of descriptive grammaticography from documentary linguistics has been criticized by several linguists (Evans 2008, Woodbury 2011), the production of educational materials does not seem to be a topic in linguistic debates, although most linguists support collaborative fieldwork and feel obliged to "give something back" to the speech community. Dobrin et al. (2009: 43) criticize the "commodification of endangered languages" as follows: "Linguists' professional obligations to field communities are often formulated in terms of transacted objects rather than through knowledge sharing, joint

engagement in language maintenance activities or other kinds of interactionally-defined achievements." Bowern (2011: 468) remarks, "Community members often also report sometimes feeling that the linguist comes in, reifies the language, turns it into a commodity, and then takes it away."

The problem with the hierarchical model is that even if a standard orthography has already been developed (Seifart 2006, Lüpke 2011), creating educational materials from a LDC may turn into a daunting task for a number of reasons:

1. the texts of the LDC represent spontaneous or elicited speech and may be difficult to understand because of repetitions, hesitation phenomena, and missing information
2. the texts may contain expressions like loan or swear words that are not acceptable for teachers
3. the people who are involved in the production of the EMs do not know how to exploit the LDC (Nathan & Fang 2009: 136)
4. the texts may not include contents or registers and genres that are considered useful or suitable for the prospective users of the planned materials

The first and the second problem can be solved by editing the transcripts, and the third one by training, though once the documentation is finished, it may be difficult for the linguist to find the time and the money to conduct training courses or help individual community members to derive EMs from the LDC. The fourth problem cannot be solved because the LDC simply does not provide any suitable resources for EMs.

2. THE INTEGRATIVE MODEL OF LANGUAGE DOCUMENTATION. The integrative model applies collaborative methods of language documentation which combine "field-work with teaching, training, and mentoring native speakers for sustainable documentation projects." (Grinevald 2003: 60). In this approach it is the indigenous language documenters who decide on the content, the purpose, and the format of the LDC, while the linguist works as their adviser in technological, organizational, and linguistic matters, explaining to them what can be done with the available resources and which kind of genres and topics would be the most suitable to begin with (Mosel 2006).

The first products of language documentation in this model are certainly not a comprehensive dictionary and recordings, transcriptions, and edited versions of conversations, but stories that are frequently told in the community or descriptions of certain activities that are considered as useful texts and are easily recorded and transcribed. Thus, from the very beginning, the team will work on building a LDC that includes educational materials for children, teachers, or the general public, but probably excludes a collection of texts that have nothing to do with the community's culture, such as translations from the contact language or elicited stories like the famous frog or pear stories (Chafe 1980, Mayer 1969, Chelliah & Reuse 2011: 427). In contrast to narratives and procedural texts, the recordings of natural conversations are too difficult to annotate and edit in the beginning of a LD project so that the production of conversational texts for learning-oriented materials can only be considered for a later stage of the project.

Linguists who depend on money from universities and scientific funding agencies may wonder if this kind of collaborative fieldwork is compatible with their professional aims and obligations. The answer is definitely yes if the compilation and annotation of the LDC meets scientific standards and provides a reliable basis for linguistic and other research (see Section 4).

3. EDUCATIONAL MATERIALS IN THE TEOP LANGUAGE DOCUMENTATION PROJECT. Teop is an Oceanic Meso-Melanesian language of the North-West Solomonic linkage (Lynch et al. 2002: 101 ff.), spoken by approximately 6,000 people in the Autonomous Region of Bougainville, Papua New Guinea. The project began in 2000 just after the civil war, the so-called Bougainville Crisis (1988–1999). During the first phase of the project (2000–2003), Ruth Saovana Spriggs, a native speaker of Teop, recorded interviews with elders of the community about customs and their personal histories, and I have been collecting legends, procedural texts, and encyclopedic descriptions with a team of highly motivated and skilled local language documenters since 2003. Currently (15 December, 2011), the Teop LDC consists of 258,957 words. The first book we published was a book of Teop legends (Magum, Enoch Horai, Joyce Maion, Jubilie Kamai, Ondria Tavagaga, with Ulrike Mosel & Yvonne Thiesen (eds.) 2007) because the legends seemed to present interesting reading materials for children and were easily recorded, transcribed, and edited. Not counting a hymn book printed in the 1950s, this is the very first book in the Teop language. It contains 40 edited versions of legends which were originally oral and narrated by 24 speakers, and it was produced in the following way:

1. recordings done by the linguist and native speakers
2. transcriptions done by native speakers, checked by the linguist, discussed with native speakers, revised and adjusted to a practical orthography developed by teachers in the 1980s, and eventually translated into English by the linguist with the help of native speakers
3. editorial work done by native speakers, checked by the linguist, discussed with the editor, revised and translated by the linguist with the help of native speakers; all writings done by hand because of the lack of electricity
4. oral and edited versions typed in Germany
5. proofreading of all legends, and minor changes made independently by two teachers; changes discussed with both teachers
6. preliminary version typed in Germany
7. final proofreading done in PNG
8. printing done in Germany because there was no printing press in the Autonomous Region of Bougainville
9. copies of the book sent to Bougainville and officially launched
10. all legends archived in PDF format in the DoBeS archive and made accessible without registration (Magum, Enoch Horai, Joyce Maion, Jubilie Kamai, Ondria Tavagaga, with Ulrike Mosel & Yvonne Thiesen (eds.) 2007)

Furthermore, the legends book was read by a native speaker and recorded. The recordings of her readings were annotated in ELAN with transcriptions and translations into English and were also archived in the DoBeS archive where they are freely accessible without registration (Magum, Enoch Horai, Joyce Maion, Jubilie Kamai, Ondria Tavagaga, with Ulrike Mosel & Yvonne Thiesen (eds.) 2007).

The editors and teachers were advised not to imitate the style of English stories but to keep as closely as possible to the original text, only remove hesitation phenomena and speech errors, and only make additions where absolutely necessary for the reader.

Editing the autobiographical narratives collected earlier by Ruth Saovana Spriggs was much more difficult than editing the legends. While the narration of legends followed a fixed story line with more or less conventionalized ways of expression, the personal narratives were embedded in casual conversations so that the editors felt that they had to make more changes to turn the transcriptions into readable texts. This is also reflected in word counts. The length of the edited legends is 98% of the length of the transcripts and, thus, almost the same, whereas the lengths of the edited autobiographical narratives is reduced to 42% of the transcripts.

When we began collecting procedural texts and encyclopedic descriptions for small specialized dictionaries on house-building, body and health, fishing, animals, and plants (Mahaka, Mark, Enoch Horai Magum, Joyce Maion, Naphtali Maion, Ruth Siimaa Rigamu, Ruth Saovana Spriggs & Jeremiah Vaabero, with Ulrike Mosel, Marcia Schwartz & Yvonne Thiesen 2010, Mosel 2011), most team members decided to write the texts straightaway without doing recordings and transcriptions. Having done the editorial work over several fieldwork sessions, they had become confident about their writing skills. All books are richly illustrated with drawings by an indigenous artist and photos which I produced, and will be published both in print and online in 2012 and 2013.

4. THE TEOP LDC AS A RESOURCE FOR LINGUISTIC ANALYSIS. The edited versions of the legends and autobiographical narratives as well as the written procedural texts and encyclopedic descriptions certainly do not represent traditional genres and indigenous registers of everyday communication. But as the data come from seven different editors and writers and have been reviewed by accepted language experts of the community, we are sure that we produced reliable language data.

An analysis of the differences between the oral legends and their edited versions showed five types of changes (Mosel 2008, forthcoming):

1. purification by the replacement of loan words
2. elaboration by the addition of words, phrases, and clauses
3. linkage of paratactic clauses by explicit coordination and embedding constructions and interlacing by raising constructions
4. compression by putting more information into a single linguistic unit, resulting in more complex structures
5. decompression by the resolution of complex structures

But the edited versions did not contain any constructions that were not found in the original transcripts.

For lexical and grammatical studies, the comparable subcorpora of oral and edited narratives show alternative ways of expressing the same content and, thus, give a fuller picture of the expressive potential of the language than the transcripts would have done by themselves. Furthermore they provide an innovative type of data for the study of the differences between spoken and written language. While the research on spoken vs. written European language varieties usually takes the structure of written varieties as the point of departure and explores how the spoken language deviates from the written one, the Teop data allow one to take the opposite perspective and study what people actually do when they put spoken texts into writing and thus develop a written language variety.

5. CONCLUDING REMARKS. This paper proposes a new perspective on the role of educational material in language documentation projects and describes how such resources can be utilized by community members as well as for linguistic research. Such community utilization can be further encouraged by workshops and other pedagogical means.

Due to the special circumstances after the civil war, the Teop language project started as a grass-roots project. We did not involve any official authorities but only worked together with teachers of the local village school without informing the wider public or bringing in further experts of documentary linguistics. But for the future we plan to conduct workshops for teachers and language activists (see Florey & Himmelmann 2009 for training strategies) and broaden our view of language documentation by learning from the related disciplines of language pedagogy and language revitalization (Hinton 2011, Nathan & Fang 2009).

REFERENCES

Bowern, Claire. 2011. Planning a language-documentation project. In Peter K. Austin & Julia Sallabank (eds.), *The Cambridge Handbook of Endangered Languages*, 459–482. Cambridge: Cambridge University Press.

Chafe, Wallace L. (ed.). 1980. *The Pear Stories: Cognitive, cultural, and linguistic aspects of narrative production*. Norwood, NJ: Ablex.

Chelliah, Shobhana L. & Willem J. de Reuse. 2011. *Handbook of Descriptive Linguistic Fieldwork*. Dordrecht: Springer.

Dobrin, Lise M., Peter K. Austin & David Nathan. 2009. Dying to be counted: the commodification of endangered languages in documentary linguistics. In Peter K. Austin (ed.), *Language Documentation and Description 6*, 37–52. London: School of Oriental and African Studies.

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

Florey, Margaret & Nikolaus P. Himmelmann. 2009. New directions in field linguistics: Training strategies for language documentation in Indonesia. In Margaret Florey (ed.), *Endangered languages of Austronesia*, 121–140. Oxford: Oxford University Press.

Grinevald, Colette. 2003. Speakers and documentation of endangered languages. In Peter K. Austin (ed.), *Language Documentation and Description 1*, 52–72. London: School of Oriental and African Studies.

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: Mouton de Gruyter.

Hinton, Leanne. 2011. Language revitalization and language pedagogy: New teaching and learning strategies. *Language and Education* 25(4). 307–318.

Lüpke, Friederike. 2011. Orthography development. In Peter K. Austin & Julia Sallabank (eds.), *The Cambridge Handbook of Endangered Languages*, 312–336. Cambridge: Cambridge University Press.

Lynch, John, Malcolm Ross & Terry Crowley. 2002. *The Oceanic Languages*. Richmond: Curzon Press.

Magum, Enoch Horai, Joyce Maion, Jubilie Kamai, Ondria Tavagaga, with Ulrike Mosel & Yvonne Thiesen (eds.). 2007. *Amaa vahutate vaa Teapu. Teop legends. Illustrated by Rodney Rasin*. Kiel: Seminar für Allgemeine und Vergleichende Sprachwissenschaft, Christian Albrechts Universität. http://www.mpi.nl/dobes/projects/teop.

Mahaka, Mark, Enoch Horai Magum, Joyce Maion, Naphtali Maion, Ruth Siimaa Rigamu, Ruth Saovana Spriggs & Jeremiah Vaabero, with Ulrike Mosel, Marcia Schwartz & Yvonne Thiesen. 2010. *A inu. The Teop-English dictionary of house building*. Kiel: Seminar für Allgemeine und Vergleichende Sprachwissenschaft, Christian Albrechts Universität. http://www.mpi.nl/dobes/projects/teop.

Mayer, Mercer. 1969. *Frog, Where Are You?* New York: Penguin.

Mosel, Ulrike. 2006. Fieldwork and community language work. In Jost Gippert, Nikolaus P. Himmelmann & Ulrike Mosel (eds.), *Essentials of Language Documentation*, 67–85. Berlin: Mouton de Gruyter.

Mosel, Ulrike. 2008. Putting oral narratives into writing – experiences from a language documentation project in Bougainville, Papua New Guinea. Paper given at Simposio Internacional Contacto de lenguas y documentación. Buenos Aires,14-15 August 2008. http://www.linguistik.uni-kiel.de/Oral%20narratives%20into%20writing.pdf (21 March, 2012).

Mosel, Ulrike. 2011. Lexicography in endangered languages communities. In Peter K. Austin & Julia Sallabank (eds.), *The Cambridge Handbook of Endangered Languages*, 337–339. Cambridge: Cambridge University Press.

Mosel, Ulrike. forthcoming. Corpus linguistic and documentary approaches in writing a grammar of a previously undescribed language In: Toshihide Nakayama & Keren Rice (eds.), Keren Practical and Methodological Issues in Grammar Writing.

Mosel, Ulrike, Jessika Reinig, Marcia Schwartz, Ruth Saovana Spriggs & Yvonne Thiesen. 2007. Das Teop Language Corpus. http://www.mpi.nl/dobes/projects/teop.

Nathan, David & Meili Fang. 2009. Language documentation and pedagogy for endangered languages: A mutual revitalisation. In Peter K. Austin (ed.), *Language Documentation and Description 6*, 132–160. London: School of Oriental and African Studies.

Seifart, Frank. 2006. Orthography Development. In Jost Gippert, Nikolaus P. Himmelmann & Ulrike Mosel (eds.), *Essentials of Language Documentation*, 275–299. Berlin: Mouton de Gruyter.

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.

Ulrike Mosel
umosel@gmx.de

16

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. 118–125
http://nflrc.hawaii.edu/ldc/sp03/
http://hdl.handle.net/10125/4525

From language documentation to language planning: Not necessarily a direct route[1]

Julia Sallabank
SOAS, University of London

In this paper I will consider how documentary linguists can provide support for community language planning initiatives, and I will discuss some issues. These relate partly to the process of language documentation: what and who we choose to document, how we define 'a language', and how we deal with language variation and change; and partly to community attitudes and dynamics.

1. INTRODUCTION. In documentary linguistics it is often thought that it is a relatively simple matter to produce materials for use by communities, such as dictionaries and storybooks, from documentary corpora. And indeed it may be a relatively simple procedure technically, given familiarity with multimedia software. But the process of negotiating which materials should be published, for what purposes and how, involves issues of language planning which are not necessarily straightforward. Moreover, the word 'community' is frequently used over-simplistically in the field of language documentation: it is important to realize that communities are not monolithic.

Language policy and planning are in turn relevant to language documentation. Language endangerment cannot be separated from its social, cultural and political causes and effects, although some linguists (e.g. Newman 2003) would like to do so. Yet linguists' very presence in a speech community may affect language attitudes and patterns of use: e.g. studying and transcribing endangered languages demonstrates that they can be learnt and written, if writing is valued and desired by community members. The process of establishing writing systems and standards is another issue which is fraught with social, cultural and political ramifications (Hinton 2003, Sebba 2007, Marquis & Sallabank 2009).

Linguists may be called on to advise or help with language planning, whether or not we are knowledgeable about potential issues and pitfalls. An 'international expert' may be invited to advise regional or national authorities on language policy. But documentary linguists are most often involved with language planning at community level as that is the level at which fieldwork is conducted.

[1] The research reported in this paper was funded by the UK Economic and Social Research Council, the Nuffield Foundation, and the Hans Rausing Endangered Languages Programme (Arcadia). I would like to acknowledge the invaluable input and collaboration of Yan Marquis in the research and analysis which informed the paper.

2. A BRIEF OVERVIEW OF LANGUAGE PLANNING. Language planning can be defined broadly as comprising any decisions or actions that affect language use, e.g. punctuation, language shift, or the choice of an (inter)national language. Language policy and planning are often conflated; along with Schiffman (1996), Hornberger (2006) and John Walsh (personal communication, November 2008), this paper views language *planning* as concrete actions or measures to implement *policies*, which are defined as decisions, positions and principles regarding language, its nature and role.

Language policies are likely to reflect prevailing language ideologies. Because some aspects of language are commonly held to be iconic (emblematic) of some aspects of identity, language policy and planning may arouse strong feelings. This paper discusses some problems with regard to the use of language documentation in language planning which may arise from ideologies and perceptions – both those of linguists and those of community members. Ideologies and perceptions also affect polices at governmental level of course, but this paper focuses on community-level language planning.

Small-scale, local, grass-roots actions in support of endangered languages are frequently characterized by more enthusiasm than planning and thus may reflect what Baldauf (1993) calls 'unplanned language planning'. They are less frequently studied or reported in the academic literature than large-scale programs or national and international policies.

Language planning falls into two main categories (Kloss 1969, Cooper 1989, Kaplan & Baldauf 1997):

- Actions to define or modify a language itself
- Actions to modify *attitudes* towards a language, or its *status* in a language ecology.

The most obvious area in which documentary materials and linguists have a role to play is in the description of a language, which is necessary for *corpus planning*: the codification, graphization, orthography, standardization of a language, as well as terminology development. Documentary materials can provide evidence-based corpora for the production of dictionaries, grammars and language learning materials. Although we wish to document and analyze the rich variations encountered among speakers, in practice it is difficult not to identify one variant as the reference model in both corpora and materials aimed at speakers. Linguists thus contribute to defining languages, and which varieties become canonical or standard, which in many cases has political ramifications.

It can be argued that documentation is necessary for *language-in-education* (or *acquisition*) planning (though as will be seen below, this does not always happen in practice). Education is the largest arena for language policy and planning, and includes medium of education, immersion, which languages are taught as school subjects, teacher training, etc. But language acquisition planning can also be carried out in less formal and more community-based ways, e.g. the master (or mentor)-apprentice programs described by Hinton (1997) and Hinton et al. (2002). Language acquisition planning may be seen as involving both corpus planning and status planning, which encompasses attempts to expand the domains in which a language is used, attempts to secure official recognition, etc.

The 'last but not least' element of language planning is *prestige planning*. This term was introduced by Haarmann (1984, 1990) to differentiate activities aimed at promoting a positive view of a language from those concerned with political status or functions: 'Not

only the content of planning activities is important but also the acceptance or rejection of planning efforts' (Haarmann 1990: 105). This stage is frequently omitted but it is essential for success: all too often measures omit to foster positive attitudes towards multilingualism, linguistic diversity, or a particular language (Fennell 1981, Dauenhauer & Dauenhauer 1998). For example, Grenoble & Whaley (2006) argue that Soviet language policy, while ostensibly supporting minority languages, led to Siberian peoples becoming passive recipients of language planning rather than active participants, and thus to lack of enthusiasm for revitalization projects.

3. IMPLICATIONS OF LANGUAGE PLANNING FOR DOCUMENTARY LINGUISTICS. If one aim of language documentation is for materials to be available for use in language planning and revitalization, it is necessary to plan for language planning when planning documentation. However, the types of linguistic events that are documented often still reflect traditional anthropological preoccupations by focusing on rituals, narratives and stories, and sometimes songs. Cultural information such as crafts and recipes may be collected, which may well be of use in future learning materials. But all too frequently interactive and child-oriented language is omitted from the documentary record. Stoll & Bickel (this volume) provide an example of child language in language documentation; however, in our project to document the indigenous language of Guernsey (Channel Islands) the youngest native speaker is 44 at the time of writing, and as part and parcel of the loss of intergenerational transmission, very few older speakers have experience of bringing up their children through Guernesiais. We have nevertheless made efforts to elicit child-rearing language, lullabies etc., e.g. I recently interviewed the second youngest native speaker and his mother together. As well as documenting traditional activities such as making crab pots, in our documentation we have made particular efforts to bring speakers together for conversations, discussions and staged events such as a traditional card game which elicited a number of swear words, another genre often omitted from the documentary record.

When planning language documentation linguists also need to consider such issues as what the materials will be used for, and who will use them (current speakers? their descendants? unrelated people?). As Holton (this volume) shows, not all potential future users and uses can be foreseen, which underlines the need for a broad corpus. For community-based planning to be possible, archives need to be locally available and accessible, bearing in mind that due to lack of resources or age, by no means all potential users of the materials will be internet users. So it is important to ascertain whether potential users make use of digital resources; and whether they (will) know how to access ELAN files or XML, for which a further level of technical skill is required. If community members do not have such access or expertise but are literate, a language documentation project needs to provide printed materials in addition to digital. In our case, consultants often have audio CD players but not computers, so we provide CDs of recordings. Mosel (this volume) gives an example of educational materials created as part of language documentation.

Documentary records need to provide resources for a variety of language planning activities, from formal lessons (a common first step in revitalization movements) to the re-establishment of intergenerational transmission. In many cases, language endangerment is not addressed until maintenance is no longer possible – in the words of Joni Mitchell, "you don't know what you've got 'til it's gone". The most important role of archived language

material may therefore be as a 'fall-back' in case there is later desire for revival, as is happening in Europe, America and Australia even 200 years after the 'last speakers'. For example, in the Isle of Man (in the Irish Sea) a group of young parents learned Manx and used it as their primary medium of socialization and child-rearing about 20 years after the death of the last traditional speaker – a remarkably successful ongoing revival which would not have been possible without adequate documentation. They then lobbied successfully for Manx-medium education provision.

It is well known that such a variety of potential users and uses necessitates 'a record of a language which leaves nothing to be desired by later generations wanting to explore whatever aspect of the language they are interested in' (Himmelmann 2006: 7, Woodbury 2003). But what is 'representative'? What and whose language practices (and idea of language) do documentary linguists represent – and who decides? As will be seen below, community-based language planning is situated in community dynamics, which need to be taken into account when deciding how we define a language itself, a 'language community', who counts as a community member, who 'owns' a language, and who gets to decide language policy.

4. SOME POTENTIAL OBSTACLES TO THE USE OF DOCUMENTARY MATERIALS IN LANGUAGE PLANNING. Language maintenance and revitalization are often grass-roots initiatives, and it is at this level that documentary linguists are best placed to provide support since they work directly with speakers. But in community-led language planning, documentary evidence may not necessarily be appreciated, believed, or taken on board.

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. Documentation which demonstrates low vitality, attrition, variation and change may be seen as promoting 'decline'. As noted by Dorian (2009), people who are viewed by the community as 'good speakers' may be highly rated for e.g. their skill in verbal arts or storytelling rather than for their knowledge of verb paradigms or rarely-used structures. Community members may see themselves as 'native speakers' (or 'language owners') without actually speaking the language well (Evans 2001). In my research into Guernesiais, the indigenous language of Guernsey, Channel islands, I have found that people who grew up hearing the traditional language may think of themselves as native speakers and therefore as fluent; but they may not use the language actively, so their production is affected by attrition (Sallabank 2010). Even if they do use the language regularly, and appear superficially fluent, they may rely on a relatively small range of formulaic constructions. Pointing out such deficiencies may not be welcome to people who are considered to be 'expert users' by their community. Grenoble (2010) recounted being accused of trying to trick consultants into making errors by trying to elicit structures and lexis outside their 'comfort zone'.

When it comes to language description, some community members may prefer 'folk linguistic' perceptions (Niedzielski & Preston 2000) or purism to documentary evidence. Like many endangered languages, Guernesiais does not have a prestige or standard dialect. Regional variation is a key feature, with a core set of iconic variables frequently cited by consultants as both valued but also as barriers to developing orthography or school curricula. However, the variation that we have found in our documentation is both more pervasive

and more complex than expected: it is not only region- and age-related, but idiolectal and related to language contact, change and attrition processes. In many cases it challenges the generally accepted regional stereotypes.

With regard to language change, we have found pervasive simplification of morphology, e.g. a reduction in verb paradigms, in many cases to just two forms without distinction between singular and plural in the first and third persons[2]:

mawjai (eat)	*mɔʒaj*	*mɔʒ*	
inf.	(2p. pl present indic.)	(all other persons)	
di (say)	*dizaj*	*diz* or *di*	*di*
inf.	(2p. pl present indic.)	(3p. pl present indic.)	(all other persons)

We have also found over-generalizations, e.g. of plural forms, and numerous features such as calques which reflect contact with English (now the main language of Guernsey), for example *nou bailli a hao* (literally 'we gave up', used for 'we retired'). We have also encountered hypercorrection and 'purism', which in Guernsey may manifest itself as convergence towards French, to which Guernesiais was formerly subaltern in a diglossic relationship, and which retains high prestige.

Such trends are typical of endangered languages (Trudgill 1983, Nettle & Romaine 2000). The dilemma for documentary linguists and language planners alike is which of these forms, if any, should form part of a 'definitive record' of a language? And which, if any, should be included in reference, learning and teaching materials? When raising such issues in Guernsey we have met with quite negative reactions from some factions in the community. There is a certain amount of unwillingness to accept the fact or inevitability of language change. It is not uncommon to hear claims that Guernesiais has not changed since the 12th century: 'We speak the language of William the Conqueror'. Many older community members (of all levels of language proficiency) have a genuine nostalgic attachment to the traditional language they heard in their youth. This may become associated with a 'discourse of the past' and an unwillingness to 'hand over' Guernesiais to new users: 'They're going to change the language to teach it – it won't be the Guernsey French we know'. There is also unwillingness to create new terminology, or to accept the need for it: '*Ya paa de naom, véyou, pour riae k'ei modern* (There aren't words, you see, for anything modern)'. In some cases this has led to denial of the need for documentation, rationalized by stating that 'there are still speakers' (although Guernesiais is not being transmitted to children) and 'it's being taught in schools' (although Guernesiais is not part of the school curriculum, so lessons are extra-curricular and taught by untrained volunteers for 30 minutes a week).

Yet despite such sentiments, some semi-speakers and a growing number of adult learners express a need for materials. Language activists may feel that language revitalization is so urgent that they may not want to wait for recordings to be transcribed, described and analyzed. So materials may be produced which are not based on a documentary corpus but on the intuitions of the 'good speakers' described above. Such materials may have prestige but may portray the language as it is perceived rather than how it is used according

[2] Abbreviations: 2p., 3p. – 2nd, 3rd person; pl – plural; indic. – indicative; inf. – infinitive.

to documentary evidence. Yet it may be claimed that further materials are not necessary (especially if corpus-based materials might challenge the 'correctness' of earlier ones).

The production of reference and learning materials for Guernesiais is also hampered by the lack of an agreed orthography, discussion of which can generate heated debates. Preferences in orthographic systems are related to purpose, but also to language ideologies. Linguists tend to prefer orthography to reflect the phonemic inventory of a language. 'Traditionalists' may prefer 'iconic' or prestige orthographies, which in Guernsey reflect the former diglossic High language, French. Language activists may want to promote *Ausbau* (Kloss 1967) in order to differentiate their language from a dominant or related one. Learners, meanwhile, need systematic, transparent spelling and would prefer a system which helps them learn – which might reflect pronunciation, etymology or morphology depending on circumstances. The resolution of such issues often does not depend on impartial assessment of which orthography is the most efficient, but on community dynamics which may be fluid and not immediately obvious to outsiders.

5. CONCLUSIONS. It is important to think about who belongs to a 'community' and which members are empowered to make decisions on language policy. Linguists often seek the collaboration of elders and traditional community leaders. But when planning the future of a language, the views of other stakeholders may be equally valid: e.g. semi-speakers, learners and heritage speakers, who are often key actors in language revitalization; and language activists and supporters, who may not learn or speak the language at all. A well-designed linguistic corpus can generate materials for a number of purposes and audiences, but given that there are usually financial and time constraints, the issue of priorities and who decides them remains.

More research is needed into how language documentation can contribute to effective language planning. Ideally, documentation would support rational decision-making by providing a corpus which demonstrates the current situation and shape of the language. But as seen above, this is not always straightforward. Given that language documentation and language planning may be taking place simultaneously, arguably part of the task of language documentation is to document the process of language planning itself in order to provide a corpus of experiences that other researchers and communities can draw on. Language planning is situated in community dynamics and is mediated by stereotypes, perceptions, personalities, vested interests and ideologies, which themselves need to be documented in order to investigate how language planning might be made more effective. Linguists and sociolinguists also need to find ways to communicate their findings effectively so that they can provide adequate advice and resources when required by language planners.

REFERENCES

Baldauf, Jr, Richard B. 1993. "Unplanned" language policy and planning. *Annual Review of Applied Linguistics* 14. 82–89.

Cooper, Robert L. 1989. *Language Planning and Social Change*. Cambridge University Press.

Dauenhauer, Nora Marks & Richard Dauenhauer. 1998. Technical, emotional, and ideological issues in reversing language shift: Examples from Southeast Alaska. In Lenore A. Grenoble & Lindsay J. Whaley (eds.), *Endangered Languages: Language Loss and Community Response*, 57–98. Cambridge: Cambridge University Press.

Dorian, Nancy C. 2009. Age and speaker skills in receding languages: how far do community evaluations and linguists' evaluations agree? *International Journal of the Sociology of Language* 200(Nov). 11–25.

Evans, Nicholas. 2001. The last speaker is dead – Long live the last speaker! In Paul Newman & Martha Ratliff (eds.), *Linguistic Fieldwork*, 250–281. Cambridge: Cambridge University Press.

Fennell, Desmond. 1981. Can a shrinking linguistic minority be saved? Lessons from the Irish experience. In Einar Haugen, J. Derrick McClure & Derick Thomson (eds.), *Minority Languages Today: A Selection from the Papers Read at the First Conference on Minority Languages at Glasgow University 8-3 September 1980*, 32–39. Edinburgh: Edinburgh University Press.

Grenoble, Lenore. 2010. Switch or shift: Code-mixing, contact-induced change and attrition. Annual Public Lecture, Hans Rausing Endangered Languages Project, School of Oriental and African Studies, London, 22 February 2010.

Grenoble, Lenore A. & Lindsay J. Whaley. 2006. *Saving Languages: An Introduction to Language Revitalization*. Cambridge: Cambridge University Press.

Haarmann, Harald. 1984. Sprachplanung und Prestigeplanung [Language Planning and Prestige Planning]. *Europa Ethnica* 41(2). 81–89.

Haarmann, Harald. 1990. Language planning in the light of a general theory of language: A methodological framework. *International Journal of the Sociology of Language* 86. 103–126.

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: Mouton de Gruyter.

Hinton, Leanne. 1997. Small languages and small language communities: Survival of endangered languages: The California Master-Apprentice Program. *International Journal of the Sociology of Language* 123. 177–191.

Hinton, Leanne. 2003. Orthography wars. Originally presented at WAIL [Workshop on American Indigenous Languages], Santa Barbara, California, 2003. Unpublished ms.

Hinton, Leanne, Matt Vera & Nancy Steele. 2002. *How to Keep Your Language Alive: A Commonsense Approach to One-on-One Language Learning*. Berkeley, CA: Heyday.

Holton, Gary. this volume. Language archives: They're not just for linguists any more.

Hornberger, Nancy H. 2006. Frameworks and models in language policy and planning. In Thomas Ricento (ed.), *An Introduction to Language Policy: Theory and Method*, 24–41. Oxford: Blackwell.

Kaplan, Robert B. & Richard B. Baldauf Jr. 1997. *Language Planning: From Practice to Theory*. Clevedon, England: Multilingual Matters.

Kloss, Heinz. 1967. 'Abstand Languages' and 'Ausbau Languages'. *Anthropological Linguistics* 9(7). 29–41.

Kloss, Heinz. 1969. *Research Possibilities on Group Bilingualism: A Report*. [Washington, D.C.]: Laval Univ., Quebec (Quebec). International Center for Research on Bilingualism. Distributed by ERIC Clearinghouse.

Marquis, Yan & Julia Sallabank. 2009. Issues in orthography development: Examples from Dgernesiais / Guernésiais / Giernesiei / Djernezié. Paper presented at workshop on Writing Systems: Analysis, Acquisition and Use 2, Institute of Education, London, 28 November 2009.

Mosel, Ulrike. this volume. Creating educational materials in language documentation projects – creating innovative resources for linguistic research.

Nettle, Daniel & Suzanne Romaine. 2000. *Vanishing Voices: The Extinction of the World's Languages*. New York: Oxford University Press.

Newman, Paul. 2003. The endangered languages issue as a hopeless cause. In *Language Death and Language Maintenance: Theoretical, Practical and Descriptive Approaches*, 1–13. Amsterdam: John Benjamins.

Niedzielski, Nancy A. & Dennis Richard Preston. 2000. *Folk Linguistics*. Berlin: Mouton de Gruyter.

Sallabank, Julia. 2010. The Role of Social Networks in Endangered Language Maintenance and Revitalization: The Case of Guernesiais in the Channel Islands. *Anthropological Linguistics* 52(2). 184–205.

Schiffman, Harold F. 1996. *Linguistic Culture and Language Policy*. London: Routledge.

Sebba, Mark. 2007. *Spelling and Society: The Culture and Politics of Orthography Around the World*. Cambridge: Cambridge University Press.

Stoll, Sabine & Balthasar Bickel. this volume. How to measure frequency? Different ways of counting ergatives in Chintang (Tibeto-Burman, Nepal) and their implications.

Trudgill, Peter. 1983. *On Dialect: Social and Geographical Perspectives*. Oxford: Blackwell.

Woodbury, Anthony C. 2003. Defining documentary linguistics. In Peter K. Austin (ed.), *Language Documentation and Description 1*, 35–51. London: School of Oriental and African Studies.

Julia Sallabank
js72@soas.ac.uk

17

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. 126–128
http://nflrc.hawaii.edu/ldc/sp03/
http://hdl.handle.net/10125/4526

Online presentation and accessibility of endangered languages data: The General Portal to the DoBeS Archive[1]

Gabriele Schwiertz
University of Cologne

Data depositories containing language documentation corpora are generally well structured, well maintained, and include large collections of many under-researched languages. However, they are not yet conceived of as resources that can be easily consulted on scientific or non-scientific questions pertaining to one of those languages. A general portal to the DoBeS archive has been created to facilitate access to the data, to attract more users to the archive, and to lower the threshold for users outside the linguistic community to access the data. The structure and the main features of this portal will be presented in this paper.

1. INTRODUCTION. When considering the exploitation of language documentation data contained in language archives, three major user groups can be identified: The speaker community, the scientific community, i.e. linguists and scholars of related disciplines, and the general public. Each of these user groups has different interests and different needs, all of which are hardly satisfied by the IMDI-tree representation of the DoBeS archive. For the community users, community portals have been created in some projects. However, in order to attract users to the archive, facilitate access to the data, and generate new user scenarios and communities, we have created a general portal to the DoBeS archive.

This portal tries to address the needs of the different user groups with different subportals where selected information specific to typical queries of those users is presented in a straightforward way and shortcuts to the relevant media are included. In the future, this portal will serve as the main entryway to the archive and as a container of information about the DoBeS projects and language endangerment in general. In this paper, I will briefly discuss features of the portal and the motivation for the current structure. A preliminary

[1] The structure of the DoBeS portal has been presented in various settings and I wish to thank everybody for their contribution: This article has benefitted greatly from input of the audiences at the DoBeS meeting 2010 in Nijmegen, from a meeting with DoBeS teams and linguistic students in Cologne, from the Potentials workshop in Leipzig, and from discussion among a dedicated group of DoBeS members. I especially would like to thank Nikolaus P. Himmelmann, Dagmar Jung, and Iren Hartmann for their help. I also wish to thank Paul Trilsbeek and the Technical Team at the MPI for Psycholinguistics in Nijmegen for the actual implementation of the website.

version of the portal can be found here: http://ems03.mpi.nl/dobes/, in due course the portal will replace the DoBeS website at http://www.mpi.nl/dobes.

2. STRUCTURE AND FEATURES OF THE DOBES PORTAL.

2.1. MAIN PAGE. The first page of the portal has two main functions: It gives a first impression of what DoBeS is, and it constitutes the gateway to the sub-portals dividing the user communities. There is an interactive map in the middle of the page with all the DoBeS projects. From here, it is possible to go immediately to the single project pages. On the left side, appetizers are presented: Videos, audio recordings, or pictures that are freely accessible can be called up. On the right side, quick links and a Google-style search-box are provided. Under the map, users are divided into three main groups: The scientific audience, e.g. linguists, ethnographers, and musicologists; the general public, i.e. journalists, community members,[2] or interested laymen; and into future-depositors. As on all pages, a constant reminder of issues of access rights, access levels, and user registration is included with links to more detailed information on these topics.

2.2. RESEARCH PORTAL. The research portal[3] provides information on what kind of data is contained in the archive and what can be done with it. Access to statistics for all the projects in the archive is provided in a table showing how many sessions there are for each language, how many include video or annotations, and at what access levels[4]. In order to allow typologists to get a first impression before deciding whether the data in the DoBeS archive is suited to their needs, a link to all grammars and sketch grammars is provided together with a link to studies deposited in the archive. Furthermore, information is given on what kinds of searches are available, and what DoBeS related tools there are. Manuals are supplied for certain usage cases, e.g. writing an MA-thesis using data in the archive. In addition, more specialized information pages for various scientific user groups are linked here.

2.3. GENERAL INTEREST PORTAL. The general interest portal[5] is aimed at providing information about language endangerment and at simplifying access to the DoBeS archive for the general public, giving an initial impression of possible points of interest it contains. Again, there are appetizers on the left, and access to an overview of archive statistics is given. A link to a collection of community portals is supplied. Information is provided about the searching and browsing functions and about registration and access. 'Highlights' of the archive, items which we hope will be interesting to the general public, are collected here; these include subtitled videos, artworks, story collections, and other appealing material.

---

[2] Some teams have created community portals for the members of the communities they work with. These include information that is relevant for this audience, often in a language other than English. However, for the general portal, it remains to be seen with increasing use of the website whether there should be a third sub-portal just for community members.

[3] http://www.mpi.nl/dobes/research

[4] The data in the archive are marked for different access levels, indicating whether everybody is allowed to access them, or some subgroup of users or whether they are not accessible right now.

[5] http://www.mpi.nl/dobes/general

There are also plans for this page to contain other generally entertaining points in the future, such as a 'word of the day' from the DoBeS languages or a photo/video gallery which can be combined in many ways with services like Facebook, Twitter, and YouTube.

2.4. THE DEPOSITORS' PORTAL. For the time being, the depositors' page[6] contains basic information for future depositors to the DoBeS archive. At some point, this page could be a more powerful interface with tutorials, tools, and possibly links to pages with an upload functionality.

2.5. THE PROJECT PAGES. Each DoBeS project has its own page. The project pages contain the same information as the former project pages of the DoBeS website. An addition provided by the new portal is a summary page with the same structure for all projects; here, statistics will be displayed as well as a very brief statement about what the project is and a note on the team. Additionally, there will be a short list of highlights contained in the respective archives, such as a large number of videos about hunting techniques, a dissertation on the language, or a topical dictionary. More detailed information about the language, the project, and the team can be accessed from here.

3. SUMMING UP. The general portal to the DoBeS archive will facilitate access to users, create new user groups, and attract an interested audience. It will hopefully turn the DoBeS archive into a resource that various user groups will consult regularly. In the process, monitoring of users and uses of the page will allow us to adapt its structure and content to suit the needs of those users. In due course, more features can be incorporated, e.g. the possibility of working with the materials online, that is adding one's own annotations to the data in the archive.

Gabriele Schwiertz
gabriele.schwiertz@uni-koeln.de

[6] http://www.mpi.nl/dobes/deposit

18

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. 129–134
http://nflrc.hawaii.edu/ldc/sp03/
http://hdl.handle.net/10125/4527

Using language documentation data in a broader context

Nick Thieberger
University of Melbourne / PARADISEC

On the one hand we have never seen as much fieldwork and recording of small and endangered languages as we have over the past decade. On the other hand linguists are now also much more aware of the need to create records that can be reused by the people we record and that will still be available for their descendants. Our own descendants, the future researchers who will use our records, will also need to be able to find and make use of our research. The fragility of digital records means we need to pay attention to their curation over time and create suitable repositories if they do not already exist. In order for these aims to be achieved, we need to establish work practices now that allow the data to move easily from creation to the archive and to community use.

1. How can language documentation data be utilized in a broader context? There are obviously many ways in which language documentation data can be used, but first they must be created in ways that permit ready reuse. That is, the data must exist in formats that allow them either to be used immediately or to be converted to a usable form without too much effort. Reworking data by hand is too time consuming, and '[t]he amount of data produced far exceeds the capabilities of manual techniques for data management.' (Borgman 2007: 6). To automate handling of linguistic data, it needs to be created in appropriate structures to begin with. This implies that we have tools that produce output in reusable forms and that linguists have training in what it means to create reusable data. We need to participate in broader initiatives in the Humanities that are leading to the development of the necessary infrastructure to house our research outputs in the longterm. Beagrie et al. (2008: 22) suggest that community-wide agreement on standards enhances the preservability of the data generated by that community. In linguistics we are fortunate to have some established standards (e.g., Leipzig glossing rules[1], OLAC[2], and IMDI[3], among others), but there is variable uptake of these standards and a corresponding lack of understanding of the reasons for using them. This paper addresses ways of using linguistic data based on an assumption that more use can be made of data that is created adhering to such standards.

[1] http://www.eva.mpg.de/lingua/pdf/LGR08_09_12.pdf

[2] http://www.language-archives.org

[3] http://www.mpi.nl/IMDI/

The creation of this data requires specialized knowledge of data formats, which should be provided by training researchers in documentation methods. Alternatively, linguists may be able to use tools that provide well-formed data and so bypass a need to really understand the underlying form of the data they are creating. This is typically the way in which most current linguistic data is created, using tools such as Elan[4] or Toolbox[5] and their output text files rather than proprietary formats as produced, for example, by Microsoft software. With training in the use of the tools, linguists can generate the kinds of material they need and at the same time allow the data to be archived and reused. Such training has been provided at organized summer schools or training workshops or intensive workshops associated with linguistic conferences (such as those run by InField[6], RNLD[7], LLL[8], ELDP[9], or DoBeS[10]).

While those working in projects funded by ELDP or DoBeS routinely receive training, the majority of language documentation is still being carried out by linguists trained in descriptive fieldwork and not in documentation methods. For example, Newman (1992, 2005) reports 34 US departments running field methods courses, most of which it can be assumed did not include documentary methods such as time-aligned transcriptions of the primary recordings, metadata catalogs of files generated in the project, annotation of files with interlinear glossed text and lexicons in open and reusable formats, or preparation of the material for long-term archiving.

However, if we take into account that there were some 230 papers at the 2nd International Conference on Language Documentation and Conservation and 180 papers at the LLL conference in 2009 there would appear to be quite a great deal of activity in language documentation projects. We can assume from this that there are at least 100 current fieldwork-based linguistic projects so, if we extrapolate back with a conservative estimate of 50 new projects per year since 1960, there should be some records of 2,500 small languages. Without a concerted effort to describe and curate the outputs of this research, many of the primary recordings prepared by these linguists have been or will be lost, if they ever existed to begin with (keeping in mind that, until recently, it was quite acceptable – and even advocated by Dixon (2006) – to use as little technology as possible, to make few or no field recordings and to make no provision for their long-term accessibility).

Some linguists have voiced concerns about the use of new technology in recording language material. The concerns range from the alleged neo-colonial nature of the use of high technology tools in third-world settings (Aikhenvald 2007) to the notion of commodification of language in an 'audit culture' (Dobrin et al. 2009) and include a critique of what they perceive as an excessive focus on the tools rather than on the language. It is tempting to ignore these arguments in the hope that they will be overtaken by the practicalities of doing fieldwork, and in the fear that reproducing them will give them more credence than they should receive. Unfortunately, it is still necessary to point out that any professional needs to

[4] http://www.lat-mpi.eu/tools/elan/manual/

[5] http://www.sil.org/computing/toolbox/

[6] http://darkwing.uoregon.edu/~spike/Site/InField_2010.html

[7] http://www.rnld.org

[8] http://www.ddl.ish-lyon.cnrs.fr/AALLED/Univ_ete/Summer_school.html

[9] http://www.hrelp.org/events/workshops/

[10] http://www.mpi.nl/DOBES/training_courses/

use current technologies for their work, and that for linguists this means learning and using new tools and methods for fieldwork and analysis.

2. HOW MUST THESE DIGITAL DATA BE STORED, REPRESENTED, AND MADE ACCESSIBLE BY THE ARCHIVES? Accessibility is based on locatability of the material in the collection. This relies on catalogs that use standard terms and are accessible via normal search mechanisms (e.g. Google, or the Open Archives Initiative (OAI)[11]). Non-compliant repositories – those whose catalogs do not conform to the normal standards (e.g., OLAC) of our research community – should be encouraged to conform. The benefit of conforming is that federated searches will include these repositories and we will thus begin, as a community of researchers, to create a dynamic documentation index of archived information about the world's languages, as provided by OLAC, and then harvested by any similar project (e.g. the Clarin Virtual Language Observatory[12] or ELCat[13]).

At the time of writing, two large-scale funding programs, ELDP[14] and DoBeS[15], have facilitated a great deal of research. The ELDP has funded 216 projects and DoBeS has funded 51. However, the ELDP archive, ELAR, contains (at the time of writing) just 110 collections, which raises the question, if a funding body like the ELDP cannot get all of its grantees to deposit in an archive in a timely fashion (or at all), how can unfunded researchers be expected to make use of an archive? One solution could be better training and raising the awareness of researchers of the need to archive, including appealing to their responsibility to make records for others to access in future. Another approach would be to make it as easy as possible to deposit in archives by use of new metadata creation tools (like, for example, ExSite9[16] or Arbil[17]).

For objects in repositories that will not be able to conform (that is, state libraries, archives, or similar institutions), a service could be built that indexes language material in these collections and assigns simple descriptors (like standard language codes) with links to the URI of the object. For example we can create a record in the PARADISEC catalog linking to an item found on the Anglican Church website that is written in Raga, a language from Vanuatu[18]. PARADISEC approached the website owners to ask if they would consider archiving their primary material, but had no response from them. Trusting that the pages on the Anglican Church website have some persistence, the links from the PARADISEC catalog will allow this item to be found in a federated search of all OAI archives.

The OLAC search page[19] provides a targeted search tool for language material, but is only as good as the collections it harvests. OLAC currently includes forty digital language

[11] www.openarchives.org

[12] http://catalog.clarin.eu/ds/vlo

[13] http://www.endangeredlanguages.com/

[14] Endangered Languages Development Programme, http://www.hrelp.org/languages/

[15] http://www.mpi.nl/DOBES

[16] ExSite9 (formerly FieldHelper) is a tool being produced in Sydney for creating standard metadata via drag and drop menus, avoiding data entry and harvesting as much information from within files as possible. A beta version is planned for mid-2012.

[17] http://www.lat-mpi.eu/tools/arbil

[18] http://www.language-archives.org/item/oai:paradisec.org.au:External-Raga

[19] http://search.language-archives.org

archives of which fourteen were active (that is, they had material deposited) within the past six months, and nineteen more were active within the past twelve months. This means seven were inactive for the past twelve months. Clearly we need more archives and more that contribute to OLAC.

Once an item has been located it should be available for use if possible (assuming that issues around rights management have been dealt with). Examples are the DOBES data sets, or PARADISEC's online collections of papers by Capell[20], Wurm[21], and Roesler[22]. By late-2012 PARADISEC will provide streaming access to most of its collection.

Accessibility also implies that analog material is digitized. While newly created linguistic records are typically digital, a great deal of legacy material exists only in analog forms and so is outside of the scope of much current language archive infrastructure. For these older materials, we need an effort of discovery and digitization as argued by Schüller, who notes that '80% of the world-wide holdings representing the cultural and linguistic diversity of mankind are not held by audiovisual archives proper' (Schüller 2004: 9), and, further, that analog recordings are in urgent need of digitization if they are to be playable at all.

3. WHAT KINDS OF USES WILL EVOLVE IN THE CONTEXT OF THE SOCIAL MEDIA?

The uses of linguistic data can be online or offline. There are two kinds of online use that need to be distinguished. The first deals with online material as the authoritative archival source. For online use of data there must be persistent location and identification that allow citation and resolution of links, which requires proper repositories with a longterm commitment to curating the material. As can be seen from the figures given earlier, there is too little use of existing language archives (and perhaps a need for more such archives to be established), so, while social media can play a role in dissemination or publicity, without long-term repositories, the data are at risk of loss. Once people start combining data from disparate sources (which could be 'mashups' or could, for example, involve correlating transcripts and media in 'compound objects'[23]), they will create new research objects that themselves may need to be identified and curated in archives (it may be that not all online interactions in small languages necessarily need to be archived in perpetuity).

The second use of online data relates to its use in presentation systems, which may (but need not) be ephemeral. Distinguishing these uses is critical as there are many examples of considerable effort being devoted to presentation systems for community use ('mobilization') which are then lost as the delivery system (which could be proprietary software or websites that are no longer maintained) becomes unusable. If the 'mobilized' material is unique, it poses real problems for longevity, but if it is derived from already archived material it is, essentially, ephemeral.

Offline use is likely to be most relevant to speakers of the languages recorded, given the lack of affordable – or indeed any – internet access. Such offline use of language records includes printed outputs and media on CD, DVD, or in computer-based (e.g. iTunes) formats. The formats in which data are initially created during fieldwork are crucial here too:

[20] http://paradisec.org.au/fieldnotes/AC2.htm

[21] http://paradisec.org.au/fieldnotes/SAW2/SAW2.htm

[22] http://paradisec.org.au/fieldnotes/ROES/web/roes.htm

[23] http://www.openarchives.org/ore/

well-formed and predictable data can be readily converted from an archival form to a deliverable form. For example, a dictionary of a language should be derived from a structured lexical data set, as in Toolbox. Similarly a set of texts for production in a book can be derived from a set of interlinear glossed texts in Toolbox. Books can now be produced relatively cheaply in publish-on-demand systems[24], with downloadable versions of the pdf file available via a suitable online repository. Media for a CD or DVD can be readily converted from high-resolution WAV or JPEG2000 files to playable formats for delivery on a CD or used in iTunes installations.

4. CONCLUSION. To conclude, having made some inroads into the production of enduring and reusable records of endangered languages, we still have a long way to go. There is a need for research into existing and emerging methods and development of tools both for creating linguistic data and then for making it useful. There is a need for data management skills to be developed among linguistic scholars so that our relatively small collections can be maintained. We need good descriptive systems (metadata) and simple systems for metadata entry as well as more repositories to hold the material. We all know of projects which have been completed and for which there are now large datasets that are not being properly maintained. The few present language archives are already stretched and cannot go looking for collections, but without such active seeking many collections will be lost.

For those outside of the present training and funding regimes there is a great need for advocacy to promote good practices in working with digital data and to bring to their attention ways of working that will make their work easier and will also have better outcomes for data sharing or reuse. This means more training in both academic and community settings and more sharing of experience and methods (using lists like RNLD, for example). Linguistic archives (e.g. PARADISEC, ELAR, and DoBeS) typically provide advice and support via their webpages and via regular training courses. There needs to be much more activity to allow new researchers to build their own collections and to assist established researchers in describing existing collections. Finally, we need to create incentives for creating the kinds of collections described here. Such incentives include academic recognition of the effort put into building and describing one's research collections and then lodging them in a suitable repository. Citing data from its archival source will also enhance the visibility of collections, and we should now, as authors, reviewers, and editors, encourage the use of such citations in academic papers.

REFERENCES

Aikhenvald, Alexandra Y. 2007. Linguistic fieldwork: Setting the scene. *Sprachtypologie und Universalienforschung - STUF (Special Issue: Focus on Linguistics Fieldwork, ed. Alexandra Y. Aikhenvald)* 60(1). 3–11.

Beagrie, Neil, Julia Chruszcz & Brian Lavoie. 2008. *Keeping Research Data Safe: A Cost Model and Guidance for UK Universities*. London: Higher Education Funding Council for England.

[24] My collection of texts in South Efate was produced in this way and is for sale online as a book here: http://www.bookshop.unimelb.edu.au/cbc/p?IS.9781921775505, or for free download as a pdf file here: http://repository.unimelb.edu.au/10187/9734.

Borgman, Christine L. 2007. *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*. Cambridge, MA: MIT Press.

Dixon, R.M.W. 2006. Dixon Accepts Bloomfield Award. SSILA [The Society for the Study of the Indigenous Languages of the Americas] Newsletter Number 236: 6.

Dobrin, Lise M., Peter K. Austin & David Nathan. 2009. Dying to be counted: the commodification of endangered languages in documentary linguistics. In Peter K. Austin, Oliver Bond & David Nathan (eds.), *Proceedings of Conference on Language Documentation and Linguistic Theory*, 59–68. London: School of Oriental and African Studies. http://www.hrelp.org/publications/ldlt/papers/dobrin_austin_nathan.pdf (21 March, 2012).

Newman, Paul. 1992. Fieldwork and Field Methods in Linguistics. *California Linguistic Notes* 23(2). 3–8.

Newman, Paul. 2005. Field methods courses in linguistics. Paper presented at the Linguistic Society of America conference on Language Documentation: Theory, Practice, and Values. July 9–11, Harvard University.

Newman, Paul. 2009. Fieldwork and field methods in linguistics. *Language Documentation and Conservation* 3(1). 113–125.

Schüller, Dietrich. 2004. Safeguarding the Documentary Heritage of Cultural and Linguistic Diversity. Language Archive Newsletter 1(3), 9–10.

Nick Thieberger
thien@unimelb.edu.au