LANGUAGE DOCUMENTATION & CONSERVATION

Vol. 10 (2016), pp. 629–640
http://nflrc.hawaii.edu/ldc
http://hdl.handle.net/10125/24720
Revised Version Received: 11 Nov 2016

Mettouchi, Amina, Martine Vanhove, and Dominique Caubet (eds.). 2015. *Corpus-based studies of lesser-described languages. The CorpAfroAs corpus of spoken AfroAsiatic languages.* Amsterdam/Philadelphia: John Benjamins. 338 pp.: 1.9 x 16.5 x 24.1 cm. ISBN: 978-90-272-0376-2. US $149.00,Hardcover.

Reviewed by Stefan Schnell, *University of Melbourne*

Linguistic research on typologically diverse and often lesser-described languages is increasingly becoming corpus-based. Many grammars of lesser-described languages draw on rich collections of more or less naturally occurring spoken or written texts (e.g. Hyslop's (2001) grammar of Lolovoli (Vanuatu) or Wegener's (2012) grammar of Savosavo (Solomon Is.)) rather than elicited isolated words, phrases, or sentences. In recent years, corpora of well- and lesser-described languages also form the basis of cross-corpus typological and theoretical studies, for instance Haspelmath et al (2014) or Seifart (2015). Corpus-based grammaticography and typological research of this kind may be seen as paving the way to a novel kind of linguistics, enabling detailed investigations of regularities of language use and its relationship to language systems. Given that corpora are accessible to other researchers, research results obtained through corpus-based studies are in principle fully accountable, thus enhancing the empirical soundness and scientific validity of linguistic research. In addition to complying with these demands of scholarly ethics, accessible published text corpora may also prove useful for other researchers, possibly in ways not foreseen by the original compiler of the corpus. An illustrative example of this kind is mentioned by Mosel (2006:53):

> Secondly, a text collection would give colleagues the opportunity to discover grammatical phenomena the linguist did not recognize, or did not have time to cover. In my comparison of Tolai and Tok Pisin (Mosel 1980), for instance, I overlooked the similarity between the Tolai particle *iat* and its Tok Pisin equivalent *yet*. I also did not describe the function of *iat* in my book on Tolai syntax (Mosel 1984) […]. But on the basis of my text edition (Mosel 1977), Sankoff (1993) was able to identify Tolai *iat* as a focus marker and relate it to Tok Pisin *yet*.

The insight that a single researcher will usually not discover everything and that corpora may reveal regularities that are not always foreseeable from a philological or descriptive perspective at any given point in time is closely related to the focus on corpus compilation in language documentation (Himmelmann 1998). Although Cox (2011) warns that language documentation and corpus linguistics draw on quite different research traditions and practices, he indicates that (parts of) text collections compiled in language documentation projects can be turned into more structured text corpora.

E-ISSN 1934-5275

The last 15 years have seen the compilation of large, diverse digitally archived text collections compiled as part of numerous language documentation projects in language documentation programs like DoBeS or ELDP. Building on these existing and further data collections, the development of modern language corpora from documentations of typologically diverse languages (see Mosel to appear) is very likely to expand over the coming years, and is likely to become one of the major developments in empirically grounded typological linguistics.

The creation of well curated and structured archived language documentations and corpora built thereupon, however, remains a standing challenge, involving tremendous amounts of academic effort (see Thieberger et al. 2016 for discussion). Hence to this day, easily accessible corpora that can be deployed systematically in corpus-based linguistic research remain relatively rare. In this context, the publication of *Corpus-based studies of lesser-described languages. The CorpAfroAs corpus of spoken AfroAsiatic languages* edited by Amina Mettouchi, Martine Vanhove, and Dominique Caubet should be a landmark contribution to the area of corpus-based typological studies. The book is one of the outcomes of a large-scale corpus development project funded by the French *Agence Nationale de la Recherche* from 2006. The book's goals are two-fold—on the one hand, it seeks to explain the design of the *CorpAfroAs* corpus, a multi-language corpus comprising sub-corpora from 15 Afroasiatic languages, in view of a number of research questions. This aspect can be regarded as something like a mine of ideas for future corpus builders. On the other hand, and more crucially given the title of the book, it aims at demonstrating the potentials of this corpus for modern linguistic research into lesser-described languages by way of illustrative sample studies based on *CorpAfroAs*. The preface, written by the three editors, describes the corpus design, and outlines its research goals and the aim of the book at hand. Each language corpus comprises one hour of spoken text, with a proportion of 60% monologues and 40% dialogues. The recordings are annotated in ELAN on seven tiers (i.e., levels of annotation). In addition to a reference (of time-aligned annotation segments), (broad phonetic) transcription, and free translation tiers, they comprise tiers for morphosyntactic words, morphemes, and morpheme-by-morpheme glosses, as well as a further tier labelled RX, which contains parts-of-speech tags as well as any further linguistic information. This latter RX tier can be seen as a unique feature of this *CorpAfroAs* corpus. Its rationale is outlined in later chapters to be discussed below. Repeatedly throughout the book, the *CorpAfroAs* is called a 'pilot corpus', hence the project is still developing further.

In this review I confine myself to discussion of the book and discussion of corpus design and content therein. As regards the corpus project as such, it is noteworthy that some corpora are (part of) the first modern documentations of respective languages. According to the authors, almost no modern text corpus from any Afroasiatic language existed at the time when the project was designed, hence the development of *CorpAfroAs* corpus is truly pioneering. Even outside of Afroasiatic linguistics, the corpus is among the first of its kind in compiling larger amounts of spoken text data. A similar corpus building project is the C-ORAL-ROM project (Cresti & Moneglia 2005), some aspects of which have had some influence on *CorpAfroAs*.

The annotation of individual texts is segmented and time-aligned along perceived intonation units, and the transcription tier renders phonological words. The definitional properties of these two units as well as the higher-order phonological units of paratone and period, and the annotation of their boundaries, are outlined by Sholomo Izre'el and Amina Mettouchi in the first chapter of *Part 1. Phonetics, phonology and prosody*, entitled 'Representation of speech in *CorpAfroAs*'. The authors then present "[…] their outcome in terms of scientific breakthroughs […]" (p 13) in stating the potentials for future research, for example that the distinction between phonological and morphosyntactic word will allow the study of sandhi and similar phenomena at the syntax/phonology interface. Likewise, prosodic segmentation enables the study of interfaces between intonation, syntax, information structure and discourse.

In the second chapter of Part 1, Bernard Caron discusses *Tone and intonation*, focussing on the relationship between lexical and grammatical tone and prosodic prominence in tone languages. His study deals exclusively with Zaar, a Chadic language with three phonemic tones (high, mid, low), the combination of two of which (high-mid, low-mid) results in two contour tones. Relating his findings to the typology of prosody in tone languages, he concludes that Zaar is a mixed language, showing both the stacking of intonation patterns over lexico-grammatical tones and intonation effects at the periphery of utterances (p 59). All patterns of tone and intonation interactions are illustrated with an example from the Zaar sub-corpus. The two chapters have somewhat different goals: while the first merely outlines the potentials of a spoken-language corpus that takes note of different prosodic units, the second actually demonstrates the deployment of the corpus by presenting findings from a small pilot study that make an immediate contribution to the typology of tone and intonation.

Part 2. Interfacing prosody, information structure and syntax comprises the two longest and most detailed chapters of the volume, with 55 pages each. Also, these two chapters are designed as larger case studies of specific linguistic phenomena that demonstrate the innovative character of the *CorpAfroAs* project, and I will discuss them in some more detail. In the first of these, *The intonation of topic and focus*, Bernard Caron, Cécile Lux, Stefano Manfredi and Christophe Pereira describe the system of prosodic marking of topic and focus and its interaction with morphosyntactic structure in four languages: Zaar, Tamasheq, Juba Arabic and Tripoli Arabic. After the concepts topic and focus are introduced in a short (2.5 pages) introductory section, four longer sections describe the basic prosodic system, including for example the prosodic properties of declarative sentences and different types of question, and the marking of (contrastive, shifted, or new) topics and focus in the four respective languages, one after the other. In addition to intonation, the authors point to morphosyntactic means of information-structure marking (where applicable) and correlations between intonation and grammatical relations subject and object, and their prosodic properties. Each structure identified for each language is exemplified with one or two examples from respective corpora; in some instances, the relevant structures are exemplified with data that is not part of the *CorpAfroAs* corpus, for instance prepositional topic phrases in Tripoli Arabic. A short concluding section

(1.5 pages) summarises the result of these pilot studies, which comprise systematic differences and commonalities between the four languages. The differences concern the systematic use of morphosyntactic and prosodic means in marking information structure categories in different languages: for example the exclusive use of the latter for 'unspecified topics' (roughly: frame setters) in opposition to all other information structure articulation differs from the interaction of intonational and syntactic means of topic and focus marking in Tamasheq. Other cross-linguistic differences concern marking in different types of question formation. The two main commonalities identified across the four languages are 1. the same bell-shape-curved intonation contour in thetic sentences, and 2. the marking of (contrastive, new) topics involving left-dislocation and an intonation boundary (although they are made up of different features across languages).

In the second chapter of *Part 2 Quotative constructions and prosody in some Afroasiatic languages*, Il-Il Malibert and Martine Vanhove test Genetti's (2011) hypothesis of a prosodic integration cline of direct speech reports, found in Dolakha Newar. For one thing, they extend the claim to other languages, namely four languages of the *CorpAfroAs* corpus, Beja, Zaar, Juba Arabic and Modern Hebrew. For another thing, they test whether variation in the degree of prosodic integration differentiates between direct and indirect speech reports. Similar to the preceding chapters, the authors give a short two-pages outline of basic concepts and terminology, and then turn to a description of the four languages in individual subsections. Each section starts with an outline of basic grammatical features, and then turns to more specific structures of reported speech and their prosodic integration, which are illustrated with examples from the *CorpAfroAs* corpus. In their summary of findings and conclusions, the authors first contrast the four language corpora according to some baseline findings concerning the prevalence of speech reports generally, and the relative proportion of direct and indirect speech reports. The Beja corpus—for which no conversations could be recorded "because of social rules of politeness and honour" (p 124), and which therefore consists only of narratives—reported speech is three times more frequent than in the corpora from Zaar and Juba Arabic, and even ten times more frequent than in the Modern Hebrew corpus. In addition to possible impacts of the type and content of narratives and conversations in the four corpora, the authors indicate that this might be explained by the fact that only Modern Hebrew has a long-established written variety, at least "[i]t is noticeable that the highest proportions of reported speech occur in the three unscripted (or very recently scripted) languages of the sample" (p 164). Likewise, the corpora show clear preferences for direct versus indirect speech reports: while direct speech reports are not attested in the Beja narratives and very rare in the Juba Arabic corpus, they account for roughly one-third of all speech reports in Modern Hebrew and Zaar. Instances of speech reports are, however, too rare in the Modern Hebrew corpus to be significant. As for the prosodic integration cline, there appear to be some differences between direct versus indirect speech reports, and across languages, for instance that in Zaar direct speech reports show a higher degree of prosodic integration than indirect ones, while in Modern Hebrew the reverse seems to hold. The authors point out that their current database

is too small to allow for clear conclusions at this stage, and that these preliminary observations need to be tested against much larger amounts of corpus data. Nonetheless, the authors propose a preliminary rudimentary typology of prosodic integration in speech reports: generally, the presence of a morphosyntactic clues correlate with less prosodic integration so that, for instance, the absence of a complementizer correlates with high prosodic integration, and vice versa. That is to say that prosody and morphosyntax complement each other across languages in the marking of speech reports. Moreover, whether the left or right boundary of a speech report is set off from adjacent discourse by prosodic means depends on the word order: in SOV languages, the speech report precedes the quotative frame, and its left boundary is prosodically marked; in SVO languages, where the quotative frame precedes the speech report, it is the right boundary that is marked off by prosodic means.

The two chapters bring new data and interesting insights to the typological discussion of information structure and quotative constructions, which both play a central role in the organisation of spoken-language discourse. It is worth pointing out here that both chapters clearly focus on *system properties* of the respective languages involved, rather than patterns of *language use*. These are illustrated by selective examples from respective corpora, and only Malibert & Vanhove provide figures regarding the relative frequencies of the basic types of construction, ie., direct and indirect speech reports. The curious and novel insights into central aspects of spoken Afroasiatic languages notwithstanding, it seems to me that a corpus like *CorpAfroAs* bears much larger potential regarding investigations of language use, and the empirical groundedness of linguistic research. To illustrate this, I shall mention only two relatively random examples, there are many more aspects to be considered: regarding the marking of information structure, it would be desirable to have some more baseline information, for instance the relative proportion of different types of information structure articulations and types of questions. The authors confine themselves to seemingly impressionistic remarks, stating for instance that a structure is 'very rare', etc. Presentation of actual figures would not only have served the purpose of clear and explicit characterisation of the discourse documented in the four corpora at hand, it would in fact also be of some relevance for our understanding of information structure patterns in natural language discourse. For instance, Lambrecht (1994) (among others) states that the topic-comment articulation (with unmarked, continuous topics) is the most basic and unmarked one in natural language discourse. We may want to ask then whether this is reflected in a high frequency of respective topic-comment constructions in actual discourse, and how this frequency compares to that of marked topic ('topicalisation') constructions involving left-dislocation? The findings may be relatively trivial, possibly simply supporting Lambrecht's (1994) hypothesis. But given the scarcity of larger corpus data from a variety of languages, merely confirming such claims appears to be a worthwhile research finding on which further critical work can build. And despite the possibility that global trends support the general view, we may nonetheless discover slight differences across corpora or texts that may lead to new exciting questions, thus ever advancing our understanding in a field that has to date often drawn on relatively small and selective amounts of text

data where lesser-described languages are concerned. Such findings may be relevant for very specific, long-standing hypotheses: for instance, if we could ask whether left-dislocation ('topicalisation') constructions appear to be used more frequently and in contexts where they do not seem warranted by information structure considerations, and would thus contribute to a critical evaluation of Givón's (1976) famous hypothesis that subject agreement evolves universally from the 'overuse' of topicalisation constructions. Likewise regarding (argument) focus, it would be interesting to see how prevalent it is, and to what extent its 'markers' are involved in contexts where we cannot identify a focus relation; this would be particularly relevant for critical assessments of focus marking analyses across languages, as discussed in a landmark paper by Matić & Wedgwood (2013).

Another question that appears to be left open in the first of the two chapters is how all the relevant instances of information structure can be retrieved from the *CorpAfroAs* corpus: we can obviously search for particles that are analysed as marking different types of focus in Juba Arabic and Tripoli Arabic, since their forms appear on the MOT tier and respective glosses on the GE tier. But where just intonation is deployed, it is less clear to me whether and how respective patterns can be retrieved from corpus annotations: although boundaries of and breaks in intonation contours are marked on the TX (broad phonetic transcription) tier, pitch peaks and lows are apparently not. Without such prosodic labelling, it appears, readers are in fact not put in a position to falsify relevant findings. This type of prosodic annotation would be very much desired. It will allow readers to retrieve all relevant instances of intonational patterns and to ascertain that relevant articulations do in fact form significant intonation contours, and will also enable quantitative analyses.

The second chapter on quotative constructions gives much more detailed information on the use of particular constructions, and also asks some corpus-linguistic questions about it, for instance when the authors indicate possible explanations for the prevalence of certain constructions in narratives versus conversations. Their conclusions are nonetheless directed towards the language systems as a whole, and one might wonder here whether some of these may also have a usage-related side, for instance, the authors find a correlation between the availability of complementizers in a given language and the degree of prosodic integration. This seems to leave room for further detailed examination of specific instances in languages where complementizers are optional, and where we might expect a correlation between the *use* of a complementizer and a lower degree of prosodic integration and vice versa. Presumably such a pattern would be the source of the typological regularities across language *systems* that the authors establish, and the availability of a multi-language corpus like *CorpAfroAs* calls unequivocally for examinations of this kind.

Part 3 Cross-linguistic comparability is concerned largely with more technical aspects of morphosyntactic analysis and cross-corpus comparability, but also—at least in part—indicates possibilities for specific research agendas. All four chapters deal primarily with glossing on the two tiers GE, which contains classical morpheme-by-morpheme glossing along Leipzig Glossing Rules (LGR) (Bickel et al. 2008), and RX, which is used for parts-of-speech tagging and also contains any other glosses

that trigger a wide range of linguistic phenomena. While the first two (*Glossing in Semitic languages* by Angeles Vicente, Il-Il Malibert and Alexandria Rarontini; *From the Leipzig Glossing Rules to the GE and RX lines* by Bernard Comrie) deal with technical and conceptual aspects of consistent glossing as such, the latter two chapters (*Cross-linguistic comparability in CorpAfroAs* by Amina Mettouchi, Graziano Savà and Mauro Tosco; *Functional domains and cross-linguistic comparability* by Zygmunt Frajzyngier and Amina Mettouchi) outline the actual deployment of this annotation design for cross-linguistic research.

The first two of these chapters show a considerable degree of overlap. Both chapters explain the motivations for morpheme-by-morpheme glossing, glossing practices for specific formal and functional categories, and the role of the RX tier (although the latter is more important in the first chapter). The two chapters are of major importance since they contribute to the scientific advancement of the *CorpAfroAs* project by making explicit and transparent in some detail the basic annotation practices. Some more editorial effort may have improved this part though, given that many readers of this book will be among the convinced already, the motivations for morphemic glossing (and its absence in the philological tradition with a much more confined scientific community) would have been sufficiently and nicely illustrated in Comrie's chapter. The practice of *CorpAfroAs* to annotate overt and covert morphological categories on two separate tiers rather than using the LGR conventions of bracketing covert ones could have been argued for more convincingly, ideally with reference to specific research projects for which this practice will make a significant difference. Some further specific assets of combining parts-of-speech tagging and other annotations on the RX tier are outlined towards the end of the first chapter, for example where it is stated that the combination of a POS tag for 'verbs' and that for 'derivative' will enable easy retrieval of derivative verb forms across corpora from different languages, which will be vital for a systematic investigation of this interesting and important area in Semitic languages. Possibly it would have been more useful for the reader to pick out some exemplary cases, and present an actual pilot study based on these glossing practices, rather than presenting a step-by-step walkthrough covering all major morphological categories in Semitic languages.

The second set of papers further illustrates the use of these annotations for cross-linguistic studies, although the last chapter in fact does not draw on *CorpAfroAs*, but rather outlines future plans for cross-corpus investigations based on a specific comparative approach. As for linguistic phenomena, the chapters cover directionals, case and gender, as well as reference. The authors present detailed descriptions of these respective systems in different Afroasiatic languages, and some quantitative findings relating to their use. They also relate these back to the annotation practices outlined in the preceding two chapters, and sketch out further corpus-based investigations. These outlines are very inspiring and may, as indicated by the authors at various points, spur some substantial research work in the future, but they also leave open some questions regarding the actual comparability of corpus annotations. Cross-corpus comparability is the major challenge in the development of corpus-based typological research and I therefore would like to take up an example from each chapter for which more

discussion may be necessary. The first concerns the *CorpAfroAs* practice of glossing language-specific categories rather than comparative concepts. This seems very practical, given that language specific glossing of this kind is often part of language documentation work anyway—and thus does not need to be done specifically for *CorpAfroAs*, except maybe for some adjustments—and that they are possibly done more easily and with the assistance of lexicon-based annotation tools (see discussion of Chanard's chapter below). But this practice also brings up the question of cross-linguistic comparability, and here the authors merely state the assumption that "[…] there is some degree of resemblance between a language-internal category and a comparative one […]" (222). While this is certainly true, the question remains whether this degree is sufficient for meaningful cross-linguistic comparison? And this I believe is ultimately an empirical question. Considering the example of case-marking, one may assume that, for example, dative or accusative case will encode arguments with similar syntactic and semantic properties across typologically diverse languages; but this is not very clear at all. A different approach is adopted by Haig & Schnell (2014): they advocate annotation of an additional layer of comparative concepts rather than language-specific formal categories, with glosses for argument functions in a comparative sense. They also include annotations of zero arguments where this seems warranted in specific discourse contexts. This further layer of annotation of course involves a lot of annotation work and related costs; if morphemic glossing can convincingly be shown to suffice for broad cross-linguistic comparison, then we should indeed prefer it over the costly creation of an additional layer of annotation. This clearly remains an open question, and it will be part of important future discussions in the area of cross-corpus typological linguistics. A similar point can be made about reference systems. Among possible forms of reference, the authors state "absence of a noun" (272) which I assume to mean something like zero anaphor. But it is not discussed in the second of these two chapters how instances of zero anaphor can be retrieved from the corpus. Since none of these chapters actually present fully-fledged studies of the phenomena involved, this is hard to evaluate. Another point regarding the sufficiency of the annotations undertaken in *CorpAfroAs* concerns the annotation of clause boundary. In their discussion of directional markers, the authors state a statistically significant co-occurrence pattern of one directional marker (affixed to the verb) with benefactive arguments. Determination of this pattern obviously requires consideration of clause boundaries, but these are not annotated in *CorpAfroAs*. How did the authors determine the relevant figures then? And how can a critical reader do so? While I would find it entirely legitimate to do these counts by hand, or in a way detached from corpus annotation similar to 'coding' in variationist sociolinguistics, for a specific study this would need to be made explicit. In the context of discussing further developments in the accountability and replicability of corpus-based cross-linguistic studies though, it may be worthwhile considering glossing of all those elements that are necessary to retrieve specific structures under discussion. In this sense, these contributions open up the ground for important discussions of practices in corpus building for cross-linguistic research, and in particular the trade-off between further annotations and sophisticated corpus exploitation.

In *Part 4 Language contact*, Stefano Manfredi, Marie-Claude Simeone-Senelle and Mauro Tosco investigate the linguistic integration of lexical borrowings and the prosodic and morphosyntactic constraints on code-switching. Their investigation illustrates the use of the RX tier as an additional layer of annotation that is open to host information of any kind. In this case, glosses for instances of borrowings and code-switching (BORR and CSW respectively) are entered on and retrievable from different corpora through searches on the RX tier. The former gloss is rarely implemented in *CorpAfroAs* though, and that has to do with one of the major findings of this study, namely that borrowings show such a high degree of linguistic integration into the recipient language systems that they are in fact not recognizable—and thus not glossed—as such (p 306). As for code switching, the authors identify two types that they show bear different prosodic properties: intersentential codeswitching—restricted more or less to Moroccan Arabic / French and Beja / Arabic bilinguals in this corpus—occurs at sentence boundaries, so that the code-switched sentence is independent as is also the preceding sentence. The code-switched sentence is also prosodically isolated and thus bears the properties of a monolingual intonation unit. Instances are stated to be very rare in the corpus, but figures are not provided. In intrasentential code-switching, the foreign elements occur towards the end of the intonation unit. They can be embedded clauses, phrases, or single words. This type of code-switching is much more common across different languages within the corpus, and also quantitatively in single corpora—however, again, no figures are given. These appear to be interesting and relevant findings regarding code-switching. The authors conclude, however, that they need to be confirmed by more systematic quantitative investigations. A noteworthy feature of this study is that the relevant features—code-switching and prosodic boundaries—are both retrievable through the annotations on the RX tier and the transcription tier (where prosodic boundaries of different types are annotated). Hence, the quantitative study envisaged by the authors can relatively easily be undertaken on the *CorpAfroAs* corpus data.

Finally in *Part 5 Information technology* Christian Chanard describes in his chapter 'ELAN-CorpA: Lexicon-aided annotation in ELAN' the development of a lexicon-based annotation tool within the ELAN software. Similar to the interlinearization process in tools like Toolbox/FLEx, the component involves a morphological parser and lexical database look-up function (pp 313–317). While this add-on to ELAN appears to improve on some details of the latter tools, in particular with regards to the notoriously problematic parser component, its most important advantage is obviously that interlinearized morpheme-by-morpheme glossing can be done immediately in ELAN. Thus, ELAN-CorpA would become a single platform for corpus building, comprising all layers of annotation, and sparing the user the import/export of annotations from/to ELAN and Toolbox/FLEx. This technical component thus greatly enhances annotation of *CorpAfroAs* texts, and since morphemic glossing is the central basis also for cross-corpus research in *CorpAfroAs*, it seems to make the corpus immediately usable for the research sketched out in the preceding chapters.

Despite some critical remarks throughout this review, I repeat here that this book is an important landmark on the road towards further developing corpus-based typol-

ogy. It outlines one of the most ambitious and systematic corpus-building projects in this context to date, and sketches out the possible implementation of its features for research in different areas of linguistics. Naturally, since one of the major characteristics of *CorpAfroAs* is that it comprises spoken texts, its potential for investigations into prosody and tone is given some more space than investigations of other phenomena heavily drawing on larger amounts of corpus data, for instance studies of discourse structure and referential choice. Although discussions of corpus building for areal and typological linguistics are naturally not resolved and completed with the appearance of this book, it can be seen as one of the first major attempts to draw together what is necessary for these developments on which future corpus builders may rely. This may lead to alternations of or additions to some of its design features once more extensive and comparable studies get underway. As indicated above, two key issues that will need further critical assessment seem to be 1. the instant retrievability and quantifiability of all structures necessary to replicate respective studies and allow for systematic analyses of language use, and 2. the interdependence of more or less extensive corpus annotation and comparability. The latter point, I believe, can be addressed only once more corpus-based studies of individual phenomena have actually been carried out and compared to findings based on other corpora with similar goals. Another aspect that will require more systematic discussion in the future is the conceptualisation of corpora as developmental entities. Rich archived documentations of diverse languages in repositories around the world essentially contain unstructured collections of texts resembling to a larger or smaller degree the overall range of communicative events characteristic of respective speech communities. Certain conceptual and practical problems notwithstanding, these collections are an unprecedented treasure trove for corpus builders who seek to convert them into more structured and usable corpora that are representative to some degree of the linguistic phenomena under investigation (whether language documentations can ever be representative of the communicative practices of a speech community appears to be questionable, see Mosel to appear for discussion). One of the key questions for future developments in this regard is what amount of text data, and of what kind, with what degree of text type variation, what layers of annotation etc. is required to arrive at a database that will allow for the investigations necessary to capture specific linguistic phenomena. Some of these aspects are indicated in parts of the book, and it may in the future be seen as a major landmark from where these further developments take their route.

References

Bickel, Balthasar, Bernard Comrie & Martin Haspelmath. 2008. *The Leipzig Glossing Rules. Conventions for interlinear morpheme by morpheme glosses*. Revised version of February 2008. Leipzig: Max-Planck-Institut für Evolutionäre Anthropologie.

Cox, Christopher. 2011. Corpus linguistics and language documentation: Challenges for collaboration. In John Newman, Harald Baayen & Sally Rice (eds.), *Corpus-based studies in language use, language learning, and language documentation*, 239–264. Amsterdam: Brill.

Cresti, Emanuela & Massimo Moneglia (eds.). 2005. *C-ORAL-ROM: Integrated Reference Corpus for Spoken Romance Languages* [Studies in Corpus Linguistics 15]. Amsterdam: John Benjamins. doi:10.1075/scl.15.

Genetti, Carol. 2011. Direct speech reports and the cline of prosodic integration in Dolakha Newar. *Himmalayan Linguistics* 10(1). 55–71. Special issue in memory of Michael Noonan and David Watters.

Haig, Geoffrey & Stefan Schnell. 2014. Annotations using GRAID (Grammatical Relations and Animacy in Discourse). Guidelines for annotators. Version 7.0. https://lac.uni-koeln.de/de/multicast-research-and-publications/.

Haspelmath, Martin & Andrea S. Calude & Michael Spagnol & Elif Bamyaci. 2014. Coding causal-noncausal verb alternations: A form-frequency correspondence explanation. *Journal of Linguistics* 50(3). 587–625.

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

Hyslop, Catriona. 2001. *The Lolovoli dialect of the North-East Ambae language, Vanuatu*. Canberra: Pacific Linguistics.

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

Matić, Dejan & Daniel Wedgwood. 2013. The meanings of focus: The significance of an interpretation-based category in cross-linguistic analysis. *Journal of Linguistics* 49. 127–163.

Mosel, Ulrike. 1977. *Tolai texts. Kivung* 10/1-2. Port Moresby: Linguistic Society of Papua New Guinea.

Mosel, Ulrike. 1980. *Tolai and Tok Pisin. The influence of Substratum on the development of New Guinea Pidgin*. Pacific Linguistic B-73. Canberra: Australian National University Press.

Mosel, Ulrike. 1984. *Tolai syntax and its historical development*. Pacific Linguistics B-92. Canberra: Australian National University.

Mosel, Ulrike. 2006. Grammaticography: The art and craft of writing grammars. In Felix Ameka, Alan Dench and Nicholas Evans (eds.), *Catching language: The standing challenge of grammar writing*, 41–68. Berlin & New York: Mouton de Gruyter.

Mosel, Ulrike. to appear. Corpus compilation and exploitation in language documentation projects. In Kenneth Rehg & Lyle Campbell (eds.), *The Oxford handbook of endangered languages*.

Sankoff, Gillian. 1993. Focus in Tok Pisin. In Francis Byrne & Donald Winford (eds.), *Focus and grammatical relations in creole languages*, 117–140. Amsterdam & Philadelphia: John Benjamins.
Seifart, Frank. 2015. Direct and indirect suffix borrowing. *Language* 91(3). 511–532.
Thieberger, Nicholas, Anna Margetts, Stephen Morey & Simon Musgrave. 2015. Assessing annotated corpora as research output. *Australian Journal of Linguistics* 36(1). 1–21. doi:10.1080/07268602.2016.1109428.
Wegener, Claudia. 2012. *A grammar of Savosavo*. Berlin: De Gruyter Mouton.

Stefan Schnell
stefan.schnell@unimelb.edu.au