LANGUAGE DOCUMENTATION & CONSERVATION

Vol. 9 (2015), pp. 344–350
http://nflrc.hawaii.edu/ldc
http://hdl.handle.net/10125/24669

Mari C. Jones. 2014. *Endangered languages and new technologies*. Cambridge: Cambridge University Press. 228 pp. ISBN: 978-1107049598. Price $99. Also available as e-Book ($79).

Reviewed by Daniel W. Hieber, *University of California, Santa Barbara*

The perceived relationship between technology and minority languages has changed drastically over the years. Linguists previously viewed technologies like television and radio as major drivers of language shift—veritable 'cultural nerve gas' (Krauss 1992:6). Today, by contrast, the role of technology is viewed primarily as a positive, enabling one. The existence of the present volume demonstrates just how far those views have shifted. The fundamental premise of this book is that, "The ever-increasing availability of new technologies, from visual to aural archiving to digitization of textual resources and electronic mapping, have the potential to revolutionize the documentation, analysis and revitalization of endangered languages for the linguist and indigenous community alike" (xiii). Of course, the editor and the collective authors do not view technology as a cure-all for language shift, and are appropriately cautionary in their suggestions for the application of technology. But the central message of this book is a hopeful one: the suite of ever-cheaper and ever-higher-quality tools available today can, with appropriate sensitivity to the cultural contexts in which they are applied and to their congruence with community goals and resources, be a powerful tool in the reemergence and flourishing of minority languages.

This hopeful message shines through most prominently in Nicholas Ostler's excellent introduction, 'Endangered languages in the New Multilingual Order *per genus et differentiam*.' Ostler argues that the forces which previously gave mega-languages—and especially English—their status are becoming less relevant: "Since linguistic dominance of this kind is always based on past social dominance (military, economic, cultural, or religious), and the social factors that favoured English-speakers over others are losing force, there is scope for change in the multilingual order of the world" (1). In support of this view, Ostler notes that the regions of the world which have shown the most rapid economic advancement this century are precisely the non-English speaking ones, and that even English's predominance on the internet is waning. Moreover, as automated translation becomes more robust, choice of language will become more a matter of preference than necessity. Ostler sees the potential for what he calls the New Multilingual Order, a world where, "English will carry on as a useful lingua franca, a support mechanism, but one that will be increasingly unnecessary. However, the direction of flow […] is increasingly towards a world where choice of language will express its inherited position and felt loyalties within the human race, even to quite small groups. In this world of aspiration, all will speak as they like, and yet the world will understand them" (12–13).

This idea of technology as the great leveler is one of several themes that recur throughout the book. In Tjeerd de Graaf, Cor van der Meer, and Lysbeth Jongbloed-Faber's chapter, "The use of new technologies in the preservation of an endangered language: The case of Frisian," they report that the cheap cost of distance-learning technologies has given Frisian language learners access to qualified language instructors, whereas prior to the prevalence of

E-ISSN 1934-5275

these technologies, hiring a language instructor would have been financially impractical due to the small number of pupils able to attend class in any given locality. Distance-learning technologies have enabled Frisian language instructors to take advantage of economies of scale, making paid language instruction economically viable (143).

In general, tools developed for well-resourced majority languages are often utilizable by minority language communities, who then do not need to undertake the fixed costs of development. The more prevalent the use of the tool among majority languages, the greater the infrastructure for it becomes, the lower its costs, and the easier it is for minority languages to utilize. Matt Coler and Petr Homola's chapter, "Rule-based machine translation for Aymara," is an excellent illustration of this process at work. As they note, even though the methods of machine translation were originally developed for well-studied languages, "there was no intrinsic reason why this should remain their focus in the long term" (10). The internet provides a repository of linguistic data on a rapidly growing number of languages, and all of this publicly-facing data is open to analysis. As Ostler notes in his introduction, "The net effect will be that the smaller 99 percent of the world's languages […] will have a corresponding opportunity to become accessible; the 'long tail' need no longer be disregarded" (2–3).

The adoption of existing technologies by minority languages is also an example of a broader pattern that emerges repeatedly throughout the book: the repurposing of a technology for uses other than its original intended ones. Anthony Scott Warren & Geraint Jennings report exactly this process in their chapter, "'Allant contre vent et mathée': Jèrrais in the twenty-first century." They describe how Les Pages Jèrrais, a website originally launched with the sole intention of making Jèrrais language material publicly available, came to function as a linguistic corpus for tracking usage, lexical variation, and neologisms, and also as a primitive spell-checker (136). Another kind of seldom-discussed repurposing that many linguists encounter is the repurposing of legacy archival materials for modern pedagogical materials or linguistic analysis. Jeffrey E. Davis makes this point especially salient in his chapter, "American Indian Sign Language: Documentary linguistic methodologies and technologies," where legacy video recordings constitute a significant component of his work to document and create pedagogical materials for American Indian Sign Language (173). Once these recordings were digitized, it was possible to conduct extensive lexical comparisons between legacy and modern sources. Thus, even legacy material in obsolete formats might still be repurposed for ends none of their creators might have imagined; one never knows what uses the data or tools you create might have. This fact, of course, should encourage caution in planning for the longevity of documentary materials, and being extra attentive to details of access rights and permissions.

Other notable examples of repurposing in this volume include: the reuse of a keyboard layout created by native Me'phaa writers to create keyboards for different operating systems (since the layout itself is a technological solution to a particular problem, and thus a type of technology) (50); the use of 6,000 tweets by fifty Frisian-speaking adults to create the beginnings of a Frisian spell-checker (149); and a set of Frisian-language DVDs which originally had limited distribution, but made their way onto YouTube, and are now used as pedagogical materials by teachers (143).

Other themes thread their way through the book as well, which is impressive given that the chapters themselves vary widely in their content and focus, and indeed in their very conception of 'technology.' With chapters on keyboards, corpus creation, machine-translation, data longevity and archiving, and many other topics, a potential reader could be forgiven for assuming that the volume lacks cohesion. But the strong contribution this book

makes to language documentation and revitalization is, in my opinion, not in the specifics of the individual projects outlined in the chapters themselves. That is not to say that the projects and ideas outlined in each contribution are not extremely valuable in their own right; it simply acknowledges the fact that not every documentary linguist is in a situation where the types of projects and technologies outlined in these chapters apply. Rather, what makes this book worth reading regardless of the relevance of the particular projects to the reader's own is the wealth of information and advice on design principles, project planning, practical and useful goal-setting, best practices in community collaboration, and accommodating community needs and cultural preferences. In the remainder of this review, I take up a few of the commonalities exhibited across different chapters, and what they have to teach us about the use of technology in language documentation and revitalization. Some of these tendencies I see as positive models to emulate, while others I offer as constructive criticism, but all of them provide valuable lessons for any project where technology plays an integral role.

The first theme is the authors' very conception of technology itself. The term 'technology' is actually quite vague, and could refer to everything from hardware to search algorithms to data formats or user interfaces, or many, many other things. Despite this, the contributors focused centrally on technology as realized through *tools*, that is, the particular programs, applications, or websites that assist users in performing certain tasks or operations. This is a decidedly end-user perspective on technology, which asks, 'What does technology allow me to *do*?' and it makes sense that linguists and community members, who are usually not themselves technologists, would be most concerned with this perspective. Technicalities and details of implementation are generally glossed over. In Aiméé Lahaussois' chapter "The Kiranti comparable corpus: A prototype corpus for the comparison of Kiranti languages and mythology," for example, the author explicitly states that, "what is advocated here is not a particular software configuration but, rather, a concept, the technical implementation of which could be realized in a number of different ways" (17–18). This is exemplary of the approach taken by the majority of the authors in this book.

While a purely conceptual understanding a given technological tool is useful, it is at the same time a bit unfortunate because linguists do ultimately have to confront the nitty-gritty details of implementation as well. To give just one example, few of the authors discuss user interfaces in any detail. However, a well-designed user interface can sometimes make all the difference between adoption of a tool or apathy towards it by researchers and community members. Hugh Paterson III makes this point extremely well in his chapter, "Keyboard layouts: Lessons from Me'phaa and Sochiapam Chinantec designs," and it is worth quoting at length:

> When language documenters and linguists build digital solutions such as keyboard layouts, they need to bear in mind that these products may have lasting effects on communities. As service providers, they have ethical and professional obligations to seek out not only solutions but great solutions. […] When linguistic and technical expertise is offered to communities of endangered language speakers and writers, we need to not only design solutions, we need also to offer well-designed solutions. Just because something is usable and useful does not mean it is desirable. When a speech community does not want to use a given input method (keyboard layout), the response should not be: 'Well, they simply don't want it enough.' Keyboard layouts are not just products, they are experiences (54).

Paterson goes on to exemplify this attention to usability with a helpful overview of design challenges in creating keyboards for minority languages.

Lahaussois' chapter on the Kiranti corpus, mentioned above, is also a good example of how writers can engage with and share details of implementation in a way that is useful to other researchers/revitalizers without being overly meticulous. It includes screenshots and descriptions of different interfaces in their database management tool, such as concordancing and side-by-side views of parallel/comparable texts. For an audience primarily focused on what tools allow them to *do*, this exposition of the tool is extremely useful, because it allows potential adopters to see precisely what they can *do* if they were to apply that technology to their own projects.

In general, however, the authors are not especially concerned to relate the implementational details behind their projects. Even Lahaussois' discussion of the Kiranti corpus, for example, does not tell us anything about what the programming world calls 'the stack,' or the suite of programming languages, operating systems, browsers, servers, etc. that are necessary to build and maintain a given product. Describing a software stack need not be tedious. Even a cursory overview can be extremely valuable for other project teams who might want to emulate a particular technological framework. Was the tool created with nothing more than JavaScript, a modern web browser, some coffee and a weekend, or did it require extensive server-side scripting and expertise in security and authentication? Does it use a data format on the backend that is already compatible with one's own data, or would using this tool require extensive manipulation and reformatting of that data? Knowledge of these facts can help project teams make more informed decisions about what is practical for them to achieve given the resources and expertise available to them, and helps them locate individuals with the proper skill sets for their project. Lahaussois for instance notes that for their project, "The goal of setting up a distributed network of databases accessible via the Internet proved too ambitious" (45). What we do not know is what tools they considered and what made those tools untenable. Similarly, Paterson notes, "The literature offers relatively little in terms of guiding principles for designers of keyboard layouts. The absence is not completely unexpected since human-computer interaction such as keyboarding is often treated and discussed as a sub-discipline of computer science or psychology [...] rather than of linguistics" (52). Thus the documentary and revitalization community as a whole would do well not to shy away from the technological details of implementation for the tools they use and create.

Moving away from constructive criticisms, one laudable theme that emerges from this volume is a focus on interoperability, i.e., the ability for tools to work with other tools, or work across different languages. An excellent example of interoperability at work is Sjief Barbiers' chapter "European Dialect Syntax: Towards an infrastructure for documentation and research of endangered dialects." This is the only chapter to discuss not just a project infrastructure, but also a broader collaborative research infrastructure, "where linguists can store and access the relevant data and where they can cooperate in the description and analysis of these data" (35). Barbiers goes on to provide some helpful implementational details illustrating a good principle for collaborative research teams to follow: decentralization of data management accompanied by centralization of search capabilities, allowing different research groups to operate independently of each other, but share their data: "the databases and tools included in such an infrastructure should not be stored on one central server. Rather, they should constitute a distributed network of databases, searchable using a common search engine (preferably via the Internet) and analysable with using [*sic*] a cartographic tool in order to visualize the geographical distribution of one or more syntactic

properties. The advantage of such a decentralized infrastructure is that every research group involved is able to maintain and update their own database independently" (43). The result is that each researcher's data is interoperable and transferable with that of others.

Interoperability is also one of the eight fundamental questions for endangered language technology projects discussed in Russell Hugo's chapter, "Endangered languages, technology and learning: Immediate applications and long-term considerations," a well-considered chapter with some excellent advice for planning the technological components of any language revitalization project. Hugo states that a primary consideration should be the avoidance of "content lock," defined as follows: "if a solution is designed and content is integrated—as much content and organization as possible should be able to be extracted and easily migrated to a future platform" (102). Interoperability therefore includes not just considerations of transfers between different tools, but different times as well.

The focus on interoperability does unfortunately also suffer somewhat from a lack of attention to implementational details. For example, while many authors advocate storing and transferring data in XML, this alone is not enough for interoperability, precisely because XML schemas can be implemented in many ways. The transcription tool ELAN (Max Planck Institute for Psycholinguistics; cf. Brugman & Russell 2004) formats its data in XML in a drastically different way than does the glossing and lexicography tool FLEx (SIL 2015). Moreover, XML, far from being a "future standard format" (102) as it was originally hailed, is being largely abandoned in favor of the far simpler format known as JSON, which is used to store and exchange data in most web-based applications today. JSON is perfectly designed for exchanging highly structured data like linguistic texts, something XML was not primarily designed to do. This means that the majority of linguistic tools, including those presented in this book, are not easily interoperable with most modern web technologies, and this is unfortunate because the future of technology most definitely lies with the web.

But Hugo here has excellent advice as well: "Complicated software development is arguably a less than ideal use of resources. Rather than seeking to 'reinvent the wheel' for each endangered language, it may be worth looking around to see whether applications that have already been created can also aid the documentation, development and distribution of learning materials for endangered language efforts" (110). Utilizing pre-existing tools is an excellent way to ensure interoperability between one's own project and others using that tool, and also helps encourage the expansion of that tool's infrastructure by putting it to new uses.

One of the project outputs mentioned in some fashion or another by every contributor is searchability. Digital technologies provide a variety of new means of searching one's data at different levels. Some of the searchable features demonstrated in the present volume include language-internal variation, geographic variation, syntactic constituency, or even basic searches on glosses, words, or part-of-speech tags. Searchability also enables the application of big data techniques to smaller corpora. Searchability is ultimately what distinguishes data from archival materials. This point is made especially clear in Bernard Bel and Médéric Gasquet-Cyrus's chapter, "Digital curation and event-driven methods at the service of endangered languages." They caution against outputs that amount to "little more than a huge and widely disseminated showcase, which is hardly useful for revitalization" (114). Mere digitization is not enough. The authors sagely point out that we don't want to wind up with the digital version of shoeboxes of fileslips sitting in our closets: "if the old tapes are merely replaced by digital recordings stored on personal computers or unconnected websites, has there been any real change?" (87).

Searchability comes with some attendant difficulties as well, namely, questions of how one searches for material across different languages, or across inconsistent glossing conventions. The impulse of the contributors and many other linguists is towards standardization. Barbiers, for example, decrying the fact that "different research groups/language areas tend to use distinct [part of speech] tags, which makes it impossible to search multiple databases using a single set of tags," argues that, "A common, standardized and well-defined tag set is therefore essential" (44). Dorothee Beermann's chapter "Data management and analysis for endangered languages," however, offers another approach. She discusses a software tool called TypeCraft, used for annotation of textual data. This includes morphological glossing of a number of different languages, each with their own traditions of grammatical terminology and analysis. How does one reconcile these varied analyses in a way that makes the corpus consistently searchable? TypeCraft solves this problem by dividing data into two classes: common data, which are standardized across corpora (such as ISO codes or glosses from the GOLD ontology), and individual data, which comprise user-defined categories. This seems a better solution than insisting on complete standardization of glosses, an approach advocated by other authors in the volume. The crosslinguistic application of grammatical categories is fundamentally an issue of linguistics and not data structures, and sits at the heart of many a debate in typology and linguistic theory. Data structures should not impose theoretical constraints on the data, and forcing standardization of glossing conventions does just that.

The story is different for metadata, however. As Barbiers also notes, "It is important that every database be enriched with standardized metadata so that the database can be selected on the basis of its properties. These metadata can include, for example, information on the language area and the dialects, dates of the recordings and profiles of consultants" (43–44). Since metadata are not dependent on the linguist's analysis or linguistic theory in the same way that morphological glosses are, and since metadata have the primary function of locating and identifying items in a collection, standardization of metadata should be strongly encouraged.

Another theme that features prominently in this volume is open access of both data and tools, while remaining sensitive to issues of access and permission. In fact, a central point in Beermann's chapter on data management is that recent technologies in some ways make managing access easier than ever before, given the ease of setting up user groups and profiles and restricting access to users in permitted groups (81). And Cecilia Odé, in her chapter on 'Language documentation and description from the native speaker's point of view: The case of Tundra Yukaghir,' shows just how beneficial such open-access tools can be. She relates the launch of a free access, interactive e-learning module about language shift focusing primarily on Tundra Yukaghir of Siberia and Mpur of West Papua, and how its broad availability has made it a valuable tool for raising awareness about language shift. Moreover, speakers of other minority languages easily related the film to their own language context and recognized the situation of the Tundra Yukaghir as analogous to their own, fostering fruitful discussions regarding language vitality, documentation, and revitalization for their own communities.

The final and perhaps most important thread that runs throughout this book is the way in which technology both furthers and in some cases makes possible increased community-academic collaboration and community involvement. Paterson summarizes this nicely: "The global levelling of information access through the Internet also enables speakers of endangered languages and academics to engage more fully with each other—rather than, as before, operating in different social circles. Roles such as 'linguist', 'language documenter'

or 'endangered language speaker', which might previously have been mutually exclusive, can therefore now be fulfilled by 'academics' and 'native speakers' alike" (50). And collaborative spaces are key to productive language learning, as noted by Hugo when discussing what makes sound pedagogy for language revitalization. He explains that, "Technology may provide additional learning time via online courses, spaces to collaborate and communicate at a distance" (98).

Taken together, the themes in this volume lend credence to Ostler's positive outlook for minority languages. The authors exemplify the way in which new technologies and the tools that stem from them can be, when appropriately leveraged, the great leveler, putting minority languages on equal footing with more dominant ones. They also show how minority language communities can co-opt tools first created for larger, more well-resourced languages, thus reducing the cost of adoption, and fostering innovative ways of leveraging preexisting tools and data for new purposes. They show how the utility of such tools is greatly enhanced by following principles of open access and interoperability, and how this in turn fosters a greater degree of collaboration and crossover between academics and community members. And in purely practical terms, every one of the authors demonstrates how much there is to be gained from enabling even simple searching across digitized data and metadata. Add to this the range of helpful advice for project planning, technology design, and strategies for maximizing adoption and use of revitalization tools sprinkled throughout the book, and the result is a volume that members of any language documentation or revitalization project would do well to read.

References

Brugman, Hennie & Albert Russell. 2004. Annotating multimedia/multi-modal resources with ELAN. In Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa & Raquel Silva (eds.), *Proceedings of LREC 2004, 4th International Conference on Language Resources & Evaluation*, 2065–2068. Paris: European Language Resources Association. http://www.lrec-conf.org/proceedings/lrec2004/.

Krauss, Michael. 1992. The world's languages in crisis. *Language* 68(1). 4–10.

Max Planck Institute for Psycholinguistics. ELAN. Nijmegan, The Netherlands: The Language Archive. http://tla.mpi.nl/tools/tla-tools/elan/.

SIL. 2015. Fieldworks Language Explorer (FLEx). http://fieldworks.sil.org/.

Daniel W. Hieber
dhieber@umail.ucsb.edu