

Linguistic Issues in Language Technology – LiLT
Submitted, October 2011

Bootstrapping the Language Archive

**New prospects for Natural Language
Processing in Preserving Linguistic
Heritage**

Steven Bird

Published by CSLI Publications

Bootstrapping the Language Archive

New prospects for Natural Language Processing in Preserving Linguistic Heritage

STEVEN BIRD, *University of Melbourne and University of
Pennsylvania*

1 Introduction

Today, for only about thirty of the world's 6,900 languages do we have non-negligible quantities of machine-readable data (Maxwell and Hughes, 2006). The recent upsurge of interest in documentary linguistics is yet to produce a million-word machine readable corpus for any endangered language, even though this quantity would be necessary for wide ranging investigation of the language once no speakers are available. The community with the most experience of creating large-scale language resources is preoccupied with the “top” 0.5% of languages, and is not yet participating in the effort to preserve the world's linguistic heritage (Abney and Bird, 2010).

There are grounds to believe that language technology in general, and natural language processing in particular, have important roles to play in creating and analyzing corpora for small languages. This goes beyond the development of data management tools to the application of natural language processing techniques to small and noisy datasets, and the design of new methods that operate within the constraints of linguistic field data. A set of seven such constraints (or “axioms for scalable work with small languages”) are presented, and suggestions for

further NLP research are related back to these axioms.

2 Computation in Language Documentation Today

Documentary linguists are early adopters of new recording hardware and new transcription and annotation tools.¹ However, few such tools stand the test of time, and much data is already moribund because the software used to create it is no longer supported (Bird and Simons, 2003).

Until recently, perhaps the most widely used computational support for language documentation was generic software which provided no support for linguistic data or workflows. In one instance, a large lexicon for a Namibian language was stored in Microsoft Word files, using various fonts and styles in order to mark various types of information. Entries were identified for study, first by using search to locate entries that met a particular requirement, then copy-paste to duplicate an entry of interest in a separate file. Modifications to an entry had to be done on all instances of that entry. Versions of the headword with and without tone marks were included in each entry to permit searches to include or exclude tone. In another instance, a large dictionary for a Nigerian language was maintained in Microsoft Word, and idiosyncratic markup such as tabs and square brackets was used to delimit the fields of an entry. It took months of painstaking work to check this markup for 28k entries, so that it could be loaded into a database.

Similar issues arise even in linguistic software such as the “Shoebox” program, in which each record (such as a lexical entry or a line of text) consists of a collection of optional and repeatable fields, with no schema (Buseman et al., 1996). For example, a Shoebox dictionary for a Cameroonian language, having about 2,500 entries, was extended by manually adding new entries containing the form of each noun with its plural prefix (whether regular or not). The records were then sorted, and with further processing there was a printed dictionary with cross references from plural nouns to their corresponding singular forms, to help readers who were unable to strip off the plural prefix for themselves before looking up the word. From the time of that extension to the dictionary, updates had to be made in two places.

These examples highlight the problem of working at the level of views or presentations of data, rather than the underlying data structures themselves. This situation is analogous to editing the HTML presenta-

¹ Evidence of this can be found in the technology reviews in each issue of *Language Documentation and Conservation*, published by the University of Hawai'i Press, and in the website of E-MELD, *Electronic Metastructure for Endangered Languages Data* at emeld.org.

tion of a database record, instead of the underlying data from which the HTML presentation is generated. Data validation is almost completely manual, and this is error-prone and leads to inconsistencies.

Frustrated with the limitations of such tools, some linguists are involved in the development of specialised software. In one mode of development, the linguist is in control. He or she might have received a grant to “archive the lexicon on the web” allowing users to search the lexicon and possibly to access images and audio clips. Sometimes, a junior graduate with no domain knowledge is hired, works in isolation, and invents local solutions for modeling and manipulating linguistic data. A variation on this theme has been to publish an interactive presentation of the data as the primary form of dissemination. Large binary files on a handsomely produced CD-ROM encapsulate a standalone version of a database application plus the linguistic content, e.g. (Csató and Nathan, 2001). In both the web and CD-ROM variations, the data has not been fully archived or disseminated. Rather, an interface to the data has been provided, one which may cease to function once it is no longer being actively supported by the project.

Notably absent from such projects are the well-understood techniques for working with large collections of structured information, such as data modelling, normalisation, and functional abstraction (Ullman and Widom, 2007). Decisions about implementation language, platform, libraries, and algorithms appear to be made at random, and may not be informed by an understanding of the full range of options, or the implications of choosing one over another.

In the second mode of linguistic software development, software engineers are in control. Prominent examples are the software development teams of the Summer Institute of Linguistics in Dallas and the Max Planck Institute in Nijmegen, which have contributed important, free tools to the community (Buseman et al., 1996, Simons, 1998, Black and Simons, 2008, Wittenburg et al., 2006). Until recently, the focus of these efforts has been software that supports field-based *descriptive* linguistics, rather than documentary linguistics.² These teams begin by eliciting requirements from their clients, i.e. field linguists. Then they develop, test and document the software, and deliver a shrink-wrapped product at the end.

The software engineering approach faces many obstacles: limited knowledge of the application domain by the programmers, prioritizing the stated needs of users and selecting an effective range of functional-

²For a survey of software for doing field linguistics, see (Antworth and Valentine, 1998).

ity, supporting the diverse levels of sophistication of users, accessing remote clients who do not always provide timely feedback on pre-releases, and reducing the significant lag time in adding functionality. These are the predictable challenges faced by most software engineering projects, but they are more acute thanks to the rapidly evolving requirements and the limited possibilities for communication.

A little-attested “third way” is *field-based* linguistic software development. An important example of this is the WeSay program, developed in Papua New Guinea and Thailand (Albright and Hatton, 2008). Developing software in a field location has its challenges, but the benefit is to have ready access to end-users and to be able to observe them using the software to do real work. Hatton and colleagues have gone a step further, and used their field location to deliver software for use by the language speakers themselves, including people with minimal western-style education.

3 Seven axioms for working with small languages

The literature on language revitalization documents a great variety of situations (Fishman, 2001a, Hinton and Hale, 2001, Grenoble and Whaley, 2006). Some small languages are supported by the state government (e.g. Hawaiian), while others are effectively undermined by practices designed to assimilate ethnic minorities, such as forced migration (Poa and Lapolla, 2007). Some small languages have stable intergenerational transfer, such as the languages of northern Vanuatu (François, 2011), while others are only spoken by a handful of elderly people and are nearly extinct (e.g. Bella Coola).³

Nevertheless, the situation of a small language is vastly different than the world’s major languages. From the standpoint of relatively well-resourced languages, we can make a variety of coarse generalizations about small languages. The most significant generalization, which hardly needs stating, is that all small languages are *under*-documented, if not completely *undocumented*. Here I wish to go beyond this observation, to a series of generalizations that might provide a starting point for the development of language technologies that could support the documentation effort.

Note that there are two important categories of under-resourced languages which we do not consider here: major world languages, and endangered dialects. For example, Hindi has about 200 million speakers, yet no million word text corpus has been published for this language. Some varieties of major world languages including English are consid-

³http://www.ethnologue.com/nearly_extinct.asp

ered to be endangered dialects – this includes “any language variety that constitutes the verbal heritage of some speech community and experiences assimilatory contact with a dominant or standard language” (Guy and Zilles, 2008, p53). The current discussion focusses on small *languages*, at least including the languages having fewer than 10,000 speakers, which comprise about half of the world’s total.

Human and Financial Resources. For most small languages, chances are that there will never be a sponsored documentation project. A few dozen fully-funded language documentation projects are started each year, but this only amounts to about 1% of the total. Many endangered and moribund languages are left out. A substantial portion of these human and financial resources are spent on the logistics such as travel, engaging with the speech community, and initial language learning. Often funding and personnel are available for language development or linguistic description, but those involved simply do not have time to take on extra tasks involved with language documentation. If they can see that the work can proceed independently, they may be more willing to provide incidental support.

Many small languages are found in developing countries with weak currencies. Thus any external funding will often go a long way. It may only be necessary to provide food and pay transportation. The guardians of the linguistic heritage may be retired and may have considerable free time to devote to language documentation work. (This also avoids the logistics of moving money, accounting, tax, and the sense that the heritage is being purchased.)

An alternative to bringing in a linguist from outside is to have the good fortune of finding a mother tongue speaker of the language who is trained as a linguist and able to work independently in documenting his/her language. In general, then, we seek approaches to language documentation which do not depend on such resources.

The following “axioms” are intended to ensure that documentary work can begin right away, while the languages are still spoken.

1. No special funding: documentation work should be supportable in the margins of existing activities.
2. No linguist: there should be significant documentation tasks that do not require professional training.

Orthography and Literacy. Rarely will the language have an orthography supported with literacy materials and programs, along with a self-sustaining local publishing industry to support and justify widespread literacy. As Bernard notes: “Five hundred years into the Gutenberg revolution, 95% of the world’s languages remain untouched

by it because, so far at least, no one has been able to turn a profit by publishing in those languages” (Bernard, 1996, p148). Creating an orthography for the express purpose of capturing the literature of a language is a vain hope, since it depends on a substantial investment of linguistic expertise (which is unavailable), and on the motivation of a speech community to master the orthography, which presupposes a body of literature in the language (but none exists). Further difficulties abound. Creating an orthography usually involves selecting one dialect over others when there is no agreed standard form of the language, a recipe for conflict (Schieffelin and Charlier Doucet, 1998). Establishing a literacy program is no guarantee that there will soon be a community of competent writers. Even when an orthography exists for a language, there may not be widespread literacy in the language. For instance, an orthography was devised for the Usarufa language in the 1950s as part of a Bible translation project (Bee and Glasgow, 1962), but today only a handful of speakers can read it, and no-one uses it regularly for their own writing. Thus, we cannot assume that there is an orthography, which means that transcriptions will probably use the orthography of the contact language, with variable spellings and inconsistent indication of word boundaries.

3. No orthography: texts and transcriptions use a phonetic alphabet or the orthography of another language

Language Clusters. Small languages are often found in clusters of typologically similar languages, with much bilingualism between adjacent languages. Oral literature collected from a cluster of languages will probably exhibit a high degree of thematic similarity, due to common history and shared material culture. Independently collected texts from across the cluster will tend to have the character of a “comparable corpus” (Fung and Yee, 1998). In light of the typological and lexical similarity between the languages, it is sometimes possible to translate texts between the languages with a combination of automatic transliteration and human editing (Weber, 1981). Bilingual speakers will often volunteer information about words and their pronunciations in different local languages, to establish the distinct identity of their mother tongue. As a consequence, it is probably very easy to elicit comparative wordlists and detect regular sound correspondences.

An outside linguist visiting the speech community must expect to take years to acquire sufficient knowledge of the language and culture to interpret it faithfully. Even if the aim is just to collect naturalistic materials, it takes time for the outsider to build trust and to identify the most appropriate individuals to work with. All such overheads are

mitigated if the linguist is able to work in more than one language of the cluster. Similarly, effort in interpreting cultural terms may be difficult for the first language, but become considerably easier thereafter.

4. Language clusters: documentation efforts can involve a group of languages, and can leverage bilingualism and shared knowledge

Knowledge Gaps. Much cultural knowledge is preserved only in the language and narratives of the oldest generation, who are often monolingual. The “ancestral code” may not be well-known by the youngest generation, speakers of the “emergent code,” who are often the ones with the best knowledge of the contact language, given their mobility and their schooling outside the language area. Much of the lexicon of the older speakers may relate to traditional practices which have fallen out of use, and may not be familiar to the only people who are available to translate the literature for the benefit of outsiders. While much of the cultural content is inevitably lost in translation (Fishman, 2001b), having a rough translation is better than nothing at all, and may still help with lexicography and with certain types of linguistic analysis. These problems are ameliorated when speakers of the ancestral code provide oral essays on key aspects of traditional culture that aid in the interpretation of the literature, and when the person who is transcribing and translating the literature lives in the speech community or has regular access to it.

A related set of problems concerns the knowledge gap between the speech community and the outside audience. How can a monolingual speaker who has never left the language area know how to interpret traditional culture to an outsider, or to someone listening to the archived materials a century later (e.g. “in our land we have only one sun”)? manifold further issues concern ownership of cultural knowledge: it may be advisable to obtain formal consent from key people who are not otherwise participating, and for them to be identified as part of a team effort.

A more subtle gap arises in the context of translation. The genre of materials that are translated from the small language (e.g. traditional narratives) may be quite different to the genre of the materials that are translated into the language (e.g. educational and religious materials). A consequence of this asymmetry is that there may be no genre where we have substantial bodies of naturally occurring text for both the source and target language.

5. Lacunae in the data: any systematization of multi-language data will contain significant gaps
6. Documentation as teamwork: rich knowledge of a language and

the ability to produce good translations will usually require different individuals

Contact Language. The contact language is likely to have rich resources, including large machine-readable text collections and lexicons. Since all the source language materials are archived, we have the option of waiting until the assumed contact language resource exists. For instance, suppose that orally translated content has not yet been transcribed. We permit ourselves to assume that the contact language will always be interpretable, and that we have decades, centuries even, to get around to doing such transcription work, or to developing the technology which automates it. In light of this, we devote our energies to ensuring that source language materials have been preserved and that they are interpretable.

7. Fully-resourced contact language: we permit ourselves to assume that all desired resources exist in the contact language

A small case study

Basic Oral Language Documentation (BOLD) is an example of a language documentation methodology which responds to the realities described above (Reiman, 2010, Bird, 2010). Oral literature is recorded in a naturalistic context, by native speakers. Selected texts are put through a further process of careful “respeaking,” where someone listens through the original recording, pauses it after each phrase, and repeats what was said with slow, careful speech. This respeaking is captured on a second recording device, and is believed to improve the interpretability of the original source. The procedure is illustrated in Figure 1.

At a later stage, the spoken text and its spoken translation are both transcribed in writing, leading to a bilingual text. Here we consider how well the BOLD methodology fits with our seven axioms for working with small languages.

1. No funding: the BOLD methodology is being used in three universities in Papua New Guinea by students in the final year of their studies; they work on their ancestral language and travel to their village locations using their existing scholarship funding; the recorders were donated by Olympus; their compensation for doing the documentation work is a course credit.
2. No linguist: the students have received a few hours of training in the BOLD methodology by the university staff member, who is not required to be a linguist; most of the students are not doing a linguistics major.



FIGURE 1 Protocol for Respeaking and Oral Translation: the operator (left) controls playback and audio segmentation; the talker (right) provides oral annotations using a second recorder

3. No orthography: no constraint is placed on the transcriptions, and they may use English orthography if the language does not have an orthography, or if the student is not literate in their mother tongue.
4. Language clusters: each of the three groups is working with at least a dozen languages, drawn from the immediate vicinity of the university and further afield; effort in equipping, training and guiding the work is shared across many languages; issues with translating cultural terms can be discussed in class.
5. Lacunae in the data: the BOLD methodology goes some way to addressing this issue simply by facilitating the collection of a large quantity of source material.
6. Documentation as teamwork: the students record elders and must discuss the meaning of cultural terms that have fallen out of use; students help each other to enter transcriptions and translations using university computers.
7. Fully-resourced contact language: the collection work is narrowly focussed on texts and translations and no effort is made to do further annotation of the sources (such as POS tagging), since we assume this can be inferred later.

4 Natural Language Processing Applications

In what ways can NLP techniques operate according to the axioms laid out in Section 3?

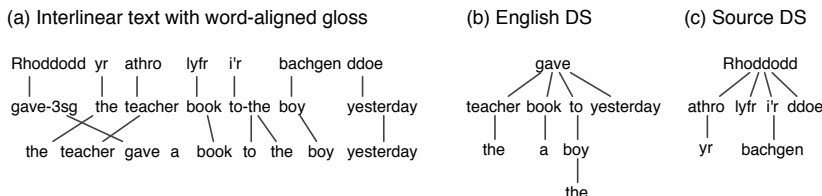


FIGURE 2 Generating parse trees for the source language, by aligning the glosses with the phrasal translation (a), parsing the translation (b), then transferring the dependency structure (DS) to the source tree (c).

It is instructive to consider some existing efforts to support language documentation and description using NLP techniques.⁴

Probst et al. (2002) created a corpus of English sentences which contains a wide range of constructions, and which are translated into the language of interest. The program is able to detect from the translations whether the language exhibits various syntactic or morphosyntactic phenomena, using an approach that is closely aligned with long-standing structuralist-style “discovery procedures.” Palmer et al. (2010) describe an active learning method which increases the efficiency of annotators in creating interlinear text. Xia and Lewis (2007) have shown how it is possible to infer syntactic structure with the help of interlinear text; they align the words of the gloss with the phrasal translation, then parse the phrasal translation, and transfer the syntactic dependencies from English back to the source language (see Figure 2). Bender (2010) has shown how we can carry forward the program of grammar engineering, applying it to small languages.

What these have in common is that they acknowledge that the linguist is a scarce resource, and try to increase the productivity of a linguist. However, this only addresses the second axiom (Section 3). If these methods were applied on a large scale, they may encounter problems that result from not considering the realities that are captured in the remaining axioms.

New opportunities for NLP techniques in language documentation

The primary task, I believe, is to collect and translate a substantial quantity of text. The justification for this is postponed to Section 5. For now I wish to suggest ways that NLP techniques might support this task. The first set concerns the text collection process, particularly the

⁴This discussion omits efforts to linguistically *analyze* small languages; here the focus is on documentation and description.

problem of creating normalized transcriptions when there is no agreed orthography.

- have a small audio collection multiply transcribed, where different speakers of the language use their own preferred transcription conventions; for each pair of transcriptions of the same phrase, automatically align the words
- learn the correspondences between the transcriptions provided by different transcribers; use the correspondences to normalize the transcriptions which have not been multiply transcribed, to suggest corrections to transcriptions, and to normalize word boundaries
- compile such transcriptions from a cluster of languages; this will be a comparable corpus
- combine information about word forms and textual distribution to infer cognates, in the broad sense of cognate meaning “words in different languages that are similar in form and meaning, without making a distinction between borrowed and genetically related words” (Kondrak, 2001)
- build a comparative wordlist combining elicited cognates and learned cognates; expand using semi-supervised learning (Abney, 2007), guided by gaps in our table, text frequency of words, confidence in the existing cognates
- elicit more source language texts by requesting human translations between local languages for valuable texts

At this point we have a collection of comparable, partially-normalized texts in several closely-related languages, along with a comparative wordlist. The required text collection work could have been done in a variety of ways, from distributed collection activities to a centralized “writers workshop.” This data can be continually expanded, by transliterating texts between the languages (combining regular sound changes with exceptional forms stored in a table), and having a human correct the outputs.

The second set of NLP tasks concerns the translation process, making the text content accessible to outsiders, and evaluating the lexical and syntactic coverage of the collection.

- elicit translations from any of the source language materials into the contact language
- include any existing translated texts (i.e. Bible translation)
- “normalize” across the different source languages, replacing identified cognates with a normalized form, or with a lexeme identifier
- pool the data from the languages and train up an alignment model

- use this to translate all the source materials into the contact language, post-edit, retrain, and repeat

5 Adequacy of a Language Documentation

What quantity and quality of archived language documentation is required in order to adequately capture a language? One yardstick would be to require that the archived materials should directly support the development of the full suite of language resources and technologies, including a treebank, a wordnet, a parser, and so forth. However, such resources depend on there being a level of linguistic analysis that will not be achieved for most languages before they fall out of use. Thus, we have no way to measure the adequacy of the archived materials while there is still time to collect more.

An alternative yardstick is that it should be possible for a linguist to bring their descriptive questions to the archive, and come away with confident answers backed up with relevant evidence. There should be enough recorded and transcribed speech that we can answer questions about the phonemic inventory, phonotactics, allophony, prosodic structure, and so forth. There should be enough text on which to base a comprehensive lexicon and a comprehensive syntactic analysis. Semantic and pragmatic knowledge should also be discoverable. To meet this notion of adequacy, the archive must take the place of the speech community.

We can guess the required size of such a collection, based on the hours of linguistic experience needed to produce a monolingual adult native speaker, which falls somewhere between ten and a hundred thousand hours, or about 100 million words. Given this quantity of data, and about twenty years to assimilate it all, we might expect a person to become highly proficient in the language, to the point where s/he could serve as a linguistic informant. Unfortunately this simple argument fails because the archived data does not capture the real world context necessary for anchoring referents and supplying plausible interpretations. Our hypothetical language learner is at a disadvantage relative to an infant who is exposed to a phrase “do you see the doggy?” simply because s/he has no chance to see the doggy.

In any case, this diagnostic requires too many resources: we would complete the archiving activity, await some future date when someone has the time to give a few years to learning the language from the archive, then try to judge how successful they were (possibly with no remaining native speakers to judge).

Instead, Abney and Bird (2010) argue that machine translation

(MT) provides the test case: can we train up an MT system to successfully translate into and out of the language? This could be evaluated while speakers are still available, and before the language documentation task is complete. This equates to L2 learning, when we build off knowledge of a native language, and may require much less material (and probably no real-world grounding either, as translations suffice).

This approach ties in nicely with existing practice for language documentation: collecting texts and translations. In the case of endangered languages – but also endangered genres of languages with millions of speakers – archived materials must be translated if they are to be interpretable in the future. Such translations double as a finding aid for people studying grammar, language usage, and culture that is expressed in language.

For languages where there is an established tradition of literacy, scholars and teachers are often working to collate and preserve the materials. For example, the Alekano language of Goroka, Papua New Guinea, has 25,000 speakers, and a reported literacy rate of 25-50%. The University of Goroka requires all students from outside the language area – the majority – to take a semester of classes in the language. University staff are working with the community to record and transcribe oral literature. The techniques discussed above, once implemented, would help them to bootstrap their language archive, producing good quality transcriptions and translations and optimizing the use of available human labor.

In time, an MT system would allow the language workers to gain a sense of the knowledge of language that has so far been captured in their collection of bilingual text, helping them prioritize their efforts in expanding the archive. This is significant, since it promises to give us an operational yardstick: we have a measure of the adequacy of a language documentation that can be used while there is still time to collect more documentation.

Acknowledgments

I am grateful to Lori Levin, Will Lewis, and Fei Xia for inviting me to present an earlier version of this paper at the ACL 2010 workshop *NLP and Linguistics: Finding the Common Ground*, and to several workshop participants for feedback. I'm grateful to Steven Abney and David Chiang for helpful input about some of the ideas discussed here. This work has been supported by the US National Science Foundation (award 0723357), the Australian Research Council (award DP0988242), and by the Firebird Foundation for Anthropological Research.

References

- Abney, Steven. 2007. *Semisupervised Learning for Computational Linguistics*. Chapman & Hall/CRC.
- Abney, Steven and Steven Bird. 2010. The Human Language Project: building a universal corpus of the world's languages. In *Proceedings of the 48th Meeting of the Association for Computational Linguistics*, pages 88–97. Association for Computational Linguistics.
- Albright, Eric and John Hatton. 2008. Wesay, a tool for engaging communities in dictionary building. In V. D. Rau and M. Florey, eds., *Language Documentation and Conservation Special Publication No. 1: Documenting and Revitalizing Austronesian Languages*, pages 189–201. University of Hawai'i Press. <http://hdl.handle.net/10125/1368>.
- Antworth, Evan and J. Randolph Valentine. 1998. Software for doing field linguistics. In J. M. Lawler and H. Aristar Dry, eds., *Using Computers in Linguistics: A Practical Guide*, pages 170–196. London and New York: Routledge. <http://www.sil.org/computing/routledge/antworth-valentine/>.
- Bee, Darlene and Kathleen Barker Glasgow. 1962. *Usarufa tone and segmental phonemes*. No. 6 in Oceanic Linguistic Monographs. University of Hawai'i Press. Republished in McKaughan, Howard (ed) *The Languages of the Eastern Family of the East New Guinea Highland Stock*, pp. 190–203, University of Washington Press, 1973.
- Bender, Emily M. 2010. Reweaving a grammar for Wambaya: A case study in grammar engineering for linguistic hypothesis testing. *Linguistic Issues in Language Technology* 3:1–34.
- Bernard, H. Russell. 1996. Language preservation and publishing. In N. H. Hornberger, ed., *Indigenous Literacies in the Americas: Language Planning from the Bottom Up*, pages 139–156. Mouton de Gruyter.
- Bird, Steven. 2010. A scalable method for preserving oral literature from small languages. In *Proceedings of the 12th International Conference on Asia-Pacific Digital Libraries*, pages 5–14.
- Bird, Steven and Gary Simons. 2003. Seven dimensions of portability for language documentation and description. *Language* 79:557–82.
- Black, H. Andrew and Gary F. Simons. 2008. The SIL FieldWorks Language Explorer approach to morphological parsing. In N. Gaylord, S. Hilderbrand, H. Lyu, A. Palmer, and E. Ponvert, eds., *Computational Linguistics for Less-studied Languages: Proceedings of Texas Linguistics Society 10*. CSLI.
- Buseman, Alan, Karen Buseman, and Rod Early. 1996. *The Linguist's Shoebox: Integrated Data Management and Analysis for the Field Linguist*. Waxhaw NC: SIL. <http://www.sil.org/computing/shoebox/>.
- Csató, Éva and David Nathan. 2001. *Spoken Karaim*. Tokyo University of Foreign Studies. CD-ROM.

- Fishman, Joshua A., ed. 2001a. *Can Threatened Languages be Saved?: Reversing Language Shift, Revisited: A 21st Century Perspective*. Multilingual Matters.
- Fishman, Joshua A. 2001b. Why is it so hard to save a threatened language? In J. A. Fishman, ed., *Can Threatened Languages be Saved?: Reversing Language Shift, Revisited : a 21st Century Perspective*, pages 1–22. Multilingual Matters.
- François, Alex. 2011. The dynamics of linguistic diversity: Egalitarian multilingualism and power imbalance among north vanuatu languages. *International Journal of the Sociology of Language* .
- Fung, Pascale and Lo Yuen Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*, pages 414–420.
- Grenoble, Lenore and Lindsay Whaley. 2006. *Saving Languages: An Introduction to Language Revitalization*. Cambridge University Press.
- Guy, Gregory R. and Ana M. S. Zilles. 2008. Endangered language varieties: Vernacular speech and linguistic standardization in Brazilian Portuguese. In K. A. King, N. Schilling-Estes, L. Fogle, J. J. Lou, and B. Soukup, eds., *Sustaining Linguistic Diversity: Endangered and Minority Languages and Language Varieties*, pages 53–66. Georgetown University Press.
- Hinton, Leanne and Kenneth Hale, eds. 2001. *The Green Book of Language Revitalization in Practice*. Emerald Group Publishing.
- Kondrak, Grzegorz. 2001. Identifying cognates by phonetic and semantic similarity. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 103–110.
- Maxwell, Mike and Baden Hughes. 2006. Frontiers in linguistic annotation for lower-density languages. In *Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora 2006*, pages 29–37. Association for Computational Linguistics.
- Palmer, Alexis, Taesun Moon, Jason Baldridge, Katrin Erk, Eric Campbell, and Telma Can. 2010. Computational strategies for reducing annotation effort in language documentation. *Linguistic Issues in Language Technology* 3(4):1–42.
- Poa, Dory and Randy J. Lapolla. 2007. Minority languages of China. In O. Miyaoka, O. Sakiyama, and M. E. Krauss, eds., *The Vanishing Languages of the Pacific Rim*, pages 337–354. Oxford University Press.
- Probst, Katharina, Lori Levin, Erik Peterson, Alon Lavie, and Jaime Carbonell. 2002. Mt for resource-poor languages using elicitation-based learning of syntactic transfer rules. *Machine Translation* 17(4):225–270.
- Reiman, Will. 2010. Basic oral language documentation. *Language Documentation and Conservation* .

- Schieffelin, Bambi B. and Rachelle Charlier Doucet. 1998. The “real” Haitian Creole: Ideology, metalinguistics, and orthographic choice. In B. B. Schieffelin, K. A. Woolard, and P. V. Kroskrity, eds., *Language Ideologies: Practice and Theory*, pages 285–317. Oxford University Press.
- Simons, Gary. 1998. The nature of linguistic data and the requirements of a computing environment for linguistic research. In J. M. Lawler and H. Aristar Dry, eds., *Using Computers in Linguistics: A Practical Guide*, pages 10–25. London and New York: Routledge. <http://www.sil.org/computing/routledge/simons/>.
- Ullman, Jeffrey D. and Jennifer Widom. 2007. *A First Course in Database Systems*. Prentice Hall, 3rd edn.
- Weber, David J. 1981. Prospects for computer-assisted dialect adaptation. *American Journal of Computational Linguistics* 7:165–177.
- Wittenburg, P., H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes. 2006. ELAN: a professional framework for multimodality research. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*. European Language Resources Association.
- Xia, Fei and William D. Lewis. 2007. Multilingual structural projection across interlinearized text. In *Proceedings of the Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 452–459.