Last Words

Natural Language Processing and Linguistic Fieldwork

Steven Bird*
University of Melbourne

March 2009 marked an important milestone: the First International Conference on Language Documentation and Conservation, held at the University of Hawai'i.[1] The scale of the event was striking, with five parallel tracks running over three days. The organizers coped magnificently with three times the expected participation (over 300). The buzz among the participants was that we were at the start of something big, that we were already part of a significant and growing community dedicated to *supporting small languages together*, the conference subtitle.

The event was full of computation and linguistics, yet devoid of computational linguistics. The language documentation community uses technology to process language, but is largely ignorant of the field of natural language processing. I pondered what we have to offer this community: "Send us your 10 million words of Nahuatl-English bitext and we'll do you a machine translation system!" "Show us your Bambara WordNet and we'll use it to train a word sense disambiguation tool!" "Write up the word-formation rules of Inuktitut in this arcane format and we'll give you a morphological analyzer!" Is there not some more immediate contribution we could offer?

Over the past 15 years, the field of computational linguistics has been revolutionized by the ready availability of large corpora. Landmark dates are the founding of the Linguistic Data Consortium (1992) and the first Workshop on Very Large Corpora (1993). While the CL community has been pre-occupied with the new-found technical capabilities for collecting and processing large amounts of data, the field of linguistics has been undergoing a revolution of its own. It is also dominated with the use of new-found technical capabilities for collecting and processing large amounts of data. However, in this case, the data comes from languages that are facing extinction.

Back in 1992, Michael Krauss, of the Alaska Native Language Center, issued the world's linguists with a wake-up call, calculating that "at the rate things are going—the coming century will see either the death or the doom of 90 per cent of mankind's languages" (Krauss 1992, page 7). He exhorted linguists to document these languages "lest linguistics go down in history as the only science that presided obliviously over the disappearance of 90 per cent of the very field to which it is dedicated" (page 10). This message was delivered at the 15th International Congress of Linguists in Quebec, and also in *Language*, the journal of the Linguistic Society of America.[2]

* Department of Computer Science and Software Engineering, University of Melbourne, Victoria 3010, Australia. E-mail: sb@csse.unimelb.edu.au.

1 http://nflrc.hawaii.edu/icldc09/.

2 The LSA has posted an FAQ containing an accessible description of the problem and its scale at http://www.lsadc.org/info/ling-faqs-endanger.cfm.

Today, endangered language documentation is part of mainstream linguistics, supported with several book-length treatments of the subject,[3] the online journal *Language Documentation and Conservation*,[4] numerous graduate courses, and funding programs in many countries. Here is the description of the U.S. NSF/NEH program, *Documenting Endangered Languages*, emphases added:[5]

> This multi-year funding partnership between the National Science Foundation (NSF) and the National Endowment for the Humanities (NEH) supports projects to develop and advance knowledge concerning endangered human languages. Made urgent by the imminent death of an estimated half of the 6000–7000 currently used human languages, this effort aims also to *exploit advances in information technology*. Funding will support fieldwork and other activities relevant to recording, documenting, and archiving endangered languages, including the preparation of lexicons, grammars, text samples, and databases. Funding will be available in the form of one- to three-year project grants as well as fellowships for up to twelve months. At least half the available funding will be awarded to *projects involving fieldwork*.

What does computational linguistics offer to a community that is exploiting advances in information technology for projects involving linguistic fieldwork with endangered languages?

The genesis of the field of computational linguistics out of the field of machine translation is well-known; this journal had a previous existence under the title *Mechanical Translation and Computational Linguistics*. The relationship between CL and MT over the past half-century has just come full circle: In March 2009 the ACL Executive Committee accepted a proposal for a new ACL Special Interest Group in Machine Translation. There can be no doubt that the multilingual information society is driving many important challenges in our discipline. However, relatively few languages have the necessary resources to participate.

Over the same half-century another strand of research has sought to use computational techniques to support linguistic fieldwork. For example, Joseph Grimes—ACL Vice President (1975)—has devoted much of his long career to studies at the interface between computational linguistics and linguistic fieldwork.[6] His NSF project with Gary Simons, called *Language Variation and Limits to Communication* (Cornell University, 1976–1978), involved building a suitcase-sized "portable" computer and lugging it around the Pacific to capture and analyze wordlists. Two decades later, my own fieldwork on tone languages of Cameroon involved a laptop computer powered by a car battery, and led to a series of "Grassfields Bantu Fieldwork" corpora published by the LDC. While the technology had improved, the modus operandi was the same: Take technology to a remote field location and bring back data, and do enough linguistic analysis in the field to ensure that the right variety and quality of data is being collected.

As if this were not challenging enough, the subsequent curation of the data is fraught with technical difficulties. It's easy to generate "endangered data" when formats, encodings, and media are so quickly obsolete (Bird and Simons 2003). Existing fieldwork tools use incompatible formats, and it is often necessary to convert data between the native formats of various tools. The experience of writing 10k lines of

3 Crystal 2000; Fishman 2001; Gippert, Himmelmann, and Mosel 2006; Grenoble and Whaley 2006; Harrison 2007.

4 http://nflrc.hawaii.edu/ldc/.

5 http://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf06577.

6 For example, Grimes (1968), http://www.ethnologue.com/show_author.asp?auth=2961.

	High tone verb: kapte *cover*		**Low tone verb: kəmte *bury***	
L	ə̀fɔ̀ ɔ̀ ˩kàptè màtsɔ̀ŋ	[-|--|---|/]	ə̀fɔ̀ ɔ̀ kə̀mtè màtsɔ̀ŋ	[-|--|---|/]
LH	fɔ̀k ɔ̀ ˩kàptè màtsɔ̀ŋ	[--|---|/]	fɔ̀k ɔ̀ kə̀mtè màtsɔ̀ŋ	[--|---|/]
HL	m̀b˩ù ɔ̀ ˩kàptè màtsɔ̀ŋ	[-|--|___|/]	m̀b˩ù ɔ̀ kə̀mtè màtsɔ̀ŋ	[-|--|---|/]
H	ǹdɔ̀ ɔ̀ ˩kàptè màtsɔ̀ŋ	[-|--|---|/]	ǹdɔ̀ ɔ̀ kə̀mtè màtsɔ̀ŋ	[-|--|---|/]
L	màfɔ̀ ɔ́ ˩kàptè màtsɔ̀ŋ	[-|-¯|---|/]	màfɔ̀ ɔ́ kə̀mtè màtsɔ̀ŋ	[-|-¯|---|/]
LH	màf˥ɔ́k ˩kàptè màtsɔ̀ŋ	[-|¯|___|/]	màf˥ɔ́k kə̀mtè màtsɔ̀ŋ	[-|¯|---|/]
HL	màp˗úŋ ŋ́ ˩kàptè màtsɔ̀ŋ	[-|¯¯|___|/]	màp˩úŋ ŋ́ kə̀mtè màtsɔ̀ŋ	[-|¯¯|---|/]
H	màlɔ́ŋ ŋ́ ˩kàptè màtsɔ̀ŋ	[-|¯¯|---|/]	màlɔ́ŋ ŋ́ kə̀mtè màtsɔ̀ŋ	[-|¯¯|---|/]

Figure 1
Tone data from Bamileke Dschang, a Grassfields Bantu language of Cameroon.

Perl scripts for manipulating fieldwork data in Cameroon was the backdrop to the development of the *Natural Language Toolkit*.[7] Clearly, with enough effort we can use computational techniques to represent and manipulate linguistic field data. Is there more we can offer?

For example, consider the tone language data in Figure 1. It represents a slice through part of an 8-dimensional tone paradigm containing 1,350 cells (Bird 2003). The address of each cell in this data cube is just a vector specifying properties like tense, mood, noun class, and lexical tones. The content of each cell is just a vector specifying a surface tone pattern using abstract pitch numbers, like 31144442. What structure could NLP techniques discover in this data? Could such analysis take place early enough to guide the data collection work?

For a long time, fieldwork has been regarded as a style of elicitation and analysis that involves an exotic language, an extended period, and an extreme location (cf. Hyman [2001]). In contrast, a new, cyber-fieldwork may be on the rise, in which the data is whatever one wants to treat as data, and where the "fieldworker" elicits data via Skype, by interrogating a sound archive, or by analyzing linguistic materials found on YouTube. However, it is hard to find cases of fieldwork that fit these stereotypes of purism and pragmatism, or what detractors might label paternalism and postmodernism. Thankfully, the real situation is more interesting. Regardless of location, language, and mode of elicitation, linguistic fieldworkers are usually immersing themselves in data, in close contact with a speech community. This may happen in the ancestral location or among a well-organized diaspora of speakers. In places where the Internet is reaching into remote places, scattered speakers of endangered languages are able to form online communities,[8] and in time this may provide another context for elicitation.

Linguistic fieldworkers are often pushing the limits of current theoretical machinery, while simultaneously experiencing the bleeding edge of digital recording and annotation technology. In the midst of this, they are eliciting and exploring a substantial quantity of primary data, where many of the descriptive categories are simply unknown or subject to revision. They may be transcribing speech when there is no existing writing system and when we don't know which sound contrasts are significant. They might be guessing word breaks and testing hunches about what particular morphemes mean. They could be puzzling over apparent inconsistencies in data from different speakers.

7 `http://www.nltk.org/`. See especially Bird, Klein, and Loper (2009, ch. 11).
8 For example, `http://www.firstvoices.com/`.

When the data is not systematized, when there is no established body of knowledge about the language, when many analytical options are available, and when every conclusion is open to question, the task becomes one of managing uncertainty—and in the meantime, avoiding an existential crisis.

(It's hard for a field linguist to explain this "fieldwork state of mind" to a computer scientist. What comes closest is the experience of debugging someone else's program. An undergraduate computing laboratory is ripe with "freely occurring programs," each one arising from a different—sometimes unrecognizable—view of a specified problem. To help someone fix their program requires that you briefly enter their world, and align your conceptual model of the problem with theirs, and point the way forward. However, this is made more difficult by the fact that you must puzzle over their code and their verbal statements, both of which may contain subtle errors. Now, scale up this experience from minutes to months!)

Migrating early pen-and-paper fieldwork onto computer is difficult, and probably fruitless. The technology gets in the way of the elicitation, and pre-occupation with systematizating the data prevents us from noticing the patterns: "premature mathematization keeps Nature's surprises hidden" (Lenat and Feigenbaum 1987, page 1177). It's probably best not to bother with linguistic software in the early stages of linguistic description.

However, things change once the descriptive notation has stabilized, and a "linguistic exploration" workflow is established. The discovery of a new word in a text may require an update to the lexicon and the construction of a new paradigm (e.g., in order to correctly classify the word). Such updates may occasion the creation of some field notes, the extension of a descriptive report, and possibly even the revision of the manuscript for an analytical paper. Progress on description and analysis gives fresh insights about how to organize existing data and it informs the quest for new data. Whether one is sorting data, or generating helpful tabulations, or gathering statistics, or searching for a (counter-)example, or verifying the transcriptions used in a manuscript, the principal challenge is computational.

Documenting and describing endangered languages presents computational linguistics with some difficult challenges. The most immediate challenge concerns linguistic data management: representing structured annotations such as interlinear text, supporting collaborative annotation, handling uncertain data, validating structure, tracking data provenance, combining human and automatic methods, and so forth. NLP techniques may enter the picture in unexpected ways. For instance, most documentation projects collect wordlists, and these typically evolve into full-fledged lexicons over time. The organization of fields within an entry is often inconsistent, yet we can recognize the structure using standard robust parsing techniques, then transform the data into a consistent structure, potentially saving months of manual effort in the process.

Once the data has some basic level of organization, the next challenge is one of simultaneously *downscaling* and *upscaling*. First, we need new techniques that work on small data sets (downscaling), with the consequence that fewer resources are spent on data collection, while permitting many more languages to be analyzed in the same timeframe (upscaling). What methods do we have that can detect structure in small, noisy data sets, while being directly applicable to a wide variety of languages? This represents uncharted territory for NLP.[9]

9 See (Palmer et al. 2009) for a promising pilot study.

This dual perspective applies to the data collection work itself. If we have just one week in a location where a language is spoken, to collect all the data we will ever have for this language, what will we do? I write this on the eve of a one-week visit to the Usarufa language area in the Eastern Highlands of Papua New Guinea, under the auspices of SIL. The language has about 1,000 speakers, and is no longer being learned by children. We will give out digital voice recorders to have people record linguistic interactions, narratives, and songs. Later, we will meet in a classroom where others will augment these recordings with voice annotations, phrase by phrase, providing a careful speech version along with translation into Tok Pisin, the language of wider communication. A handful of speakers who are literate in the language will transcribe a small portion of the collection. The resulting corpus, it is hoped, will be adequate to support future analysis and revitalization work. If it is possible to collect a useful corpus in the space of a week (downscaling) then it will also be possible to apply such methods to many other languages (upscaling). In this way, limited resources are deployed efficiently in a breadth-first approach to language documentation.

Apart from technical challenges, there is also an important sociological challenge to create maximally interoperable language analysis software. To imagine this can be done simply by adopting common file formats, or by operating an in-house software development lifecycle using project funds, or by invoking the XML family of buzzwords is to misunderstand the nature of the problem. Instead, we need to foster new research collaborations involving computational linguists and field linguists, leading to new understanding about how to collect and analyze corpora of data from endangered languages. We need to nurture a community to share in the development of tools, formats, interfaces, data repositories, query systems, machine learning techniques, visualization methods, and so forth. We need to collaborate on a global federated database of language data, permitting Web-based collaborative annotation of primary linguistic data, continuously expandible and fully exportable for local processing.[10] Everything should be available under open source and open content licenses, fostering a Web-scale ecosystem in which geographically distributed computational linguists, field linguists, and the speakers of endangered languages themselves are united in their efforts to document and describe the world's languages.

We live during a brief period of overlap between the mass extinction of the world's languages and the advent of the digital age. What can we do—as individuals and as a professional association—as we wake up to this global linguistic crisis? Recently, we have seen that national bureaucracies have been able to take unprecedented steps in the face of the global economic crisis; are we less agile? If the economic motivation for language technology research has lost some of its luster, what do we have to lose?

So, shall we eke out an incremental existence, parasitic on linguistic theories, language corpora, and machine learning algorithms developed by others? Are we content to tweak parameters and deliver results that are surpassed at next year's meeting, while important sources of new data are falling silent? It's time that we focused some of our efforts on a new kind of computational linguistics, one that accelerates the documentation and description of the world's endangered linguistic heritage, and delivers tangible and intangible value to future generations. Who knows, we may even postpone the day when these languages utter their *last words*.

10 The Open Language Archives Community (`http://language-archives.org`), the World Atlas of Language Structures (`http://wals.info`), and the Rosetta Project (`http://rosettaproject.org`) represent significant early steps in this direction.

References

Bird, Steven, editor. 2003. *Grassfields Bantu Fieldwork: Dschang Tone Paradigms*. Linguistic Data Consortium. LDC2003S02, ISBN 1-58563-254-6.

Bird, Steven, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with* Python. O'Reilly Media, Sebastopol, CA. `http://www.nltk.org/book`.

Bird, Steven and Gary Simons. 2003. Seven dimensions of portability for language documentation and description. *Language*, 79:557–582.

Crystal, David. 2000. *Language Death*. Cambridge University Press, Cambridge, UK.

Fishman, Joshua A., editor. 2001. *Can Threatened Languages be Saved?: Reversing Language Shift, Revisited: A 21st Century Perspective*. Multilingual Matters, Clevedon, UK.

Gippert, Jost, Nikolaus Himmelmann, and Ulrike Mosel, editors. 2006. *Essentials of Language Documentation*. Mouton de Gruyter, Berlin and New York.

Grenoble, Lenore and Lindsay Whaley. 2006. *Saving Languages: An Introduction to Language Revitalization*. Cambridge University Press, Cambridge, UK.

Grimes, Joseph E. 1968. Computer backup for field work in phonology. *Mechanical Translation and Computational Linguistics*, 11:73–74.

Harrison, K. David. 2007. *When Languages Die: The Extinction of the World's Languages and the Erosion of Human Knowledge*. Cambridge University Press, Cambridge, UK, pages 15–33.

Hyman, Larry M. 2001. Fieldwork as a state of mind. In Paul Newman and Martha Ratliff, editors, *Linguistic Fieldwork*. Cambridge University Press, Cambridge, UK.

Krauss, Michael E. 1992. The world's languages in crisis. *Language*, 68:4–10.

Lenat, Douglas B. and Edward A. Feigenbaum. 1987. On the thresholds of knowledge. In *Proceedings of the 10th International Conference on Artificial Intelligence*, pages 1173–1182.

Palmer, Alexis, Taesun Moon, and Jason Baldridge. 2009. Evaluating automation strategies in language documentation. In *Proceedings of the NAACL HLT 2009 Workshops on Active Learning for Natural Language Processing*, pages 36–44, Boulder, CO.