# **Comput*EL*: The use of computational methods in the study of endangered languages**

*Original workshop proposal (edited for presentation)*

Contemporary efforts to document the world's endangered languages—often going under the rubric of *documentary linguistics*—are dependent on the widespread availability of modern recording technologies, in particular digital audio and video recording devices and software to annotate the recordings that such devices produce. However, despite well over a decade of dedicated funding efforts aimed at the documentation of endangered languages, the technological landscape that supports the work of those involved in this work remains fragmented, and the promises of new technology remain largely unfulfilled. Moreover, the efforts of computer scientists, on the whole, are mostly disconnected from the day-to-day work of documentary linguists, making it difficult for the knowledge of each group to inform the other. On the one hand, this deprives documentary linguists of tools making use of the latest research results to speed up the time-consuming task of describing an underdocumented language. On the other hand, it severely limits the ability of computational linguists to test their methods on the full range of the world's linguistic diversity.

The most salient emblem of this situation is almost certainly the continued and widespread use of the Shoebox/Toolbox tool[1] for lexical and text data management. This tool was first developed in the 1980s and filled a crucial need for field linguists at the time. However, it is based on an obsolete approach that does not allow for a proper means of data validation (see Robinson et al. (2007) for an overview). Nevertheless, no successor has definitively shifted it from its position as a tool of choice among documentary linguists, even though there is no computational reason for this. The problems here seem largely cultural: The endangered language, computer science, and software developer communities have not yet been able to organize in a way which would allow a tool with the same core functionality to be developed with greater computational sophistication that would also fit as smoothly into the documentary linguist's workflow.

The workshop proposed here seeks to address this state of affairs by bringing together papers exploring the use of computational methods to facilitate the documentation and study of endangered languages. It is being supported by funding from the National Science Foundation Award No. 1404352 and will be followed by a one-day closed meeting where the same issues will be considered by invited participants meeting in breakout groups.

Despite the concerns listed above, recent efforts do indicate that there is significant potential in collaboration between computational linguists (and other computer scientists) and linguists working on endangered languages. The results of Palmer et al. (2009), for instance, suggest that machine labeling and active learning can make the process of textual analysis of low-resource languages more efficient. In another vein, Bender (2008) demonstrates that state-of-the-art tools in grammar engineering can be applied at a relatively low cost to new languages that are typologically divergent from those that primarily informed their design. Moreover, new models of data collection based on the ubiquity of low-cost, networkable devices with recording capabilities, such as smartphones, show the extent to which the barriers to collecting significant amounts of primary data have fallen in recent years (Bird 2010), and it has similarly been found that the pairing of crowdsourcing and machine translation techniques can yield useful results for low resource languages in a short time frame (Lewis et al. 2011). Research along these latter lines, in particular, indicates that computationally-driven advances in the documentation of the world's languages may

[1] http://www-01.sil.org/computing/catalog/show_software.asp?id=79

need to rely as much on clever engineering and user-interface solutions as on methods for processing language data developed within computational linguistics proper, in a manner parallel to efforts in other domains that have considered how new online services can be used to facilitate computational linguistic research (Snow et al. 2008). The potential of all of these developments has not gone unrecognized, as evidenced by the recent efforts of the NSF-funded AARDVARC project[2].

A different set of activities within the documentary linguistics community involving the increasing use of open standards for encoding language data is also significant in this regard. For instance, in the last decade, standardized XML formats have become more widely used to encode text annotations and lexical data (see, e.g., Palmer & Erk (2007)). This facilitates the reuse of documentary materials. Even in the absence of the use of such standards, significant results have been achieved in gathering structured data from materials placed on the web (Lewis & Xia 2010). As more data becomes available in standardized forms, there will only be increased potential for building new kinds of language resources (Abney & Bird 2010).

Despite these positive signs, it is clear that more concentrated efforts are needed if the full promise of computational research on endangered languages is to be realized, with computational methods allowing documentary linguists to work more effectively and documentary data representing a more diverse array of languages becoming available for use by computer scientists. This workshop, therefore, seeks to provide a forum for papers: (i) examining the use of specific methods in the analysis of data from low-resource languages, with a focus on endangered languages, (ii) proposing new models for the collection and management of data in endangered language settings, and (iii) considering what concrete steps are required to allow for a more fruitful interaction between computer scientists and documentary linguists. Its intention is not merely to allow for the presentation of research on these topics but also to help build a community of computational and documentary linguists who are able to effectively pair together to serve their common interests.

## References

Abney, S. & S. Bird. 2010. The human language project: Building a universal corpus of the world's languages. In *Proceedings of the 48th Annual Meeting of the ACL*, 88–97. ACL.

Bender, E. M. 2008. Evaluating a crosslinguistic grammar resource: A case study of Wambaya. In *Proceedings of ACL-08: HLT*, 977–985. ACL.

Bird, S. 2010. A scalable method for preserving oral literature from small languages. In G. Chowdhury, C. Khoo & J. Hunter (eds.), *The role of digital libraries in a time of global change: Twelfth International Conference on Asia-Pacific Digital Libraries (ICADL 2010)*, 5–14. Berlin: Springer.

Lewis, W. D., R. Munro & S. Vogel. 2011. Crisis MT: Developing a cookbook for MT in crisis situations. In *Proceedings of the sixth Workshop on Statistical Machine Translation (WMT-2011)*, 501–511. ACL.

Lewis, W. D. & F. Xia. 2010. Developing ODIN: A multilingual repository of annotated language data for hundreds of the world's languages. *Literary and Linguistic Computing* 25. 303–319.

Palmer, A. & K. Erk. 2007. IGT-XML: An XML format for interlinearized glossed text. In *Proceedings of the Linguistic Annotation Workshop (LAW '07)*, 176–183. ACL.

Palmer, A., T. Moon & J. Baldridge. 2009. Evaluating automation strategies in language documentation. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, 36–44. ACL.

Robinson, S., G. Aumann & S. Bird. 2007. Managing fieldwork data with Toolbox and the Natural Language Toolkit. *Language Documentation & Conservation* 1. 44–57.

Snow, R., B. O'Connor, D. Jurafsky & A. Y. Ng. 2008. Cheap and fast—but is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP '08)*, 254–263. ACL.

[2] http://linguistlist.org/aardvarc/