Informal Statement on Open Linguistic Data Issues

Doug Cooper, CRCL

doug.cooper.thailand@gmail.com, *http://sealang.net*

This is just a brief statement on points I raised for this meeting.

Issue: Common statement on open access to federally funded linguistics resources

Argument: US federal funders in the NSF, NEH, USED (Dep't of Ed), and elsewhere support language-data production through a variety of programs. However, regulations that govern these programs don't unambiguously address questions about reuse of data, and legislation currently under debate (such as the *Federal Research Public Access Act of 2009*) is primarily directed at peer-reviewed published articles, not data sets.

I think it would be useful to have some back-and-forth with funding agency reps on the most practical and effective approach to making provisions for open access to project results be an integral part of grant solicitations and RFPs. For example, such access could be seen as an integral part of most grants' "dissemination" requirement, or the agencies could make it clear that requests for public-interest data reuse will be approved automatically.

Outcome: A statement from this working group ("open access good; no access bad!") is not the goal – rather, we want to help agency representatives find the wording they need to introduce or strengthen their own messages to federal grantees.

Issue: Agenda on "open linguistic data" for low-affluence languages

Argument: About 50 languages, give or take, probably have enough economic impact to support the development of linguistic data purely on the basis of anticipated open-market benefits. (Why 50? Number of languages in Google Translate: 51; number with GLP[1] [combined GDP of all speakers] >11 digits: 44.) Even doubling that to account for research/academic production of resources, for the great majority of languages concerns like scaling parallel corpora to support machine translation is not a pressing issue.

My concern is that a group devoted to defining an agenda for open linguistic data ought to consider, and if appropriate help to set, what this might mean for the 'other' 5,950 (or 6,950, or whatever the current estimate is) languages.

Programs like REFLEX-LCTL[2] have addressed the issue for one group of languages (typically with extensive written resources) from the computational linguistics side, and a variety of initiatives from NSF, NEH, HRELP, DoBeS, etc. support work on languages whose written traditions are limited or nonexistent, often from the vantage point of endangered language documentation.

Outcome: Targets for data acquisition / analysis / open access that are reasonable for languages large and small, and are useful from both the computational and documentation perspectives.

[1] *Hammarström, Harald* 2009. "A Survey of Computational Morphological Resources for Low-Density Languages." Chapter Two of *www.cs.chalmers.se/~harald2/phd.pdf*.

[2] *Simpson, Heather* et al 2008. "Human Language Technology Resources for Less Commonly Taught Languages" *www.lrec-conf.org/proceedings/lrec2008/workshops/W10_Proceedings.pdf*

Issue: Best practices for Web APIs for open linguistic data and services

Argument: In coming years there is likely to be greater reliance on providing data and functionality via Web services, as an alternative to archiving / exchanging full data sets. Interchange of fixed language resources will be extended by the availability of increasingly rich language *services*. The current paradigm of open linguistic data (standardized markup supports reusable tools) will be extended by APIs that abstract underlying structure, and provide fine-grained access to point-of-use data and services.

At present, Web APIs for linguistic data are limited indeed; I know of no widely used API for corpus-oriented data or services, and even dictionary services are primitive (e.g. DICT's RFC 2229, see www.dict.org) or reflect specific data implementations (e.g. APIs for WordNet access).

Outcome: What are the abstract resources and services that a linguistic data API might provide for dictionaries, text / bitext / audio / video / image corpora? For other linguistic data (e.g. grammars)?

Issue: Alternative model for long-term open access to Web services and resources

Argument: The world's primary repositories are LDC and ELRA, which have also done an excellent job of defining the purpose and process of linguistic resource management. Both organizations are funded by subscription; it may now be an appropriate time to seek a funding mechanism that doesn't (usually) entail prohibiting further redistribution of resources; a la DoBeS (Volkswagen Foundation, www.mpi.nl/DOBES/), or existing federal models (e.g. Library of Congress *Digital Collections* or the National Archives).

A second concern is ensuring persistent access to Web services / dynamically served data provided via Web APIs. Although repositories (and the Internet Archive) mirror static data in sites, they do *not* generally mirror servers or provide services – when the original server goes down, the API is lost.

Technical solutions are readily available; e.g. packaging a site as a so-called LAMP (for Linux, Apache server, MySQL, and Perl/Python/PHP; Windows-oriented WAMPS also exist) lets services be instantiated in the future, *if* we can guarantee that an appropriate architecture / operating system can be run. Virtual machines (e.g. as provided by Amazon ec2, or at home by VMware) address this problem; they emulate hardware and allow installation of long-gone OSs.

Outcome: Two issues need to be addressed. First, one can only speculate for how many years it will be possible to boot a circa 2010 machine. I'd guess at least two decades, but it's not really my area. Secondly, who can commit to providing such services (LDC? LoC? DoBeS? Google?), what will it cost, and how will funding be assured?

Issue: Linguistics publication as an essential element of open linguistic data

Argument: Access to publication is of central importance to linguistics; large bodies of data are found not in books and monographs, but in articles, conference proceedings, festschrifts, and other collections, in a legacy that stretches back for more than a century. This material is often poorly indexed and difficult to discover; when it is aggregated by major US and European suppliers, it may be unaffordable or inaccessible in the countries in which the languages are spoken.

Outcome: Include access to publication in consideration of open linguistic data.