LAP, LICHEN, and DASS – Experiences combining data and tools

Opas-Hänninen, Lisa Lena
lisa.lena.opas-hanninen@oulu.fi
University of Oulu, Finland

Juuso, Ilkka
ijuuso@ee.oulu.fi
University of Oulu, Finland

Kretzschmar, William A. Jr.
kretzsch@uga.edu
University of Georgia, USA

Seppänen, Tapio
tapio@ee.oulu.fi
University of Oulu, Finland

The Linguistic Atlas team at the University of Georgia (LAP) and the LICHEN research team at the University of Oulu have been investigating the application of advances in information engineering to humanities scholarship, in particular methods for managing and mining large-scale linguistic databases. Our aim has been to bring together the linguistic and technological expertise from Oulu and Georgia in order to develop practical solutions for common problems. In this paper we will show the results of this cooperation--including the Digital Archive of Southern Speech (DASS), the pilot product for LAP-LICHEN released in 2009--and discuss our experiences and lessons learned.

The LAP audio archive, amounting to 7000 hours of interviews, is an unparalleled resource for study not only of the common language of the US but for its culture more generally, stories of daily life in America. Study of LAP interviews so far has taken advantage only of small bits of transcribed data extracted from the full interview. The large, untranscribed bulk of LAP interviews consists of the speakers' accounts of their lives and their families, their occupations and their diversions, their houses and their land. Along with direct questioning and conversational passages, in a quarter of the thousands of audio files so far processed these stories take the form of narratives of at least one minute of continuous speech. To preserve and to make this audio archive available puts the people back into what has seemed to some scholars to be a dry academic exercise of collecting words. The LAP team has always appreciated the personal, individual nature of each interview, as well as the way that the interviews can represent American culture; with DASS, we can now share that appreciation much more broadly with both the academic community and with the public.

DASS is a collection of 64 interviews from the Linguistic Atlas of the Gulf States (LAGS) selected by LAGS Director Lee Pederson. Four interviews come from each of the sixteen regional LAGS sectors. Within each sector there is one speaker from each Atlas Type: folk (largely uneducated and insular), common (moderate education and experience), cultivated (higher education and/or participation in high culture). One African American speaker was selected from each sector, and folk, common, and cultivated African American speakers are distributed across the sectors. Speakers cover a wide range of ages and social circumstances. Over 400 hours of audio files are provided both as large uncompressed .wav files (useful for acoustic phonetic processing) and as thousands of small .mp3 files for general listening. Files are indexed according to subject matter and speaker, according to a set list of 40 topics. Metadata and finding aids for particular topics and kinds of speakers are provided, including search tools and a GIS function. Together the DASS data and the LICHEN tools comprise about 200 GB of data, provided on a portable USB drive. The interviews were digitized and processed by the LAP team at the University of Georgia with assistance from a grant from the National Endowment for the Humanities (PW-50007, "Digitization of Atlas Audio Recordings", with Opas-Hänninen as partner). The first phase of the LICHEN project was lead jointly by Opas-Hänninen and Seppänen and funded by a grant from the Emil Aaltonen Foundation (2006-2008, with Kretzschmar as an international collaborator). The University of Oulu and the University of Georgia drew up legal agreements regarding copyrights and the distribution of the software with the data. The DASS/LICHEN package is distributed by the LAP.

DASS is only the beginning, however. The research team at the University of Oulu has developed LICHEN as an electronic framework, i.e. a type of toolbox, which handles multimodal data. The toolbox has been developed using two sets of data as testbeds, namely the Oulu Archive of Minority Languages in the North containing samples from the Kven, Meänkieli, Veps and Karelian languages, as well as DASS. Some of the data from minority languages exists as video, which the toolbox handles along with audio and text. We are now working on a transcription tool, so that audio and video materials can be provided with textual representations aligned with the sound and video. Finally, we are rebuilding LICHEN as a Web-enabled framework, so that users can access our language and cultural materials remotely in line with the movement for the creation of public corpora (see, e.g., Kretzschmar, Anderson, Beal, Corrigan, Opas-Hänninen, and Plichta 2006; Kretzschmar, Childs, and Dollinger 2009).

LAP-LICHEN cooperation began in 2004, when Kretzschmar, Opas-Hänninen, Anderson, Beal, and Corrigan met in Newcastle, to follow up on conversations about public corpora begun earlier. The group found that there were common problems, standards, best practices for the corpora managed by those attending, and agreed to prepare a presentation at ICAME in 2005 to highlight the possibility for shared methods and joint actions (later published as the 2006 programmatic article). The LAP-LICHEN collaboration bloomed as a result. Grants were obtained for cooperation: the LICHEN model was included in a large NEH proposal for archival digital audio processing for LAP, and conversely LAP was adopted as the large-scale test for the enhancement of LICHEN at Oulu. An operational version of LICHEN that might be used for LAP was available and demonstrated at DH2007 (Urbana). Further development occurred in conjunction with DH2008 (Oulu), leading to testing of LICHEN on LAP materials in Georgia in Fall 2008. As a result of these steps, the collaborators rewrote the specifications for what the program needed to do in February 2009, as substantial bodies of archivally-processed LAP sound files became available for LICHEN development and testing. The collaborators decided that the key requirements for the specification arose from not only the characteristics of the data (e.g. the structure consisting of interviews, reels and clips with metadata and multimedia on each level) and the desired uses of the framework (e.g. search, browse, and view possible audio selections on a map with GIS), but also from the fact that both tools development and the final stages of data preparation were taking place simultaneously. The data had to be available as flat files usable through any regular file browser, and the sheer scale of the data ruled out the possibility of creating duplicate files inside the program structure as originally designed. The software needed to make use of the existing files and file structure, and so the task of combining the data and the tools became an exercise in conforming tools to data with as little effect on the data itself as possible. To this end, the collaborators developed two methods for bringing in the data and its associated metadata: 1) a general-purpose parser that could traverse file and folder structures and evaluate regular expression patterns to parse metadata from the file and folder names, and 2) a mechanism for re-formatting standard spreadsheet documents into XML documents for use by the tools. The existence of the tools affected the preparation of the data in two ways: 1) some incorrectly named files and folders had to be renamed to conform to the agreed format, and 2) all text files had to be converted from the Microsoft Word format into a plain text format for viewing from within the developed tools. Both changes were also beneficial to the data collection itself, improving consistency within the data and file support across different platforms. In turn, the developed tools provide access into the database through browsing of the data by natural entities such as interviews, topics and geographic location (as opposed to just files and folders) and queries leveraging the full potential of all the metadata fields.

The DASS product was launched in April 2009 at the SECOL conference (New Orleans), and public distribution began in the summer of 2009 after resolution of the legal issues of the collaboration with university authorities at Georgia and Oulu. Even given the close collaboration between developers, arriving at legal language that suited the university authorities proved to be quite difficult: the Georgia lawyers thought they could be more laissez-faire because the product was unlikely

to generate substantial monetary returns, while the Oulu authorities were more focused on retaining the university's rights of ownership. In the end, separate rights language had to be included for materials developed at Georgia and for LICHEN development at Oulu.

As might be expected, integration of a large-scale database with multimedia display functions, while still maintaining the high degree of usability necessary for public access, has turned out to be more difficult than accomplishing any of the separate tasks. And we are not finished. We are currently building a transcription tool to incorporate into the LICHEN framework, so that the public (as well as professional researchers) can contribute transcriptions of audio files. We also want to increase the Geographic Information System (GIS) functionality into the LICHEN toolbox in order to support map-based selection and visualization schemes, surpassing those implemented on the existing Atlas Web site. Finally, we want incorporate the CATMA concordancing tools, currently being built by a collaborating research team at the University of Hamburg. All of these goals, we trust, will make LAP and other data accessible, not just on portable media, but on the web with LICHEN.

References

Kretzschmar, William A. Jr., Anderson, Jean, Beal, Joan, Corrigan, Karen, Opas-Hänninen, Lisa Lena, Plichta, Bartek (2006). 'Collaboration on Corpora for Regional and Social Analysis'. *Journal of English Linguistics*. **34**: 172-205.

Kretzschmar, William A. Jr., Childs, Becky, Dollinger, Stefan (2009). 'Creating Public Corpora: Accessibility, Copyright and Enhancement, and Human Subjects and Metadata'. *Workshop offered at NWAV 38*. Ottawa, October 22-25, 2009.

LAP – Linguistic Atlas Projects; includes LAGS and DASS. `http://us.english.uga.edu/`.

LICHEN – The Linguistic and Cultural Heritage Electronic Network. `http://www.lichen.oulu.fi/`.