Alveo, a Human Communication Science Virtual Laboratory

Dominique Estival
U. of Western Sydney
d.estival@uws.edu.au

Steve Cassidy
Macquarie University
steve.cassidy@mq.edu.au

Abstract

We give a hands-on demonstration of the Alveo Virtual Laboratory, a new platform for collaborative research in human communication science (HCS). Funded by the Australian Government National eResearch Collaboration Tools and Resources (NeCTAR) program, Alveo involves partners from a range of disciplines: linguistics, natural language processing, speech science, psychology, as well as music and acoustic processing. The goal of the platform is to provide easy access to a variety of databases and a range of analysis tools, in order to foster inter-disciplinary research and facilitate the discovery of new methods for solving old problems or the application of known methods to new datasets. Alveo integrates a number of tools and enables non-technical users to process communication resources (including not only text and speech corpora but also music recordings and videos) using these tools in a straightforward manner.

1 Introduction

Alveo provides easy access to a range of databases relevant to human communication science disciplines, including speech, text, audio and video, some of which would previously have been difficult for researchers to access or even know about. The system implements a uniform and secure license management system for the diverse licensing and user agreement conditions required. Browsing, searching and dataset manipulation are also functionalities which are available in a consistent manner across the data collections through the web-based Discovery Interface.

2 Alveo Tools and Corpora

The first phase of the project, from December 2012 to June 2014 (Estival et al. 2013) saw the inclusion of data collections contributed by the project partners (see the list of partners in the Acknowledgments section). Some of these were already well-known, e.g. 1, 3 and 9, but some had been difficult of access or not available, e.g. 2, 5, 6, 7, 8.

1. PARADISEC (Pacific and Regional Archive for Digital Sources in Endangered Cultures: http://paradisec.org.au): audio, video, text and image resources for Australian and Pacific Island languages (Thieberger et al. 2011).
2. AusTalk (https://austalk.edu.au): audio-visual speech corpus of Australian English (Burnham et al. 2011).
3. The Australian National Corpus (https://www.ausnc.org.au) (Cassidy et al. 2012) comprising: Australian Corpus of English (ACE); Australian Radio Talkback (ART); AustLit; Braided Channels; Corpus of Oz Early English (COOEE); Griffith Corpus of Spoken English (GCSAusE); International Corpus of English (ICE-AUS); Mitchell & Delbridge corpus; Monash Corpus of Spoken English (Musgrave and Haugh 2009).
4. AVOZES, a visual speech corpus (Goecke and Millar 2004).
5. UNSW Pixar Emotional Music Excerpts: Pixar movie theme music expressing different emotions.
6. Sydney University Room Impulse Responses: environmental audio samples which, through convolution with speech or music, can create the effect of that speech or music in that acoustic environment.
7. Macquarie University Battery of Emotional Prosody: sung sentences with different prosodic patterns.
8. Colloquial Jakartan Indonesian corpus: audio and text, recorded in Jakarta in the early 1990's (ANU).
9. ClueWeb, a dataset consisting of 733,019,372 English web pages collected between 10/02/2012 and 10/05/2012 (lemurproject.org/clueweb12).

Through the web-based Discovery interface (see Figure 2) the user can select items based on the

Dominique Estival and Steve Cassidy. 2014. Alveo, a Human Communication Science Virtual Laboratory. In *Proceedings of Australasian Language Technology Association Workshop*, pages 104−107.

results of faceted search across the collections and can organise selected data in Items Lists. Beyond browsing and searching, Alveo offers the possibility of analysing and processing the data with a range of tools. In the first phase of the project, the following tools were integrated within Alveo:

1. EOPAS (PARADISEC tool) for interlinear text and media analysis.
2. NLTK (Natural Language Toolkit) for text analytics with linguistic data (Bird, Klein, and Loper 2009).
3. EMU, for search, speech analysis and interactive labelling of spectrograms and waveforms (Cassidy and Harrington 2000).
4. AusNC Tools: KWIC, Concordance, Word Count, statistical summary and analysis.
5. Johnson-Charniak parser, to generate full parse trees for text sentences (Charniak and Johnson 2005).
6. ParseEval, to evaluate the syllabic parse of consonant clusters (Shaw and Gafos 2010).
7. HTK-modifications, a patch to HTK (Hidden Markov Model Toolkit, to enable missing data recognition. (http://htk.eng.cam.ac.uk/).
8. DeMoLib, for video analysis (http://staff.estem-uc.edu.au/roland/research/demolib-home/).
9. PsySound3, for physical and psycho-acoustical analysis of complex visual and auditory scenes (Cabrera, Ferguson, and Schubert 2007).
10. ParGram, grammar for Indonesian (Arka 2012).
11. INDRI, for information retrieval with large data sets (http://www.lemurproject.org/indri/).

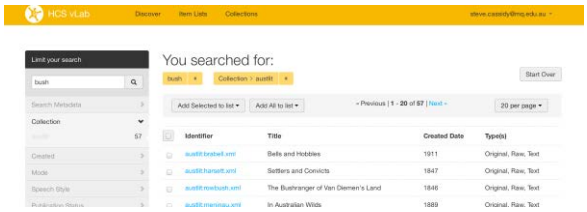

Figure 1: Screenshot of the Alveo Web interface

3 Alveo Architecture

Most of these tools require significant expertise to set up and one of the Alveo project goals is to make this easier for non-technical researchers. The Alveo Workflow Engine is built around the Galaxy open source workflow management system (Goecks et al. 2010), which was originally designed for use in the life sciences to support researchers in running pipelines of tools to manipulate data. Workflows in Galaxy can be stored, shared and published, and we hope this will also become a way for human communication science researchers to codify and exchange common analyses.

A number of the tools listed above have been packaged as Python scripts, for instance NLTK based scripts to carry out part-of-speech tagging, stemming and parsing. Other tools are implemented in R, e.g. EMU/R and ParseEval. An API is provided to mediate access to data, ensuring that permissions are respected, and providing a way to access individual items, and 'mount' datasets for fast access (Cassidy et al. 2014). An instance of the Galaxy Workflow engine is run on a virtual machine in the NeCTAR Research Cloud, a secure platform for Australian research, funded by the same government program (nectar.org.au/research-cloud). Finally, a UIMA (Unstructured Information Management Architecture) interface (Verspoor et al. 2009) has been developed to enable the conversion of Alveo items, as well as their associated annotations, into UIMA CAS documents, for analysis in a conventional UIMA pipeline. Conversely annotations from a UIMA pipeline can be associated with a document in Alveo (Estival et al. 2014). Figure 2 gives an overview of the architecture.

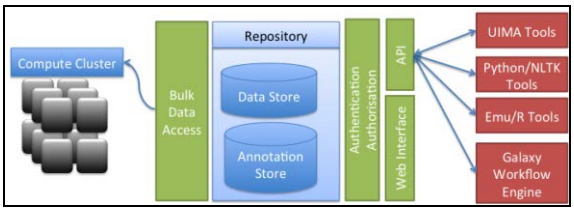

Figure 2: The architecture of the Alveo VL

4 User Acceptance Testing

Alveo was designed and implemented in partnership with Intersect, a commercial software development company specialised in the support of academic eResearch. This partnership afforded extensive professional support during development, using the Agile process (Beck et al. 2001) as well as thorough regression testing and debugging. In other projects of this type, Intersect provide User Acceptance Testing (UAT) or managed the UAT process in-house. For the Alveo project, user testing was the main way in which the academic partners were involved in the pro-

ject. The central team at the lead institution oversaw the creation of the tests, distributed the tests and monitored the results.

Some testers were Linguistics students with no computing background, some were Computer Science students with limited linguistic knowledge. At some sites, the testers were Research Assistants who had worked on the tools or corpora contributed by their institutions, while others were the tool developers themselves. This variety of backgrounds and skills ensured coverage of the main domains and functionalities expected of the Alveo Virtual Lab. Some sites had undertaken to conduct large amounts of testing throughout the development, while other partners only chose to perform limited or more targeted testing, with commitments varying from 10 to 200 hours. Over 30 testers participated at various times during of the project and a total of more than 300 hours has been spent on testing during Phase I.

For each version of the system during development, a series of tests were developed. The first tests were very directive, giving very specific instructions as to what actions the user was asked to perform and what results were expected for each action. Gradually the tests became more open-ended, giving less guidance and gathering more informative feedback. The latest round of testing asked Testers to log in and to carry out a small research task. Some of the early tests, have become tutorials provided on the Alveo web page and are now available as help from within the Virtual Lab. We will use these as the basis for the hands-on demo.

5 Conclusion

One of the conditions of success of such a project is that the platform be used by researchers for their own projects and on their own data. The organisation of the User Acceptance Testing, requiring partners to contribute during the development, and providing exposure to the tools and the datasets to a large group of diverse researchers is expected to lead to a much wider uptake of Alveo as a platform for HCS research in Australia. Alveo is now open to users outside the original project partners. We will also continue to explore further interactions with complementary frameworks, such that the data and annotation storage available in Alveo can be enhanced via processing and tools from external services to supplement the functionality that is currently directly integrated.

We hope that by presenting Alveo to the Australian NLP community, we will encourage researchers to consider using Alveo as a potential repository for their data and as a platform to conduct new analysis. Alveo is already used in teaching a Computational Linguistics course at Monash University and we would encourage more instances of such educational use of the platform. Finally, we would like to invite students as well as researchers in HCS fields to propose tools and corpora which they would like to use in their own research for future inclusion in Alveo.

Acknowledgements

We thank NeCTAR for its financial support during Phase I and all the project partners (University of Western Sydney, RMIT, Macquarie University, Intersect, University of Melbourne, Australian National University, University of Western Australia, University of Sydney, University of New England, University of Canberra, Flinders University, University of New South Wales, La Trobe University, University of Tasmania, ASSTA, AusNC Inc. NICTA) for their on-going contributions to the project.

References

Arka, I. Wayan. 2012. "Developing a Deep Grammar of Indonesian within the ParGram Framework: Theoretical and Implementational Challenges " 26th Pacific Asia Conference on Language, Information and Computation.

Beck, Kent, Mike Beedle, Arie van Bennekum, Alistair Cockburn, Ward Cunningham, Martin Fowler, James Grenning, Jim Highsmith, Andrew Hunt, Ron Jeffries, Jon Kern, Brian Marick, Robert C. Martin, Steve Mellor, Ken Schwaber, Jeff Sutherland, and Dave Thomas. 2001. Manifesto for Agile Software Development. http://agilemanifesto.org/.

Bird, Steven, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python - Analyzing Text with the Natural Language Toolkit*: O'Reilly Media.

Burnham, Denis, Dominique Estival, Steven Fazio, Felicity Cox, Robert Dale, Jette Viethen, Steve Cassidy, Julien Epps, Roberto Togneri, Yuko Kinoshita, Roland Göcke, Joanne Arciuli, Marc Onslow, Trent Lewis, Andy Butcher, John Hajek, and Michael Wagner. 2011. "Building an audio-visual corpus of Australian English: large corpus

collection with an economical portable and replicable Black Box." Interspeech 2011, Florence, Italy.

Cabrera, Densil , Sam Ferguson, and Emery Schubert. 2007. "'Psysound3': Software for Acoustical and Psychoacoustical Analysis of Sound Recordings." International Community on Auditory Display.

Cassidy, Steve, Dominique Estival, Timothy Jones, Denis Burnham, and Jared Burghold. 2014. "The Alveo Virtual Laboratory: A Web Based Repository API." 9th Language Resources and Evaluation Conference (LREC 2014), Reykjavik, Iceland, 26-31 May 2014.

Cassidy, Steve, and Jonathan Harrington. 2000. "Multi-level Annotation in the Emu Speech Database Management System." *Speech Communication* 33:61–77.

Cassidy, Steve, Michael Haugh, Pam Peters, and Mark Fallu. 2012. "The Australian National Corpus : national infrastructure for language resources." LREC.

Charniak, Eugene, and Mark Johnson. 2005. "Coarse-to-fine n-best parsing and MaxEnt discriminative reranking." 43rd Annual Meeting on Association for Computational Linguistics.

Estival, Dominique, Steve Cassidy, Peter Sefton, and Denis Burnham. 2013. "The Human Communication Science Virtual Lab." 7th eResearch Australasia Conference, Brisbane, Australia, October 2013.

Estival, Dominique, Steve Cassidy, Karin Verspoor, Andrew MacKinlay, and Denis Burnham. 2014. "Integrating UIMA with Alveo, a human communication science virtual laboratory." Workshop on Open Infrastructures and Analysis Frameworks for HLT, COLING 2014, Dublin, Ireland.

Goecke, Roland, and J.B. Millar. 2004. "The Audio-Video Australian English Speech Data Corpus AVOZES." 8th International Conference on Spoken Language Processing (INTERSPEECH 2004 - ICSLP), Jeju, Korea.

Goecks, Jeremy, Anton Nekrutenko, James Taylor, and The Galaxy Team. 2010. "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences." *Genome Biology* 11 (8):R86.

Musgrave, Simon, and Michael Haugh. 2009. "The AusNC Project: Plans, Progress and Implications for Language Technology." ALTA 2009, Sydney.

Shaw, Jason A., and Adamantios I. Gafos. 2010. "Quantitative evaluation of competing syllable parses." 11th Meeting of the Association for Computational Linguistics. Special Interest Group on Computational Morphology and Phonology, Uppsala, Sweden.

Thieberger, Nick, Linda Barwick, Rosey Billington, and Jill Vaughan, eds. 2011. *Sustainable data from digital research: Humanities perspectives on digital scholarship. A PARADISEC Conference*: Custom Book Centre. http://ses.library.usyd.edu.au/handle/2123/7890.

Verspoor, Karin, William Baumgartner Jr, Christophe Roeder, and Lawrence Hunter. 2009. "Abstracting the types away from a UIMA type system." *From Form to Meaning: Processing Texts Automatically*:249-256.