An Intelligent Search Infrastructure for Language Resources on the Web

ARC Special Research Initiative (E-Research) SR0567353
Chief Investigators: Timothy Baldwin, Steven Bird, Baden Hughes

Initiative Description - Public Version (January 2006)
http://lt.csse.unimelb.edu.au/projects/langsearch

Aims, Significance and Background

Language occupies a central role on the web: most content is expressed in a given language, and most access takes place via natural language input and interfaces. Today, investigation of human language in all its forms depends on access to this vast store of language data. In particular, linguists and language technologists annotate and analyze this data and develop new language resources including grammars, dictionaries, and a raft of new technologies for automatic translation, information extraction, question answering, and so forth. As this new documentation is disseminated on the web, and as the new technologies are in turn deployed on the web, a further round of collection and processing is enabled, closing the loop. For instance, a collection of Japanese text with an aligned English translation can be used for translation studies, for adding examples to bilingual dictionaries, and developing translation systems. These resources can then be used for new purposes, e.g. to provide English speakers access to content stored in Japanese text, or to provide Japanese learners of English with more authentic example sentences.

In the first five years of the web English content was dominant. Then in mid-2000, the combined content from all other languages exceeded English for the first time; the growth of this non-English content continues to outstrip the growth of English content. Most striking of all has been the emergence of websites run by or behalf of small indigenous communities, containing texts, dictionaries and other linguistic resources. Many of these languages are endangered, and it is important to preserve this heritage for the benefit of future scientific and humanitarian work. Here, the web enables low-cost mass publication, uniting distributed communities of language speakers and scholars into a critical mass.

In the Australian context, interest in multilingual web content is significant, owing to major waves of immigration from Europe, Asia, and the Middle East. As an example, the City of Melbourne website [9] highlights the cultural and linguistic diversity of the city's inhabitants, and states that they have come to Melbourne from no fewer than 130 countries. The Australian Bureau of Statistics (ABS) reports that at 30 June 2002, Australia's overseas-born residents comprised 4.6 million people, 23% of the total estimated resident population (19.6 million) [3, 4]. Correspondingly, ABS reports that 522 languages are spoken in Australia according to the 2001 Census [2]. In recent years, the source of inbound migration has changed markedly from a majority English-speaking migrants to a majority non-English speaking. This growing multilingual diversity is reflected in the popular media. The Australian Broadcasting Commission (ABC) [1] provides radio and television broadcasts and web content in 18 languages, while the Special Broadcasting Service (SBS) [24] provides content in more than 60 languages. The National Ethnic and Multicultural Broadcaster's Council [17] identifies 500 organisations producing more than 1700 hours per week of broadcast content in over 100 languages.

This situation is in stark contrast with the state of Australia's indigenous languages. The most recent estimates by McConvell and Thieberger [16] indicate Australia's indigenous linguistic diversity is in the order of 250-300 languages identified since settlement. More recently however, the number of indigenous languages spoken in Australia has declined rapidly, to the extent that more than 50% of these original languages are now extinct, and with another 25-30% expected to disappear with the current generation of speakers. Hence the collection of language data for Australia's indigenous languages is of considerable urgency, and much linguistic data is being published online by researchers and indigenous communities in an effort to address the decline.

This methodology of publishing and collating material on the web for the purposes of linguistic re-

search and language technology development has been very successful. This success is due in large measure to the quality of web search engines and the ability of users to adapt to the exigencies of keyword query. However, therein lies a critical weakness: as the web grows, resource discovery has become a hit-and-miss affair; it is easy for users to be inundated with irrelevant resources, and to miss important resources because they did not try enough combinations and translations of query terms. A recent attempt to address this problem has been OLAC, the Open Language Archives Community [21, 8]. OLAC applies new technologies in digital libraries from the Open Archives Initiative [20] to support a worldwide virtual library of language resources. OLAC users can search over 30,000 resources in over 30 language archives simultaneously, using keyword or fielded search over the stored metadata, using either a customised search engine [13] or through a DP9 Gateway accessible to Google. However, OLAC has two major shortcomings of its own. First, it can only be used to search resources which have already been catalogued. Second, most of the resources, once found, cannot be accessed because they are not available on the web.

The shortcomings of the available resource discovery technologies are most easily understood with the help of some hypothetical scenarios. We assume that the user is a language researcher, teacher, or learner, and is searching for specific resources:

Scenario 1: Finding resources for a specified language. A user searches the web using the name of the language. However, they experience low precision as this language name is also a normal word in other languages. The also experience low recall, since there are a variety of spellings for the language name, and since most texts in the language do not explicitly identify the language anyway. Attempts to refine the search by limiting its scope to a country (e.g. `site:.ar`) do not prune the result set to a manageable size, and they eliminate some of the most useful resources created by the diaspora. Additional keywords like *dictionary* must be entered in several languages (e.g. *wörterbuch, diccionario*), requiring more effort for limited returns.

Scenario 2: Finding resources by proximity. A research project is investigating some geographical region like the Afghanistan/Pakistan borderland, and requires information on the linguistic situation including numbers of languages, population per language, literacy level, and available language resources. Web search using the name of the political region only turns up materials about companies and government agencies. Another project is investigating Australia's Western Desert Languages, and the user discovers that it is necessary to search for each of a dozen language names (and variant spellings) separately, a tedious and error-prone process.

Scenario 3: Finding examples of a linguistic construction. A user wants to find examples of sentences that contain multiword expressions that match a specified template, such as verb-particle constructions involving the word "up" (e.g. "... put the team up"). Searching just on the word "up" turns out to be fruitless. The user picks a verb at random and tries using Google's starred expressions, e.g. "put * * up". This finds a handful of examples but gives no sense of which verb-particle constructions most often involve "up". By using automatic means to identify instances of particular constructions in the web corpora collected in this project [5, 6, 15], and mapping the results onto a language-universal annotation schema of syntactic construction types [7], we are able to support such queries.

Aims: This initiative will develop new infrastructure for language research: namely a language-aware search engine. It will use existing technologies and software components developed by the investigators, and deployed using the infrastructure at Melbourne University's Advanced Research Computing Centre. In particular, we will develop:

Language Crawler: This will selectively obtain text from the web, with an emphasis on the languages of economic, scientific and cultural interest to Australia. This material will be stored in a centralized repository for the purposes of indexing, annotation and preservation.

Metadata Creation: Each resource will be automatically classified, and an OLAC metadata record

will be generated which identifies language and linguistic resource type. Texts will be annotated using robust language technologies to permit sophisticated indexing and retrieval.

Language Archive: High quality resources, including dictionaries and texts from endangered languages, will be archived to ensure they remain accessible. As it accrues, this content will provide important seed data to the crawler.

Language-Aware Search Engine: This will permit users to enter language names, location names, linguistic constructions, or specify regions on a world map, returning URLs.

The significance of the proposed research lies in its integration of existing infrastructure developed by the investigators with the high-performance computing infrastructure at the University of Melbourne, for the high-profile application of language-aware web-search. High performance computing will play three key roles in the project, for large-scale processing, storage, and access. First, we will process hundreds of gigabytes of text in hundreds of languages, classifying each item by language and linguistic type, applying annotations which are expensive to compute, and indexing the results. Second, we will host the primary data, along with annotations and large indexes, requiring terabytes of storage. Third, we will provide public access to the materials using our own local search engine, which will need to support multiple simultaneous accesses to the data, collating and summarizing the results for end-users.

The project will also be significant on account of the service it provides to a potentially large community of people interested in language, including researchers, teachers, and students. The project will provide them with a completely new window on the web, enabling them to quickly discover and access resources in and about language.

Outline of the Proposed Initiative

The research and development will integrate existing data sets and software components which have been developed by the investigators and their colleagues. These are all based on state-of-the-art web technologies, international standards in language engineering, and international best practices in metadata creation. The project will be undertaken during the period January 2006 – December 2006. Project activities will be structured into four stages, each corresponding to one of the project goals, and each lasting three months:

1. Language Crawler: This will use pre-existing tools to selectively obtain text from the web which has been identified as being of potential linguistic interest. The language of each such document will be identified through a combination of code-string analysis and character n-gram analysis [14, 23, 10], and those documents authored in languages of economic, scientific and cultural interest to Australia identified. This material will be stored in a centralized repository for the purposes of indexing, annotation and preservation.

This task will consist of a number of discrete sub-tasks. First, we will acquire seed URIs pertaining to language resources (using existing software developed by CI Hughes). Next, we will crawl for web content linked to or from seed URIs (using the Internet Archive's Heritrix software). Following this, we will acquiring the web content for the crawled scopes, again using Heritrix. These first 3 steps will be conducted in a broad context (ie. for all languages around the world) and then narrowed further to focus crawling activity on particular languages relevant to Australia. Having acquired both the URLs and their associated content, we will then deploy the crawler to identify actual language resources. Finally, we will ascertain if the resource found should be archived locally, or merely indexed remotely.

The majority of this module requires integration of existing components and data, previously developed by the CIs. In particular, we already have seed linguistic data URIs from prior projects; seed lexical data in other languages (drawn from the Aboriginal Studies Electronic Data Archive (ASEDA),

SBS Language Services, the SIL Ethnologue, and the Rosetta Project); an existing web query aggregator for the Google, Yahoo, MSN and A9 search engines; web query expansion techniques for linguistic data and for geospatial concepts; heuristics for discovering primary linguistic data inside structured documents; and experimental data concerning different search strategies for linguistic data on the web.

The work items in this module will be carried out in quarter 1 (January - March 2006).

2. Metadata Creation: Each resource will be automatically classified, and an OLAC metadata record will be generated which identifies language and linguistic resource type. Texts will be annotated using robust language technologies to permit sophisticated indexing and retrieval.

This module consists of a number of discrete sub-tasks, namely: language identification, which will in turn be used to seed the Heritrix crawler; linguistic type identification - classifying language resources into various data-oriented categories; coverage identification - identifying the linguistic or geospatial scope cognate to a given resource (including a third party extension to OLAC metadata which will allow the encapsulation of geospatial data with a language resource, using the Dublin Core coverage element).

Again, this module will extend existing work (both software and data), including data from ASEDA, the SIL Ethnologue and SBS Language Services; and software developed by CI Hughes for automated metadata enrichment.

The resulting URLs, together with their OLAC metadata will continue to be provided to a number of international collaborators for mining interlinear text (Lewis), mining grammatical concepts (Langendoen, Bender and Flickinger) and building other corpora (Scannell), all of whom will contribute their outputs as OLAC records.

The work items in this module will be carried out in quarter 2 (April – June 2006).

3. Language Archive: High quality resources, including dictionaries and texts from endangered languages, will be archived to ensure they remain accessible. As it accrues, this content will provide important seed data to the crawler.

Again, this module leverages existing components and data including the OLAC customization of e-prints software and the OLAC harvester infrastructure.

The sub-tasks of this module include: obtaining resources to be archived as an output of work in quarter 2; create OLAC metadata record for each new resource, using technologies refined in quarters 1 and 2; publishing resource catalogue as an OLAC repository; making content available as "collections" in the Linguist's Search Engine [22].

The work items in this module will be carried out in quarter 3 (July – September 2006).

4. Language-Aware Search Engine: This will permit users to enter language names, location names, linguistic constructions, or specify regions on a world map, returning URLs.

In this module we integrate other existing components and data, including OLAC Search Engine[13], OLAC Archive Report cards[12], SIL Ethnologue including spatial data, the Getty thesaurus of geographic names, presentation forms of OLAC metadata including (X)HTML (allowing the embedding of OLAC metadata in web pages) and RDF (allowing interoperability and discovery), together with accumulated knowledge of user requirements from query logs for OLAC and LINGUIST search engines and the Linguist's Search Engine.

In particular, the features of the search engine will include: a language id search operator `lang:`; OLAC metadata record search and display; map search and display (including spatial proximity search, location name search, location keyword `loc:`); annotation search and display (including part-of-speech tagging and syntactic parsing); linguistic proximity search and display; query processing to identify which terms are language names, location names, and so forth; search for "similar resources by location", incl. similarity by type, language, subject and score; display the distribution

of resources by location based on the existing OLAC vocabularies (e.g. show lexicons in Kenya, show resources collected by Pike in Mexico); and automated enrichment of harvested metadata (e.g. add country data).

This module represents the culmination of earlier integration efforts to deliver a new, unique service allowing users to interact with language data in a manner not previously possible.

The work items in this module will be carried out in quarter 4 (October – December 2006).

Infrastructure

The University of Melbourne's Advanced Research Computing Centre (MARCC) is a centrally-funded entity which provides computing infrastructure and research support for projects across the university. The University has identified MARCC as the key enabler for e-Research within the University. This initiative will use this existing infrastructure to enable compute-intensive tasks which are either easily parameterised, parallelised or which demand large CPU and/or memory, such as language identification, metadata creation, linguistic annotation.

We will employ the following MARCC facilities:

- a 16-node cluster (P4 2.4Ghz, 512Mb RAM)
- a 24-node cluster (dual P4 2.4Ghz, 1Gb RAM)
- two high performance compute servers (quad 2.8Ghz Xeons, with 8Gb RAM)
- 4Tb of network-accessible storage

The high level computational middleware deployed over these resources is NorduGrid's Advanced Resource Connector (ARC) [18, 19]. NorduGrid ARC is a production grade confederation-oriented cluster management system which has been used extensively in the international e-science context (for example in CERN's ATLAS Data Challenge experiments [11] as well as by the CI's. NorduGrid ARC provides an extensible compute brokering mechanism which can be instantiated via a graphical user interface, via a command line, or via an application programming interface. The low level computational middleware on which NorduGrid is built is the Globus toolkit version 2.4 [25] which complies with the recommendations of this Special Research Initiative for middleware preference.

Expected Outcomes

The initiative will provide a new research tool – a language-aware web search engine – which will be useful to a very wide range of web users, from academics to students, organizations within national and local government, and members of the general public, who are looking for information in and about language. Our focus on the immigrant and indigenous languages of Australia will ensure coverage for the material most in demand. The automatic classification and annotation of content will support a wide range of access modes, e.g. constraining the result of a search using a language keyword (`lang:Japanese`) or syntactic construction identifier.

Alignment with National Research Priorities: This project is aligned with Research Priority 3: Frontier Technologies for Building and Transforming Australian Industries, and on the following priority goals:

1. *'Frontier technologies' (enhanced capacity in frontier technologies to power world-class industries of the future and build on Australia's strengths in research and innovation).* This project will integrate state-of-the-art components for web data-mining and language analysis to provide a suite of research tools for accessing language data.

2. *'Smart information use' (improved data management for existing and new business applications and creative applications for digital technologies).* This project will enable a vast quantity of language data to be catalogued and explored in entirely new ways, in support of research and teaching in a wide variety of disciplines connected with language.

The project also contributes to Research Priority 4: Safeguarding Australia, in connection with the following priority goal:

3. *'Understanding our region and the world' (enhancing Australia's capacity to interpret and engage with its regional and global environment through a greater knowledge of* languages*, societies, politics and cultures).* This project will provide language-aware web search, permitting researchers from a variety of disciplines to perform focussed search for material which may be of strategic or operational significance at the national and regional levels. It will produce data collections on a par with those developed by the Linguistic Data Consortium at the University of Pennsylvania (where CI Bird is a senior researcher) for the DARPA Translingual Information Detection, Extraction and Summarization (TIDES) program, and the DARPA Surprise Language program.

Alignment with Objectives of ARC Special Research Initiative:

This project is aligned with a number of objectives of the ARC Special Research Initiative in E-Research. In particular, the focus of this initiative is the provision of new methods of access to very large data collections consisting of language data on the web, and enabling this data to be shared through a language-aware search engine. The project brings together a multidisciplinary team of investigators, and through extension with collaborators in the national and international contexts covering linguistics, computer science, anthropology, cultural institutions, language technology, geospatial technologies etc.

Through the extensive use of the University of Melbourne's Advanced Research Computing ICT infrastructure, this project will increase the insitutional return on investment for these facilities. This initiative will extend a range of existing open source software tools, and develop new software components as required. In keeping with the open source philosophy, these new tools, as well as the services enabled by their deployment will be released back to the community. This initiative also demonstrates the application of grid computing technologies, both as enabling technologies in computationally complex processing large amounts of data but also as the invisible underlying fabric of end user services in the case of the language-aware web search engine. Additionally, this initiative further fosters cross-institutional collaboration on a national and international scale through intersection with cognate research projects.

References

[1] Australian Broadcasting Commission. Australian Broadcasting Commission website. `http://www.abc.net.au`.

[2] Australian Bureau of Statistics. *Census 2001*. Australian Bureau of Statistics, 2001.

[3] Australian Bureau of Statistics. *3105 Australian Historical Population Statistics*. Australian Bureau of Statistics, 2004.

[4] Australian Bureau of Statistics. *3412 Migration, Australia*. Australian Bureau of Statistics, 2004.

[5] T. Baldwin. The deep lexical acquisition of English verb-particle constructions. *Computer Speech and Language, Special Issue on Multiword Expressions*, to appear.

[6] T. Baldwin and F. Bond. Learning the countability of English nouns from corpus data. In *Proceedings of the 41st Annual Meeting of the ACL*, pages 463–70, Sapporo, Japan, 2003.

[7] E. M. Bender, D. Flickinger, J. Good, and I. A. Sag. Montage: Leveraging advances in grammar engineering, linguistic ontologies, and mark-up for the documentation of underdescribed languages. In *Proceedings of the Workshop on First Steps for the Documentation of Minority Languages: Computational Linguistic Tools for Morphology, Lexicon and Corpus Compilation, LREC 2004*, Lisbon, Portugal, 2004.

[8] S. Bird and G. Simons. The open language archives community: an infrastructure for the distributed archiving of language resources. *Literary and Linguistic Computing*, 18:117–128, 2003.

[9] City of Melbourne. City of Melbourne website. `http://www.cityofmelbourne.vic.gov.au`.

[10] D. Elworthy. Language identification with confidence limits. In *Proceedings of the 6th Annual Workshop on Very Large Corpora*, pages 94–101, Montreal, Canada, 1998.

[11] P. E. et al. Atlas data-challenge 1 on nordugrid, 2003.

[12] B. Hughes. Metadata quality evaluation: Experience from the open language archives community. In *Proceedings of the 7th International Conference on Asian Digital Libraries*, pages 320–329. Springer-Verlag, 2004.

[13] B. Hughes and A. Kamat. A metadata search engine for digital language archives. *D-Lib Magazine*, 11 (2), February 2005.

[14] G. Kikui. Identifying the coding system and language of on-line documents on the internet. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING '96)*, pages 652–7, Kyoto, Japan, 1996.

[15] A. Korhonen. *Subcategorization Acquisition*. PhD thesis, University of Cambridge, 2002.

[16] P. McConvell and N. Thieberger. *State of Indigenous Languages in Australia - 2001*. Australia State of the Environment Technical Paper Series (Natural and Cultural Heritage), Series 2. Department of the Environment and Heritage, 2001.

[17] National Ethnic and Multicultural Broadcasting Council. National ethnic and multicultral broadcasting council website. `http://www.nembc.org.au`.

[18] NorduGrid Consortium. NorduGrid Advanced Resource Connector. `http://www.nordugrid.org`.

[19] O Smirnova et al. The nordugrid architecture and middleware for scientific applications. In *Proceedings of the International Conference on Computational Science (ICCS 2003)*.

[20] Open Archives Initiative. Open Archives Initiative website. `http://www.openarchives.org`.

[21] Open Language Archives Community. Open Language Archives Community website. `http://www.language-archives.org`.

[22] P. Resnik and A. Elkiss. The linguist's search engine: Getting started guide, 2004. `http://lse.umiacs.umd.edu/lse_guide.html`.

[23] P. Sibun and J. C. Reynar. Language determination: Examining the issues. In *Proceedings of the 5th Annual Symposium on Document Analysis and Information Retrieval*, pages 125–35, Las Vegas, USA, 1996.

[24] Special Broadcasting Service. Special Broadcasting Service website. `http://www.sbs.com.au`.

[25] The Globus Alliance. Globus website. `http://www.globus.org`.