EMERGING TECHNOLOGIES OF ELASTIC CLOUDS AND TREEBANKS: NEW OPPORTUNITIES FOR CONTENT-BASED AND DATA-DRIVEN LANGUAGE LEARNING

Robert Godwin-Jones
Virginia Commonwealth University

Creating effective electronic tools for language learning frequently requires large data sets containing extensive examples of actual human language use. Collections of authentic language in spoken and written forms provide developers the means to enrich their applications with real world examples. As the Internet continues to expand exponentially, the vast "cloud" of Web pages created provides a nearly inexhaustible and continuously updated language bank, particularly in English. The issue remains, however, of how to make practical use of large amounts of data for language learning, given storage and data processing demands. Recently, new developments in storage virtualization and distributed computing offer practical solutions, as demonstrated by Amazon's Elastic Computer Cloud and SimpleDB. At the same time, the move to XML encoding of language corpora and text collections provides the compatibility and interchange which has hampered their practical exploitation for language learning. Tools are also being created to facilitate the transformation of text collections into more usable formats, particularly into syntactically annotated corpora called treebanks. These developments offer opportunities for content-based language learning in particular.

GROWTH IN CONTENT-BASED LANGUAGE LEARNING

Rich data collections are especially important for development of learner focused language applications. In recent years there has been a sharp increase in the development of language learning tools for specific learner populations. Not surprisingly, this has been most in evidence in Europe, as the European Union has continually added new member nations bringing with them additional official languages. The EU Europa Web site list 171 different projects in the area of content-based language learning that have, since 1999, earned the "European Language Label", awarded for creative applications in language learning. A number of these projects have been created with funding supplied by EU grant programs, including Lingua, Leonardo, and Socrates. Most involve the creation of electronic tools and multimedia and increasingly are using the Web for delivery. Many are designed for use in either instructor-led or self-study settings, or both.

The EU site highlights a variety of projects in language learning for special purposes, including such diverse targeted areas as agricultural workers, apprentices, architectural workers, automotive workers, building maintenance workers, computer scientists, construction workers, customs officers, dock workers, entrepreneurs, hospital patients, insurance industry workers, isolated rural inhabitants, teachers, prison officers, the unemployed, and young immigrants. Some projects are even more narrowly focused, such as French for racing apprentices, Polish for missionaries, or English for ski lift cashiers. The largest number of projects targets the hospitality sector, where the need for multi-lingual workers is evident. The VIRTEX project was recently awarded first place in the European Language Label competition and is designed for workers in the hotel and restaurant industries learning English or German. Originally a CD-ROM project, it now incorporates a rich set of online tools, including streaming video.

Several of the vocational language projects make use of a full-fledged virtual learning environment. The EUROVOLT project, which offers vocationally-oriented language learning in a variety of languages for many industries, is implemented in Moodle and makes extensive use of new media and collaborative tools. It also incorporates language e-portfolios. Interesting projects in this area also include BeCult and Online VoCAL/Weblingua, both of which have richly developed tools and media.

Not enough information on the projects listed above is given by their Web sites to know to what extent they make use of word sets or data collections. An example that shows the benefit of word sets for content-based language learning is the Academic Word List (AWL)for English, developed by Averil Coxhead. The 570 words on the list (sub-divided into ten categories) were compiled from a corpus of 400 written academic texts. It excludes the most common 2000 English words. The list targets students entering an English-speaking university and provides an efficient base on which to create language learning exercises such as matching or cloze. The AWL Highlighter offers a nice example of the benefits of having such a list: it allows users to enter an arbitrary text, which is then parsed for AWL words and returned as a new document with the AWL items in bold, allowing students to work with the words in context. This helps guide the students to focus on vocabulary likely to be found in the text repeatedly, rather than learning items that are unlikely to be encountered again.

Content-based language learning is inherently learner-centered, focusing as it does on the specific context in which the target language will be used. It also lends itself well to task-based learning activities. Many of the projects targeting language for special purposes are built around real-life scenarios, often delivered through digital video clips, as an example from the Virtex project demonstrates. The students watch a real or simulated conversational exchange or an on-the-job interaction and are provided with comprehension aids such as full/partial transcripts, isolated audio playback, cultural notes, or lists of idiomatic expressions. Students are then asked to use the expressions from the dialogues in on-line exercises, written assignments, or group work. The importance of vocabulary development in content-based language learning necessitates that the vocabulary items chosen are those needed by the learners. Developing content-specific word lists in the manner of AWL would be highly beneficial, assuming enough texts can be found to build a specialized corpus.

One of the advantages of having a corpus to draw from is the possibility of using concordances as a vocabulary and grammar learning tool. Concordances are not effective for all learners, but for many motivated students it can provide a means for working with language structures through real world use. Students using concordances can be asked to reflect on areas such as inflections and collocations involving core vocabulary for the areas they are studying. Since the materials are tailored specifically for students' needs, it is more likely that such efforts will be successful. Some interesting examples of the use of concordances are collected by Bernd Rüschoff based on workshops and other sources. Tom Cobb's lextutor enriches the use of concordances by linking the found items to the on-line WordNet dictionary. Wordnet is a large lexical database of English that was first made public in 1991 and has since inspired similar collections in other languages.

LANGUAGE CORPORA AND XML

The percentage of Web-based vocabulary and discrete grammar exercises based on language corpora is quite low. There are many understandable reasons for this, including lack of access to appropriate corpora, incompatibility of the data with authoring tools, ignorance of how to incorporate data sets, and the need to focus on vocabulary prioritized in textbooks. The process could be made considerably easier for the average language instructor if available tools interfaced more readily with language corpora or text collections. Many popular tools for creating Web-based exercises, such as Hot Potatoes, allow for importation of text files for creating cloze or gap exercises. However, they do not allow for retrieval and incorporation of texts from large data sets or concordances. This situation is largely a by-product of the proprietary format in which language corpora and text collections have traditionally been encoded. Data with idiosyncratic encoding schemes and interfaces does not lend itself to searching or sharing. In many cases tools created in conjunction with the data have not been designed to be interoperable.

Fortunately, the widespread use of XML for encoding corpora and text collections is moving towards a resolution of this problem. XML has become the *de facto* standard for encoding of language corpora. XML recommends itself because of its platform independence, extensibility, and widespread acceptance

by software companies and researchers. Standardizing text encoding in XML greatly facilitates data interchange. Since structural and semantic information about a text is separated from its presentation in XML, the same encoded text can be displayed in multiple ways, using CSS style sheets or XSLT transformations. With the advent of XML as the preferred system for representation of corpus resources, existing tools have been modified to work with XML, while new applications have been created that are designed to be XML ready. The Linguist's Toolbox, for example, now features export to XML. The text searching software, Xaira, designed to be used with the British National Corpus, has been re-written as a general purpose XML search engine with full Unicode support. The Unicode editor CLaRK has been designed specifically to work with XML. Language archives can now be submitted to OLAC (Open Language Archives Community) by uploading a single XML file containing the necessary metadata information about the resource. Tools for the semi-automatic annotation of corpus data are being developed, such as @nnotate from the University of Saarland. DepAnn is a treebank creation tool, which uses Tiger-XML, the accepted standard for treebank encoding. EULIA, from the University of the Basque Country, provides a graphical Web interface for editing annotated corpora. These kinds of tools will become increasingly important as language data sets increase in size, since manually annotating texts to create treebanks is a slow and expensive process.

One of the most widely used XML encoding schemes for text archives is TEI, Text Encoding Initiative. A new version of the TEI Guidelines was released in November, 2007. It offers a number of enhancements, including more support for manuscript descriptions and better support for multimedia and graphics. Additionally, a Web application called Roma has been developed which provides a visual editor for working with TEI. An example of the power and versatility of TEI is the Henry III Fine Rolls project, from the British National Archives. These are fiscal and administrative records in Latin from the 13th century. The site provides user-friendly access to graphic representations of the original parchment rolls, as well as the original texts, translations, and notes/annotations. TEI allows the Henry III project to be included in general searches and to be easily referenced within other text projects. Version 4 of the Perseus Digital Library, a collection of classics texts, also uses TEI encoding and adds a set of XML-based Web services which allow for chunking larger texts into smaller units, as well as for sophisticated morphological analysis.

NEW OPTIONS FOR DATA STORAGE AND PROCESSING

Projects that house discrete, well-defined collections of texts can usually manage storage and delivery resources using traditional options, namely one or more servers housing a database, a Web server, and any associated Web services. If the site is popular, redundant servers might be needed. However, if the project is unusually large, such as the American National Corpus, being created as an American English cousin to the British National Corpus, the traditional project paradigm may not suffice. This is particularly the case if the goal is not just to deliver static text selections, but to allow for dynamically generated resources selected by sophisticated search, retrieval, and concatenating options such as are available with the Perseus project. In this scenario, there are significant demands in terms of processing which may well overwhelm the traditional setup for a text repository.

In recent years, some new options have emerged which make it easier to set up and manage a large-scale text project. The technical means have been available for some time to enable load balancing and parallel processing, but traditionally such systems have been difficult to create and run and tended to be so expensive as to be beyond the means of most academic projects. Today, through tools and services originating with Google and Amazon, there are ways for programmers without experience with parallel or distributed systems to use the resources of a large distributed environment to achieve high performance with off-the-shelf PC's that are linked together. Large Web companies such as these, as well as Yahoo and eBay, have established developer outreach programs, through which they hope to drive more users to visit their site. As part of that program, these companies provide application programming interfaces (APIs)

which instruct developers on how to write Web applications that take advantage of their sites and services.

Google's MapReduce is one example which has generated considerable interest. It is a programming model and associated code library for processing and generating large data sets. The design simplifies the process of enabling multiple computers to process information and then collect back the results centrally. MapReduce assigns program instructions to multiple computers to be accomplished in parallel. It breaks down the calculations into two steps. In step one (the "map" function) a key/value pair is processed, providing a set of intermediate results. In step two (the "reduce" function), these intermediate results themselves are merged to compute a final answer. An example of MapReduce from a Google developer presentation shows how the phrase "to be or not to be" would be processed in the MapReduce model:

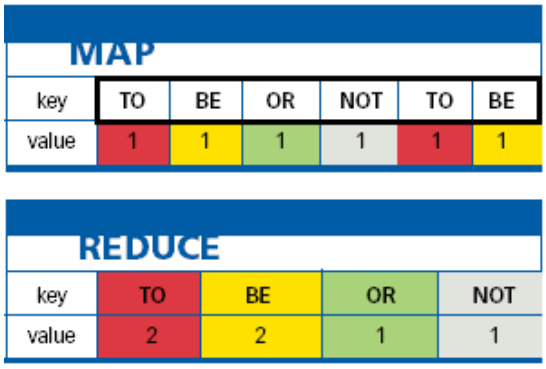

Figure 1: MapReduce processing of "to be or not to be"

This seems very simple, and it is, but by extending the process to several levels of analysis (i.e. further mapping of reduced results) it allows for very complex calculations to be broken down into simple steps. The general technique can be applied to many analytical problems.

MapReduce includes its own middleware that automatically breaks down computing jobs, doles out tasks to multiple computers, and collects the results. It also creates duplicate copies of each map-and-reduce function, finds idle machines to which to assign the tasks, and tracks the results. The worker machines load their individual piece of data processing, do the work, and notify the master machine when the work is completed ("mapped") and ready to be collected ("reduced"). If a machine freezes or breaks down, the master re-assigns that task after a specific period of not being able to communicate with the worker. The process is used by Google in many different ways, including machine translation between languages.

While MapReduce itself is proprietary to Google, an open source implementation, Hadoop, which implements the MapReduce method, has been released. Recently, the *New York Times* used Hadoop as the basis for creating a system to serve up archived newspaper articles. It needed to implement a large-scale operation as the decision had been made to make all the *Times* archives from 1851 to 1980 publicly available for free. In addition to Hadoop, the project was implemented using several Web services available through Amazon, namely Amazon Simple Storage Service (S3) and the Amazon Elastic Compute Cloud (EC2). S3 is an archive storage service that uses the same scalable system as is implemented in Amazon's retail site. EC2 is a computing service on which one can load and run applications. Both use a standard Web services interface, as does the recently announced Amazon SimpleDB, a database service. Collectively, these services provide the ability to store, process and query data sets residing on the Internet. Traditionally, this would require a relational database (such as Oracle or mySQL) and a dedicated database administrator. In contrast, the Amazon system is designed to be relatively easy to use. While it is not free, its pricing is low enough that it may be cheaper than operating a home server, let alone setting up a cluster-based computing environment. The Amazon services used by the *New York Times* work well not only with text and graphics, but with other media as well. For

example, CastingWords, a podcasting transcription service, stores audio files and transcribed text on S3. Clearly, this could be an interesting option for large-scale language projects.

LEARNING OBJECTS REPOSITORIES AND METADATA

One could envision something like the Harvard Text Annotator, an authoring tool for creating online glossed texts, running under Amazon and serving up vast quantities of on-the-fly annotated texts culled from Internet sources. For such a project to be successful, however, more than just text searching would have to be possible, even if sophisticated search options are available. Items collected in large data sets also need accompanying metadata to allow for more efficient narrowing of searches. This is important as well for finding and retrieving structured language learning resources, often labelled "learning objects" (LO). The OLAC metadata set implements a consensus approach among language corpora researchers. However, the modified Dublin Core metadata used in OLAC does not fulfil all the needs for materials to be used in language learning. One project that moves in this direction is the FLORE learning objects repository (LOR) for teaching and learning French. FLORE takes advantage of the French/English CanCore Learning Resource Metadata Initiative, a collaborative Canadian project, itself based on the IEEE Learning Object Metadata (LOM) standard. FLORE leverages a number of the LOM elements to provide additional information important for judging the appropriateness of a resource for language learners, including level of language proficiency and type of language learning environment targeted (i.e., immersion, self-study, etc.) The FLORE project is noteworthy also because it supports the Open Archive Initiative's Metadata Harvesting Protocol (OAI-MHP), which allows FLORE's metadata records to be shared with other repositories and to allow its metadata records to be linked directly with other systems.

There are, in fact, more and more collections of learning resources on the Web. A recent study features an extensive international listing. However, relatively few of the LORs include standard metadata such as that provided by OAI-MHP. The GLOBE initiative (Global Learning Objects Brokered Exchange) is an effort to move repositories in this direction. The CORDRA project is also attempting to standardize LO encoding. Including standard identifying information with learning resources would help enormously in making searches across multiple data sets, known as federated searches, faster and more efficient. Federated searches for learning objects are now available from LOR sites such as Merlot and Ariadne (which even include searching of sites such as Flickr and YouTube), but the search results are inconsistent and incomplete and do not allow for advanced search options.

A language learning LOR that exemplifies best practices in this area is the L_2O project out of the University of Southampton. This is a collaborative project building on the work of the eLanguage group, which produced a set of lessons for English for Academic Purposes. The L_2O project has been generating reusable LOs created mostly from existing materials. The project has developed a metadata set based on the LOM, but which adds contextual information important for language learning such as accent/region and subtitles/transcript. It complements the work done in this area by the FLORE group. The tagged LOs are retrievable from the project's repository, CLARe (Contextualized Learning Activity Repository). CLARe is currently being expanded to include social networking tools such as tag clouds and ratings.

A related project, MURLLO, has begun to develop a user-friendly LO editor. One of the features that would be helpful to see included in both LO editors and repositories is support for RSS feeds. The required information for the feeds could be automatically collected from the LO metadata and used by teachers or learners to be notified whenever new learning resources in targeted areas become available.

Developing easy-to-use tools for LO editing is a high priority if there is an expectation that subject matter experts such as language teachers create the resource, rather than it being created by technical specialists. A Swiss project from the University of Zurich is developing a tool for use with its LO model known as eLML. One of the better-known open source LO editors, eXe, has recently released a new version available for Windows, Mac, and Linux. A commercial LO editor, the SoftChalk LessonBuilder, is also

about to see a new version with additional features including more support for multiple languages. These editors support SCORM, an LO standard that originated with the U.S. Department of Defense but which has recently been transferred to a new international organization, LETSI, Learning-Education-Training Systems Interoperability. These and other editors will likely support the new IMS standard, Common Cartridge. This is a project designed to combine e-learning standards including SCORM, LOM, and IMS QTI (Question and Test Interoperability), along with other Web services, to create a fully developed learning module which can be imported into learning management systems such as moodle or Blackboard. In the US, it is generating considerable interest as an electronic alternative to traditional textbooks. This is also the thrust of the new Digital Marketplace initiative, an outgrowth of the Merlot project based at California State University. This has been hailed as a possible model for a "national digital marketplace," advanced recently by a US government study on the price of textbooks. The Global Text Project and wikibooks are non-commercial efforts in this direction.

RESOURCE LIST

Content-based Language Learning

- E-Lingua European project for learning the language of hotel service and management
- BeCult European language project for students in hospitality industries
- EUROVOLT European Vocational Online Language Teaching and Language learning via a VLE
- VIRTEX Project for Hotel and Catering
- Education & Training Programs in the EU List of projects related to content-based language learning
- MapReduce: Simplified Data Processing on Large Clusters By Jeffrey Dean and Sanjay Ghemawat

Corpora and Data-driven Language Learning

- The Compleat Lexical Tutor Example of data-driven language learning
- WordNet Lexical database for English
- Wordnets in the world Wordnets in multiple languages
- Academic Word List From Averil Coxhead
- Sample Exercises Data-driven Language Learning examples
- Best Practice Recommendations for Language Resource Description For language archives
- Penn Treebank Online Searchable tagged corpora in English
- Xaira Corpus search engine
- The Linguist's Toolbox and XML Technologies By Chris Hellmuth, Tom Myers & Alexander Nakhimovsky
- CanCore Metadata system from Canada
- Raise and Rise Example of learning object from wisc online
- The Open Archives Initiative Protocol for Metadata Harvesting OAI guidelines
- Harvesting Issues About implementing OAI metadata harvesting
- Open Archives Initiative Metadata Harvesting Project University of Illinois project
- Exposing information resources for e-learning Combining OAi and IMS metadata harvesting
- Digital Repositories Specification From IMS
- Real-time demonstration of interoperability between Learning Object Repositories Interoperability demonstration involving the ARIADNE network and the FIRE federation

- CORDRA Content Object Repository Discovery and Registration/Resolution Architecture
- GLOBE Global Learning Objects Brokered Exchange
- Federated Search Through Ariadne
- Sharing Language Learning Objects Example walk through the technological and pedagogical 'process models' of L2O project
- A Typology of Learning Object Repositories By Rory McGreal
- EML eLesson Markup Language for creating structured eLessons using XML
- Common Cartridge: e-Learning Made Easy IMS Standard in place of textbooks
- LETSI The international group now in charge of the SCORM standard
- TEI Text Encoding Initiative

Distributed Computing

- Self-service, Prorated Super Computing Fun! NY Times archive use of Amazon S3
- hadoop Open source implementation of MapReduce
- MapReduce The Google white paper
- Running Hadoop MapReduce on Amazon EC2 and Amazon S3 From Amazon development services
- S3 Amazon Simple Storage Service
- EC2Amazon Elastic Compute Cloud
- SimpleDB Amazon data base service
- Windows Live Web services from Microsoft