EMERGING TECHNOLOGIES
Tools and Trends in Corpora Use for Teaching and Learning

Bob Godwin-Jones
Virginia Commonwealth University

INTRODUCTION

Language corpora have long been exploited for language instruction. Vocabulary lists for learners, for example, have been generated from corpora, and word counts derived from corpus analysis have helped in defining goals for vocabulary acquisition. Dictionary and textbook creators have used corpora extensively. In recent years, the move to the use of authentic language materials in language pedagogy has enhanced the role collections of spoken or written language can play in language learning. Corpora are, after all, huge storehouses of real language use. The interest in languages for special purposes further favors the use of corpora, as a means to identify the specific language components to be taught. Technology enhancements have made corpora more widely available, as well as provided more powerful tools for their use. In particular, the Internet is playing a steadily growing role in the dissemination of corpora and corpus-based teaching materials. Corpora are no longer the exclusive domain of lexicographers and computational linguists.

ACCESS TO CORPORA

Corpora are of interest today to professionals in a wide variety of fields, from ethnologists to telecommunication conglomerates. Creating a language corpus is a major undertaking, both time-consuming and expensive. This is all the more the case for collections which include multiple languages and/or audio/video recordings. Given the cost and the growing interest, it makes little sense for corpora not to be made widely accessible. In fact, there have been a large number of corpora in many different languages which have become available over the Internet in the last few years. Good starting points for finding them are Michael Bohman's Corpus Linguistics page, the Linguistic Exploration page (at the LDC - Linguistic Data Consortium) or the Tractor page (the "Telri Research Archive of Computational Tools and Resources"). These pages in many cases link to direct corpus access, including a number of parallel corpora of particular interest in translation studies and language learning for specific purposes. There are as well a substantial number of text collections of literary works in a variety of languages. Some include comprehension aids and annotations for use in language learning.

As the number of language archives grows, locating the specific resources needed for a project will become more problematic. One can only go so far with lists of Web links (even when annotated) or traditional Web searching. There is a recently launched international project, the Open Language Archives Community (OLAC), to build an infrastructure linking language archives of all types together. OLAC builds on the Open Archives Initiative and on the Dublin Core Metadata Initiative. The Dublin Core project began in 1995 to develop conventions for resource searching on the Web. OLAC uses the core 15 elements of the Dublin Core and extends them through the use of qualifiers to fit the needs of the language community. The use of a controlled vocabulary of descriptors should allow more efficient searching of archives.

The consistent use of meta-data in language resources is likely to become of growing importance in the language community. There has not been a standard way to include information about a resource, such as the participants in an interview included in a corpus (i.e., age, nationality, first language, education, etc.). Such information is typically included in a "header" which is either part of the resource file itself or stored separately. Because of the different ways such meta-information has been stored, there has been a proliferation of tools and approaches for the user to access that information. It would be very helpful for

both researchers and users to have a common approach to resource description, not only for corpora, but for all language resources such as text collections, lexicons, grammar tutorials, multimedia files, Web lessons, and so forth. This would in turn facilitate the development of universal tools.

ENCODING AND ANNOTATION

Standardization, or at least inter-operability, is needed not only in resource description but, of course, also in the encoding and annotation of language resources. Increasingly, corpus creators have moved from proprietary systems to standard-based approaches. Given the effort behind corpus creations and the longevity of most corpora, the challenge is to design an environment which is adaptable over time as technologies evolve. It also needs to be flexible enough to be extensible to include classification categories for which a need may arise in the future.

In recent years the Text Encoding Initiative (TEI) has provided a standard used by a large number of language and literature resources. TEI uses SGML ("specialized general markup language") and provides for an extensive header containing meta-data. The header information is included within the annotation files. While the most widespread use has been in the encoding of literary texts, there is also an extensive list of projects using TEI encoding in corpora in a variety of languages. The TEI standard is part of the Corpus Encoding Standard (CES) proposed by the EAGLES group ("Expert Advisory Group on Language Engineering Standards"). CES specifies a minimal encoding level to be "standard" and provides encoding specifications for linguistic annotation. TEI is also used in the MATE project ("Multilevel Annotation Tools Engineering"), designed for the encoding and annotation of spoken dialogue corpora. The TEI standard, however, has some drawbacks as well. SGML is highly complex (as experienced by anyone having tried to decipher the intricacies of the TEI header), and SGML documents are not directly accessible from standard Web browsers. While extensible, customizing the TEI for an individual project is a daunting enterprise. Some projects have done so, such as the BBAW digital dictionary of German, adding custom headers in separate files.

In fact, there are a number of advantages to a "stand-off" data architecture in which the annotations and meta-data are stored in separate files from the data itself. This allows for considerable flexibility in adding and changing the annotation categories and information as needed, without having to revise the data files themselves. The encoding system that lends itself the best to doing that is XML ("extensible markup language"), the widely acclaimed successor to HTML and slimmed-down version of SGML. Recent Web browsers have native support for XML documents, but more importantly, there are standards and methods for transforming XML documents on the fly into a variety of formats. For a corpus, annotation can be stored in separate XML documents from the data itself, which are linked in hypertext to the documents. XML enables such linking to be one-way or two-way, useful for parallel corpora. A number of the most recent corpus projects are beginning to use XML, which in fact is being supported by the EAGLES group in an XML version of CES (XCES). There is also an XML version of TEI forthcoming.

One of the advantages of XML is that there need not be uniformity in the precise tags used, as long as there is an available description of each tag. Through XSLT ("extensible style language transformations"), information from XML documents can be retrieved and reformatted in a variety of ways, providing a powerful means for delivering data to a variety of users and browsers. Of course, a common data model for language resources would make it much easier to standardize access. Points (discrete objects) and spans (strings of objects) must be identified and tagged, with a common level of granularity (i.e., detail), and a means provided of identifying structure, class membership, and inheritance. There have been several large-scale projects, such as Tipster, to provide such a data model. The Atlas project also aims to provide an extensible architecture for linguistic annotation, through use of an "annotation graph model".

RETRIEVAL TOOLS

A common (or at least exchangeable) data model would facilitate the use and development of tools for corpus extraction. In the past, new tools were often developed for the processing of each new corpus created. New projects needed to budget time and money not only to data collection but also to creating an encoding/annotation system as well as a set of tools for accessing the data. Many of these tailor-made systems replicated functionality available elsewhere but not useable due to differences in software, platform or data architecture. Fortunately, there are tool projects underway which are designed with reusability as a major goal. They tend to use a modular, building blocks approach, rather than a monolithic all-or-nothing design, allowing for more flexible use as well as future extensibility. Among such projects are GATE ("General Architecture for Text Engineering") which sets as its goal a set of infrastructure tools for natural language processing which can accommodate models written for a variety of programming and scripting languages. The Multext project, similarly, encompasses a series of projects whose goals are to develop standards and specifications for the encoding of corpora and to develop tools and resources using these standards. Multext projects are underway in at least 18 different languages.

One of the other positive developments in the area of tools is the Natural Language Software Registry (NLSR), which collects and makes available over the Web detailed information on a wide variety of natural language processing software, including annotation tools (taggers, parsers), speech analysis, machine learning, evaluation tools, corpus analysis, translation, etc. The fourth edition of the NLSR provides for both browsing and searching, using a taxonomy based on the "State of the Art in Language Technology", edited by G.B. Varile and A. Zampolli. Many of the tools listed in the Registry are Internet-based, which is increasingly the case in tool creation. Most use Web forms to provide an access interface, as in sample collections in French,German, Spanish, Chinese, or Japanese. An interesting approach is provided by a service from the University of Leeds, which accepts email to amalgam-tagger@comp.leeds.ac.uk containing English text, which is then parsed and tagged for parts of speech and sent back by email.

OUTLOOK ON LANGUAGE LEARNING

One of the more frequently used tools in working with corpora for language learning are concordances. A concordance is an alphabetical listing of words in a text or collection of texts, together with the contexts in which they appear. Typically concordances are in KWIC format ("key word in context") in which each word is centered in a fixed field, and each occurrence of the word is listed on a separate line. Good concordancers do more than simply index words to lines, they can sort in a variety of ways, search for collocations, and produce extensive statistics. Concordances have been used extensively in literary studies and stylistic analysis, but less frequently in language learning. An extensive linguistic corpus is a gold mine of authentic language use and mining that through KWIC concordances can provide students with multiple contexts from which to learn new vocabulary. An interesting example of this use of concordances is providing contextual help in the reading of second-language texts. This approach seems to work best when students try computer-aided contextual inferences first (through the concordance) which can then be confirmed through on-line dictionary access. Concordances can also be very useful in providing assessment items. Cloze exercises, for example, can easily be generated from KWIC concordances.

Corpora, of course, can provide much more than just lexical information; they are invaluable in supplying syntactical examples. One of the caveats in using corpora in this way, is that for the most part corpora have been created for research purposes, rather than for language learners and as a consequence may not supply the needed information. Not all corpora, for example, are annotated for syntactic functions. Most of the parallel corpora available are restricted to narrow, often technical, language uses, thus making them less useful for contrastive analysis or translation studies. Such corpora can on the other hand be

invaluable in language learning for special purposes. There have been experiments using syntactically annotated corpora in providing grammar help for learners. The Cytor project at the University of Lancaster showed interesting results in providing students access to concordances, which led to improvement in their categorization of part-of-speech distinctions. This kind of activity provides a means of putting research tools into the hands of students, and working towards shifting some of the responsibility for learning on to their shoulders.

An area of significant interest to language educators are collections of recorded speech preserved as audio or video. This adds an entirely new dimension to corpora, with the addition of gestures, intonation, and facial expressions, but also adds a challenge in terms of encoding and annotation. There are several projects underway to help in establishing standards for such resources. The ISLE Meta Data Initiative is seeking to create a standard for meta-data description of multimedia language resources. The Talkbank project is an interdisciplinary project hosted by Carnegie Mellon University to provide standards and tools for human (and animal) communication. EUDICO ("European Distributed Corpora Project"), from the Max Planck Institute, is looking at ways to categorize and search collections of annotations on digital video and audio recordings.

One of the corpus needs for developers of CALL applications is for collections of non-native speech. Large corpora of transcribed speech data from language learners, for example, could be very useful in efforts to improve the understanding of the speech patterns of language learners necessary for interactive voice applications. There are databases of telephone speech available (from LDC) in a variety of languages. The European Science Foundation Second Language Data Bank consists of data obtained over a 3-year period for adult migrant workers in five European countries with a focus on language learning in the absence of formal instruction. Clearly, creating such non-native language collections is a huge task, complicated by the fact that there should be separate databases for different kinds of non-native speakers (according to country of origin, amount and nature of language exposure, nature of need for language ability, etc.). The needs of the telecommunication industry for reliable voice-based applications might be helpful in finding funding for such large-scale projects. It would be useful as well to have a corpus of email messages, from both natives and non-native, to provide a basis for evaluating the transformation of language through technology, and how that might affect language teaching and learning.

A significant impediment in the use of corpora in teaching and learning is the form in which most corpora are stored. Most are annotated in SGML and housed in large Unix servers. It most cases, it is not practical to store such large amounts of data locally. Thus access is provided remotely, which may present performance issues. The other barrier, of course, is the proliferation of different formats for accessing corpora and the bewildering array of tools available. The growth in Web access to corpora and tools is helpful, but often the interfaces are poorly designed. The corpus linguistics community has recognized this issue, as well as the need for greater consideration of teaching needs in corpus design, and the situation looks likely to improve in the future.

RESOURCE LIST

General Corpus Information

- Language Software Helpdesk from the Language Technology Group (Edinburgh)
- Corpora List archive in Hypermail excellent source of up-to-date info on corpora
- Multilingual Theory & Technology from Xerox
- Corpus Linguistics Michael Barlow's extensive listing

Corpora Access

- projects using the TEI
- English Language Corpora and Corpus resources from the British National Corpus

- Corpora, Text Resources good list from Kiat Lab (Japan)
- CobuildDirect Corpus Access Information commercial site with trial access available
- TRACTOR Network of multilingual resources corpora in multiple languages listed
- COMPARA Portuguese-English parallel translation corpus
- COSMAS access to the Mannheim corpus of German
- LAPT&DA access to special vocabulary lexica in German (Erlangen)
- Digital Dictionary of the 20th Century German BBAW project
- Archives for Language Documentation and Description from the University of Pennsylvania
- Linguistic Exploration list of resources from the Uuniversity of Pennsylvania
- Web EuroWordNet Interface access to multilingual lexical knowledge bases
- European Literature - Electronic Textscomprehensive listing

Standards and Projects

- Open Language Archives Community
- TEI Text Encoding Initiative
- MATE Multilevel Annotation, Tools Engineering
- The GATE project ambitious project for building a NLP infrastructure (Sheffield)
- XML from the W3C (World Wide Web Consortium)
- XSLT from the W3C (World Wide Web Consortium)
- EAGLES Expert Advisory Group on Language Engineering Standards
- The XML Cover Pages - Home Page excellent resource list by Robin Cover
- Multext large-scale corpora and tools project from the Centre National de la Recherche Scientifique (France)
- Talkbank multimedia database project from Carnegie Mellon University
- Synchronized Multimedia Integration Language from the W3C
- Survey of the State of the Art in Human Language Technology
- EAGLES/ISLE Meta Data Initiative
- Corpus Encoding Standard part of the EAGLES initiative
- XCES XML version of CES
- Tipster main site
- Tipster Architecture info
- ATLAS: A Flexible and Extensible Architecture for Linguistic Annotation
- EUDICO European Distributed Corpora Project

Corpus Retrieval Tools

- Concordancers FTP downloads
- TACT (Text Analysis Computing Tools) DOS Concordancer from the University of Toronto
- LTG Software tools for text processing (including XML) from Edinburgh
- On-line corpus analysis Web-based concordance generator (in German) for texts in French, Italian and Spanish
- Software Tools for NLP list from Kita Lab (Japan)
- NLSR Natural Language Software Registry
- CRATER tools and resources for multilingual corpus work

Teaching and Learning

- Teaching and Language Corpora article in *ReCALL* by T. McEnery and A. Wilson (PDF)
- Tutorial: Concordances and Corpora Web-based introduction by Catherine Ball (Georgetown)
- Corpora in the Teaching of Languages and Linguistics
- Can the rate of lexical acquisition from reading be increased? case study in concordance use in reading

- Pruebas de PHP-KWIC Web-based concordance general for Spanish texts (in Spanish)
- Corpus of Historical and Modern Spanish Web-based access to large Spanish corpus (from Mark Davies)
- VLC Web Concordancer search options in Chinese, English, French, Japanese, as well as parallel texts