ALLC/ACH 2004
(Abstracts for papers or poster presentations should be 750-1500 words in length)

Title: A Generic Application for Corpus Management and Administration

Keywords: corpus management, corpus administration, database

Our corpus project is building a digital collection of both written and spoken texts. The corpus is a publicly available resource, mounted on and searchable via the Web.
This paper will describe the corpus management and workflow administration methods that the project has developed and the technologies used. We believe that the structures we have created to manage the different parts of the administration of the project are the basis for a re-usable, generic package for scholars building an online corpus from new linguistic materials.

While every existing corpus has, of course, developed its own methods of ensuring the correct procedures are followed, this knowledge is usually not available to other projects in their early development or is too specific to a particular area of language research. We have been approached by several projects who have seen or heard of our system and are planning to build multilingual and multi-media corpora. The paper will describe two case studies involving disparate data in various media and our proposed pilot studies on methods of creating an application suitable for their use, and extension to a generic package.

Our corpus is synchronic, a 'snapshot' of the languages used in here in recent years, and a monitor corpus which will be continually updated. It includes written texts, sound recordings, video recordings, and transcriptions of sound from the latter two. It also contains extensive sociolinguistic metadata. We are processing readily available materials (c.1m words) and will identify the gaps and find or commission new materials to fill them. We intend that the corpus will contain upwards of 4,000,000 words, at least 20% spoken. This will form a valuable language research tool in its own right, and will offer a structure flexible enough to expand and accommodate future sub-corpora.
Sociolinguistic metadata are held in the following categories: resource type, text type, setting, medium, audience, text details, author/speaker details and copyright information. For example, the author and speaker categories contain information on gender, age, geographic region, education, occupation, languages spoken etc. The author/speaker's parents' information is also recorded. Standards described by the Text Encoding Initiative Guidelines and Dublin Core have been applied and the metadata categories have been decided upon after extensive research into the requirements both for useful language research and the legal necessities for publication of the data.

In creating the corpus structure we have included database functions to control the workflow. Administrative functions range from contact management through to document entry and associated metadata manipulation. These are functions that any project creating a corpus must perform. Having these functions built into the system means that, for example, copyright law compliance is enforced. The system generates reports to facilitate the administration and management of each key stage. The data storage method is more advanced than simple flat files and multiple concurrent users are fully supported, aiding large-scale data entry. Document contents are accessed via the unified interface, alleviating the requirement for file naming and directory organisation.

The corpus system is divided in two sections: administration and online search. With such valuable data being held in the system it was important for us to choose open, standards-based solutions where possible. Although more advanced and proprietary software is available, we chose to configure a Linux server with MySQL RDBMS, Apache and PHP, as this configurations give the greatest degree of flexibility and portability. To reduce training of our team MS Access was chosen to provide the front-end to the administrative functions. This allows interaction with the whole Office suite to provide mail merge and data interoperability.

The search system uses a subset of the administrative data. Only those documents that are complete and have all their relevant permissions are allowed into the public data set. As an added security measure this consists of a separate database, but this is not a necessity. Access to documents is provided via a two tier search mechanism. Commonly used criteria are the basis for a basic search interface, those users who wish to explore the dataset in more depth can choose an advanced query tool giving access to the whole range of metadata. Word/phrase search can be combined with a metadata search, matching documents provide word highlighting to show context of match.

Following approaches from scholars proposing to start building digital corpora it became apparent that a subset of our corpus management system (concentrating on administration and workflow) would be very useful to others. Two such projects are detailed below.

As our data structure is tailored and grew up with the particular requirements for our corpus, it will be necessary to develop an abstracted model of the corpus management system. This will allow standard data objects and types to be used freely inside the established model. As open source software is used there are many different software and hardware platforms available as host to any system developed.

First steps to integration would involve identifying each document type and their metadata and contents. In addition the interrelationship of different objects (e.g. author to document) must be established before the data can enter our framework. Sample data would be identified to test the maximal set of possibilities available to verify expected operation. Any search front-end will necessitate a higher degree of customisation to match the specific project.

The first corpus in our pilot study is a corpus of transitional dialects**.** 198 individuals were recorded for approximately 1 hour each (c. 120 gb of data). We know from our discussions with the scholars who have created this data that the abstracted model referred to above would be of use to them in creating an online corpus.

The abstracted model would not deal with all the issues particular to this corpus and we have identified some areas requiring investigation. For example, the dialects concerned contain speech sounds which are not represented by characters in current Unicode sets. Transcription of the data will require the creation of new Unicode code points and glyphs to allow reproduction of the particular dialects in use. Until this is resolved methods for searching this data will have to be investigated. Storage and encoding are easily solved for the researchers local use but these methods may not be available to end users online.

The second corpus in our pilot is based on recent Parliamentary elections. The data collected are a rich mix of media (text, images, sound and video recordings, and mixed media documents such as web pages). An investigation into which metadata are generic for all media types must be conducted as any search must be capable of scanning all media types and present a unified result. Integrating the various media types into the unified front-end will present different challenges for each media type, as would display of these media types. Upon search completion quick access to related documents would be very useful e.g. identify other documents by the same author / same location.

We believe our corpus management software can give new projects an easy to use framework on which to base their system. Collaboration with other projects will enhance the system as we identify new data types and modules to include in the framework. As these issues are solved this will feed back to the abstracted model and provide ready made solutions for future projects. As any collaborative effort would have the same base structure it would make integration of data from many different corpora straightforward, this may be of particular use to researchers.

Most literature on corpus projects concentrates on either the content, the encoding of the content or the research results of the use of the corpus. We have found little on the management and administration of a corpus project. Researchers often do not recognise initially the amount of time needed to develop good management procedures and the complexity required to control the process from contact management through document entry, metadata manipulation to publication. We believe this discussion will be use to us, to the scholars planning the corpora mentioned above and to others in the ALLC/ACH community.

Bibliography

Sperberg-McQueen, C. M., and Lou Burnard (ed.) (1994), *Guidelines for Text Encoding and Interchange*, Chicago and Oxford: Text Encoding Initiative

Bonhomme, P. and Laurent, R. (1997) *XCORPUS - Version 0.2: A Corpus Toolkit Environment: User Manual*, http://www.loria.fr/projets/XCorpus/manual/

Brew, C. and Moens, M., *Data-Intensive Linguistics*, HCRC Language Technology Group, The University of Edinburgh, http://www.ltg.ed.ac.uk/

Gibbon, D., *Ubiquitous multilingual corpus management in computational fieldwork*, http://coral.lili.uni-bielefeld.de/LangDoc/EGA/Presentations/ubicorpus_lrec2002.pdf