The Language Resource Archive of the 21st Century

Martin Wynne

Oxford Text Archive, Oxford University Computing Services, 13 Banbury Road, Oxford, UK – OX2 6NN

Abstract

What will an archive of language resources look like in the future? It is to be expected that developments in computer technology will have an impact on the nature of language resources which will be created in the future. A projection current trends into the future helps us to see that there will be more multimedia and multilingual resources. It is also likely that increasing internet bandwidth will lead to a more distributed architecture whereby resources are accessed remotely rather than held locally. This will also facilitate the development of *virtual corpora*, whereby temporary, *ad hoc*, collections of texts can be assembled for a specific analysis. Increasingly it will become the norm to extract information from resources held in the archive, rather than downloading the corpus, installing software to analyse it with and getting them to work together.

It can therefore be predicted that although archives will continue to have an important role in the preservation of resources, other roles will develop or grow in importance, as archives adapt to allow the creation of virtual corpora and online access to resources, and become centres of resource creation expertise, metadata validation and resource discovery. This paper discusses the new directions envisaged by the Oxford Text Archive (OTA), and in particular its current initiatives to improve the service provided for the community of academic linguistics researchers in the UK.

1. Introduction: the archive today

What will an archive of language resources look like in the future? This paper discusses the new directions envisaged by the Oxford Text Archive (OTA), and in particular its current initiatives to improve the service provided for the community of linguistics researchers.

The OTA is a long-established facility supported by Oxford University, and based within the Humanities Computing Unit. The archive holds several thousand electronic texts, in a variety of languages. The holdings include electronic editions of works by individual authors, standard reference works such as the Bible, monolingual and bilingual dictionaries, and a range of language corpora. They have recently been made more easily discoverable, identifiable and accessible through the participation of the OTA in the Open Language Archives Community (OLAC) initiative. OTA resource descriptions are now delivered to the OLAC portal, so that users looking for language resources have only to send a query to the central 'virtual data provider' in order to find some basic information about the resources which are available in a multitude of archives.

The OTA has been working as a *service provider* for the national Arts and Humanities Data Service (AHDS) since 1996, to support academics working in all areas of literary and linguistic studies in the UK.

The OTA seeks to address three key areas:

- collect, catalogue, preserve, and redistribute digital resources of interest to those working in literary and linguistic studies within the UK's Higher and Further Education communities;
- develop appropriate licensing conditions and technical mechanisms for the effective distribution of such resources;
- promote good practice in the creation and use of such resources in research and teaching.

Following a study of the OTA's subject coverage and Collections Policy, it was decided that additional support should be offered in the subject area of Linguistics.

A consultation process with active members and with key organisations in the UK academic linguistics community is being undertaken, in order to identify users and analyse their needs, and to establish the directions in which the OTA's service needs to develop, in order to provide a better service to a wider user base.

A fresh and forward-looking approach is proposed whereby existing and traditional notions of the nature of resources and online services are reappraised and new directions are sought. It is not taken for granted, for example, that the delivery of entire text corpora to individual users is the most useful model. In this paper, some current and future trends in the development of language resources for linguistic research are identified, and an attempt is made to see how these trends will impact on the nature of archives like the OTA.

2. Resources of the future

From observation of current trends, it is anticipated that in the forthcoming period language resources are likely to display the following characteristics:

- Multilinguality
- Multimedia
- Dynamic content
- Distributed architecture
- Virtual corpora
- Online extraction of linguistic information
- Connections with web searching

Below, these features of language resources are investigated in more detail. Then, section 3 examines the possible consequences of the changes in language resources for today's archives and for the archives of the future.

2.1. Multilingual resources

Human Language Technology (HLT) research is increasingly focused on the exploitation of multilingual resources, for the development of tools for software and text localization, machine-aided translation, speech synthesis and recognition, language recognition and a variety of text processing tasks. Archives such as ELDA holds many such resources, and TRACTOR explicitly states that the acquisition and development of such resources is their chief goal (Wynne 2001).

2.2. Multimedia resources

The economic importance of spoken language resources and the role they play in the development of speech technology is well recognized. Increasingly audio-visual data is also being exploited, particularly by linguists keen to capture as much of the context of speech events as possible, but also by researchers into human-computer interaction. Developments in storage media and communications bandwidth make such resources increasingly usable and distributable.

2.3. Dynamic content

Websites, email communications, news groups, chat rooms and electronic publications of various types are increasingly objects of linguistic study, and are important sources of texts. These new media blur the traditional distinctions between static and dynamic, time-based and non-time-based, published and unpublished material. New strategies for the capture of relevant metadata are needed in this area. This is particularly important so that the user of the resource knows what they are getting, and also so that research can be reproducible.

2.4. A distributed architecture

Modern corpus linguistics was made possible by the advent of the computer, because it became possible to store large amounts of text in electronic form and use computers to count and sort words and extract concordances and collocation information. In the first stages, mainframe computers were used and users needed access to the computing facilities. At this level of technology, corpora were relatively small, one million words being the norm, and annotation and analysis were done in a painstaking manual fashion.

The next development was the rapid growth in the availability of computing power and disk space, often on powerful desktop personal computers but also still on mainframe or shared workstation environments. As a result, computers got bigger, and user-friendly generic computer programs for extracting concordances and wordlists were developed. At this stage, the user of language resources typically downloads the corpus to his machine, installs a program to analyse it, then tweaks the program and/or the corpus markup to get the program and the corpus to work together, and finally performs the analysis.

The next leap forward would seem to involve developments in computer networks. Already the massive amounts of electronic texts which are now available mean that capture of data is not the problem which it once was, although the reliability of texts downloaded from the internet, and the capture of metadata mean that there is still a role for the corpus and for the archive.

It is also the case that higher bandwidths make online processing faster and more reliable. With these developments, the older, though still dominant, model outlined above (of downloading the corpus and the tools to the desktop) will become obsolete, as it becomes redundant to replicate the digital data in a local copy, as the processing can now be done over the network at an acceptable speed, and the need to install a local copy of the software is also thus obviated. Developments in the Grid and eScience are moving towards a more distributed model of resource sharing, whereby large amounts of data can quickly be shared across an ultra-high speed network.

On the other hand, the still dramatic and rapid increases in computing power and disk storage which are still taking place mean that there are factors supporting the local processing model, and this will not go away overnight, and will continue to be important.

However, as the distributed model becomes a reality, issues of content and software interoperability will come to the fore.

2.5. Virtual corpora

In the traditional model of the language resource, the resource (typically a corpus) is carefully prepared, by taking a sample of the population of texts of which it aims to be representative, and is encoded and annotated in ways which make it amenable for linguistic research. The value and the reusability of the resource are therefore dependent on a bundle of factors, such as the validity of the design criteria, the quality and availability of the documentation, the quality of the metadata and the validity and generalisability of the research goals of the resource creator.

A more useful general model might be the archive of electronic texts whereby the user creates a collection of texts (the corpus) on an ad hoc basis according to the values of one or more metadata categories, such as 'all 17th century English fiction' or all 'Bulgarian newspaper texts'. In this way a large archive, which need not contain any corpora, but only electronic texts, can be exploited as a corpus linguistic resource. It is considered that this type of *virtual corpus* would be of great value to the OTA's users and a system that will make this possible is under development.

A projection further into the future of this idea, seen in tandem with the developments in the distributed architecture of processing and online working outlined above, enables us to foresee the possibility of the selection of texts for inclusion in the virtual corpus from different archives (or other locations). In theory, the desired texts could be identified and located through a meta-archive (where resource metadata records from various archives are harvested and collected) and the processing of the corpus (e.g. word counts, concordances, collocation extraction) run on the disparate texts as a single corpus.

There are two preconditions for the accurate exploitation of such virtual corpora. The first is adequate and standardized metadata, so that the required texts could be identified in a reliable fashion. The second is a certain level of content interoperability, so that the tools could extract the required information from the different texts. At present these not inconsiderable problems are being addressed at the level of the Oxford Text Archive holdings. Tools are being develop to convert texts in various formats to conformance to a common XML DTD. It is not clear to the author how they the second problem of textual encoding could be addressed at a more global level.

The various formats of legacy data, the technical limitations of SGML-based encoding systems, the over-abundance of standards, the intrinsic scientific value of heterogeneity, political rivalries: these are all significant problems. It is the opinion of the author that the 'search for the perfect metalanguage' has not yielded encouraging

results so far, and is unlikely to do so in the foreseeable future.

However, this does not mean that interoperability is impossible. The solution is in the metadata. If the encoding is clearly described, then resources which can be used for a particular application can be selected.

2.6. Online extraction of linguistic information

As well as the distributed architecture of resources, the tools for the extraction of linguistic information from these resources may also be part of the distributed architecture. In fact, it is already the case that the user does not necessarily need to install a program on this own computer. For example, the *lookup* program which subscribers can use to access the Bank of English runs on the server where the corpus is held, and downloads concordances and other data specified by the user. There are also websites which run server-side applications to deliver data to the user. However, in these cases, the user is limited to using the program which is installed by the operater of the server where the resource is held, and inevitably this program will have limited functionality. A more flexible and less restrictive approach would be possible is the user (on computer A) was able to access the corpus (on computer B) and run programs installed at another site (on computer C) which possesses the required functionality.

The prerequisite for this type of computing is a high level of interoperability between computer systems, networks, software and text markup. It is not yet clear that this could be achieved on any significant scale. It seems logical that the first stage will be a distributed architecture for the access to the resources (i.e. online queries and virtual corpora), and the distributed processing will be a later potential development.

2.7. Searching for content and searching for form

The ubiquity of electronic texts on the internet has led to much speculation about the potential use of the web as corpus. But internet search engines are designed to search for content on the web which is *about* a given search term. So searching for the term 'holocaust' will give tend to give you hits on websites with content about the historical event which is usually referred to by the term 'holocaust'. Hits on URLs, titles and headings containing the search term and on keywords in the metadata will be prioritized.

Language researchers however are more likely to want to find occurrences of the word form 'holocaust' in texts, in order to see how the word is used. They would most likely want to prioritise occurrences in the body of texts, not in the metadata.

A search engine which is designed to search for use of word forms in running text would be a useful tool for linguists. This could be added as advanced search option to existing search engines. However this would not overcome the problem of the identification and classification of the texts from which examples would be drawn. In short, one would expect to find lots of occurrences of the search term in non-native speaker texts, in texts in the wrong language, with a weighting towards text types of a computational and technical nature, etc..

The Webcorp project is an interesting place to investigate some of these problems (http://www.webcorp.org.uk/).

3. Archives of the future

3.1. Archives today

There are currently different types of archive of language resources, which may be roughly characterised as follows:

- Etext repositories;
- Repositories of tools;
- Commercial archives for the HLT community;
- Non-commercial "shareware" archives for HLT;
- Non-commercial "freeware" archives for academic research.

The changes in technology and resources can be expected to have a differential impact on these different types of archive. Indeed, this categorization of the current state of affairs can itself be expected shift and blur under the impact of the changes in the nature of the resources which they hold.

The *raisons d'être* of current archives are created by the research scenario outlined above (the "traditional" model) whereby resources are created for the purposes of a particular piece of linguistic research, deposited in an archive for preservation and distribution, and then users who wish to reuse the resource come along, download a copy for their own research. The nature of the current archive is therefore a reflection of the nature of language resources now and in the past, and includes the following key roles:

- Preservation;
- Distribution;
- Resource discovery.

It should also be noted, that some UK research funding bodies insist on the archiving of electronic resources which are the product of projects for which they have provided funding, thus providing a further reason for the existence of archives.

Given the changes in language resources which are described above, how will these be reflected in changes in the archive of the future?

3.2. Archives tomorrow

The centralized discovery of archives, resources and individual texts will become possible through initiatives like OLAC and will become increasingly important.

Standardized metadata will be a precondition for resource discovery initiatives of this type and also for all distributed services.

Textual quality control and textual metadata will become more important than textual availability, given the ubiquity of electronic text on the web. Twenty-five years ago the OTA was set up in response to a need for somewhere to gather and distribute electronic text. Now there is a need for reliable text of known provenance and characteristics which can be used for linguistic research.

The distinction made above between etext repositories and language resource archives will become blurred as the creation of virtual corpora becomes a reality. A repository of interoperable electronic texts can become the source of countless corpora.

In the future it is anticipated that the value of the archive will derive mainly from the following services which it will seek to provide:

1. A place to make virtual corpora and connect to tools for text processing, information extraction and conversion (whether the texts or the tools are held in the archive or elsewhere);
2. An aid to resource discovery;
3. Preservation of resources;
4. A source of expertise in resource creation and exploitation;
5. Quality assurance and validation, especially quality of metadata

There are tendencies noted above which may work against these roles for the archive. The development of a more distributed architecture for text analysis software would work against point 1 above in the longer term. Also, meta-archives such as proposed by OLAC would mean that the individual resource provider could provide metadata records directly to the meta-archive and circumvent the need for an archive to act as a repository, although point 4 above is relevant here as the role of the archive becomes more important to guarantee the stability and availability of the resource and the validity of the metadata. And in any case, at least in the short term, the easiest way to guarantee the inclusion and ongoing support of your metadata in a meta-archive is to deposit it in a reliable participating archive.

A further objection which may be levelled is that the necessity for preservation (point 3 above) is obviated by web archives like Wayback Machine (http://www.archive.org/). It has not however been demonstrated that this is a viable mechanism for preservation of digital resources, and in any case, it cannot guarantee the preservation of relevant metadata.

It can perhaps be said in summary that the single most important role of an archive such as the OTA in future is quality assurance of content and metadata.

3.3. Copyright

If copyright law as it currently exists in the developed world remains the same, it will continue to be a considerable burden and a barrier to linguistic research. In general, corpus builders are unable to take complete published texts and make them available for linguistic research. Some corpora, such of the Bank of English, have managed to acquire the desired texts (and continue to do so), but at the cost of the non-availability of the corpus, except through the lookup tool provided. Other corpora have to restrict themselves to older, out of copyright works or unpublished texts of various types. The British National Corpus was able to negotiate the use of fragments of texts. None of these solutions are ideal.

While no simple solution offers itself to the apparent conflict between, on the one had, intellectual property rights and the economic viability of electronic publishing and, on the other had, the free availability of high quality, published texts for academic research, the virtual corpus model may offer a way ahead. If the texts can still be held by the resource owners, but made available for querying online by various tools, then at least it becomes merely a technical problem of authorising the tools and ensuring the security of the texts.

4. Future work at the OTA

In the light of the above discussion of current and future trends, this section summarises and outlines some of the new directions in the archiving and delivery of digital language resources which are being explored at the OTA.

4.1. More new technologies

The OTA are keenly aware that the present period is seeing a measure of convergence of technologies and standards in several related fields which have in common the goal of delivering linguistic content through electronic means. These include relatively established technologies such as electronic publishing, internet search engines, electronic delivery of language reference and translation services, as well as emerging technologies such as ebooks, the mobile internet, digital libraries and the semantic web. The distribution of corpora, or of the linguistic knowledge embedded in corpora, is clearly related to these larger-scale, industrial developments. The OTA keeps a close watch on, and sometimes participates in, initiatives in the development of practice in these areas. The archive currently offers conversion “on the fly” to XML, SGML and ASCII formats for the downloading of resources, and we hope to offer migration to other formats in the near future.

Furthermore, there is the potential for opportunities to make use of, or even influence, emerging metadata and text encoding standards to make different types of electronic texts more easily usable for language research.

4.2. New resources

In the forthcoming period, there will be a concerted attempt to develop a new collections policy for language resources. Accession of digital resources created by and for the UK academic community in literary and linguistic subject areas remains at the forefront of our priorities. Furthermore, priority is still placed on the preservation of resources, with a particular emphasis on the new strategies needed to deal with multimedia, multimodal and dynamic resources.

There is a constant drive to acquire new resources such as corpora for the archive, particularly through identifying resource creation and enhancement projects in the UK. Although this has been identified above as the “traditional model”, this does not mean that such resources are not being produced and used in ever greater numbers today. It is also a priority to make the existing holdings more easily available and accessible, through the development of new, improved cataloguing procedures and a rights management strategy. The functionality of the website and the catalogue search mechanisms are also being upgraded.

4.3. New Services

As has been described above, the new areas in which language resource archives need to be active nowadays are the provision of more powerful and flexible online services, validation of metadata, resource discovery and sharing expertise in resource creation. These are being developed in addition to the more traditional roles of accession, cataloguing, preservation and distribution of resources.

In order to deliver the desired level of online access to the resources, with the possibility of building virtual corpora (temporary, *ad hoc* collections of texts for analysis), a concerted effort to deal with inconsistencies in text formats, markup and metadata is being applied.

The OTA has developed a checklist and a procedure for in-house use in the validation of resource metadata and content.

The OTA is participating in initiatives that assist in and enable resource discovery, such as the Open Archives Initiative (OAI), the Open Language Archives Community (OLAC) and the Arts and Humanities Portal in the UK.

One key activity in the area of developing and promotion of expertise and best practice is the commissioning and publication of *Developing Linguistic Corpora: a Guide to Good Practice*. This will be a practical guide to designing and creating a corpus, and will feature chapters written by experienced practitioners in the field. Progress in the preparation of this publication will be reported on in future publications.

5. Bibliography

Wynne, M. (2001). An archive for all of Europe: the TRACTOR initiative. In Workshop Proceedings: Sharing Tools and Resources at the 39th Annual Meeting and 10th Conference of the European Chapter of the Association for Computational Linguistics (pp. 59-62), Toulouse, France: CNRS.