Do we need annotated corpora in the era of the data deluge?

Martin Wynne
University of Oxford, UK
E-mail: martin.wynne@oucs.ox.ac.uk

Abstract

Language corpora were originally developed as datasets for linguistic research, in a world where researchers rarely had access to machine-readable language data. Corpus linguistics subsequently developed methodologies based on discrete, bounded datasets, created to represent certain types of language use, and studied as exemplars of that domain. The growth of the field and advances in technology meant that corpora became bigger and more plentiful and various, with huge reference corpora for a vast range of languages and time periods, and numerous specialist corpora. Researchers in many other fields have found that corpora are rich repositories of data about not only language but also culture, society, politics, etc.. Annotation has been widely used, initially by linguists, but also by researchers in other fields to categorize, interpretate and analyse texts, and to aid information retrieval and data linking.

Nowadays, the enormous wealth of digital language data at our fingertips brings the role of the corpus into question. Born-digital data, along with large-scale digitization of historical texts, are delivering the cultural products of the both the present and the past directly to our desktops. We can relatively easily make bespoke datasets for different research questions. The boundaries between the corpus and other type of data are becoming blurred. Can we still justify spending our time carefully crafting and annotating corpora today?

There is a danger that adding annotation is too time-consuming and costly, and forces us into using smaller and older datasets, and only starting our research after they have been painstakingly annotated. Furthermore, there is a lack of generic software which can make use of annotations, and so we build separate web interfaces for each annotated corpus. This leads to the now-familiar problem of the creation of digital silos - isolated from other corpora and tools, limited in functionality, and each with a different interface. Thus it can be argued that adding annotation to a corpus also adds to the problem of fragmentation of digital resources in the humanities.

These new difficulties come along at a time when we continue to wrestle with longstanding problems with annotation:

- How do we avoid the circularity of adding annotations and then counting and analysing them ourselves - “finding the Easter eggs that you hid in the garden yourself”?
- Will two humanists ever agree on the relevance or usefulness of any given annotation scheme, or the accuracy of any instantiation of it?
- Can automatic annotation be accurate enough to be useful to the humanities scholar?
- Do we risk ignoring the actual data itself by focussing on our interpretations and annotations?

The era of the data deluge poses additional questions. Do we now need to explore how far can we go in our research without carefuly crafted, annotated corpora? Or do we need better and faster tools to add automatic annotations to the deluge of data? Or will initiatives such as CLARIN make it easier to use a wide range of annotated corpora on common platforms?