Constructing E-Language Corpora: a focus on CorCenCC (The National Corpus of Contemporary Welsh)

Dawn Knight
Centre for Language and Communication Research, Cardiff University,
2 Column Drive, CF10 Cardiff, UK
E-mail: knightd5@cardiff.ac.uk

Abstract

Digital communication in the age of 'web 2.0' (that is the second generation of in the internet: an internet focused driven by user-generated content and the growth of social media) is becoming ever-increasingly embedded into our daily lives. It is impacting on the ways in which we work, socialise, communicate and live. Defining, characterising and understanding the ways in which discourse is used to scaffold our existence in this digital world is, therefore, emerged as an area of research that is a priority for applied linguists (amongst others). Corpus linguists are ideally situated to contribute to this work as they have the appropriate expertise to construct, analyse and characterise patterns of language use in large-scale bodies of such digital discourse (labelled 'e-language' here). Indeed, an increasing amount of e-language corpora are being developed to allow us to investigate e-language use.

This presentation discusses some of the methodological, technical, practical and ethical considerations and challenges faced in the construction of e-language corpora. It will outline, for example, some of the approaches used when planning the construction of e-language corpora including: obtaining consent; approaches to sampling, collecting and anonymising data; sourcing and attributing metadata, as well as some reflections on constructing a corpus infrastructure.

Discussions will be contextualised with reference to the Economic and Social Research Council (ESRC) and the Arts and Humanities Research Council (AHRC)-funded CorCenCC corpus (Corpws Cenedlaethol Cymraeg Cyfoes - The National Corpus of Contemporary Welsh) project. CorCenCC will be the first large-scale corpus of Welsh representative of language use across communication types, including 2 million words of e-language and 4 million words each of spoken and written language. CorCenCC will be open-source and freely available for use by professional communities and anyone with an interest in language. Bespoke applications and instructions will be provided for different user groups. The corpus will enable, for example, community users to investigate dialect variation or idiosyncrasies of their own language use; professional users to profile texts for readability or develop digital language tools; to learn from real life models of Welsh; and researchers to investigate patterns of language use and change.