Grappling with Shakespeare's words: maximizing historical corpus-based approaches

Jonathan Culpeper and Amelia Joulain-Jay
(Lancaster University, UK)

This paper reports some of the work being undertaken in the context of the Encyclopaedia of Shakespeare's Language Project (see Shakespearelang, n.d.). Funded by the Arts & Humanities Research Council (AHRC), this project brings the corpus approach into the heart of Shakespearean studies and more generally Early Modern English. It affords fresh insights into Shakespeare's use of language at multiple levels – words, phrases, semantic themes, character profiles and more. In particular, it reveals what Shakespeare's language meant to the Elizabethans through the analysis of millions of words written by his contemporaries. The main output of the project will be a two-volume Encyclopaedia, published in paper and online. This paper focuses on Volume 1, which essentially is a corpus-based dictionary of Shakespeare's words. We elaborate on (a) the problems we encountered and solutions we adopted, and (b) how our corpus-based approach improves on current Shakespearean language scholarship.

Working on historical data brings with it familiar problems for the corpus researcher. We will briefly elaborate on the following problematic areas and explain how we tackled them:

(1) Spelling variation. Spelling variation is not, of course, unique to historical data. In the present-day world of global electronically mediated communication, authors are hardly conforming to one spelling standard. Nevertheless, spelling variation is particularly marked in Shakespeare's period, partly because standardised spelling was only just beginning to emerge, but also because of printing practices (e.g. line justification was largely achieved by adding in extra consonants or the letter <e>, or using the wider <y> instead of <i>). Our solution was to use VARD (VARiant Detector) (Baron, n.d.), software which can help identify and standardise historical spelling variation. But all is not plain sailing. One key issue is: what do you regularize to? This is not a problem when regularizing spelling today: there are standardized spellings, such as British English or American English, that one can deploy. But in Shakespeare's period there was no such thing. In general, our solution was to use data in Early English Books Online (EEBO-TCP) to establish the most frequent spelling variant, and then use that variant as our 'standard'. However, a downside of this is that one can end up with a less than transparent regularized form, from the perspective of today's reader. An example is *a clock*, which was hugely more frequent than *o clock* in this period. We handled this issue in the dictionary through using cross-references and supplying critical information about spelling variants.

(2) Part of speech tagging and EModE. The CLAWS part-of-speech annotation system works well for present-day English (see CLAWS, n.d.), and has been adapted for Early Modern English. Rayson et al. (2007) found it to perform at 85% accuracy for Shakespearean texts. However, when a dictionary's headwords are based on a list of 'taglemmas' (i.e. lemma + POS tag), accuracy is critical. Just

one example of a problem is the word *blest*, which is variously tagged. In *which not to have been blest withall*, it is erroneously tagged as JJ (adjective). We handled this issue in two ways: we instituted a number of "fixes" to CLAWS (often simply to the lexicon), and we did a manual post-check on our core Shakespearean data.

(3) Data and genre. Data is almost always a problem for historical corpus linguistics, because what survives is at best a reduced and patchy record of the total linguistic output of any period. Given that our project aim is to place Shakespeare in context by comparing his works with those of other writers, we needed a large quantity of comparative data. Fortunately, we have seen the advent of the transcribed 1.2 billion-word EEBO-TCP, approximately 321 million words of which span the period 1580-1640. However, this data lacks a full classification of genre. In historical work, genre is perhaps the key notion for accessing the stylistic flavour of an expression - whether it is formal, colloquial, literary, informational, and so on. Consequently, we instituted a classification scheme for the 1560-1640 period, largely based on the existing titles of works (in effect, their self-classifications).

The problems illustrated above are certainly not unique to our project. What is unique to our project is our approach to solving the over-arching problem of bringing together all the relevant information generated during these preparatory phases. Each dictionary entry is to be based on multiple pieces of information – information about spellings, part-of-speech, collocates, genre distribution, social distribution (e.g. male/female; high rank/low rank), and more – and, moreover, there are multiple information sets – Shakespeare's plays, his poetry, the Folios, the Quartos, our comparative corpus of playwrights and the EEBO-TCP. Extraction of the information is not the difficulty; we will largely be using CQPweb (Hardie, 2012) for this. The problem is more one of resources: it would take a team of researchers an inordinate amount of time, well beyond the bounds of project funding, to manually extract each piece of information and then make sense of the whole. We need a way of automatically pooling the bulk of the information, presenting it to researchers in a palatable fashion, and allowing them *in situ* to construct an interpretative summary that will constitute a dictionary entry. Our solution was to construct a database, accompanied with a user-friendly interface, for use by our team of lexicographers.

The database consists of two major parts. One part contains unchanging data, organised around taglemmas (lemma and POS tag pairing, see also above): for each taglemma identified in Shakespeare's First Folio, a series of information is automatically extracted from CQPweb and loaded into the database. This information includes overall frequency and dispersion of the taglemma in the First Folio, but also frequency and dispersion within sub-categories relevant to the First Folio, such as text genre, gender of characters, social status of characters, regularised morphological forms, and original spelling variants. Also included are overall and sub-category frequency and dispersion figures for these taglemmas in EEBO-TCP (specifically 1560-1640) and in a corpus containing plays by Shakespeare's contemporaries.

The second part of the database contains data generated by the lexicographers, including definitions, examples, cross-references and comments about their observations of the data. The user-friendly interface facilitates the generation of this content by providing access to the unchanging data stored in the database, as well as facilities for uploading manually generated data. Beyond providing access to stored data, the interface also provides other crucial functions for the project, such as helping facilitate collaboration between users (e.g. via the sharing of comments), providing version control, and helping with error- and inconsistency- checking (e.g. by providing a mechanism for selecting and updating cross-references).

To illustrate both the use of the database and interface, and the scholarly contributions of the project, we present two case studies chosen to maximize diversity:

(1) A more grammatical word: *I*. Though typically omitted from Shakespearean dictionaries, presumably on the assumption that its meaning has not changed or that it does not contribute much to understanding Shakespeare, analyses of collocates reveal that it is key in revealing character states, thoughts and feelings, as well as doing interpersonal work. It also turns out that Shakespeare had a penchant for *I*, relative to his contemporaries, at least in certain constructions, and used it to bolster particular types of characters (e.g. Desdemona in *Othello*).

(2) A more lexical word: *good*. Shakespearean dictionaries seem overwhelmed by the 2,711 instances of the word *good*, something that seems to be reflected in their widely varying accounts of the word. We will show how corpus-based analyses improve on those accounts, and actually provide support for one of the older accounts, Onions (1986/1911), in placing the usage he referred to as "conventional epithet" in pole position.

References

Baron, Alistair (n.d.). *VARD – about.* URL: http://ucrel.lancs.ac.uk/vard/. Last accessed: 22 Dec. 2016.

CLAWS (n.d.). *CLAWS Part-of-speech Tagger for English.* URL: http://ucrel.lancs.ac.uk/claws/. Last accessed: 22 Dec. 2016.

Hardie, Andrew (2012). CQPweb—combining power, flexibility and usability in a corpus analysis tool. *International journal of corpus linguistics*, *17*(3), 380-409.

Onions, Charles T. (1986/1911) (2nd edn.). *A Shakespeare Glossary.* (Enlarged and revised by Robert D.Eagleson). Oxford: Clarendon Press.

Rayson, Paul, Archer, Dawn, Baron, Alistair, Culpeper, Jonathan and Nick Smith (2007). Tagging the Bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora. In *proceedings of Corpus Linguistics 2007, July 27-30, University of Birmingham, UK*.

Shakespearelang (n.d.). *Encyclopaedia of Shakespeare's Language Project.* URL: http://wp.lancs.ac.uk/shakespearelang/. Last accessed: 22 Dec. 2016.