A New Wordtree Corpus Interface

Hilary Nesi

Department of English and Languages
Coventry University, UK

Paper presented at the Asia Pacific Corpus Linguistics Conference,
University of Auckland, 16-19 February 2012

Keywords: interface; collocation, concordance, lexico-grammatical pattern, visualization

Introduction

The most used and cited corpora in the public domain tend to be those which enable the easiest and fastest extraction of relevant data. Potentially useful corpora are often underused because they lack online interfaces, or because their interfaces were originally designed for lexicographers and information scientists rather than for language teachers, language learners, and researchers in applied linguistics.

This paper reports on work in progress for a publicly funded project to develop a more usable and effective visual corpus interface. It is currently being demonstrated as an interface to the BAWE corpus, utilising BAWE TEI-encoded header information to enable users to create and compare their own bespoke corpus subcomponents.

Existing text visualization tools

Interactive visualizations, such as those created by Clark (2008) and McCandless (2010), allow users to zoom in on specific details within large amounts of information, and encourage browsing because they are easy to navigate and are visually attractive. Doyle (2010) provides an overview of the visual displays in current corpus query tools. WordSmith Tools (Scott 2004), for example, can display collocation patterns and dispersion plots, and the BYU interface for COCA has interactive distribution charts. Such displays are limited by the fact that they are based around tables and grids, however, and Doyle (2010: 161) argues that more effective visualisations can be achieved by transforming data into visual objects, for example in the form of word cluster diagrams, also known as ‘word clouds’ (Clark, 2008), ‘tag clouds’ (Alex, undated) or ‘Wordles’ (Feinberg, undated). Word cluster diagrams group words within a text, using word size to indicate frequency and positioning and color schemes to indicate collocational relationships. They are generally used for decorative rather than for educational purposes, however. TextArc (Paley 2009) adds interactivity to the word cluster diagram, using moving coloured lines to link words and collocations. Paley (2009) and Doyle (2010) claim that this stimulates users to explore works of literature, but nevertheless the use of TextArc as a corpus linguistic tool seems limited because the lexical connections are not quantified, and there does not seem to be any facility to gather and store the data that is fleetingly revealed.

Word tree visualisations are essentially interactive Key Word in Context (KWIC) concordance lines grouped into tree-like branching structures. Although standard concordancing programs can make collocational patterns more apparent by sorting words preceding or following the search term, it still takes skill to interpret concordance lines effectively, and novice concordance users often tend to scan the lines horizontally for meaning rather than vertically for lexico-grammatical patterns. Students quickly lose interest in classroom concordancing tasks, as Thurston and Candlin (1998:278) point out. A tree structure provides the same amount of information on screen but reduces the amount of text; the tree is obviously not a piece of continuous prose, and the patterns are immediately apparent when the tree is read horizontally. Moreover the interactive ‘zoom in’ feature makes word tree searches more visually exciting than concordancing.

Our project takes as its starting point the IBM Many Eyes Word Tree visualization (Wattenberg and Viégas, 2008). This works well, but is limited by the fact that text can only be uploaded in a single file of no more than 5MB, and the tree does not display any statistical information. Figure 1
shows a screenshot of a Many Eyes Word Tree for ‘although’, using the BAWE subcorpus of assignments belonging to the ‘Critique’ genre family (the entire BAWE corpus is too large for Many Eyes).

Figure 1: ‘although’ in a Many Eyes Word Tree

Our Word Tree specifications

Many Eyes Word Tree visualisations were shown to experienced corpus linguists, novice researchers, and language learners as a starting point for our project. We then set about developing a more flexible and elaborate interface in response to their feedback. Some of their main requirements for our new interface were increased data capacity, the facility to create subcorpora, word frequency, distribution and dispersion statistics (both raw and normalised), and access from the screen to the original text.

This paper will demonstrate a beta version of our new Word Tree, and discuss some of the issues surrounding its development. The blog for the project is http://cuba.coventry.ac.uk/wordtree.

References

Corpus of Contemporary American English (COCA). http://corpus.byu.edu/coca