"What kind of neighbourhood is that?" Helping learners explore the semantic neighbourhood of words and phrases.

Stephen Jeaco (Xi'an Jiaotong-Liverpool University, China)

In a language learning setting, concordancers are particularly useful for showing differences between patterns of different lexical items. Language learners may explore grammatical and lexical patterning in concordance lines or draw on results from measurements like collocation. One aspect of patterning which could be important but is hard for learners to explore, is the way similar lexical items may typically be used in rather different semantic contexts. Being able to recognise the typical semantic contexts of synonyms and words which have a common translation could be a useful step in distinguishing between words with a similar meaning. For words which seem to have more than one sense or more than one use in different domains, information about the patterning of the semantic contexts could also help language learners find distinctions. These kinds of patterning could also help language learners explore and uncover hidden connotation-like qualities which words may have resulting from common use with other words. An overview of the historical development of semantic prosody and related theories, including the influences of work by Sinclair, Louw, Stubbs, Hoey and others, is provided by Stewart (2010). Although there are some differences between conceptions and definitions of semantic prosody, semantic association and semantic preference (Hoey, 2005), they all include the possibility for emotive charging of words through their frequent use in specific contexts. In terms of the pedagogical implications of work in this area, Stewart argues there is a need for "serious improvements" in descriptive works for English language learning (2010, p. 263). The need to address the gap in English dictionary resources in China has also been highlighted as a high priority (Ping-Fang & Jing-Chun, 2009). In cross-linguistic work, it has been shown that the tendencies of semantic prosody of similar items across different languages can be quite different (Xiao & McEnery, 2006). Lack of access to information about typical uses of words and collocations and lack of access to information about tendencies of semantic prosody in dictionary resources and other descriptive works could account for some of the difficulties non-native speakers face in trying to produce language which matches the expectations and conventions of their intended audience.

With some assistance, learners can use concordancers to explore corpus examples with these questions in mind. When comparing words or phrases (particularly when the results can be viewed side-by-side on screen), some aspects of semantic association should be fairly clear. Language teachers would probably not want to specifically teach the linguistic terminology of semantic prosody, but visualizations such as collocation clouds often provide sufficient evidence of how some words tend to be associated with positive or negative collocates. More detailed analysis allows for narrower categorizations or groupings. Hunston introduces ways in which a hidden meaning of words may be deduced through analysis of concordance lines, an also provides examples of how different meanings of words can be observed by looking at patterns in the co-text (2002). When researchers are looking at data like these, there is a need for them to consider possible bias in the corpus because of the kinds of texts from which it was built, to consider "resonances of intertextuality" and to be aware that observational evidence

should usually be interpreted as "often" but "not always" (Hunston, 2007). When language learners are looking at data like these, they may need encouragement to open up to the possibility of looking for semantically-related items in the nearby context, and they may also need also assistance in judging the strength of any attitudinal meaning. The question arises as to whether computational approaches could assist with this.

Within the field of Information Retrieval and Natural Language Processing, there are an increasing number of electronic resources for Sentiment Analysis holding information about associations between words and attitudes, emotions, etc. The General Inquirer lexicon, the Dictionary of Affect in Language, SentiWordNet, and WordNet-Affect are four widely used resources in this field (Devitt & Ahmad, 2013). Other resources include LIWC (Pennebaker, Chung, Ireland, Gonzales, & Booth, 2007) and USAS (Rayson, Archer, Piao, & McEnery, 2004). When making use of these resources, computational methods are often concerned with attempting to mark texts or text extracts as being associated with particular attitudes or used for searches within a specified semantic environment. Some applications of these resources are more related to information retrieval or monitoring of content; others provide additional tools for corpus linguistic analyses. When the object to be explored is a text (or a collection of texts), for example, *WMatrix* (Rayson, 2008) makes comparisons very clear by showing results of key semantic domains for one using the other as a reference corpus. However, when the object to be explored is a word, a collocation, or a pair of words with a similar meaning, semantic tags could also be used to provide summary information about tendencies the search query to be used in specific semantic contexts.

The Prime Machine (Jeaco, 2015) is a concordancer which was developed for students of English and their teachers as a resource for language learning. As well as search support features, it has several ways of encouraging learners to explore aspects of the environment of the lexical items they are interested in, and to compare these side by side with similar items. This paper reports on some further developments of *The Prime Machine*, which now draws on semantic tagging data to provide two kinds of information to learners regarding the neighbourhood of the words from their search query. They can then interact with these data in three ways: through tables or clouds of strongly related semantic tags; through tables or graphs showing the proportions of concordance lines within positive or negative environments; and through filtering and comparing concordance lines.

The first way in which semantic tagging is used in the new version is by marking sentences in the corpus based on semantic tagging data from USAS (Rayson, et al., 2004). This is operationalized by counting only those semantic tags which meet a threshold of repetition within +/-1 sentence. In the concordancer, this wider context is easily accessible for each concordance line and it is also worth noting that while collocation information about words within windows of several words is certainly very important, Hoey has demonstrated how associations between words may occur (and be of importance) in wider contexts too (2014). Links based on thresholds of 2-8 repetitions are stored in the database, meaning that some fine-tuning is available. Since many items may have multiple semantic tags, this provides a straight-forward (albeit limited) means of "disambiguation", as only after a threshold number of items in the sentence have been found to share the same semantic tag will they be counted. The software uses log likelihood contingency

tables for each word and each collocation stored in the database (shown in Table 1), by creating a sub-corpus of sentences marked with each specific semantic tag. Figure 1 shows a screenshot of the cloud of results for *due to*.

Table 1: Semantic Tag Contingency Table

Sub-corpus of sentences marked with a specific semantic tag	The rest of the corpus
A = Count of word (or words in a collocation)	B = Total count of word (or words in a collocation) – A
C = Count of all words in sentences with the semantic tag	D = Total corpus size – C

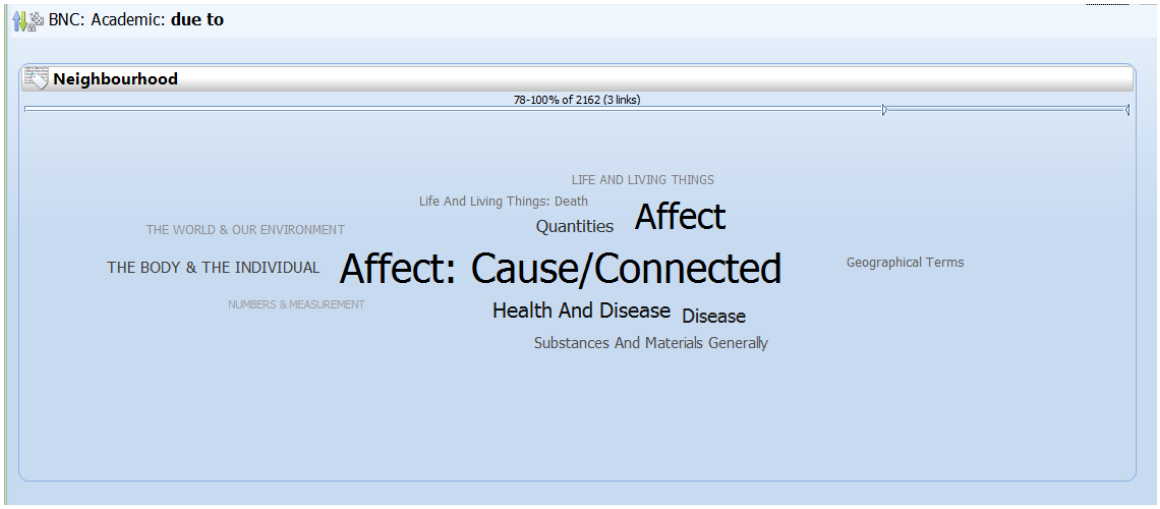

Figure 1: Screenshot of the Neighbourhood Cloud for *due to* in the BNC: Academic Sub-corpus.

The second way semantic tagging is used is based on a narrower window (+/- 4 words), drawing on specific USAS tags associated with positivity or negativity and two lists of positive and negative words: 24 semantic tags from USAS (such as "A1.1.2 Damaging And Destroying" and "A1.4 Chance, Luck"), and the two other lists from as combined resources derived from the General Inquirer lexicon (GI, 2000), the NRC emotion lexicon (Mohammad & Turney, 2012) and a Loughran-McDonald wordlist (Bodnaruk, Loughran, & McDonald, 2015). Each word in the corpus is marked according to whether it occurs near these items. Since words on the list will always be marked, a further flag is used to indicate whether at least one other positive/negative item also occurs. The markers are then processed like the other features of lexical priming (as in earlier versions of the software): proportions are stored for each lexical item and the norms for the corpus, and a log likelihood measure is used to determine whether or not icons appear in the application to encourage users to explore this aspect. When looking up words in the corpus, if the positive or negative relationship is strong, the icon link used to explore this pops out on the icon bar. Figure 2 shows the graph of the results for *due to*, with the storm

icon at the bottom of the screen indicating it is strongly related to negative environments.

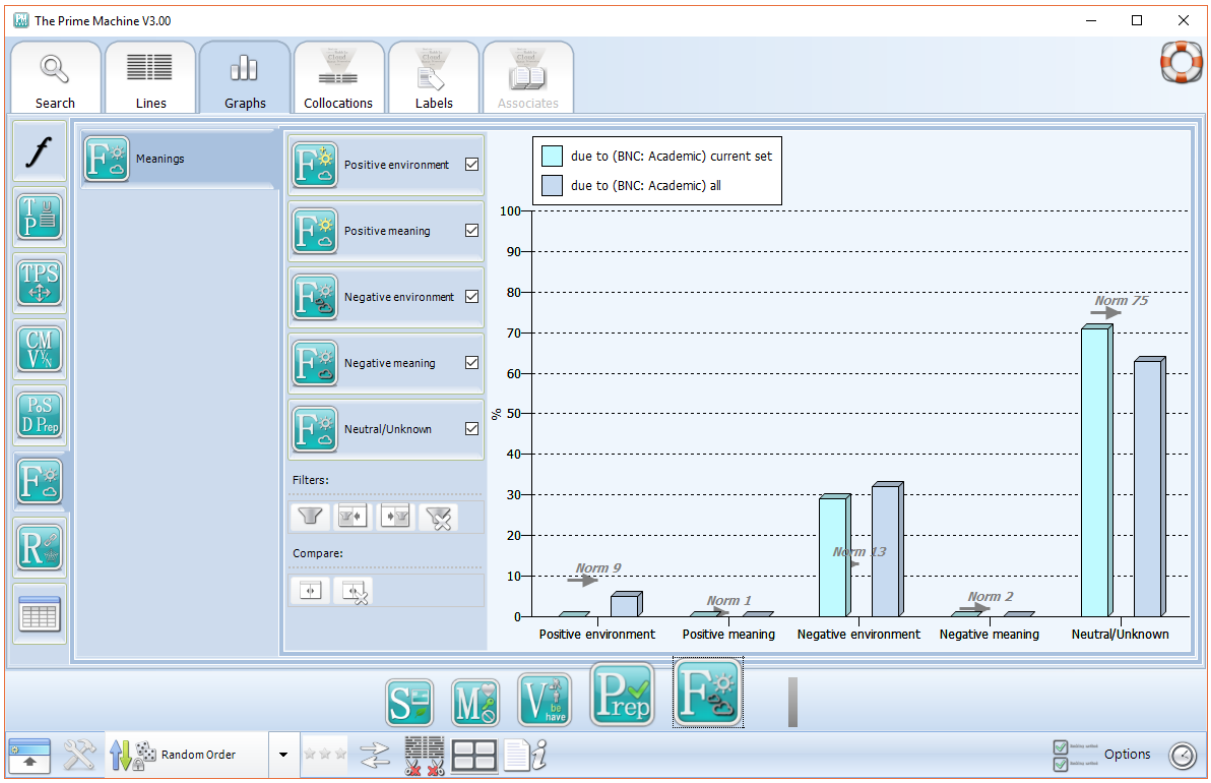

Figure 2: Screenshot of the Graph Showing the Proportion of Concordance Lines Marked in Positive and Negative Environments for *due to* in the BNC: Academic Sub-corpus.

It is argued that this approach provides a means of helping language learners explore pertinent features of the neighbourhood in the concordance lines, giving a helping hand for their concordance line analysis and ultimately for their insights and awareness of patterns of language use. Details about access to the software will be available from www.theprimemachine.com.

References

Bodnaruk, A., Loughran, T., & McDonald, B. (2015). Using 10-K Text to Gauge Financial Constraints. *Journal of Financial & Quantitative Analysis, 50*(4), 623-646. doi: 10.1017/s0022109015000411

Devitt, A., & Ahmad, K. (2013). Is there a language of sentiment? An analysis of lexical resources for sentiment analysis. *Language Resources & Evaluation, 47*(2), 475-511.

GI. (2000). General Inquirer: URL: http://www.wjh.harvard.edu/~inquirer/homecat.htm.

Hoey, M. (2005). *Lexical Priming: A New Theory of Words and Language*. London: Routledge.

Hoey, M. (2014). Words and their neighbours. In J. R. Taylor (Ed.), *Oxford Handbook of the Word*. Oxford: Oxford University Press. Advance online publication. doi: 10.1093/oxfordhb/9780199641604.013.39.

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Hunston, S. (2007). Semantic prosody revisited. [Article]. *International Journal of Corpus Linguistics, 12*(2), 249-268.

Jeaco, S. (2015). The Prime Machine: a user-friendly corpus tool for English language teaching and self-tutoring based on the Lexical Priming theory of language. Unpublished Ph.D. dissertation, University of Liverpool. Retrieved from https://livrepository.liverpool.ac.uk/2014579/

Mohammad, S. M., & Turney, P. D. (2012). Crowdsourcing a word–emotion association lexicon. *Computational Intelligence, 59*.

Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. (2007). The development and psychometric properties of LIWC2007. Retrieved 18 June 2015, from http://homepage.psy.utexas.edu/homepage/faculty/pennebaker/reprints/liwc2007_languagemanual.pdf

Ping-Fang, Y., & Jing-Chun, C. (2009). Semantic prosody: A new perspective on lexicography. *US-China Foreign Language, 7*(1), 20-25.

Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics, 13*(4), 519-549.

Rayson, P., Archer, D., Piao, S., & McEnery, T. (2004). The UCREL semantic analysis system. Paper presented at the Beyond Named Entity Recognition Semantic Labeling for NLP Tasks Workshop, Lisbon, Portugal.

Stewart, D. (2010). *Semantic Prosody: A Critical Evaluation*. New York: Routledge.

Xiao, R., & McEnery, T. (2006). Collocation, semantic prosody, and near synonymy: A cross-linguistic perspective. *Applied Linguistics, 27*(1), 103-129.