Lexical Network Potentials Based on Co-occurrence Patterns: A Preliminary Analysis of Graded Readers

Naoki Sugino[1], Noriko Aotani[2], Simon Fraser[3], and Yuya Koga[4]

[1]College of Information Science and Engineering, Ritsumeikan University,

[2]Faculty of Education, Tokai Gakuen University,

[3]Institute for Foreign Language Research and Education, Hiroshima University,

[4]School of Interdisciplinary Mathematical Sciences, Meiji University

gwisno@is.ritsumei.ac.jp, aotani@tokaigakuen-u.ac.jp,

fraser@hiroshima-u.ac.jp, yuya.koga@gmail.com

Abstract

As a preliminary study to investigate if, and how, extensive reading can facilitate learners' efforts to organise their lexical knowledge, this study extracts co-occurrence patterns of words in reading materials, and visualises them as a network. This network of co-occurring words can be regarded as a learning potential implicitly presented to the learners. Six graded readers (three expository texts and three narratives) are used to compile two small-sized corpora. Association analysis is applied to each corpus to extract pairs of words that co-occur in sentences with sufficient salience. The extracted patterns are then submitted to network analysis in order to visualise the network structures and compare metrics that characterise each of these structures.

Keywords

lexical network, graded readers, association analysis, co-occurrence patterns, Gephi

Introduction

L2 vocabulary learning through extensive reading has been a vigorously researched topic. However, these studies have focused mainly on an increase in learners' vocabulary size. Similarly, analyses of reading materials have concentrated on coverage and repetition of the words at various frequency levels. To our knowledge, little attention has been directed to how extensive reading might help learners organise their lexical knowledge. As a preliminary study to fill this gap, we will attempt to present lexical network potentials inherent in graded readers. Employing association analysis, co-occurrence patterns of words in sentences are extracted from each book, which is followed by explication of how the patterns evolve as the total number of the running words increases. This explication is made possible by the use of *Gephi* (Bastian, Heymann, & Jacomy, 2009), an open-source software package for network analysis and visualisation. The properties of the networks will be reported, and pedagogical implications will be drawn.

1 The organisation dimension of lexical knowledge

One of the most comprehensive schematisations of learners' lexical knowledge is the one proposed by Nation (2001) with nine categories for each of the receptive and the productive skills. Meara (1996), however, had pointed out that this kind of schematisation would be impractical in developing a reliable test to measure the learner's vocabulary development. Instead, he proposes two dimensions, viz., size and organisation, and suggests that once the vocabulary has sufficiently developed in size, the organisation dimension becomes of more significance, differentiating just knowing a large number of words from knowing them as an organised system.

As the measurements that captures how the lexical knowledge as a whole is organised, Meara (1996: 46) suggests, referring to Kiss (1968), that application of graph theory would reveal key features of a network. However, except for some pioneering studies (e.g., Meara, 2008; Wilks & Meara, 2002; Vitevitch, 2008), applications of graph theory are still limited in number and range.

2 The present study

Six books, all at the third of the six stages (the 1000 headword level) on the same graded reader scheme, are placed under scrutiny. Three narratives and three expository texts are used to investigate if differences are observed between the two genres as in Gardner (2004).

In this preliminary study, we will focus on the methodological consideration of combining association analysis and network analysis in elucidating learning potentials and demands in reading materials.

Association analysis is a methodology in data mining that detects salient relationships, represented as association rules, hidden in large data sets. The salience of a relationship is determined in terms of its *support* and *confidence*. *Support* indicates the probability of a particular pair of items co-occurring in the whole data set, whereas *confidence* shows the probability of a particular item co-occurring with another item. The extension packages ‘arules’ and ‘arulesViz’ in R are employed in the present study. In the network analysis, following Vitevitch (2008), three metrics, viz., average path length, clustering coefficient, and degree distribution, will be used to compare the network structures.

3 Results

For illustrative purposes, let us briefly report on the lexical network obtained from one of the expository reading materials. The total word count of the material is 10,489 words. The text contains 737 sentences, and co-occurrences of words in each sentence are examined.

Association analysis extracted one hundred association rules, among sixty five words, as are represented by the nodes and the links in Figure 1. Six modularity classes are identified in this material, as indicated by the colors of the nodes, two of which comprises approximately 75% of the co-occurring pairs, while their size represents *degree*, the number of connections.

The organisation characteristics inherent in the reading materials are thus explicated and visualised by the use of association analysis and network analysis.

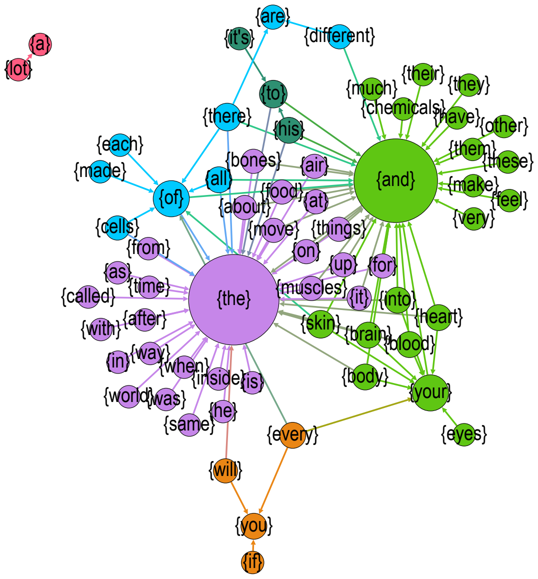

Figure 1: Lexical Network of a Graded Reader

References (doi omitted)

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. *International AAAI Conference on Web and Social Media*, *8*, 361–362.

Gardner, D. (2004). Vocabulary input through extensive reading: A comparison of words found in children’s narrative and expository reading materials. *Applied Linguistics*, *25*, 1–37.

Kiss, G.R. (1968). Words, associations, and networks. *Journal of Verbal Learning and Verbal Behavior*, *7*, 707–713.

Meara, P. (1996). The dimensions of lexical competence. In G. Brown, K. Malmkjaer, & J. Williams. (Eds.), *Performance and competence in second language acquisition* (pp. 35–53). Cambridge, England: Cambridge University Press.

Meara, P. (2006). Emergent properties of multilingual lexicons. *Applied Linguistics*, *27*, 620–644.

Nation, I.S.P. (2001). *Learning vocabulary in another language*. Cambridge, England: Cambridge University Press.

Vitevitch, M. S. (2008). What can graph theory tell us about word learning and lexical retrieval? *Journal of Speech, Language, and Hearing Research*, *51*, 408–422.

Wilks, C., & Meara, P.M. (2002). Untangling word webs: graph theory and the notion of density in second language word association networks. *Second Language Research*, *18*, 303–324.