GLANCE Visualizes Lexical Phenomena for Language Learning

Mei-Hua Chen[*], **Shih-Ting Huang**[+], **Ting-Hui Kao**[+], **Sun-Wen Chiu**[+], **Tzu-His Yen**[+]
[*]Department of Foreign Languages and Literature, Hua Fan University, Taipei, Taiwan, R.O.C. 22301
[+]Department of Computer Science, National Tsing Hua University, HsinChu, Taiwan, R.O.C. 30013
{chen.meihua,koromiko1104,maxis1718,chiuhsunwen, joseph.yen}@gmail.com

Abstract

Facilitating vocabulary knowledge is a challenging aspect for language learners. Although current corpus-based reference tools provide authentic contextual clues, the plain text format is not conducive to fully illustrating some lexical phenomena. Thus, this paper proposes GLANCE [1], a text visualization tool, to present a large amount of lexical phenomena using charts and graphs, aimed at helping language learners understand a word quickly and intuitively. To evaluate the effectiveness of the system, we designed interfaces to allow comparison between text and graphics presentation, and conducted a preliminary user study with ESL students. The results show that the visualized display is of greater benefit to the understanding of word characteristics than textual display.

1 Introduction

Vocabulary is a challenging aspect for language learners to master. Extended word knowledge, such as word polarity and position, is not widely available in traditional dictionaries. Thus, for most language learners, it is very difficult to have a good command of such lexical phenomena.

Current linguistics software programs use large corpus data to advance language learning. The use of corpora exposes learners to authentic contextual clues and lets them discover patterns or collocations of words from contextual clues (Partington, 1998). However, a huge amount of data can be overwhelming and time-consuming (Yeh et al., 2007) for language learners to induce rules or patterns. On the other hand, some lexical phenomena seem unable to be comprehended fast and directly in plain text format (Koo, 2006). For example, in the British National Corpus (2007), "however" seems more negative than "but". Also, compared with "but", "however" appears more frequently at the beginning of a sentence.

With this in mind, we proposed GLANCE[1], a text visualization tool, which presents corpus data using charts and graphs to help language learners understand the lexical phenomena of a word quickly and intuitively. In this paper, we focused on five types of lexical phenomena: polarity, position, POS, form and discipline, which will be detailed in the Section 3. Given a single query word, the GLANCE system shows graphical representations of its lexical phenomena sequentially within a single web page.

Additionally we believe that the use of graphics also facilitates the understanding of the differences between two words. Taking this into consideration, we introduce a comparison mode to help learners differentiate two words at a glance. Allowing two word input, GLANCE draws the individual representative graphs for both words and presents these graphs in a two-column view. The display of parallel graphs depicts the distinctions between the two words clearly.

2 Related Work

Corpus-based language learning has widened the perspectives in second and foreign language education, such as vocabulary learning (Wood, 2001). In past decades, various corpus-based reference tools have been developed. For example, WordSmith (Scott, 2000), Compleat Lexical Tutor (Cobb, 2007), GRASP (Huang et al., 2011), PREFER (Chen et al, 2012).

Recently, some interactive visualization tools have been developed for the purpose of illustrating various linguistic phenomena. Three exam-

[1] http://glance-it.herokuapp.com/

Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, pages 34–37,
Baltimore, Maryland, USA, June 27, 2014.

ples are Word Tree, a visual concordance (Wattenberg and Viégas, 2008), WORDGRAPH, a visual tool for context-sensitive word choice (Riehmann et al., 2012) and Visual Thesaurus, a 3D interactive reference tool (ThinkMap Inc., 2005).

3 Design of the GLANCE System

The GLANCE system consists of several components of corpus data visualization. We design and implement these visualization modules separately to ensure all graphs are simple and clear enough for users to capture and understand the lexical phenomena quickly.

In this paper, we use the d3.js (Data-Driven Documents) (Bostock et al., 2011) to visualize the data. The d3.js enables direct inspection and manipulation of a standard document object model (DOM) so that we are able to transform numeric data into various types of graphs when fitting these data to other visualization tools. In this section, we describe the ways we extract the data from the corpus and how we translate these data into informative graphs.

3.1 Data Preprocessing

We use the well-formed corpus, the BNC, to extract the data. In order to obtain the Part-of-speech tags for each text, we use the GENIA tagger (Tsuruoka et al., 2005) to analyze the sentences of the BNC and build a list of <*POS-tag*, *frequency*> pairs for each word in the BNC. Also the BNC contains the classification code assigned to the text in a genre-based analysis carried out at Lancaster University by Lee (2001). For each word, the classification codes are aggregated to a list of <*code*, *frequency*> pairs.

3.2 Visualization of Lexical Phenomena

Polarity

A word may carry different sentiment polarities (i.e., positive, negative and objective). To help users quickly determine the proper sentiment polarity of a word, we introduce the sentiment polarity information of SentiWordNet (Baccianella et al., 2010) into our system. For each synset of a word, GLANCE displays the polarity in a bar with three different colors. The individual length of the three parts in the bar corresponds to the polarity scores of a synset (Figure 1).

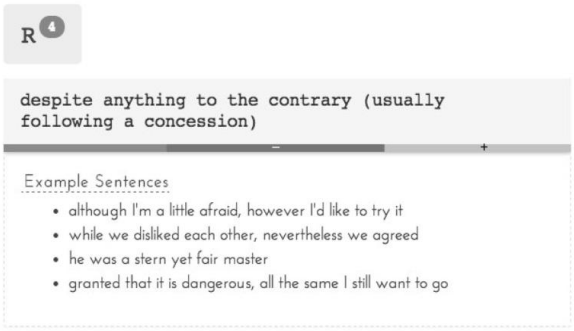

Figure 1. Representation of sentiment polarity

Position

The word position in a sentence is also an important lexical phenomenon. By calculating the word position in each sentence, we then obtain the location distribution. GLANCE visualizes the distribution information of a word using a bar chart. Figure 2 shows a plot of distribution of word position on the x-axis against the word frequency on the y-axis.

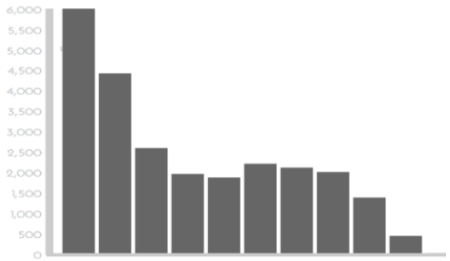

Figure 2. Distribution of word position

Part Of Speech (POS)

A lexical item may have more than one part of speech. Knowing the distribution of POS helps users quickly understand the general usage of a word.

GLANCE displays a pie chart for each word to differentiate between its parts of speech. We use the maximum likelihood probability of a POS tag for a word as the arc length of the pie chart (Figure 3).

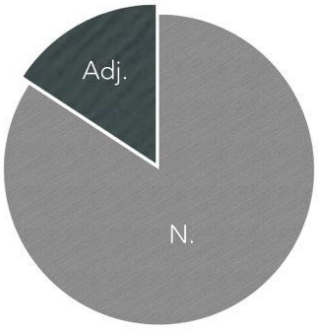

Figure 3. POS representation

Form

The levels of formality of written and spoken language are different, which also confuse language learners. Pie charts are used to illustrate the proportion of written and spoken English of individual words as shown in Figure 4.

We derive the frequencies of both forms from the BNC classification code for each word. The arc length of each sector is proportional to the maximum likelihood probability of forms.

Figure 4. Form representation

Discipline

Similar to language form, the discipline information (e.g., newspaper or fiction) was gathered from the BNC classification code. The relations of the disciplines of a word are presented using a sunburst graph, a radial space-filling tree layout implemented with prefuse (Heer et al., 2005). In the sunburst graph (Figure 5.), each level corresponds to the relation of the disciplines of a certain word. The farther the level is away from the center, the more specific the discipline is. Each level is given equal width, but the circular angle swept out by a discipline corresponds to the frequency of the disciplines.

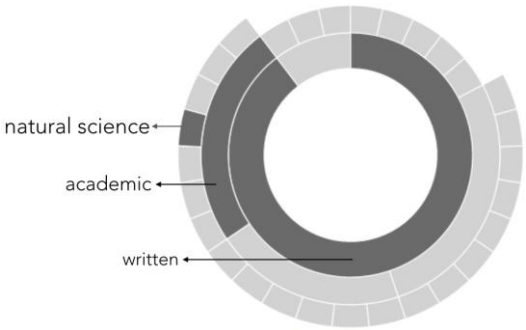

Figure 5. Discipline relations

4 Results

4.1 Experimental Setting

We performed a preliminary user study to assess the efficiency of our system in assisting language learners in grasping lexical phenomena. To examine the effectiveness of visualization, we built a textual interface for comparison with the graphical interface.

Ten pre-intermediate ESL college students participated in the study. A total of six pairs of similar words were listed on the worksheet. After being introduced to GLANCE, all students were randomly divided into two groups. One group was required to consult the first three pairs using the graphical interface and the second three pairs the textual interface, and vice versa. The participants were allowed a maximum of one minute per pair, which meets the goal of this study of quickly glancing at the graphics and grasping the concepts of words. Then a test sheet containing the same six similar word pairs was used to examine the extent of students' word understanding. Note that during the test, no tool supports were provided. The student scored one point if he gave the correct answers to each question. In other words he would be awarded 6 points (the highest number of points) if he provided all the correct answers. They also completed a questionnaire, described below, evaluating the system.

4.2 Experimental Results

To determine the effectiveness of visualization of lexical phenomena, the students' average scores were used as performance indicators. Students achieved the average score 61.9 and 45.0 out of 100.00 after consulting the graphic interface and textual interface respectively. Overall, the visualized display of word characteristics outperformed the textual version.

The questionnaire revealed that all the participants showed a positive attitude to visualized word information. Further analyses showed that all ten participants appreciated the position display and nine of them the polarity and form displays. In short, the graphical display of lexical phenomena in GLANCE results in faster assimilation and understanding of word information. Moreover, the participants suggested several interesting aspects for improving the GLANCE system. For example, they preferred bilingual environment, further information concerning antonyms, more example sentences, and increased

detail in the sunburst representation of disciplines.

5 Conclusion and Future Work

In this paper, we proposed GLANCE, a text visualization tool, which provides graphical display of corpus data. Our goal is to assist language learners in glancing at the graphics and grasping the lexical knowledge quickly and intuitively. To evaluate the efficiency and effectiveness of GLANCE, we conducted a preliminary user study with ten non-native ESL learners. The results revealed that visualization format outperformed plain text format.

Many avenues exist for future research and improvement. We attempt to expand the single word to phrase level. For example, the collocation behaviors are expected to be deduced and displayed. Moreover, we are interested in supporting more lexical phenomena, such as hyponyms, to provide learners with more lexical relations of the word with other words.

Reference

Baccianella, S., Esuli, A., & Sebastiani, F. (2010, May). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *LREC* (Vol. 10, pp. 2200-2204).

Bostock, M., Ogievetsky, V., & Heer, J. (2011). D^3 data-driven documents. *Visualization and Computer Graphics, IEEE Transactions on*, *17*(12), 2301-2309.

Chen, M. H., Huang, S. T., Huang, C. C., Liou, H. C., & Chang, J. S. (2012, June). PREFER: using a graph-based approach to generate paraphrases for language learning. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP* (pp. 80-85). Association for Computational Linguistics.

Cobb, T. (2007). The compleat lexical tutor. *Retrieved September*, *22*, 2009.

Heer, J., Card, S. K., & Landay, J. A. (2005, April). Prefuse: a toolkit for interactive information visualization. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 421-430). ACM.

Huang, C. C., Chen, M. H., Huang, S. T., Liou, H. C., & Chang, J. S. (2011, June). GRASP: grammar- and syntax-based pattern-finder in CALL. In *Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 96-104). Association for Computational Linguistics.

Kyosung Koo (2006). Effects of using corpora and online reference tools on foreign language writing: a study of Korean learners of English as a second language. PhD. dissertation, *University of Iowa*.

Lee, D. Y. (2001). Genres, registers, text types, domains and styles: Clarifying the concepts and nevigating a path through the BNC jungle.

Partington, A. (1998). *Patterns and meanings: Using corpora for English language research and teaching* (Vol. 2). John Benjamins Publishing.

Riehmann, P., Gruendl, H., Froehlich, B., Potthast, M., Trenkmann, M., & Stein, B. (2011, March). The NETSPEAK WORDGRAPH: Visualizing keywords in context. In *Pacific Visualization Symposium (PacificVis), 2011 IEEE* (pp. 123-130). IEEE.

Scott, M. (2004). WordSmith tools version 4.

The British National Corpus, version 3 (BNC XML Edition). 2007. *Distributed by Oxford University Computing Services on behalf of the BNC Consortium*. URL: http://www.natcorp.ox.ac.uk/

ThinkMap Inc. (2005). Thinkmap Visual Thesaurus. Available from http:// www.visualthesaurus.com

Tsuruoka, Y., Tateishi, Y., Kim, J. D., Ohta, T., McNaught, J., Ananiadou, S., & Tsujii, J. I. (2005). Developing a robust part-of-speech tagger for biomedical text. In *Advances in informatics* (pp. 382-392). Springer Berlin Heidelberg.

Wattenberg, M., & Viégas, F. B. (2008). The word tree, an interactive visual concordance. *Visualization and Computer Graphics, IEEE Transactions on*, *14*(6), 1221-1228.

Wood, J. (2001). Can software support children's vocabulary development.*Language Learning & Technology*, *5*(1), 166-201.

Yeh, Y., Liou, H. C., & Li, Y. H. (2007). Online synonym materials and concordancing for EFL college writing. Computer Assisted Language Learning, *20*(2), 131-152.