Behav Res
DOI 10.3758/s13428-010-0042-z

Contextual self-organizing map: software for constructing semantic representations

Xiaowei Zhao · Ping Li · Teuvo Kohonen

Abstract In this article, we introduce a software package that applies a corpus-based algorithm to derive semantic representations of words. The algorithm relies on analyses of contextual information extracted from a text corpus—specifically, analyses of word co-occurrences in a large-scale electronic database of text. Here, a target word is represented as the combination of the average of all words preceding the target and all words following it in a text corpus. The semantic representation of the target words can be further processed by a self-organizing map (SOM; Kohonen, *Self-organizing maps,* 2001), an unsupervised neural network model that provides efficient data extraction and representation. Due to its topography-preserving features, the SOM projects the statistical structure of the context onto a 2-D space, such that words with similar meanings cluster together, forming groups that correspond to lexically meaningful categories. Such a representation system has its applications in a variety of contexts, including computational modeling of language acquisition and processing. In this report, we present specific examples from two languages (English and Chinese) to demonstrate how the method is applied to extract the semantic representations of words.

Keywords Distributed semantic representation · Sematic vectors · Corpus analysis · Contextual self-organizing map

Language scientists have long debated how to faithfully represent the complex semantic relationships of words in one or multiple languages. Connectionist researchers have also been concerned with how to represent semantics accurately in their models. The development of formal mechanisms to capture semantics faithfully and accurately will thus not only lend insights into the nature of the lexical system of natural languages, but will also have significant implications for understanding the nature of the mental representation of meaning (the *mental lexicon*; Bonin, 2004) and its processing and acquisition (see Li, Burgess, & Lund, 2000).

During the early days of connectionist modeling of language, different lexical attributes of words were often represented in the so-called *localist* fashion. That is, a node randomly picked by the modeler in the target lexical pool was assigned a numerical value to represent the meaning of a word (or other linguistic aspects of the word, such as sound). In this fashion, the activation of a node could be unambiguously associated with the meaning of a unique word that the node was supposed to represent, and the strength of the activation could be taken as the indicator of how well the concept was represented (Plunkett & Elman, 1997).

In the last two decades, significant progress has been made in deriving semantic representations of words in a *distributed* fashion. In contrast to localist models, where there is one-to-one mapping between concepts and units, distributed representations rely on a global pattern of activations across a common set of units (all dimensions of a vector). Different activation patterns represent different

Electronic supplementary material The online version of this article (doi:10.3758/s13428-010-0042-z) contains supplementary material, which is available to authorized users.

X. Zhao (✉)
Department of Psychology, Emmanuel College,
400 The Fenway,
Boston, MA 02115, USA
e-mail: xiaoweizhao@gmail.com

P. Li
Pennsylvania State University,
University Park, PA, USA

T. Kohonen
Aalto University,
Aalto, Finland

Published online: 24 November 2010

words, and overlap among these distributional patterns is permitted. The larger the overlap, the more similar the words are in the representation. Researchers have developed several methods to derive distributed semantic representations of words, and these methods can be roughly classified into two groups: feature-based representation and corpus-based representation (Riordan & Jones, 2010).

In a typical feature-based model, a word's meaning is represented by a vector, and each dimension of this vector represents a possible descriptive feature/attribute of the concept. The value of each dimension could be 0 or 1, indicating the absence (0) or presence (1) of a particular feature for the target word. For example, the representations of *dove* and *hen* are very similar, except for one dimension representing the flying feature (see Ritter & Kohonen, 1989, for a detailed representation of 16 animals based on 13 attributes). In this type of model, empirical data are often used to help generate the features describing the meanings of words. For example, participants in a study by Li and MacWhinney (1996) were asked to evaluate whether particular features could be applied to 160 English verbs; participants in a study by McRae, Cree, Seidenberg, and McNorgan (2005) were instructed to generate features associated with 541 concrete English nouns, and then a norm with more than 2,500 dimensions based on the generated features was constructed.

Corpus-based distributed models, in contrast, build semantic representations of a word through co-occurrence statistics in large-scale linguistic corpora. The underlying hypothesis is that two words should have similar meanings or belong to similar lexical categories if they often occur in similar contexts. The idea that linguistic context determines word meaning has been championed by linguists since de Saussure (1916/1977) and has led to fruitful explorations in child language research (Maratsos & Chalkley, 1980) and computational modeling (Elman, 1990). In the last decade or so, there have been several corpus-based distributed representation models, including the Hyperspace Analogue to Language (HAL; Burgess & Lund, 1997), Latent Semantic Analysis (LSA; Landauer & Dumais, 1997), the Word Co-occurrence Detector (WCD; Li & Farkas, 2002), Bound Encoding of the Aggregate Language Environment (BEAGLE; Jones & Mewhort, 2007), and TOPICS (Griffiths, Steyvers, & Tenenbaum, 2007). Some of the models rely on calculating the word–word co-occurrence frequencies (e.g., HAL, WCD, and BEAGLE), while others rely on calculating the co-occurrence matrix of target words with their surrounding context (e.g., sentences, paragraphs, or essays, as in LSA). The resulting representation of a target word in each of these models is usually a high-dimensional vector with each dimension denoting a linguistic entity (word or passage). The value of a dimension is often determined by a function of the co-occurrence frequency of the target word with the linguistic context. Here, the representation vectors can be thought of as points in a high-dimensional hyperspace, and the number of dimensions usually increases as the corpus size increases. Often, dimension reduction methods are used to make the computation more tractable (e.g., singular value decomposition in LSA [Landauer & Dumais, 1997] or random mapping in WCD [Li & Farkas, 2002]).

Comparing different approaches

The various approaches discussed above each have their own advantages and limitations. With a one-to-one mapping between concepts and units, the localist representation clearly has simplicity and efficiency and has brought great computational success for simulating language processing. However, the one-node–one-concept representation is subject to the criticism that it lacks linguistic and psychological reality (Jacquet & French, 2002). For example, one cannot simulate similarity-based semantic priming effects with the localist representation, given that similarities among concepts are not encoded in the representation when concepts have been assigned random values in the model.

In contrast to the localist representation, an obvious advantage of the feature-based distributed representation is its ability to incorporate real-world referents and perceptual cues in human experience. However, this method also has its limitations. First, it is relatively subjective, given that the investigators often need to hand pick a list of features that can be associated with the words solely on the basis of their own experience. Second, it does not work very well with abstract and closed-class words, since the descriptive features for these words are often hard to define and evaluate. Finally, given that each word in the lexicon needs to be evaluated by several participants, this approach is obviously time-consuming and resource-demanding, and it cannot scale easily to very large lexicons.

The most salient advantage of distributed representations based on corpus analysis is that they can be computationally derived from a text corpus on a very large scale, and automatically, without human intervention. Indeed, the method overcomes each of the three disadvantages mentioned above for feature-based representations. As was noted in the previous section, several corpus-based representation models have been developed. These models have been applied to various contexts, including the study of priming effects in both monolingual (Landauer & Dumais, 1997) and bilingual (Zhao & Li, 2009a) language processing and the study of vocabulary growth in children's language acquisition (Li, Zhao, & MacWhinney, 2007).

However, despite significant progress made with this approach, the availability of these different representations to

the larger research community is not yet within easy reach. Often, investigators want to apply these methods on their own corpus to meet the particular goals of their research. For example, researchers studying language development may want to investigate developmental changes in semantic representations based on their own collection of children's speech across different ages. Other linguists may want to construct semantic representations of words in a particular language that has not been extensively studied. In these contexts, researchers may not have the necessary formal training in computer programming to transform their needs into executable codes. Indeed, one of the authors of this report frequently receives requests from researchers regarding use of the semantic vectors developed by our team or of the program code used to derive the vectors.

Therefore, the purpose of this article is to provide researchers with a powerful and easy-to-use software package to efficiently derive distributed semantic representations based on corpus analysis. Specifically, we have used a computational algorithm called the *contextual self-organizing map* to derive the distributed representational vectors. Here, we illustrate use of the software with examples from both English and Chinese and show that the method can capture representations of thousands of common words in the two languages with a high degree of accuracy for word meanings and lexical categories. By applying the algorithm to two languages, we demonstrate the generalizability of the program. On the one hand, English is the language on which most previous work on semantic representations has been based, and many English corpora are available to investigators. On the other hand, Chinese represents a popular language significantly different from English and other Western languages (e.g., Chinese does not have inflectional endings like *-ing* or *-ed* in English) but is less studied in the area of semantic representations. By illustrating our program with these two widely used languages, we seek to demonstrate to researchers the usefulness of this program and hope that they will be able to use it for other languages of interest with little or no change of the code.

Our package also includes an optional function to display the derived semantic representations on a 2-D map (see the Method section). Indeed, visualizing semantic relations among words or concepts has been useful in psycholinguistics since the introduction of the hierarchical network model (Collins & Quillian, 1969) and the spreading activation theory (Collins & Loftus, 1975). A recent development in this direction was to project the semantic associations of tens of thousands of words onto networks with small-world structures (Steyvers & Tenenbaum, 2005). We hope that this visualization procedure can help researchers check the validity and accuracy of the representations derived from our software.

Below we discuss the contextual self-organizing map package (CTM_PAK) and the procedure by which researchers can download and use it.

Method

The algorithm

The contextual self-organizing map is a distributed representational model of word meaning developed by Ritter and Kohonen (1989). It is largely a corpus-based algorithm relying on statistical regularities of word–word co-occurrences, but it also has the potential to include feature-based components in the representation. The original algorithm was based on artificially generated three-word sentences randomly formed from a list of nouns, verbs, and adverbs (e.g., *Dog drinks fast*). However, it can easily be applied to large-scale corpora consisting of real human conversations, as discussed below.

Once a corpus has been selected as the basis for deriving semantic representations, it needs to be preprocessed for further analysis. First, the corpus should be digitized and transcribed into numerical indices. Particularly, depending on the level of detail in the representation, only words with certain frequencies of occurrence in the target corpus will be assigned index numbers, and all other words will be treated as noise in the construction of semantic representations. This is because words with very low frequencies of occurrence may not be adequate or important in the representation of the target words. It is also worth noting that an extra step of preprocessing is needed for certain languages (e.g., Chinese) in which written words are not delimited by spaces, as in English or many other alphabetical languages. In such "nondelimiting" languages, it is difficult to automatically judge the correct boundaries of words, although good programs do exist to aid researchers in this regard—for example, the ICTCLAS (http://ictclas.org/), an automatic word tagging and classification system for Chinese. The basic idea behind this process is to first manually segment a small but representative text into words (based on a well-constructed dictionary) and then to use the statistical rules derived from this well-segmented text as the "guidelines" for word segmentation in other novel data. In the present study, the Chinese corpora that we use are already lexically segmented.

After preprocessing, each word in the target lexicon is assigned a vector r_i randomly drawn from a high-dimensional space,[1] where i is the numerical index of a

[1] The vector can also be manually generated based on empirical data. In this way, the virtues of both feature-based and corpus-based representations can be combined in the final semantic representations, as has been done by Li, Farkas, and MacWhinney (2004) and Li et al. (2007).

word. The vectors are then normalized into unit length. The purpose of this process is to guarantee the isotropy and independence of the vectors in the constructed hyperspace by creating a spherically symmetric distribution of the vectors around the origin. As mathematically proven by Ritter and Kohonen (1989), vector isotropy and independence are crucial to the representational accuracy of the algorithm and can be obtained by assigning the vectors random numbers following Gaussian distributions with a mean of 0 and standard deviation 1. Practically, a simpler function with random numbers uniformly distributed between –1 and 1 can often yield similar results. We provide a switch in our package so that investigators can select either function they desire, although the results reported in the next section were derived based on the simpler, uniformly distributed vectors.

Investigators can also decide on the length (number of dimensions) of a random vector based on their needs; for simplicity, we set the default value in our program to be 100, a number we consider large enough to make semantically distinguishable representations in the constructed hyperspace while keeping the computation efficient and tractable.

The co-occurrence matrix for all target words is derived by calculating how many times a word directly follows or precedes another word in the corpus. Put differently, we can use a trigram window to go through the digitized corpus and count the co-occurrence frequency of each word in the middle of the trigram with its two closest neighbors. Based on the co-occurrence matrix, we can obtain each word's *average context vector* X_k, which is the combination of the average of all the words preceding the target word k and the average of all the words following it, as shown in (1):

$$X_k = [\mathrm{MEAN}\{r_{j(k)-1}\}, \mathrm{MEAN}\{r_{j(k)+1}\}]. \quad (1)$$

Here, $j(k)$ is a possible position of the target word k in the corpus, $r_{j(k)-1}$ or $r_{j(k)+1}$ is the random vector of the word occurring at the position $j(k) - 1$ (preceding the target word) or $j(k) + 1$ (following the target word). The function MEAN{} computes the average across all possible positions. Given that the random vectors have 100 dimensions in our program, X_k is a 200-dimension vector, and it should be further normalized to unit length. The result of this procedure is the semantic representation of the target word, which is then ready to be used by investigators.

The semantic representations derived as just described can be further input to a self-organizing map (SOM; Kohonen, 2001), an unsupervised neural network model that provides efficient data extraction and representation methods. A SOM usually consists of a 2-D topographic map for the organization of input representations, where each node is a unit on the map that receives input via the input-to-map connections. The SOM algorithm starts out by identifying all the incoming connection weights to each and every unit on the map, and for each unit, compares the combination of weights (weight vector) with the combination of values in the input pattern (input vector). If the unit's weight vector and the input vector are similar or identical by chance, the unit will receive the highest activation and is declared the winner (the best-matching unit, or BMU). Once a unit becomes highly active for a given input, its weight vector and those of its neighboring units are adjusted such that they become more similar to the input, and hence will respond to the same or similar inputs more strongly the next time. This process continues until all of the input patterns elicit specific response units in the map. With the help of this topography- preserving feature, the SOM is able to project the statistical structure implicit in the input onto a 2-D space, such that words with similar meanings cluster together, forming groups that correspond to lexically meaningful categories. The larger the difference between two input patterns, the larger the Euclidean distance will be on the SOM between the activated units corresponding to the two input patterns.

There are two main goals of using the SOM in the analysis process. First, by examining the map, investigators can check the validity and reliability of the derived representations, particularly if the representations have captured the semantic similarities and differences among words. Second, the resulting semantic map can serve as a medium for building more complex models of language processing (e.g., Li et al., 2007; Mayor & Plunkett, 2010; Miikkulainen, 1997; Zhao & Li, 2007). After proper training, the distributional representation of a word can usually be projected onto a unique unit on the map (the BMU), and as such, the semantic map provides a robust system for representing the dynamic properties of concepts and categories, combining the virtues of both localist and distributed representations. In other words, nodes on the map generally retain one-node–one-concept representations, allowing investigators to map them to items in the mental lexicon; at the same time, the nodes are grouped together based on their lexical similarities, allowing for the simulation of behavioral patterns such as semantic priming between words.

The software package

Our contextual self-organizing map package is available for researchers to download at http://sites.google.com/site/xiaoweizhao/tools or http://cogsci.psu.edu/, or from the Psychonomic Society supplemental archive. Once downloaded, the file CTM_PACK.zip should be unpacked into a working directory for it to run (a recommendation from the

authors is C:\CTM_pack\). The program code is written in MATLAB (The MathWorks, Inc., 2009) and runs under the MATLAB environment. The Readme.txt file provides a detailed description on the use of the program.

There are three main functional modules of the package, each of which is handled by an executable file under the MATLAB environment.

First, **Cooccur_.m** is the lexical co-occurrence detector that learns to approximate transitional probabilities between neighboring words in the text. The co-occurrence matrix of target words is saved in an output file and ready to be used for further construction of semantic representations. In addition, an optional HAL-based representation (Burgess & Lund, 1997) can be generated by this module using the derived co-occurrence matrix.[2]

Second, **Contextual_.m** is the core program of CTM_PAK, which reads through the co-occurrence matrix and generates semantic representations based on the contextual map algorithm described earlier. To increase the computational flexibility and usability of the package, the program can save the output in two different formats, the default MATLAB format and the ASCII text format, so that the output can be easily imported to other neural network packages such as SOM_PAK (www.cis.hut.fi/), DISLEX (http://nn.cs.utexas.edu/?dislex), or Emergent (http://grey.colorado.edu/emergent/).

Third, **Batch_context.m,** an optional routine, can be used to test the performance of the derived semantic representation.[3] To run this optional function, one needs to further install an additional program called SOM_Toolbox, which is a MATLAB-based open-source software package for SOM and can be obtained from www.cis.hut.fi/projects/somtoolbox/ (Vesanto, Himberg, Alhoniemi, & Parhankangas, 2000). The optional routine constructs a SOM network based on batched training (Kohonen, 2001), and the semantic representation of each target word is sent to the network as an input. After training, each word's BMU on the map is labeled, and the labeled words can show the lexically meaningful categories. A U-matrix figure (Ultsch & Siemon, 1990) can be drawn to display the boundaries among different categories.

Finally, a step-by-step demonstration (CTM_demo.m) is provided in the package to facilitate researchers' understanding of the package.[4] To run this demo, simply type "CTM_demo" in the MATLAB environment.

Results

We tested the CTM_PAK on three text corpora. First, we derived the semantic representations of English words from a small corpus based on the book of *Grimm's Fairy Tales*, downloaded from the website of the Gutenberg Project at www.gutenberg.org (a demo of this example is also provided in the package under the folder \example_E\). Second, we tested our program on two other Chinese corpora that have been widely used in the Chinese language research community. One is the Beijing Normal University Corpus (henceforth the BNU corpus; see Shu, Chen, Anderson, Wu, & Xuan, 2003), an electronic database that contains all of the text from the elementary school textbooks used in Beijing. The other is the MCRC corpus (Modern Chinese Research Corpus; Sun, Sun, Huang, Li, & Xing, 1996), which is an electronic collection of text material from newspapers, novels, magazines, TV shows, folktales, and other text materials from modern Chinese media. Both corpora are lexically segmented (note that Chinese texts are not word-delimited; see the Method section).

English corpus

The corpus of *Grimm's Fairy Tales* includes 104,724 word tokens with 5,387 unique word types. We selected this corpus because a typical school-aged English-speaking child should know most of the words in it. Although the size of the corpus is relatively small, we can still get quite accurate semantic representations of English words, as discussed below.

In this example, we first chose 2,443 of the 5,387 different word types as the basis to construct the word co-occurrence matrix, and subsequently the semantic representations. These 2,443 words occur more than twice in the corpus; words with an occurrence frequency less than or equal to two were treated as noise. After the semantic representations of these words were constructed according to the procedure described in the Method section, they were further input to a SOM for verification of lexical similarities among the words. For purposes of brevity and legibility, only the most common 300 words in the corpus were

[2] Users can set up a few parameters in the corresponding configuration file for each function so that they can fine-tune the output—for example, size of the target lexicon, window size for co-occurrence counting, vector normalization on or off, and optional outputs such as HAL-based representations.

[3] Some other multivariate statistical procedures can also be applied to test the validity and accuracy of the representations, such as principal components analysis (PCA), multidimensional scaling (MDS), or hierarchical cluster analysis.

[4] Here we provide an index file for an illustrative sample of a Chinese corpus (BNUall.ind) and a lexicon file with the target word types listed in descending order of their occurrence frequency in the sample corpus (BNUall.lex.txt). Users should prepare their own index file and lexicon file if they use other corpora.

displayed on the SOM, as shown in Fig. 1. Among the 300 words in the target lexicon were 67 nouns, 92 verbs, 34 adjectives, 9 quantifiers, and 98 closed-class words (including 36 pronouns, 24 prepositions, 3 articles, and some other words). The SOM had 3,000 nodes (50 × 60) and rectangular neighborhoods with an initial radius parameter of 25, which gradually decreased to 1 as training progressed. The map was trained for 200 epochs (i.e., each of the 300 target words was presented to the network 200 times). The resulting distribution of the target words on the map is shown on Fig. 1.

As can be seen here, the network captures the semantic similarities among the trained words, and lexically meaningful categories have emerged on the map as a result of the training. For instance, most of the verbs are clustered on the left side of the map, while the nouns are grouped in the bottom-right part. The pronouns are grouped together in the top center of the map, and the articles are close to them. Pronouns and articles, along with prepositions in the top-right corner, form a big area of closed-class words on the top of the map (see also Honkela, Pulkki, & Kohonen, 1995, for similar results).

It is worth noting that our program can further capture fine-grained features of the target language. For example, in English, verbs can be morphologically marked to indicate tense, aspect, mood, or voice, sometimes with the help of auxiliary verbs. It is clear that our derived semantic representations of the verbs capture this feature—the verb area on the map can be roughly split into three subcategories: pure verb stems, auxiliary verbs, and the inflected past tenses or past participles.

BNU corpus

Our first training set in Chinese for the package, the BNU corpus, includes 126,000 word tokens with 11,233 unique word types. The size of this corpus is also relatively small, but it provides a good working platform for the study of semantic representations of Chinese words from a developmental perspective, given that the materials are clearly marked for elementary school children (grades 1–5) who learn Chinese writing (see Xing, Shu, & Li, 2004, for an application of the BNU corpus in a SOM-based model of Chinese children's character acquisition).

From this corpus, we chose 2,340 of the 11,233 different word types as the basis to construct the word co-occurrence matrix, and subsequently the semantic representations. These 2,340 words occur more than five times in the BNU corpus; words with an occurrence frequency less than or equal to 5 were treated as noise. Similar to the English example, only the most common 500 words in the corpus were displayed on a SOM, which had 3,000 nodes (50 × 60) and rectangular neighborhoods, with the radius parameter gradually decreasing from 25 to 1 during training. Among the 500 words in the target lexicon, there were 165 nouns, 132 verbs, 124 closed-class words, 31 adjectives, 29 classifiers, and 19 quantifiers. The map was trained for

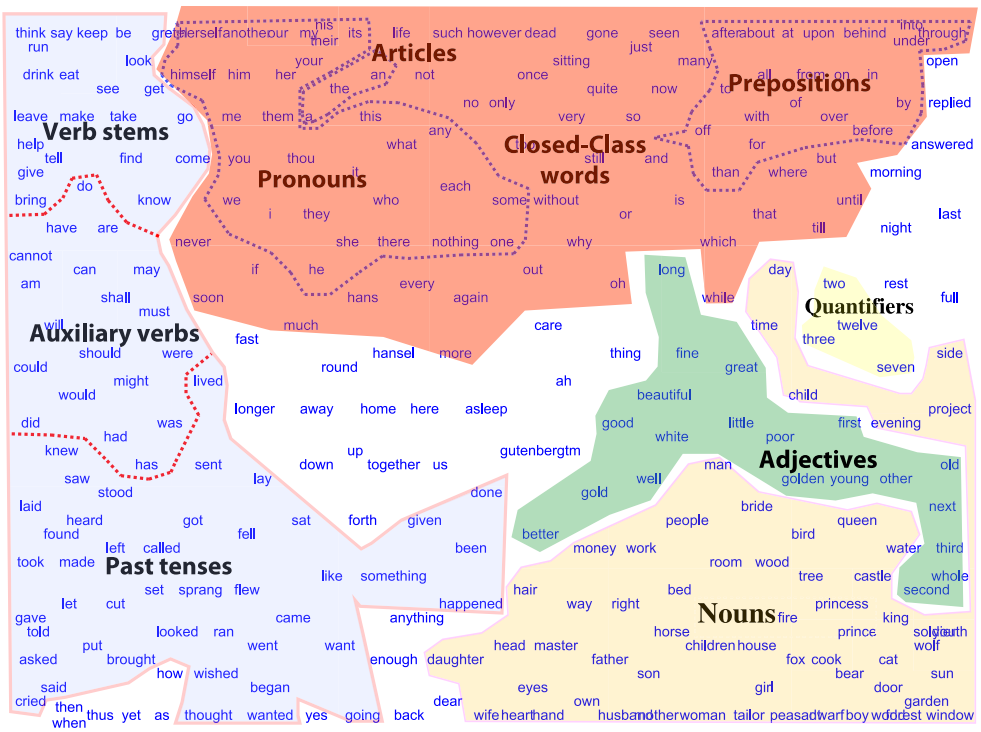

Fig. 1 Structured semantic representation in a self-organizing map (SOM) model after 200 epochs of training on the numerical representations of the 300 most common English words derived from a small-scale corpus (an edition of *Grimm's Fairy Tales*). The map size is 50 × 60, and the different lexical clusters are marked by different colors and shades

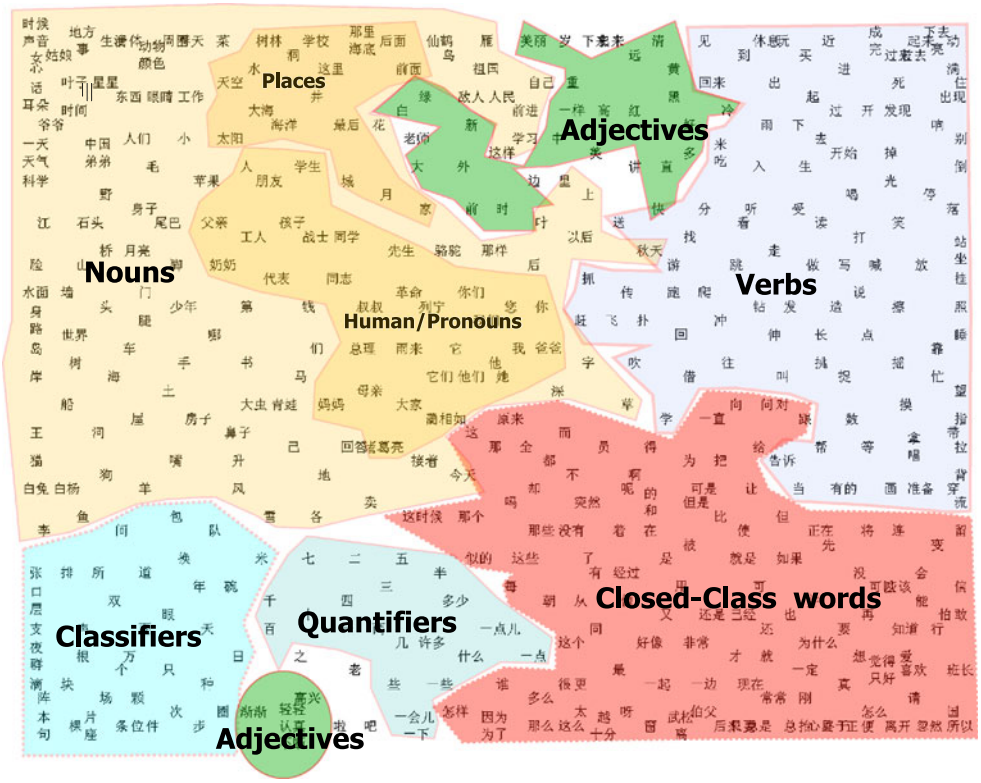

Fig. 2 Structured semantic representation in a self-organizing map (SOM) model after 200 epochs of training on the numerical representations of the 500 most common Chinese words derived from a small-scale corpus (the BNU corpus). The map size is 50 × 60, and the different lexical clusters are marked by different colors and shades

200 epochs, and the resulting distribution of the target words on the map is shown in Fig. 2.

Like the results from the previous English example, the network captures the semantic similarities among the trained words. For instance, most of the closed-class words are clustered at the bottom-right corner of the map, while the verbs and nouns are grouped in the top-right and top-left regions, respectively. A unique feature of the Chinese language is that it has nominal classifiers that categorize objects according to similarities along such dimensions as length, volume, shape, and orientation (Huang & Liao, 2001). Figure 2 shows that most classifiers are clustered at the bottom-left corner of the map, indicating that our method captures this important linguistic characteristic of Chinese. In addition, most quantifiers, which typically occur with classifiers in speech, can be found adjacent to the classifiers on the map. Upon further close examination, one can also see other lexical subcategories of the Chinese lexicon in the derived representations; for example, both nouns referring to people and pronouns are located in the center of the map, and nouns referring to places are grouped together.

To further test the validity and reliability of our method for representing the meaning of Chinese words, we used a *k*-nearest-neighbor (k-NN) classifier to examine the accuracy of semantic representations on the derived contextual map. k-NN is a quantitative method for classifying an object by assigning it to the class most common among its *k* closest neighbors in a feature space such as SOM (Duda, Hart, & Stork, 2000). In particular, we used a 5-NN classifier to examine if a word was clustered together with its peers belonging to the same lexical category. If the label assigned by the 5-NN classifier was consistent with the actual category of the word, we determined that the word was correctly classified. We then calculated the correct classification rates for each of the six lexical categories. High rates would imply that the words belonging to the same category tended to group together. To complete this process, we created five different versions of the semantic representations for each of the 500 words and fed these representations to five different SOMs with the same learning parameters discussed above. Figure 3 presents the means and standard deviations for the classification rates of the six lexical categories on the five maps.

Figure 3 shows that, except for the adjectives, all the other five lexical categories achieved high classification accuracy according to the k-NN. It is unclear, however, why the adjectives are outliers. Perhaps this is because the boundaries between adjectives and other lexical categories are often unclear in Chinese. In particular, unlike in Western languages, either an adjective or a verb in Chinese can serve the predicate role in the same sentence (i.e., distinct meanings map to the same grammatical form). The overall classification rates weighted by the number of words in each category reached an average of 72% (with a standard deviation of 1.2%), showing that 72% of the 500 target words tended to cluster together with their peers in

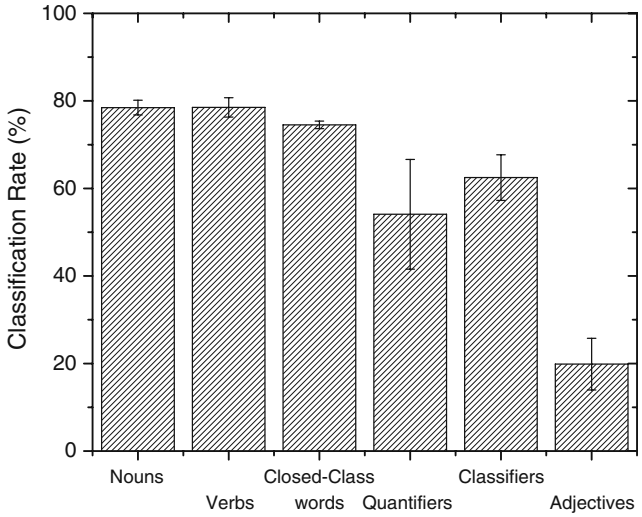

Fig. 3 Rates of correct classification of words in the six major lexical categories by a 5-NN classifier. Each data point was calculated based on five trials. Error bars indicate standard deviations

the same category on the trained maps. This quantitative result is consistent with our visual examination of the map discussed above (Fig. 2).

We also wanted to study whether the accuracy of semantic representations increases as the contextual map processes an increasing amount of text. We therefore conducted a separate experiment based on a portion of the BNU text (text for first graders). All of the training parameters were set to be identical to those of the experiment based on the entire corpus for all five grades. Due to the much smaller corpus size, there were only 550 words with occurrence frequencies greater than 1, and only 220 words with frequencies greater than 5. The 220 words included 72 nouns, 61 verbs, 46 closed-class words, 13 adjectives, 7 classifiers, and 3 quantifiers. A 5-NN classifier again was applied on five resulting maps to test the classification rates in these categories. The average overall classification rate was 46.5% (with a standard deviation of 3.9%), which indicates that only 46.5% of the 220 target words were grouped together with their peers in the same category. An independent-samples t test was conducted on the overall classification rates of the two groups of contextual maps (all grades vs. first grade), which indicated significantly better lexical clustering in semantic representations derived from the text for all grades rather than the text for first grade only (72% vs. 46.5%), $t(8) = 13.83$, $p < .001$. This analysis shows that as more information becomes available to our contextual map, the organization of the lexical categories in the derived semantic representations becomes more meaningful and transparent. Thus, the accuracy of representation clearly increases with the amount of data learned by the contextual map, which led us to test the contextual map on a larger database in Chinese, the MCRC corpus.

MCRC corpus

As mentioned earlier, the MCRC corpus is a database of modern Chinese, including material from newspapers, novels, magazines, TV shows, folktales, and so on, from modern Chinese media. As a corpus widely used in the Chinese language research community, MCRC can provide researchers with a representative sample of everyday language use in China. The total size of the corpus is about 1.5 million word tokens with 48,952 different word types, which is much larger than the previous two corpora. We wanted to see if this corpus could provide us with a basis to derive a more accurate semantic representations of more Chinese words, which may reveal finer semantic structures (e.g., people, vehicles, furniture, animals, etc.) under large lexical categories (e.g., nouns and verbs).

To conduct this experiment, we used 2,834 different word types as the basis for constructing the co-occurrence matrix; all of the word types had frequencies of occurrence of 50 or more within the corpus. In contrast to the experiment based on the BNU corpus, the derived semantic representations of all 2,834 words were fed to a SOM, which had 1,500 nodes (30 × 50) and rectangular neighborhoods, with the radius parameter gradually decreasing from 25 to 1 during training. Given the larger lexicon size, the map was trained for 400 epochs. Figure 4a presents the SOM's representation of the target words on the map.

Even for such a large lexicon with a size of 2,834 words, we can still see the emergence of the major lexical categories of Chinese words, as illustrated by areas with different colors on the map. As in the previous experiment with the BNU corpus, we again applied a 5-NN classifier on five resulting maps to test the classification rates over six large lexical categories. The average overall classification rate was 67.49% (with a standard deviation of 0.9%), showing that roughly 1,900 of the 2,834 target words were grouped together with their peers in the same lexical category. Similar to the results based on the BNU corpus from all grades, the results here show that nouns, verbs, classifiers, and closed-class words had good clustering rates (77.06%, 64.63%, 75.1%, and 68.67%, respectively), while the performance for quantifiers and adjectives was less ideal (43.5% and 34.5%, respectively). Considering the large size of the lexicon in question (and the increased likelihood of words being assigned to different lexical categories),[5] we can say that this overall rate of classification is highly satisfactory.

[5] We analyzed the most easily confused 419 words, which are words spanning two or more lexical categories in the corpus and serving in the syntactic role of the less dominant lexical category more than 10 times within the corpus. Our analysis showed that these words had a correct classification rate of 44.29% (SD = 0.9%), considerably below the average level.

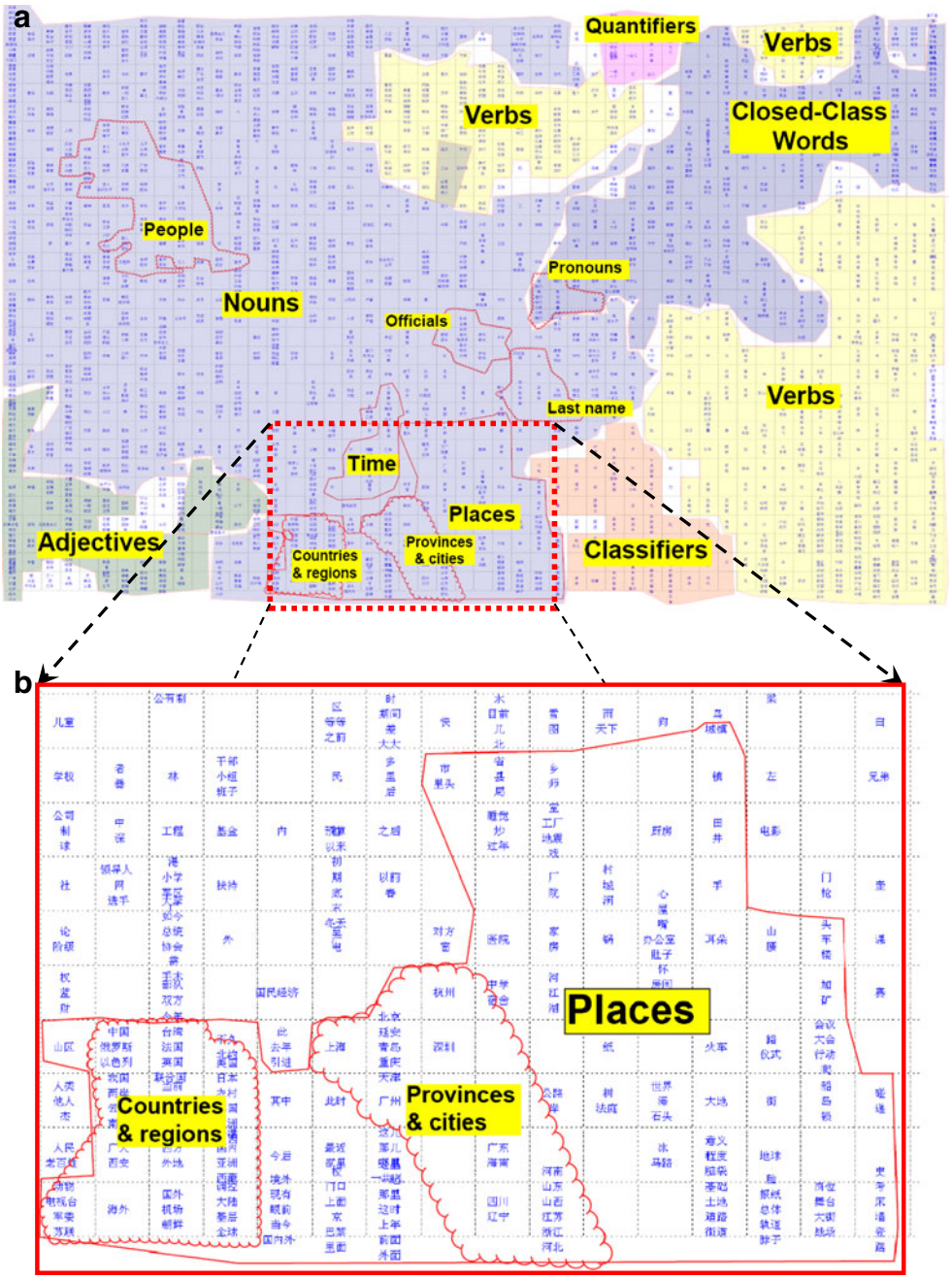

Fig. 4 Structured semantic representation in a self-organizing map (SOM) model after 400 epochs of training on the numerical representations of the 2,834 most common Chinese words derived from a large-scale corpus (the MCRC corpus). The map size is 30 × 50. (**a**) The entire map. (**b**) Expanded view of a portion of the map, demonstrating that CTM_PAK can capture the fine semantic structure of subcategories within larger lexical categories

More important, as predicted, our method could capture the fine-grained details of semantic structures in many smaller lexical subcategories using a larger corpus like MCRC. For example, the nouns related to places/locations are grouped together in the middle portion of the lower part on the map, close to words representing time concepts. Upon closer inspection of the nouns referring to places or locations (Fig. 4b), we can find even smaller subcategories. For example, the names of the provinces (such as *Szechwan*, *Canton*, etc.) and cities (such as *Beijing*, *Shanghai*, etc.) are grouped together, and the names of countries and geographical regions (such as *China*, *Russia*, *Israel*, *France*, *England*, *Asia*, *Europe*, etc.) are clustered together. Also, these specific names of places are located close together in the left part of the subarea, a bit farther from the general nouns referring to places (such as *city*, *village*, *town*, *county*, *province*, *hole*, *hill*, *street*, *road*, etc.), which are located in the right and top parts of the area. Similar fine lexical structure within categories can also be found elsewhere on the map—for example, proper names and titles of officials are shown in the middle center of the map.

Combining our analyses based on the MCRC and BNU corpora, we can say that as more information becomes available to our model, the organization of lexical categories becomes more structured in the derived semantic representations. Metaphorically, one can think of the scenario of capturing lexical semantic knowledge as filling in the holes in a Swiss cheese: Initially there may be more holes (less organized lexical knowledge) than cheese (clear representation of meaning), but the holes get filled up gradually, as more materials are incorporated and the co-occurrence context expands. This type of experience-based organization of lexical categories is consistent with the emergentist view in human cognitive development: Representational structures emerge as a result of the dynamic interaction between the learning system (the brain) and the environment (Li, 2009).

A second goal of our experiment using the MCRC corpus was to test the speed of our program package with deriving semantic representations from large-scale corpora with sizes comparable to real-world learning situations (i.e., on the scale of millions of word tokens). Using the entire MCRC corpus (1.5 million word tokens) with the most common 10,072 words indexed (frequency > 8 in the corpus), we ran our program several times to systematically test the amount of time needed for deriving representation vectors of target lexicons with sizes ranging from 500 to 10,072.[6] The experiments were conducted on a Dell desktop with 4 GB memory and an Intel Core2 Duo CPU (clock speed = 2.66 GHz). The results, based on 10 trials for each lexicon size, can be found in Fig. 5, which shows that our program can generate the relevant semantic representations in a matter of minutes. Even with a lexicon of 10,072 words, the average running time of our program was 424.28 s (just over 7 min), with a standard deviation of 1.15 s. This speed is very encouraging and shows that our contextual SOM package is computationally efficient and accessible to investigators who have only basic computational resources.

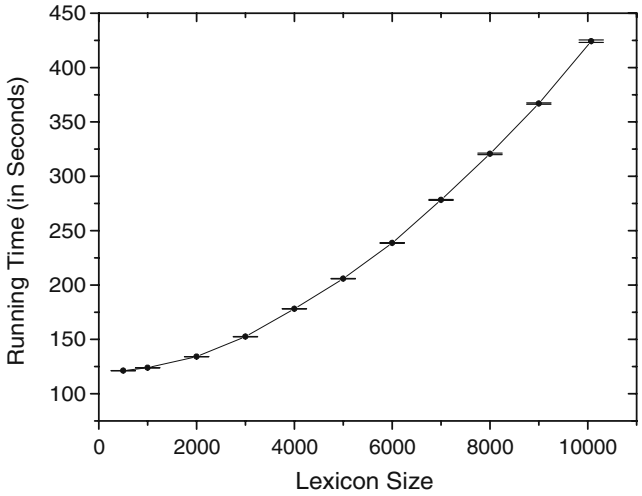

Fig. 5 Running time (in seconds) of CTM_PAK as a function of the size of the target lexicon, based on 1.5 million words from the MCRC corpus. The data for each lexicon size are based on 10 trials. Error bars indicate standard deviations

Conclusions

In this report, we introduced a fast and easy-to-use computational software package for deriving semantic representations, and we illustrated the software with electronic text corpora from English and Chinese. The representation system is based on the contextual self-organizing map algorithm, which relies on analyses of contextual information extracted from a text corpus—specifically, analyses of word co-occurrences in a computerized database. Our software captures the semantic properties and lexical structures of up to thousands of words with a high degree of accuracy and allows computational modelers to derive and use semantic representations in their models. The software package has a good degree of generalizability, allowing it to be used with various corpora with different scales and contents. It can derive quite accurate semantic representations, even with databases that are relatively small in size (e.g., *Grimm's Fairy Tales* and the BNU corpus), and it can also quickly derive semantic representations of thousands of words in large-scale corpora with millions of word tokens (e.g., the MCRC corpus).

As we have demonstrated, our software for semantic representations can be applied to quite different languages, such as English and Chinese. For researchers who want to study languages other than these two, they need make little or no change to the code, but need only have a good text corpus of the relevant language, digitized and preprocessed as described earlier in the Method section. Thus, our software can provide a convenient tool for linguists, psychologists, and other cognitive scientists who are interested in cross-linguistic and comparative studies of languages. For example, because the typical linguistic contexts for the use of so-called *translation equivalents* may be quite different across languages, it might be interesting to study whether some translation equivalents are represented differently for different languages in terms of our vectors. The contextual analyses our software generates may help researchers make some testable predictions on this issue. In addition, one could compare contextual information extracted from children's and adults' speech and identify whether vectors for the

[6] The time for training the SOM itself (conducted by the optional routine of Batch_context.m) was not included in the measurements, since that relies on the computational efficiency of the SOM_Toolbox.

same words in the same language differ between children and adults, and between children at different developmental stages.

We expect some further developments of the package in the near future, including the addition of a graphical user interface. Such refinements to the program should allow for better and easier access to the system for a large number of researchers. We hope that, along with other programs we have developed (e.g., PatPhon for phonological representations: Li & MacWhinney, 2002; Zhao & Li, 2009b), CTM_PAK can provide full-scale representations of the lexicon that serve as accurate linguistic inputs for computational language models.

Author Notes Preparation of this article was made possible by a grant from the National Science Foundation (BCS-0642586) to P.L. and by a faculty discretionary research grant from Colgate University to X.Z. during the 2009–2010 academic year. We thank Hua Shu and Jianfeng Yang for providing the BNU corpus, and Hongbing Xing for providing the MCRC corpus. X.Z. also thanks Zachary Helft, who assisted in the preparation of some of the figures.

References

Bonin, P. (Ed.). (2004). *Mental lexicon: Some words to talk about words*. Hauppauge, NY: Nova Science.

Burgess, C., & Lund, K. (1997). Modeling parsing constraints with high-dimensional context space. *Language and Cognitive Processes, 12*, 177–210. doi:10.1080/016909697386844

Collins, A., & Loftus, E. (1975). A spreading-activation theory of semantic processing. *Psychological Review, 82*, 407–428. doi:10.1037/0033-295X.82.6.407

Collins, A., & Quillian, M. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning & Verbal Behavior, 8*, 240–247. doi:10.1016/S0022-5371(69)80069-1

De Saussure, F. (1977). *Course in general linguistics* [*Cours de linguistique générale*] (Ed. C. Bally & A. Sechehaye, Trans. W. Baskin). Glasgow: Fontana/Collins. (Original work published 1916)

Duda, R., Hart, P., & Stork, D. (2000). *Pattern classification* (2nd ed.). Wiley.

Elman, J. (1990). Finding structure in time. *Cognitive Science, 14*, 179–211.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review, 114*, 211–244. doi:10.1037/0033-295X.114.2.211

Honkela, T., Pulkki, V., & Kohonen, T. (1995). Contextual relations of words in Grimm tales, analyzed by self-organizing maps. In F. Fogelman-Soulié & P. Gallinari (Eds.), *Proceedings of International Conference on Artificial Neural Networks, ICANN '95, vol II* (pp. 3–7). Nanterre, France: EC2.

Huang, B., & Liao, X. (Eds.). (2001). *Modern Chinese* (3rd ed.). Beijing: Higher Education Press.

Jacquet, M., & French, R. M. (2002). The BIA++: Extending the BIA+to a dynamical distributed connectionist framework. *Bilingualism, 5*, 202–205. doi:10.1017/S1366728902223019

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review, 114*, 1–37. doi:10.1037/0033-295X.114.1.1

Kohonen, T. (2001). *Self-organizing maps* (3rd ed.). Berlin: Springer.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211–240. doi:10.1037/0033-295X.104.2.211

Li, P. (2009). Lexical organization and competition in first and second languages: Computational and neural mechanisms. *Cognitive Science, 33*, 629–664.

Li, P., Burgess, C., & Lund, K. (2000). The acquisition of word meaning through global lexical co-occurrences. In E. V. Clark (Ed.), *Proceedings of the thirtieth annual child language research forum* (pp. 167–178). Stanford, CA: Center for the Study of Language and Information.

Li, P., & Farkas, I. (2002). A self-organizing connectionist model of bilingual processing. In R. Heredia & J. Altarriba (Eds.), *Bilingual sentence processing* (pp. 59–85). Amsterdam: Elsevier. doi:10.1016/S0166-4115(02)80006-1

Li, P., Farkas, I., & MacWhinney, B. (2004). Early lexical development in a self-organizing neural network. *Neural Networks, 17*, 1345–1362. doi:10.1016/j.neunet.2004.07.004

Li, P., & MacWhinney, B. (1996). Cryptotype, overgeneralization, and competition: A connectionist model of the learning of English reversive prefixes. *Connection Science, 8*, 3–30. doi:10.1080/095400996116938

Li, P., & MacWhinney, B. (2002). PatPho: A phonological pattern generator for neural networks. *Behavior Research Methods, Instruments, & Computers, 34*, 408–415.

Li, P., Zhao, X., & MacWhinney, B. (2007). Dynamic self-organization and early lexical development in children. *Cognitive Science, 31*, 581–612.

Maratsos, M., & Chalkley, M. (1980). The internal language of children's syntax. In K. E. Nelson (Ed.), *Children's language* (Vol. 2). New York: Gardner Press.

Mayor, J., & Plunkett, K. (2010). A neurocomputational account of taxonomic responding and fast mapping in early word learning. *Psychological Review, 117*, 1–31. doi:10.1037/a0018130

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods, 37*, 547–559.

Miikkulainen, R. (1997). Dyslexic and category-specific aphasic impairments in a self organizing feature map model of the lexicon. *Brain and Language, 59*, 334–366. doi:10.1006/brln.1997.1820

Plunkett, K., & Elman, J. (1997). *Exercises in rethinking innateness: A handbook for connectionist simulations*. Cambridge, MA: MIT Press.

Riordan, B., & Jones, M. (2010). Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science*. doi:10.1111/j.1756-8765.2010.01111.x

Ritter, H., & Kohonen, T. (1989). Self-organizing semantic maps. *Biological Cybernetics, 61*, 241–254.

Shu, H., Chen, X., Anderson, R., Wu, N., & Xuan, Y. (2003). Properties of school Chinese: Implications for learning to read. *Child Development, 74*, 27–47. doi:10.1111/1467-8624.00519

Steyvers, M., & Tenenbaum, J. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science, 29*, 41–78. doi:10.1207/s15516709cog2901_3

Sun, H. L., Sun, D. J., Huang, J. P., Li, D. J., & Xing, H. B. (1996). Corpus for modern Chinese research. In Z. S. Luo & Y. L. Yuan (Eds.), *Studies of the Chinese language and characters in the era of computers* (pp. 283–294). Beijing: Tsinghua University Press.

Ultsch, A., & Siemon, H. P. (1990). Kohonen's self-organizing feature maps for exploratory data analysis. In B. Angeniol & B. Widrow

(Eds.), *Proceedings of the International Neural Network Conference, INNC'90* (pp. 305–308). Dordrecht, The Netherlands: Kluwer.

Vesanto, J., Himberg, J., Alhoniemi, E., & Parhankangas, J. (2000). *SOM Toolbox for MATLAB 5* (Tech. Rep. A57). Helsinki University of Technology, Neural Networks Research Center.

Xing, H. B., Shu, H., & Li, P. (2004). The acquisition of Chinese characters: Corpus analyses and connectionist simulations. *Journal of Cognitive Science, 5*, 1–49.

Zhao, X., & Li, P. (2007). Bilingual lexical representation in a self-organizing neural network. In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th annual meeting of the cognitive science society* (pp. 755–760). Austin, TX: Cognitive Science Society.

Zhao, X., & Li, P. (2009a). *Cross-language priming in L1 and L2: A computational study.* Paper presented at the 39th Annual Meeting of the Society for Computers in Psychology, Boston, MA.

Zhao, X., & Li, P. (2009b). An online database of phonological representation for Mandarin Chinese monosyllables. *Behavior Research Methods, 41*, 575–583.