Computational Linguistics and Chinese Language Processing
Vol. 17, No. 2, June 2012, pp. 55-68

Using Kohonen Maps of Chinese Morphological Families to Visualize the Interplay of Morphology and Semantics in Chinese

Bruno GALMAR*

Abstract

A morphological family in Chinese is the set of compound words embedding a common morpheme, and Self-organizing maps (SOM) of these Chinese morphological families can be built. Computation of the unified-distance matrices for the SOMs allows us to perform semantic clustering of the members of the morphological families. Such semantic clustering sheds light on the interplay between morphology and semantics in Chinese. We studied how the word lists used in a lexical decision task (LDT) (Chen, Galmar, & Su, 2009) are mapped onto the clusters of the SOMs. We showed that this mapping is helpful to predict whether repetitive processing of members of a morphological family would elicit a satiation in an LDT - habituation - of both morphological and semantic units of the shared morpheme. In their LDT experiment, Chen, Galmar, and Su (2009) found evidence for morphological satiation but not for semantic satiation. Conclusions drawn from our computational experiments and calculations are in accordance with the behavioral experimental results in Chen *et al*. (2009). Finally, we showed that our work could be helpful to linguists in preparing adequate word lists for behavioral study of Chinese morphological families.

Keywords: Self-Organizing Maps, Computational Morphology and Semantics.

1. Introduction

In this paper, we call a morphological family the set of compound words embedding a common morpheme. Hence, the compound words in Table 1, which all contain the morpheme ‘明’ (míng) as a first character, belong to the morphological family of ‘明’.

* Institute of Education, National Cheng Kung University, Tainan, Taiwan
E-mail: hsuyueshan@gmail.com

Table 1.Some examples of the明 morphological family (Chen, Galmar, & Su, 2009).

明朝	明天	明白	明確	明星	明亮
Ming Dynasty	tomorrow	to understand / clear	explicit	star	bright

In Chinese, the meaning of a morpheme can be either transparent or opaque to the meaning of the compound word embedding it. For example, the common morpheme in Table 1 "明" can mean (*clear)* or (*bright)* and is transparent to the meaning of "明星" (*star*) but rather opaque to the meaning of "明天" (*tomorrow*). If some members of a morphological family are semantically similar, one could advance as a reason for such a similarity that these members are transparent to the same meaning of the shared morpheme. Most Chinese morphemes are polysemous (Chen & Chen, 2000). Hence, in theory, *transparent members* of a morphological family could belong to different semantic clusters whose centers would be the different meanings of the shared polysemous morpheme.

This paper is aimed primarily at using computational linguistics methods to perform semantic clustering of the members of the morphological families. Such a clustering is used to predict the results of a behavioral Lexical Decision Task[1] (LDT) designed by Chen, Galmar, & Su (2009) to study the phenomenon of morphological satiation in Chinese.

In visual word recognition, morphological satiation is an impairment of morphological processing induced by repetitive exposure to the same morpheme embedded in different Chinese compound words (Chen *et al*., 2009; Cheng & Lan, 2009). Chen, Galmar, and Su (2009) posited that morphological satiation is due to habituation of the morphological unit of the repeated morpheme. This is represented in Figure 1 by Diagram (a).

As a morpheme is thought to be a meaningful unit, it is logical to consider whether a semantic satiation (Kounios, Kotz, & Holcomb, 2000; Smith & Klein, 1990; Tian & Huber, 2010), an impairment of semantic processing causing a temporary loss of the meaning of the common morpheme, would occur concomitantly with morphological satiation.[2] In other words, the satiation observed by Chen *et al*. (2009) could have two loci: a morphological locus and a semantic locus, as represented in Figure 1 by Diagram (d).

A morphological satiation could also have its loci of satiation on the links between the morphological, lexical, and semantic units, as represented in Figure 1 by Diagrams (b) and (c). We quickly can rule out the possibility of a locus on the link between morphological and

[1] An LDT is a behavioral task for which subjects have to identify whether presented visual stimuli are words or non-words.

[2] If most of the members of a morphological family used in an experimental task are transparent to the same meaning of the shared morpheme, the same semantic units of the shared morpheme are repeatedly accessed and finally habituate - satiation diagram (d) -. Therefore, there could be a semantic satiation in addition to morphological satiation.

lexical units, as represented by Diagram (b). The reason is that, in a LDT, this link is changing at each presentation of a new two-character word. The morphological unit of the repeated morpheme constitutes one fixed endpoint of the morphological/lexical link but the over endpoint is always changing.

The present work of semantic clustering focuses on clarifying by computational means whether morphological satiation would probably have a sole morphological locus (Diagram (a)) or whether it would have both a morphological and semantic locus (Diagram (d)). The behavioral LDT experiment results in Chen *et al*. (2009) point to the existence of a sole morphological locus.

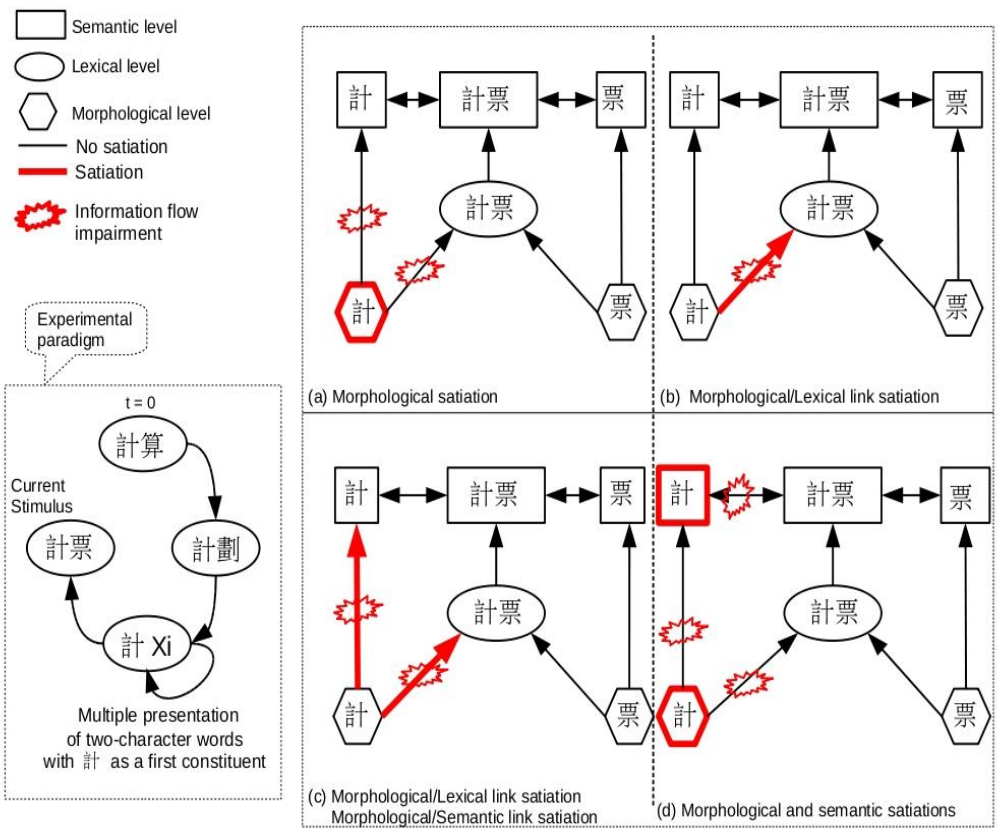

Figure 1. Different possible loci of satiation for morphological satiation.

2. Rationale of our Approach

As human subject agreement for semantic clustering tasks is low (Jorgensen, 1990), computational corpus-based semantic clustering was thought to be a valuable and complementary experimental approach compared to a behavioral one with human subjects.

A corpus of written texts is a human artifact, its content is relevant to the human reader and, from a cognitive psychology standpoint, a corpus does embed a subset of organized human semantic knowledge and is worthy to be studied in computer simulations as a pure abstract semantic memory stripped out of sensory and motor representations.

In natural language processing, proponents of the "bag of words" approach simplify each document internal structure to a set of words and use a whole corpus to build a matrix of co-occurrence of the words corpus (Landauer & Dumais, 1997). Computational methods, such as Latent Semantic Analysis (LSA), take as input such a high dimensional matrix and reduce its dimensionality to form a vector space of the documents and words (Landauer, McNamara, Dennis, & Kintsch, 2007). This space embeds only an associative kind of semantic information[3]: words that co-occur in the same documents or that have common co-occurrents are close associates.

For a news corpus, the association can often be of the situational type. For example, "Father Christmas" will be a close associate of "department store," as there are many news reports around Christmas about the bustling agitation in department stores full of "Father Christmas".[4] In cognitive science and AI, it is said that the two terms "Father Christmas" and "department store" belong to a common memory frame, a frame being defined by Minsky as "*a data-structure for representing a stereotyped situation*" (Minsky, 1974).

In the present work, we follow a "bag of words" approach by first building a term document matrix (TDM). Then, Self-Organizing Maps (SOMs) and associated unified-distanced matrices (called U-matrix thereafter) are built from the TDM. The SOMs and U-matrices serve to visualize semantic clusters in a morphological family on a 2D hexagonal grid of bins (Kohonen, 2001).

On the SOMs, a semantic cluster is made of members of a morphological family that have been fitted into the same bin of the grid and into contiguous bins which are close neighbors - according to the U-matrix information - in the original high dimensional space. SOMs have been used successfully to capture associative semantic relationships between words in corpora. Closer to the present approach, SOMs have been used to study the developmental aspect of vocabulary acquisition in Chinese (Li, 2001, 2009; Li, Farkas, &

[3] Semantic information also can be, for example, of the categorical or featural types.

[4] This example is borrowed from Galmar & Chen (2010b).

MacWhinney, 2004; Zhao & Li, 2008). Zhao, Li, and Kohonen (2010) studied the clustering of subsets of the most common Chinese words along both linguistic and semantic dimensions. Kohonen and Xing (2011) computed the SOMs of different linguistic classes for Chinese and studied how word frequency modulates the clusters on the SOMs. Our study is the first one to use SOMs to study the interplay between morphology and semantics in Chinese compounds words sharing a common morpheme, *i.e.* to study the semantics of morphological families. Previous studies on the semantics of morphological families (Galmar & Chen, 2010a; Galmar & Chen, 2010b) followed a supervised approach that relied upon etymological knowledge of Chinese morphemes to identify the meaning of a morpheme in a given compound word. The present work followed an unsupervised approach, and the goal is meaning discrimination through clustering rather than meaning identification. We chose the SOM algorithm to ensure that our work can be replicated thanks to the widespread availability of SOM packages and toolboxes and because it has not yet been applied to our specific research topic.

3. The Corpus and the Term Document Matrix (TDM)

3.1 The Academia Sinica Balanced Corpus

We used the Academia Sinica Balanced Corpus (ASBC), a five million word annotated corpus based on Chinese materials from Taiwan, mostly newspapers articles. The corpus is made of roughly 10000 documents of unequal length.

We removed the foreign alphabet words and most of the Chinese functional words from the corpus. We kept POS tag information to allow differentiation between different grammatical instances of the same word[5] (Galmar & Chen, 2010).

3.2 The Term Document Matrix (TDM)

The TDM was built using the *TermDocumentMatrix* function of the R package *tm* (Feinerer, 2008) with a self-customized Chinese tokenizer. The TDM is a 136570 terms * 9179 documents matrix.

The TDM was weighted:

1. Using the classical term frequency-inverse document frequency (TfIdf) weighting scheme for both local and global weighting of the terms in the TDM (Landauer & Dumais, 1997). We used the function *weightTfIdf* of the package *tm* (Feinerer, 2008).
2. Using a weighting scheme at the document level to reduce the effect of the size difference between documents:

[5] Some of the Chinese words can have up to 5 different POS tags [10].

$$\log_2\left(\frac{Max_{Document\_size}}{Document\_size}+1\right) \quad (1)$$

Each document of the TDM is a genuine article of the ASBC corpus and is considered as a semantic unit. More weight is given to small documents of the ASBC corpus. A complete justification for such a decision is given in Galmar & Chen (2010). Briefly, one can say that the gist of a news article is easier to extract from a very short article than from a very long one for a human reader due to attention capacity limitations.

4. The Self-Organizing Maps

For a given morphological family, the rows corresponding to the members of the family in the TDM were extracted. The extracted rows constitute a submatrix of the TDM. From this submatrix, an SOM is built using the *Batch map algorithm* (Kohonen, 2001). The U-matrix (Ultsch & Siemon, 1990) is computed to assess how close members fitted to contiguous bins (bins are hereafter called units) on the SOM are in the original high-dimensional space (hereafter called input data space).

4.1 The batch version of the SOM algorithm

As all of the data - the TDM - can be presented to the SOM algorithm from the beginning of learning, the batch version of the SOM algorithm (called "Batch Map") is used instead of the incremental learning SOM algorithm. The batch SOM is very similar to the k-means (Linde-Buzo-Gray) algorithm (Kohonen, 2001).

Our SOM defines a mapping from the input data space $\Re^n$ of observation samples onto a hexagonal two-dimensional grid of N_u units. Every unit *i* is associated with a *reference vector* $m_i \in \Re^n$. The set of units located inside a given radius from unit *i* is termed *neighborhood set* N_i.

The Batch Map algorithm can be described as follows (Kohonen, 2001, p. 139-140; Kohonen, Oja, Visa, & Kangas, 2002, p. 1360).

1. Initialize the N_u reference vectors by taking the first N_u observation samples.
2. For each unit i, collect a list L_i of copies of all those observation samples whose nearest reference vector belongs to N_i.
3. Update the value of each reference vector m_i with the mean over L_i.
4. Repeat from Step 2 a few times.

The Batch Map presents a main advantage over the incremental learning version of the SOM algorithm (Kohonen, 2001; Fort, Letremy, & Cottrell, 2002) in that no learning rate

parameter has to be specified. To double-check the computed batch SOM's representativeness of the input data space, we followed the recommendation of both Fort, Letremy, and Cottrell (2002) and Kohonen (2001) to compare organization in the Batch Map and in the incremental learning SOM.

We used the code in the R package *class* (Venables & Ripley, 2002) for the batch SOM to build the SOMs. For the morphological family 計 (jí), used as an illustrative example in Section 5, we built the SOMs on a 7*8 hexagonal grid of 56 bins. The size of the SOM was determined by experimental testing while ensuring a lower quantization error (Kohonen, 2001).

4.2 The Unified-Distance Matrix

We reused and modified the code in the R package *kohonen* (Wehrens & Buydens, 2007) to build the U-matrix for the Batch Map and to plot a grey-level map superimposed on the SOM map. The U-matrix is the distance matrix between the reference vectors of contiguous units. On the grayscale SOMs, contiguous units in a light shade on the SOM are representative of existing clusters in the input data space. Contiguous units in a dark shade draw boundaries between existing clusters in the input data space (Ultsch & Siemon, 1990).

5. Results

We present the results for the study of the 計 (jí) morphological family.[6] This Chinese morpheme has two main meanings: (1) to count, to calculate or (2) to plan, to scheme. The study was limited to the members in the ASBC corpus embedding 計 as a first character. The SOM map of these members is noted SOM_{93} and is shown in Figure 2.

At the first level, the map is divided in two zones: a dark shaded one – the upper part of the map - and a light shaded one. Most of the words belong to the light shaded zone. Among the diverse existing clusters, we note that:

- Cluster C_1 mainly gathers word sharing and other words related to the first meaning of 計.
- Cluster C_2 gathers three words related to the frame *taxi* in the same unit.
- Cluster C_3 includes many words belonging to two contiguous units in a light shade. We decided to recompute a Batch Map SOM for the members in these two units to zoom in and have a clearer map of these members. The map is shown in Figure 3.

[6] Others examples are also given in the script file – available upon request - to create and plot the SOMs presented in the present paper.

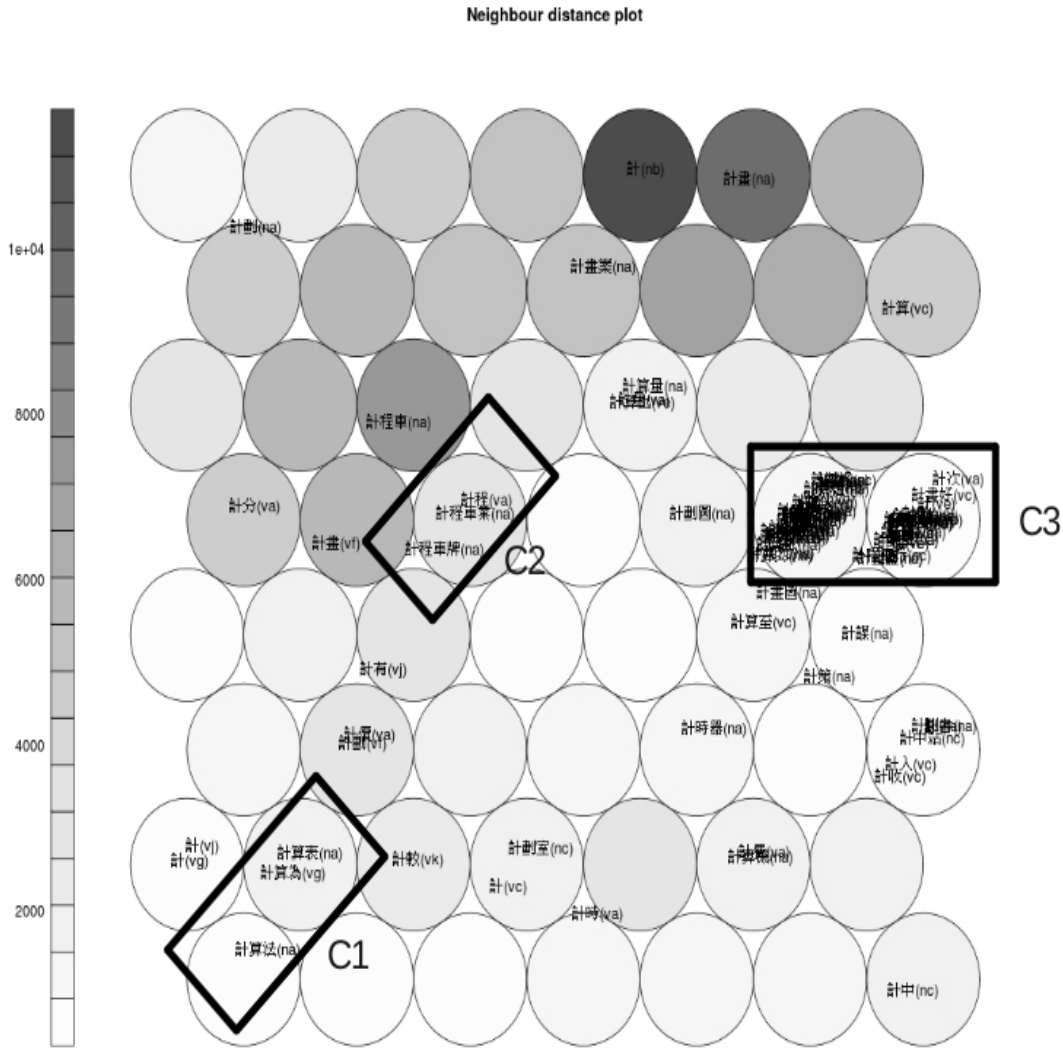

Figure 2. SOM$_{93}$ of the 計 (jì) morphological family.

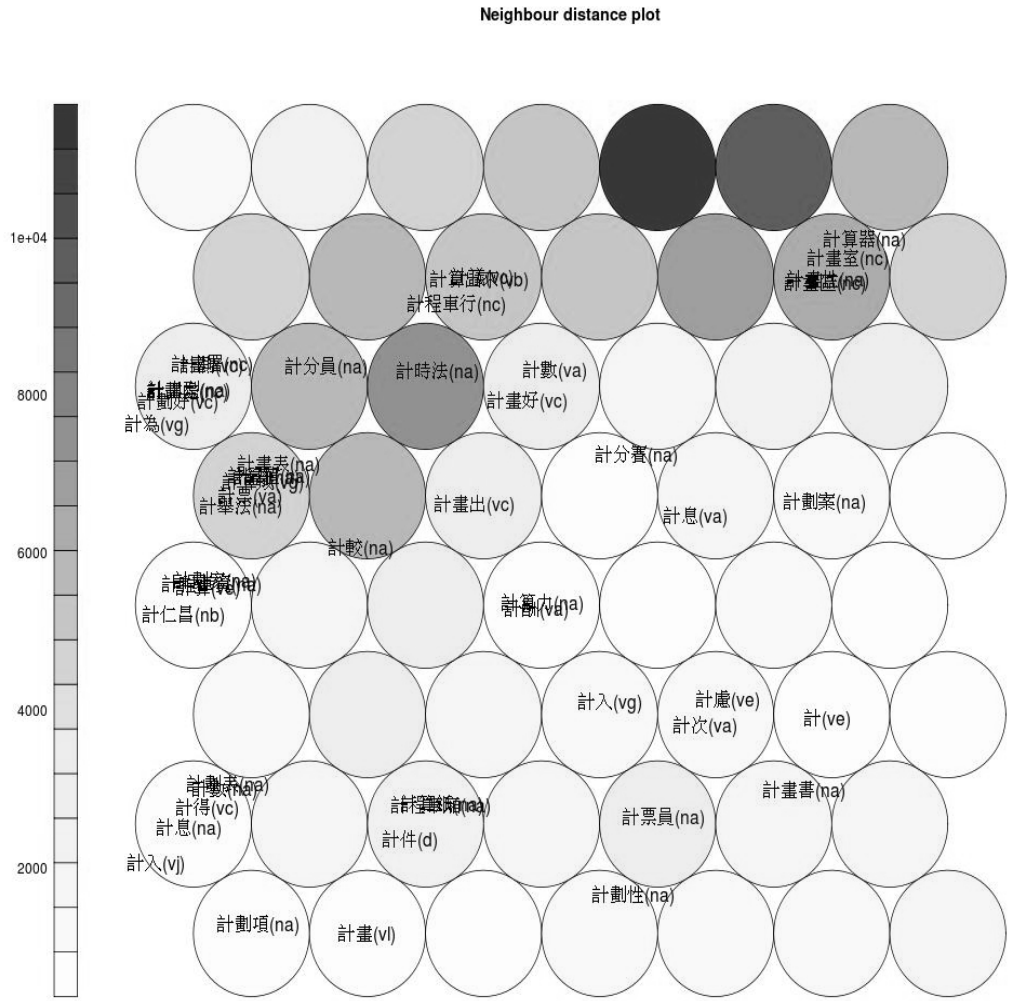

Figure 3. SOM for Cluster C3 in Figure 2.

Figure 4 shows only the 13 words used by Chen *et al.* (2009) in one block of their LDT experiment.[7] Some of the words have two POS tags, so the total number of data points represented in Figure 4 is 17.

[7] One word of the original experiment not existing in the ASBC corpus is missing here.

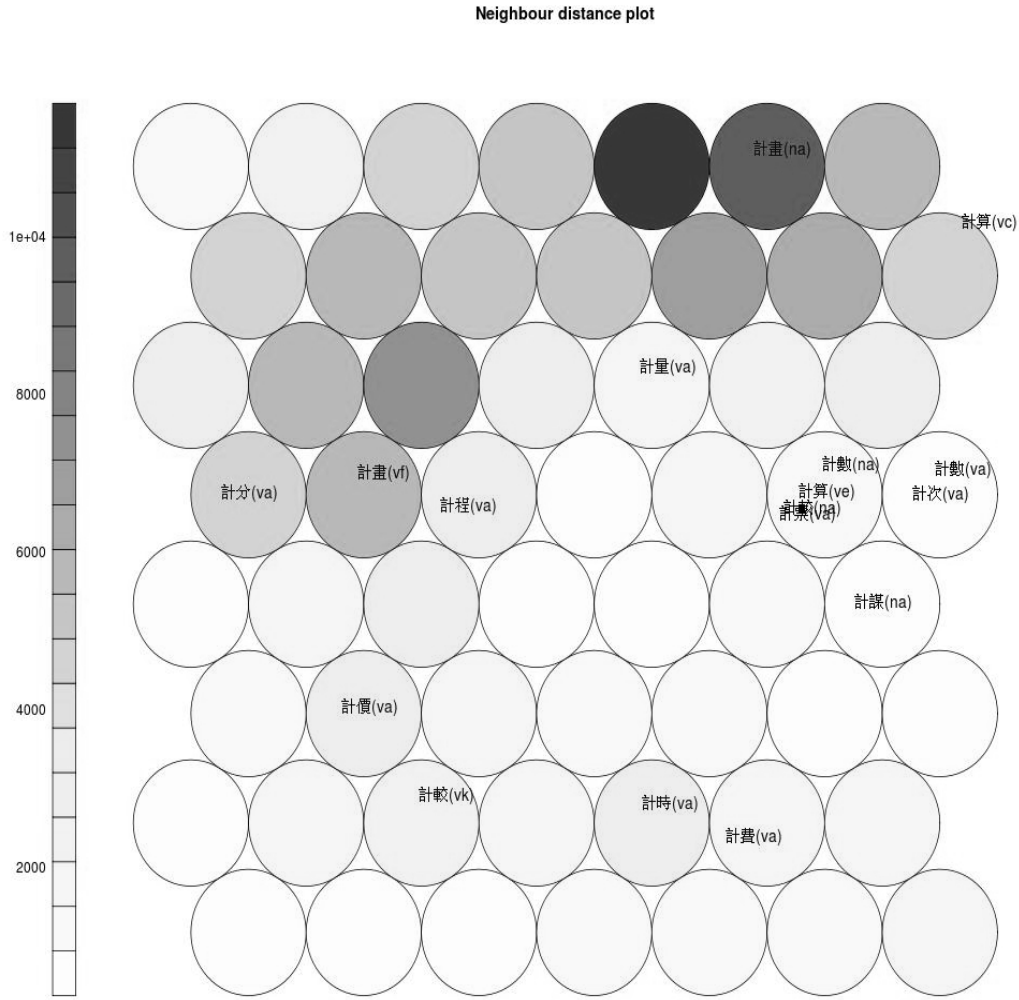

Figure 4. SOM of the members of the 計 morphological family used in Chen et al. (2009).

Clustering is observed easily with such a few words. Three contiguous units in a light shade form the unique big cluster with a total of six different words. In the latest experimental research on semantic satiation, Tian and Huber (2010) found that after five or seven repetitions of a given word, the word's meaning starts to be satiated. From two to five repetitions, there is semantic priming - behavioral enhancement in semantic tasks - and more repetitions are the realm of semantic satiation.

If, in Chen *et al.* (2009)'s lexical decision task (LDT), these six[8] words occur successively, there should be semantic satiation. In Chen *et al.*'s LDT, the 13 words in Figure 3 were randomly mixed with 13 non-words. Non-words, being meaningless, should not contribute significantly to satiate the semantic units of the different meanings of the shared morpheme. Therefore, from the analysis of our SOM, we predict that there could be a preliminary sign of semantic satiation only in the case where the 6 members of the big cluster occur successively in the 26-word list - we call this the best case.

To compute the probability of this best case, we need to calculate two numbers:

1. N_a, the number of distinguishable arrangements of n=26 words of which 6 - belonging to our big cluster - constitute a first set S1 and the 20 remaining ones constitute another set S2. The order of occurrence of the 6 words of S1 does not matter; therefore, the words of S1 are considered to be of a same type T1. For the same reason, words of S2 are of a same type T2, different from type T1.

$$N_a = \frac{26!}{6!20!} = 230230 \tag{2}$$

2. The number of distinguishable arrangements of 6 successive occurrences of S1 words[9] in a 26-word list: 21.

The probability *p* of the best case is given by dividing the number of distinguishable arrangements of 6 successive occurrences of S1 words by the number of distinguishable arrangements of n=26 words made of the two types T1 and T2.

$$p = \frac{21}{230230} \approx 9*10^{-5} \tag{3}$$

This best case has a very low probability, so subjects in Chen *et al.* would almost always be given a 26-word list that do not warrant - according to our analysis - elicitation of semantic satiation.

Hence, we agree with Chen *et al.* that there was no semantic locus of satiation in their experiment. On the other hand, we refine Chen *et al.*'s conclusions by advancing that one could prepare specific experimental word lists that would maximize the probability of observing semantic satiation.

6. General Conclusion

By visualizing the SOMs augmented with neighboring distance information from the U-matrix, one can observe whether semantic clusters exist in a morphological family and how the

[8] Tian and Huber (2010) found satiation effects after six repetitive accesses to a word's meaning.

[9] Order of occurrence of the S1 words does not matter.

experimental data in Chen *et al.* (2009) is mapped to these clusters.

Conclusions drawn from our computational experimental results are in accordance with Chen *et al.*'s behavioral experimental results revealing the absence of a semantic satiation while morphological satiation occurs. Nevertheless, we propose that semantic satiation theoretically could be elicited with specifically arranged word lists for Chen *et al.*'s experiment. Such lists have a very low probability of occurrence when a random assignment of words is used to prepare experimental word lists. Therefore, the present work shows the necessity of preparing adequate experimental word lists based on computational semantic clustering - as shown here - or human norms of semantic similarity if available.

7. Future Directions

Alternatives to SOMs, such as GTM (Bishop, Svensen, & Williams, 1998), exist and could be used for comparison purposes with the present results.

8. Code to generate the SOMs from the ASBC corpus

The source code and R command lines are available upon request in a script file. In order to run the whole script file from the very beginning, one needs the Academia Sinica Balanced Corpus (ASBC). The ASBC has to be purchased.[10]

Acknowledgements

This work was supported by a doctoral fellowship grant (NSC100-2420-H-006-007-DR) awarded to Bruno Galmar.

References

Bishop, C. M., Svensén, M., & Williams, C. K. I. (1998). GTM: The generative topographic mapping. *Neural computation*, 10(1), 215-234.

Chen, J. Y., Galmar, B., & Su, H. J. (2009). *Semantic Satiation of Chinese Characters in a Continuous Lexical Decision Task.*

Chen, K. J., & Chen, C. (2000). *Automatic semantic classification for Chinese unknown compound nouns.*

Cheng, C. M., & Lan, Y. H. (2011). An implicit test of Chinese orthographic satiation. *Reading and Writing*, 24(1), 55-90.

Cottrel, M., Fort, J., & Letremy, P. (2005). *Advantages and drawbacks of the batch Kohonen Algorithm.*

[10] Contact the Academia Sinica (中央研究院語言所).

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text mining infrastructure in R. *Journal of Statistical Software,* 25(5), 1-54.

Galmar, B., & Chen, J. (2010). Identifying Different Meanings of a Chinese Morpheme through Latent Semantic Analysis and Minimum Spanning Tree Analysis. *International Journal of Computational Linguistics and Applications,* 1(1-2), 153-168.

Galmar, B., & Chen, J. Y. (2010). Identifying different meanings of a Chinese morpheme through semantic pattern matching in augmented minimum spanning trees. *The Prague Bulletin of Mathematical Linguistics,* 94(-1), 15-34.

Jorgensen, J. C. (1990). The psychological reality of word senses. *Journal of Psycholinguistic Research,* 19(3), 167-190.

Kohonen, T. (2001). *Self-Organizing Maps*: Springer.

Kohonen, T., Oja, E., Simula, O., Visa, A., & Kangas, J. (1996). Engineering applications of the self-organizing map. *Proceedings of the IEEE,* 84(10), 1358-1384.

Kohonen, T., & Xing, H. (2011). Contextually self-organized maps of chinese words. *Advances in Self-Organizing Maps,* 16-29.

Kounios, J., Kotz, S. A., & Holcomb, P. J. (2000). On the locus of the semantic satiation effect: Evidence from event-related brain potentials. *Memory & cognition,* 28(8), 1366-1377.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review,* 104(2), 211.

Landauer, T. K., McNamara, D. S., Dennis, S. E., & Kintsch, W. E. (2007). *Handbook of latent semantic analysis*: Lawrence Erlbaum Associates Publishers.

Li, P. (2001). *A self-organizing neural network model of the acquisition of word meaning.*

Li, P. (2009). Lexical organization and competition in first and second languages: Computational and neural mechanisms. *Cognitive science,* 33(4), 629-664.

Li, P., Farkas, I., & MacWhinney, B. (2004). Early lexical development in a self-organizing neural network. *Neural Networks,* 17(8-9), 1345-1362.

Minsky, M. (1974). A framework for representing knowledge. *AIM-306*.

Smith, L., & Klein, R. (1990). Evidence for semantic satiation: Repeating a category slows subsequent semantic processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 16(5), 852.

Tian, X., & Huber, D. E. (2010). Testing an associative account of semantic satiation. *Cognitive psychology,* 60(4), 267-290.

Ultsch, A., & Siemon, H. P. (1990). *Kohonen's Self Organizing Feature Maps for Exploratory Data Analysis.* Paper presented at the Proceedings of the International Neural Network Conference (INNC' 90).

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S*: Springer verlag.

Wehrens, R., & Buydens, L. M. C. (2007). Self-and super-organizing maps in R: the Kohonen package. *Journal of Statistical Software,* 21(5), 19.

Zhao, X., & Li, P. (2008). *Vocabulary development in English and Chinese: A comparative study with self-organizing neural networks*.

Zhao, X., Li, P., & Kohonen, T. (2010). Contextual self-organizing map: software for constructing semantic representations. *Behavior Research Methods*, 1-12.