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

Bruno GALMAR
Institute of Education
National Cheng Kung University
hsuyueshan@gmail.com

Abstract

A morphological family in Chinese is the set of compound words embedding a common morpheme. Self-organizing maps (SOM) of Chinese morphological families are built. Computation of the unified-distance matrices for the SOMs allows us to perform a semantic clustering of the members of the morphological families. Such a semantic clustering shed light on the interplay between morphology and semantics in Chinese. Then, we studied how the word lists used in a lexical decision task (LDT) [1] are mapped onto the clusters of the SOMs. We showed that such a mapping is helpful to predict whether in a LDT repetitive processing of members of a morphological family would elicit a satiation - habituation - of both morphological and semantic units of the shared morpheme. In their LDT experiment, [1] found evidence for morphological satiation but not for semantic satiation. Conclusions drawn from our computational experimentations and calculations are concordant with [1] behavioral experimental results. We finally showed that our work could be helpful to linguists to prepare adequate word lists for the 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. So, the compound words in Tab. 1 which have all the morpheme '明' as a first character belong to the morphological family of '明'.

Table 1. A subset of words belonging to the morphological family of 明 [1].

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 Tab.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 a same meaning of the shared morpheme. Most of Chinese morphemes are polysemous [2]. 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 aims primarily at using computational linguistics methods to perform a semantic clustering of the members of the morphological families. Such a clustering is thereafter used to predict the results of a behavioral Lexical Decision Task[1] (LDT) designed by [1] to study the phenomenon of morphological satiation in Chinese.

In visual word recognition, morphological satiation is an impairment of morphological processing induced by a repetitive exposure to a same morpheme embedded in different Chinese compound words [1][3]. [1] posited that morphological satiation is due to habituation of the morphological unit of the repeated morpheme. This is represented on Fig. 1 by diagram (a).

As a morpheme is thought to be a meaningful unit, it is logical to consider whether a semantic satiation [4][5][6] - 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 [1] could have two loci: a morphological one and a semantic one as represented on Fig. 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 on Fig.1 by the diagrams (b) and (c). We can quickly rule out the possibility of a locus on the link between morphological and lexical units as represented by the 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.

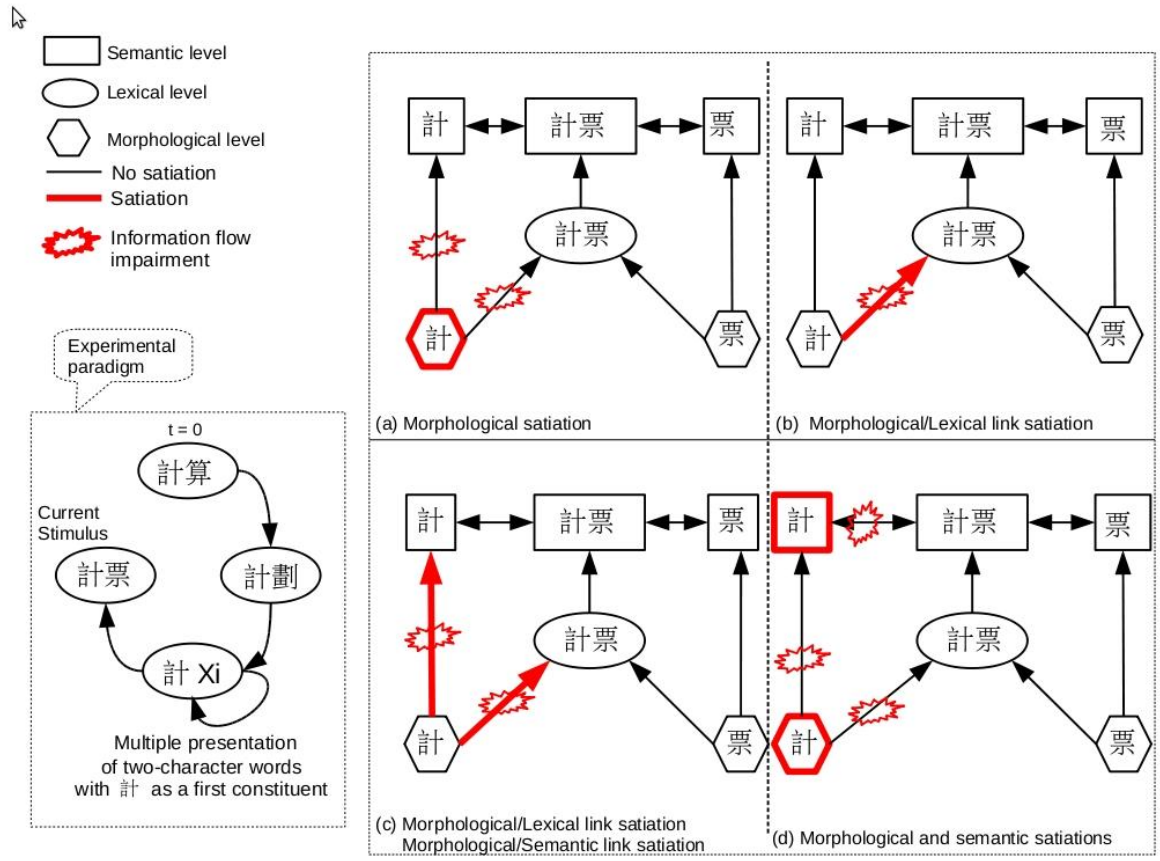
The present work of semantic clustering focuses on clarifying by computational means whether morphological satiation would probably have a sole morphological locus - diagram (a) - or whether it would have both a morphological and semantic locus - diagram (d) -. [1] behavioral LDT experiment results pointed to the existence of a sole morphological locus.

---

[1] A 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 a 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.

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



2. Rationale of our Approach

As human subjects agreement for semantic clustering tasks is low [7], 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 therefore 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 [8]. Computational methods as Latent Semantic Analysis take as input such a high dimensional matrix and reduce its dimensionality to form a vector space of the documents and words [9]. This space embeds only an associative kind of semantic information[3]: words that co-occur in the same documents or which have common

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

co-occurrents are close associates.

For a news corpus, the association can often be of the type situational. 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”*" [11].

In the present work, we do follow a `bag of words' approach by firstly 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 the U-matrices serve to visualize semantic clusters in a morphological family on a 2D hexagonal grid of bins [12].

On the SOMs, a semantic cluster is made of members of a morphological family which have been fitted into a 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, [13][14][15][16] have used SOMs to study the developmental aspect of vocabulary acquisition in Chinese. 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.

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 words annotated corpus based on Chinese materials from Taiwan, mostly newspapers articles. The corpus is made of roughly 10000 documents of unequal length.

We removed from the corpus the foreign alphabetic words and most of the Chinese functional words. We kept POS tags information to allow differentiation between different grammatical instances of a same word[5] [10].

3.2 The Term Document Matrix (TDM)

The TDM was built by using the *TermDocumentMatrix* function of the R package *tm* [17] with a self-customized Chinese tokenizer. The TDM is a 136570 terms * 9179 documents

[4] This example is borrowed from [10]

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

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 [8]. We used the function *weightTfIdf* of the package *tm* [17].
2. using a weighting scheme at the document level to reduce the effect of the size difference between documents:

$$\log_2\left(\frac{\text{Max\_document\_size}}{\text{Document\_size}} + 1\right) \quad (1)$$

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 [10]. Briefly, one can say that for a human reader due to attentional capacity limitations, the gist of a news article is easier to extract from a very short article than from a very long one.

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, a SOM is built using the *Batch map algorithm* [12]. The U-matrix [18] is computed to assess how much members fitted to contiguous bins - bins are thereafter called units - on the SOM are close in the original high-dimensional space - thereafter called input data space -.

4.1 The batch version of the SOM algorithm

As all 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 [12].

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.

From [12, pp139-140] and [19, p1360], the Batch Map algorithm can be described as follows:

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 [12][20]: 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 [20] and [12] to compare organization in the Batch Map and in the incremental learning SOM.

We used the code in the R package *class* [21] for the batch SOM given by [22] to build the SOMs on a 7*8 hexagonal grid of 56 bins.

4.2 The Unified-Distance Matrix

We reused and modified the code in the R package *kohonen* [23] to build the U-matrix for the Batch Map and to plot a grey-level map superimposed to the SOM map. The U-matrix is the distance matrix between the reference vectors of contiguous units. On the grayscale SOMs, contiguous units in 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. Contiguous units in a dark shade draw boundaries between existing clusters in the input data space [18].

5. Results

We present the results for the study of the 計 (ji2) morphological family[6]. This Chinese morpheme has two main meanings: (1) to count, to calculate (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 on Fig 2.

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

- Cluster C_1 mainly gathers word sharing and other words related to meaning 1 of 計.
- Cluster C_2 gathers in a same unit three words related to the frame *taxi*.
- 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 on Fig. 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 4 shows only the 13 words used by [1] in one block of their LDT experiment[7]. Some of the words have two POS tags so that the total of the data points represented on Fig.4 is 17.

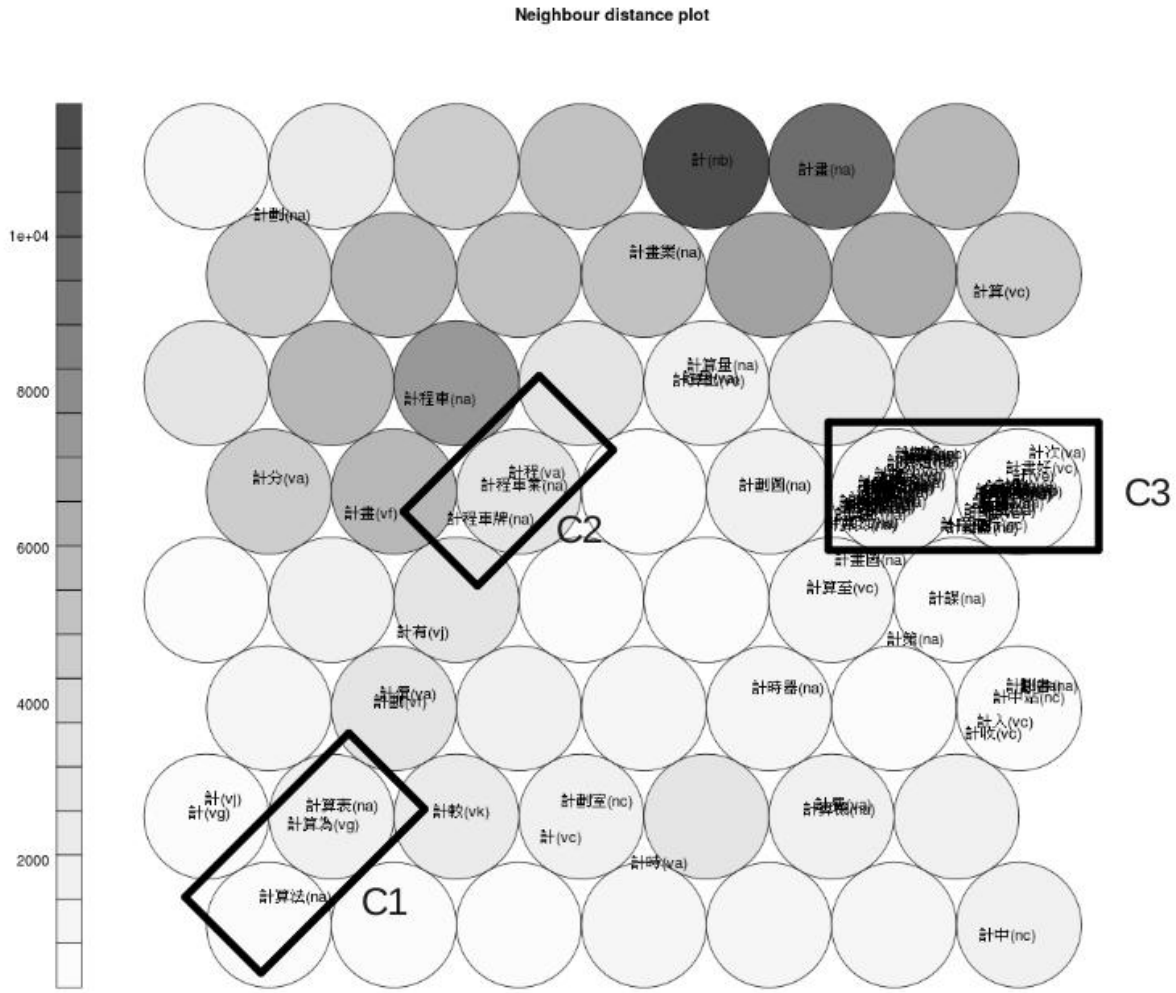


Figure 2. SOM_{93} of the 計 morphological family.

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

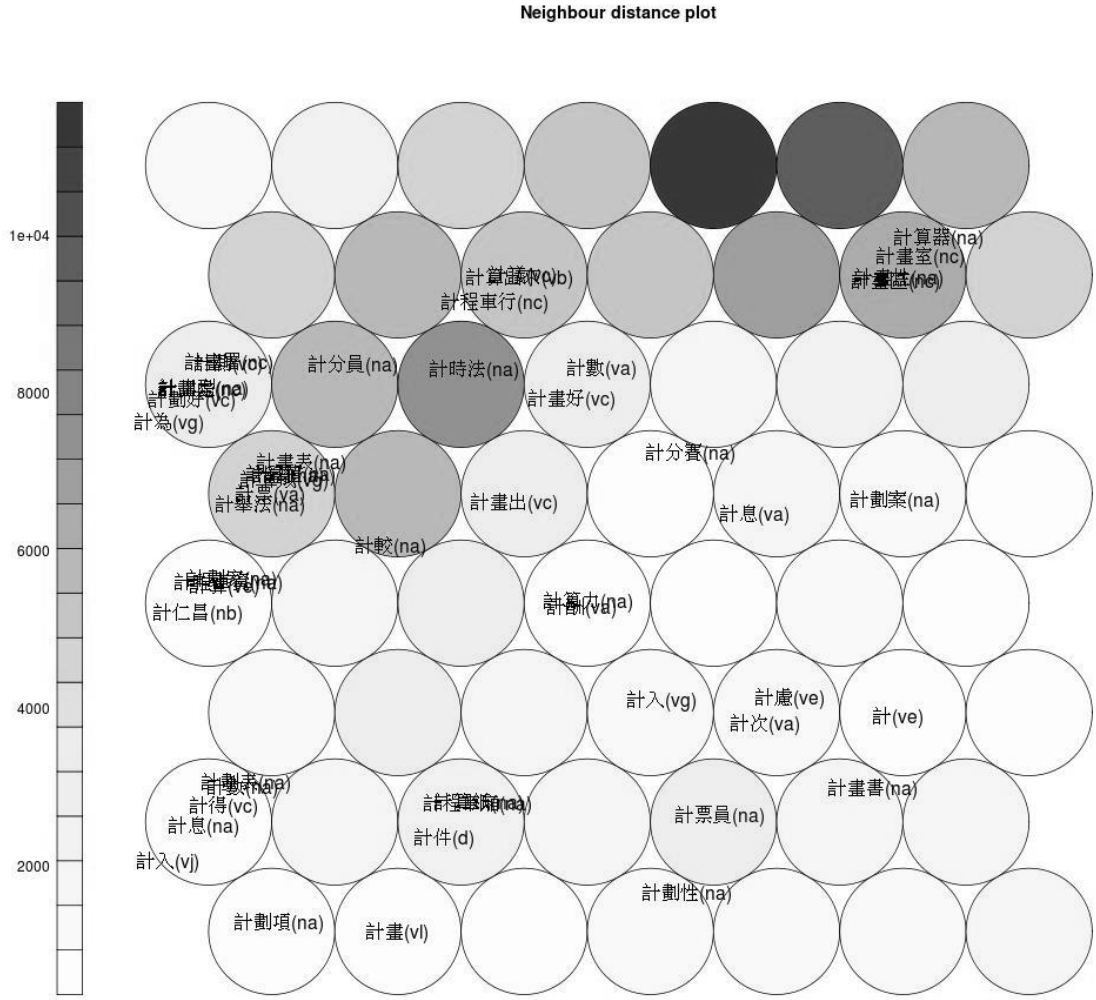


Figure 3. SOM for cluster C3 in Fig. 2

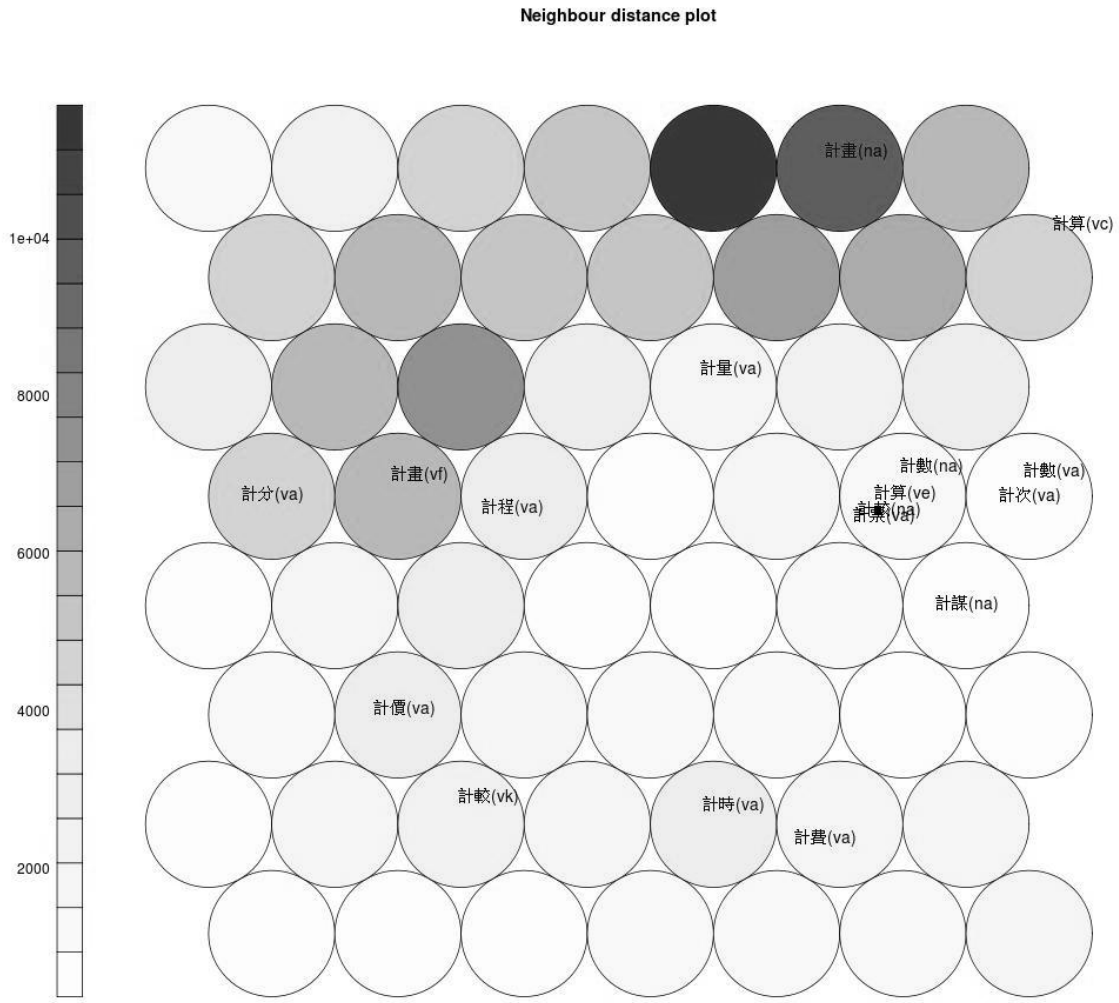


Figure 4. SOM of the members of the 計 morphological family used in [1] experiment.

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 6 different words. In the latest experimental research on semantic satiation, [6] found that after 5 or 7 repetitions of a given word, the word's meaning starts to be satiated. From 2 to 4 repetitions, there is semantic priming - behavioral enhancement in semantic tasks - and more repetitions are the realm of semantic satiation.

If in the [1] lexical decision task (LDT), these 6 words occur successively, there should be semantic satiation. In [1] LDT, the 13 words in Fig. 3 were randomly mixed with 13 non-words. Non-words being meaningless should not contribute to satiate significantly the semantic units of the different meanings of the shared morpheme. Therefore, from the

analysis of our SOM, we predict that only in the case were the 6 members of the big cluster occur successively in the 26 words list - we call it the best case -, there could be a preliminary sign of semantic satiation.

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 and therefore the words of S1 are considered as being of a same type T1. For the same reason, words of S2 are of a same type T2, different of type T1.

$$N_a = \frac{26!}{6!20!} = 230230 \qquad (2)$$

2. the number of distinguishable arrangements of 6 successive occurrences of S1 words[8] in a 26 words 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.

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

This best case has a very low probability so that subjects of [1] experiment would almost always be given a 26 words list that do not warranty - according to our analysis - elicitation of semantic satiation.

Hence, in one hand, we agree with [1] that in their experiment there were no semantic locus of satiation. On the other hand, we refine [1] conclusions by advancing that one could prepare specific experimental word lists which 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 experimental data in [1] is mapped to these clusters.

Conclusions drawn from our computational experimental results are concordant with [1] behavioral experimental results revealing the absence of a semantic satiation while morphological satiation occurs. However, we proposed that semantic satiation could theoretically be elicited with specifically arranged word lists for [1] experiment. Such lists have a very low probability of occurrence when random assignment of words is used to

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

prepare experimental word lists. Therefore, the present work showed 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 [24] - 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[9].

References

[1] J.-Y. Chen, B. Galmar, and H.-J. Su, "Semantic satiation of Chinese characters in a continuous lexical decision task," in *The 21st Annual Convention of the Association For Psychological Science*, 2009.

[2] K. Chen and C. Chen, "Automatic semantic classification for Chinese unknown compound nouns," in *Proceedings of the 18th conference on Computational linguistics-Volume 1. Association for Computational Linguistics*, 2000, pp. 173–179.

[3] C. Cheng and Y. Lan, "An implicit test of Chinese orthographic satiation," *Reading and Writing*, pp. 1–36, 2009.

[4] L. Smith and R. Klein, "Evidence for semantic satiation: Repeating a category slows subsequent semantic processing," *Learning, Memory*, vol. 16, no. 5, pp. 852–861, 1990.

[5] J. Kounios, S. Kotz, and P. Holcomb, "On the locus of the semantic satiation effect: Evidence from event-related brain potentials," *Memory and Cognition*, vol. 28, no. 8, pp. 1366–1377, 2000.

[6] X. Tian and D. Huber, "Testing an associative account of semantic satiation," *Cognitive Psychology*, 2010.

[7] J. Jorgensen, "The psychological reality of word senses," *Journal of Psycholinguistic Research*, vol. 19, no. 3, pp. 167–190, 1990.

[8] T. Landauer and S. Dumais, "A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge," *Psychological review*, vol. 104, no. 2, pp. 211–240, 1997.

[9] T. Landauer, D. McNamara, S. Dennis, and W. Kintsch, *Handbook of latent semantic analysis*. Lawrence Erlbaum, 2007.

[10] B. Galmar and J. Chen., "Identifying different meanings of a Chinese morpheme through

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

semantic pattern matching in augmented minimum spanning trees," *The Prague Bulletin of Mathematical Linguistics*, vol. 94, 2010.

[11] M. Minsky, "A framework for representing knowledge," AIM-306, 1974.

[12] T. Kohonen, *Self-Organizing Maps, 3rd Edition*, Berlin, Heidelberg, 2001.

[13] P. Li, "A self-organizing neural network model of the acquisition of word meaning," in *Proceedings of the 2001 Fourth International Conference on Cognitive Modeling*, July 26-28, 2001 George Mason University, Fairfax, Virgiania, USA. Lawrence Erlbaum, 2001, p. 90.

[14] P. Li, I. Farkas, and B. MacWhinney, "Early lexical development in a self-organizing neural network," *Neural Networks*, vol. 17, no. 8-9, pp. 1345–1362, 2004.

[15] X. Zhao and P. Li, "Vocabulary development in English and Chinese: A comparative study with self-organizing neural networks," in *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, 2008, pp. 1900–1905.

[16] P. Li, "Lexical Organization and Competition in First and Second Languages: Computational and Neural Mechanisms," *Cognitive Science*, vol. 33, no. 4, pp. 629–664, 2009.

[17] I. Feinerer, "tm: Text mining package, 2008," UR L http://CRAN. R-project. org/package= tm. R package version 0.3-3.

[18] A. Ultsch and H. Siemon, "Kohonen's self organizing feature maps for exploratory data analysis," in *Proceedings of the International Neural Network Conference (INNC'90)*, 1990, pp. 305–308.

[19] T. Kohonen, E. Oja, O. Simula, A. Visa, and J. Kangas, "Engineering applications of the self-organizing map," *Proceedings of the IEEE*, vol. 84, no. 10, pp. 1358–1384, 2002.

[20] J. Fort, P. Letremy, and M. Cottrell, "Advantages and drawbacks of the Batch Kohonen algorithm," in *10th European Symp. On Artificial Neural Networks*. Citeseer.

[21] W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S*, 4th ed. New York: Springer, 2002, iSBN 0-387-95457-0. [Online]. Available: http://www.stats.ox.ac.uk/pub/MASS4

[22] W. Venables, B. Ripley, and W. Venables, *Modern applied statistics with S-Plus*. Citeseer, 1998.

[23] R. Wehrens and L. Buydens, "Self- and super-organising maps in r: the kohonen package," *J. Stat. Softw.*, vol. 21, no. 5, 2007. [Online]. Available: http://www.jstatsoft.org/v21/i05

[24] C. Bishop, M. Svensen, and C. Williams, "GTM: The generative topographic mapping," *Neural computation*, vol. 10, no. 1, pp. 215–234, 1998.