Computational Models to Estimate Semantic Transparency of English Compounds and Chinese Words using Latent Semantic Analysis

Hsueh-Cheng Wang[1], Li-Chuan Hsu[2], Yi-Min Tien[3], and Marc Pomplun[1]
[1] Department of Computer Science, University of Massachusetts at Boston
[2] Department of Psychology, Chung Shan Medical University, Taiwan
[3]Graduate Institute of Neural and Cognitive Sciences, China Medical University, Taiwan

Semantic Transparency

- A compound word is a word composed of at least two free constituents that refer to a new concept. Transparency Category:
  - Opaque-Opaque (OO, honeymoon)
  - Transparent-Opaque (TO, staircase)
  - Opaque-Transparent (OT, dragonfly)
  - Transparent-Transparent (TT, farmland)
- Transparency ratings are often subjective and vary across raters.
- How are compound words represented in the mental lexicon and how does a rater access meanings and decide “T” or “O”?

Latent Semantic Analysis (LSA)

- A computational model may be a way to average across subjective differences of estimating semantic transparency.
- LSA is a method to represent the meaning of words by statistical computations applied to a text corpus (Landauer & Dumais, 1997).
- A term-to-document co-occurrence matrix is established from a corpus. Then singular value decomposition (SVD) is used to reduce the dimensions of the original matrix.
- The meaning of each term is represented as a vector in semantic space.
- Represent mental lexicon using LSA (see Jones & Mewhort, 2007)

English Compounds

Model 1: Whole word vs. each of its constituents.

	Whole Word	1st Constituent	2nd Constituent	LSA C1	LSA C2
OO	honeymoon	honey	moon	0.03	0.01
TO	staircase	stair	case	0.57	0.07
OT	dragonfly	dragon	fly	0.12	0.43
TT	farmland	farm	land	0.55	0.67

- Semantic space of English: an LSA web site is freely available (http://lsa.colorado.edu/, accessed September, 2010; see Dennis, 2007).
- Access the meaning of the compound and each of its constituents. Each word, irrespective of how many meanings or senses it has, is represented by a single vector. However, when a word is used in different contexts, context appropriate word senses emerge (Kintsch, 2002).
- LSA value for any two terms (either compound or its constituents) ranges between -1 and 1, but rarely goes below 0.

Chinese Words

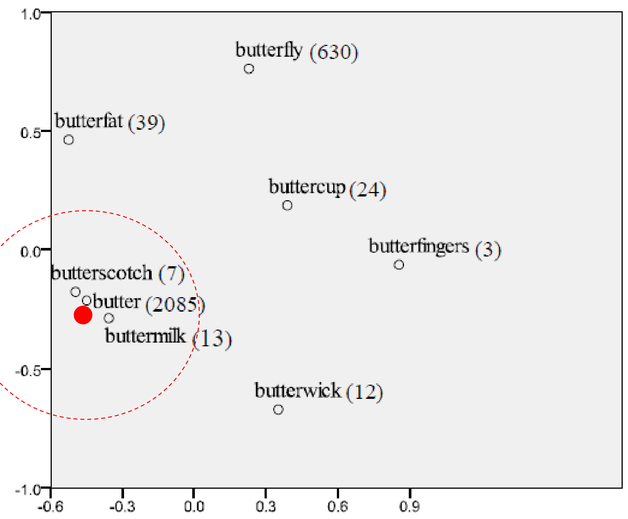
Model 1: Whole word vs. each of its constituents.

	Whole Word	1st Character	2nd Character	LSA C1	LSA C2
OO	壽司 (sushi)	壽 (age)	司 (in charge of)	0.06	0.01
TO	字母 (letter)	字 (character)	母 (mother)	0.50	0.06
OT	垂死 (almost dead)	垂 (hanging)	死 (dead)	0.08	0.28
TT	球場 (ball court)	球 (ball)	場 (court)	0.56	0.39

- Semantic space of Chinese: our previous studies (Wang et al., 2010; Chen, Wang, & Ko, 2009).
- Unlike English, Chinese words are written without spaces in a sequence of characters. The concept of a word is not as clearly defined in Chinese as it is in English.
- Not all characters are stand-alone words in corpus, and therefore the representation of character in mental lexicon may be different.
- A character might be shared by many words, but the meaning of the character and those words may not be consistent.

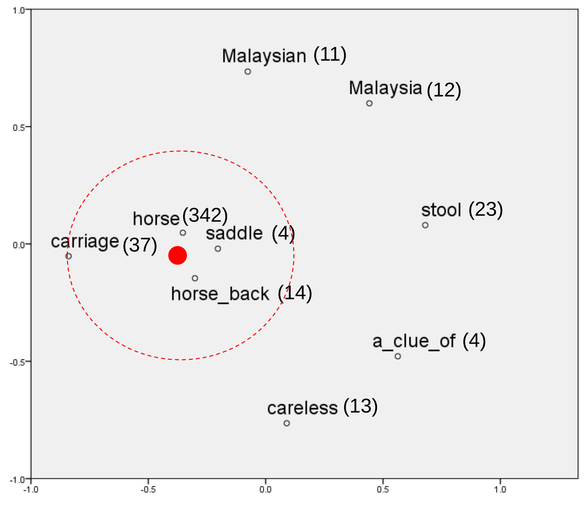
Model 2: Whole word vs. dominant meaning of each of its constituent

	butter	butterfly	buttercup	butterfingers	buttermilk	butterscotch	butterfat	butterwick
butter	1							
butterfly	0.04	1						
buttercup	0.09	0.09	1					
butterfingers	0	-0.05	-0.06	1				
buttermilk	0.44	-0.01	0.12	0.01	1			
butterscotch	0.45	0.05	-0.02	0.02	0.35	1		
butterfat	0.12	-0.04	0.04	0	0.11	0.16	1	
butterwick	-0.01	0.01	0.12	-0.03	0.09	0.03	0.04	1



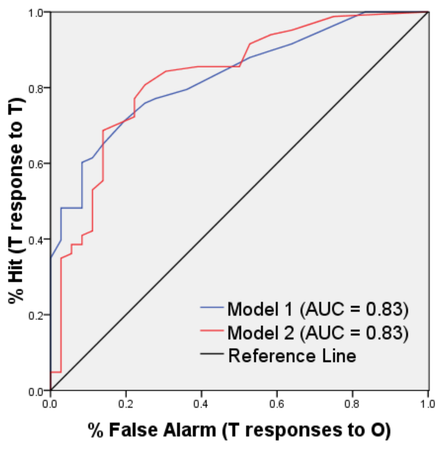
1. Find words containing a constituent that a rater possibly activates.
2. Compute LSA value between each word pair.
3. Hierarchical clustering algorithm and a given threshold
4. The cluster with the highest sum of word frequency is considered the dominant meaning.
5. Compute LSA value between compound and the dominant meaning (a single vector representing the words in the cluster).

	馬	馬背	馬鞍	馬車	馬虎	馬桶	馬腳	馬來	馬國
馬 (horse)	1								
馬背 (horse back; back)	0.83	1							
馬鞍 (saddle; saddle)	0.74	0.74	1						
馬車 (carriage; car)	0.17	0.07	0.03	1					
馬虎 (careless; tiger)	-0.02	-0.04	-0.01	-0.04	1				
馬桶 (stool; tub)	-0.05	-0.04	-0.03	0.01	0.10	1			
馬腳 (a clue of; foot)	0.00	0.04	-0.05	0.01	0.13	0.02	1		
馬來 (Malaysian; come)	0.08	0.06	0.04	0.04	0.00	-0.09	-0.03	1	
馬國 (Malaysia; country)	0.03	-0.01	-0.03	0.03	0.01	-0.04	0.02	0.15	1

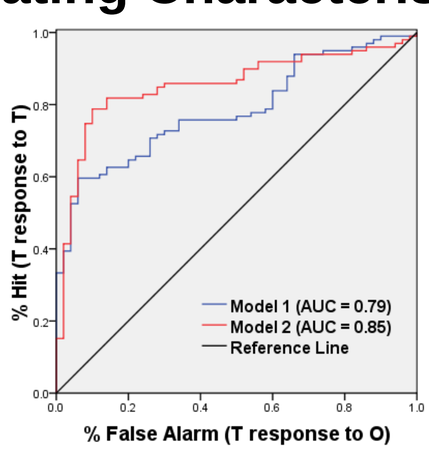


- The representation of meaning of a character may be different from an English constituent.
- Model 2 assumes that a rater accesses the dominant meaning of a character when the character has multiple (inconsistent) meanings.
- Model 2 takes the polysemy of a character into account and works even when a character is not a stand-alone word.
- Model 2 is especially useful for the Chinese language.

Evaluations - Receiver Operating Characteristic (ROC)



- We reanalyzed the transparency rating materials in Frisson et al. (2008)
- ROC analysis (Green & Swets, 1966) was performed and the area under the curve (AUC) was used as measurement.
- The results of Model 1 and Model 2 are compatible.
  - The representation of a constituent is close to the computed dominant meaning.
  - Human raters access the dominant meaning to perform transparency judgments.



- We reanalyzed the transparency rating materials in Tsai (1994) and Lee (2007).
- Model 2 outperforms Model 1 and is a better approach for predicting transparency of characters of two-character Chinese words.
  - The polysemy of a character may affect the accuracy of Model 1, given that there is no context for the transparency judgment task.
  - Some constituents are not stand-alone words.
- Chinese readers might learn the polysemy of characters implicitly from polymorphemic words.

Discussion and Conclusions

- The most important outcome of the current study is to offer a different perspective and an opportunity to represent mental lexicon and examine the lexical processing, which may reflect the polysemy of constituents and how raters access meanings.
- Corroborating evidence from two different languages was presented, and Model 1 is suggested for English and Model 2 is suggested for Chinese.
- The selection of a threshold might be involved in the transparency judgments by human raters and each participant might have a different threshold for the “cut-off” of opacity.
- The results could be adapted to further Chinese reading research using eye movements to examine how Chinese words are accessed, whether as a single entity in the mental lexicon or via its characters during a transparency rating task (no context) and natural reading (with context).