Resolving Linguistic Ambiguity in the Chinese Writing System

Cheng-Yu Hsieh[1]

Supervisors: Prof Kathleen Rastle[1], Prof Marco Marelli[2,3], Prof Chris Watkins[4]

[1] Department of Psychology, Royal Holloway, University of London, Egham, UK

[2] Department of Psychology, University of Milano–Bicocca, Milan, Italy

[3]NeuroMI, Milan Center for Neuroscience, Milan, Italy

[4] Department of Computer Science, Royal Holloway, University of London, Egham, UK

Background: Ambiguity in Chinese

- Chinese characters, as a unit of meaning (*morpheme*), are mostly embedded in two-character compound words, which is the most common word type in Chinese.
- Characters often convey more than one meaning (*morphemic ambiguity*).

C1 'huā' 'píng' C2

花瓶

花錢 花心 花生

- There are no explicit word boundaries in Chinese script.

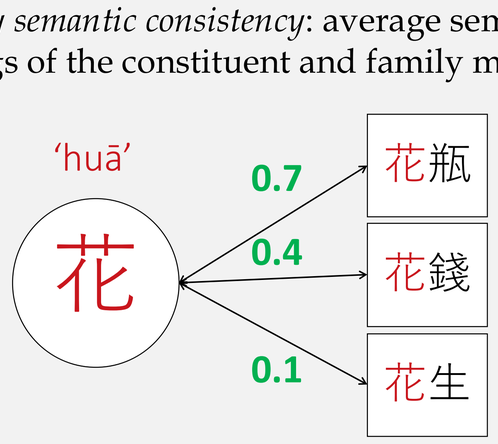
這些花生長得很好。

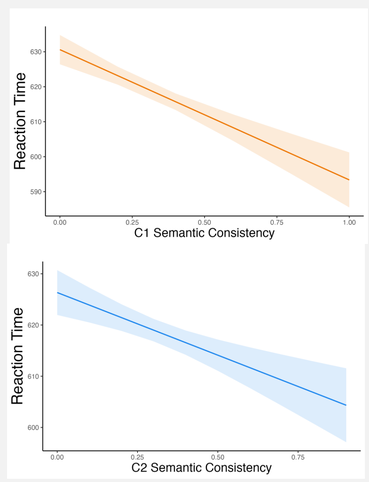
Previous Attempts Using Traditional DSMs

Does the consistency with which constituents convey meaning matter?

Metrics derived from distributional semantic models (DSMs)

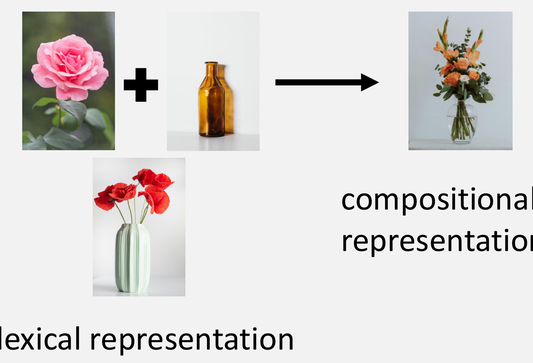
- *Family semantic consistency*: average semantic relatedness between the meanings of the constituent and family members





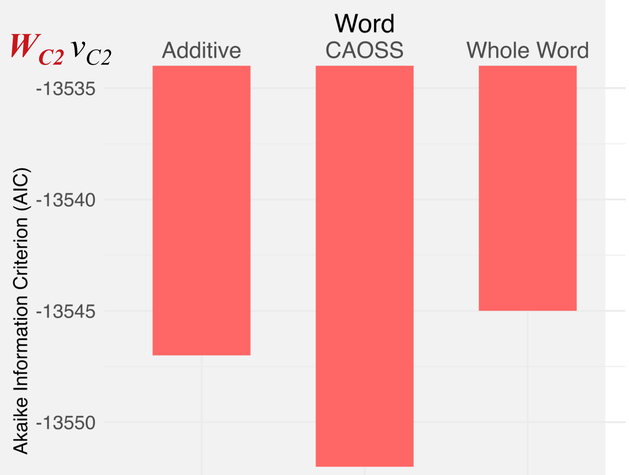
How meanings of individual characters are combined?

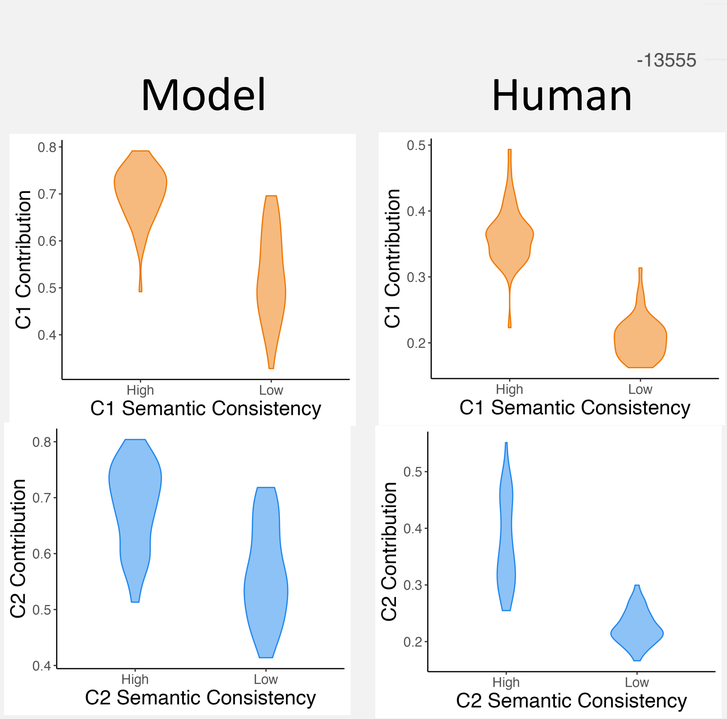
- Compositional distributional semantic models (DSMs)
  - *Additive model:* $c = v_{C1} + v_{C2}$
  - *CAOSS model:* $c = \boldsymbol{W_{C1}} v_{C1} + \boldsymbol{W_{C2}} v_{C2}$

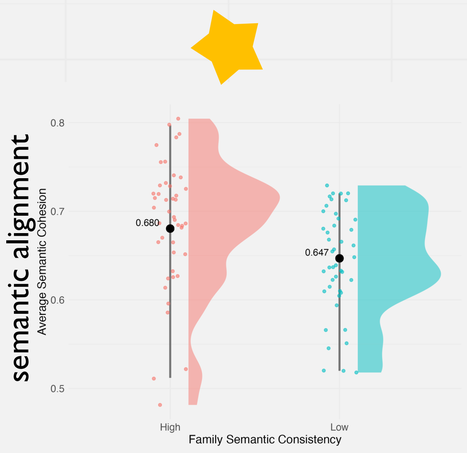


compositional representation

lexical representation







Semantic Consistency Modelling

JEP:LMC *PBR*

Proposal I

Understanding constituents' meaning in novel compounds using contextualised word embeddings from Bidirectional Encoder Representations from Transformers (BERT)

Cosine similarity between contextualised embeddings		*flower garden*	*flower market*	*spend money*	*expend/ expenses*
flower garden	花園	1	**0.9149**	0.8584	0.8814
flower market	花市		1	0.8451	0.8502
spend money	花錢			1	**0.9164**
expend/ expenses	花費				1

'hua'

花罐

'hua'

花金

Proposal II

Predicting word boundaries in Chinese sentences using contextualised word embeddings from BERT

Left-Branching

日光燈

sunlight lamp

Right-Branching

超音速

exceed speed of sound

美國選手在男子滑雪比賽中有望蟬聯冠軍。
(Translation: It's possible that the US athletes will win the men's ski competition again)

Standard segmentation predefined by the researchers (word boundaries marked by slashes; Yan et al., 2010).
美國/ 選手/ 在/ 男子/ 滑雪/ 比賽/ 中/ 有望/ 蟬聯/ 冠軍/。

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18
	美	國	選	手	在	男	子	滑	雪	比	賽	中	有	望	蟬	聯	冠	軍
S	0	1	0	1	1	0	1	0	1	0	1	1	0	1	0	1	0	1
P1	0	0	0	1	1	0	1	0	1	0	1	1	0	1	0	1	0	1
P2	0	1	0	1	1	0	1	0	0	0	1	1	0	1	0	1	0	1
	0	0.5	0	1	1	0	1	0	0.5	0	1	1	0	1	0	1	0	1

Fig. 2 An illustration of the coding scheme, Following Yan et al. (2010), we used "0" to indicate "not a word boundary" and "1" indicate "a word boundary", C1 to C18 refers to the character position in the sentence. The row "S" refer to the "standard" segmentation predefined by the researchers, The row "P1" and "P2" refer to the segmentation by the two hypothetical participants. The points of deviation from the standard are underlined. The bottom row is the agreement score reported in the present CWSE corpus

(Tsang et al., 2025, *Behavior Research Methods*)

Implication

- Redress the overreliance on English and alphabetic writing systems in reading research;
- Underscore the important role of semantic nature in reading and writing;
- Improve literacy education in Chinese-speaking regions;
- Develop interventions for typically developing children and children with literacy difficulties.