ICPEAL 18

Computing Chinese character ambiguity based on the variability of word formations

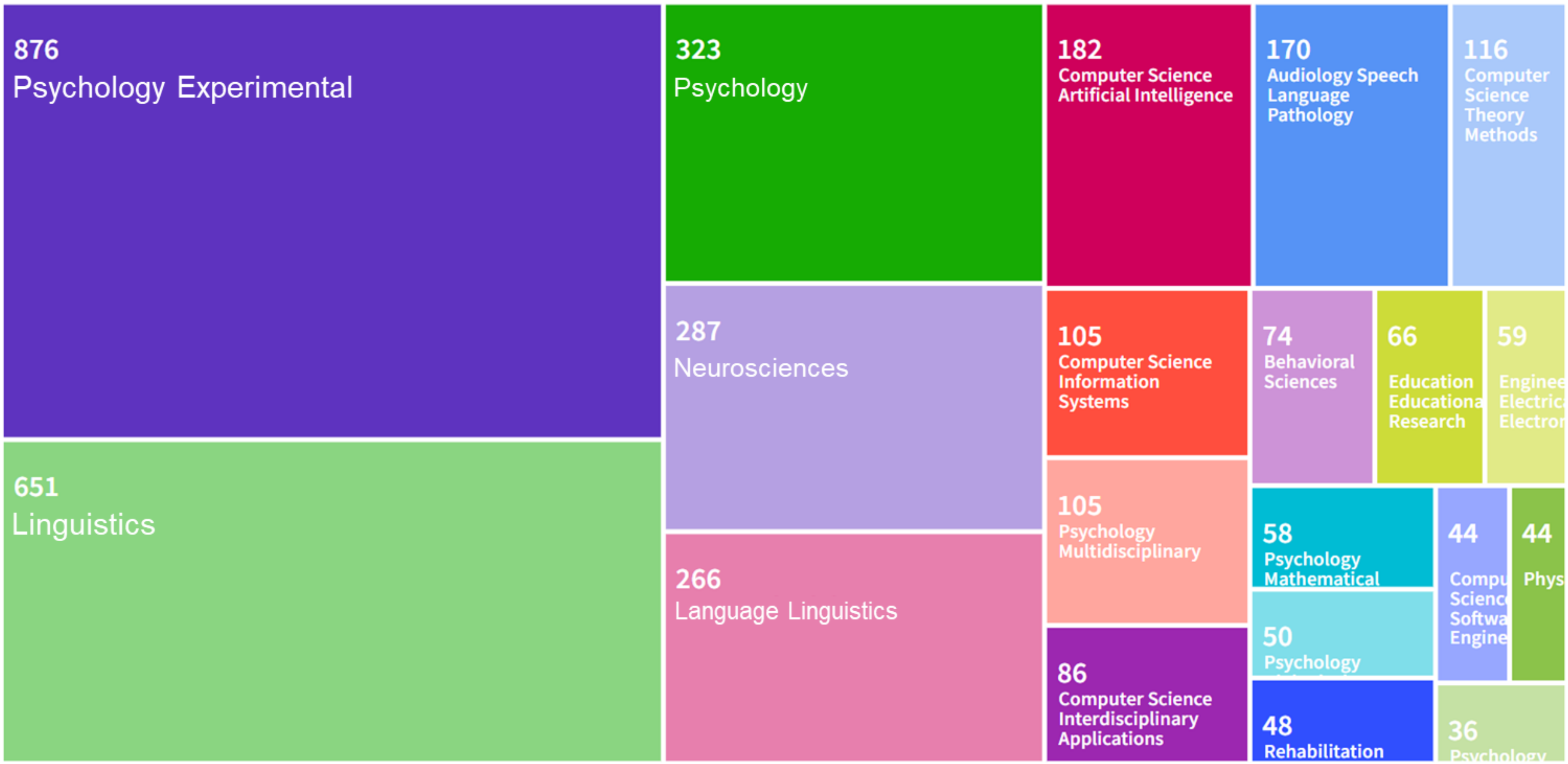
Tianqi Wang[1,2], Xu Xu[2]

[1] Speech Science Laboratory, The University of Hong Kong
[2] School of Foreign Languages, Shanghai Jiao Tong University

What is lexical ambiguity?

- **Lexical ambiguity:** one single word form with more than one meaning
  - Ubiquitous in all human language (Youn et al., 2016)
  - Enables the expression of a near-infinite set of ideas with a small finite lexicon (Piantadosi et al., 2012; Ramiro et al., 2018)
  - Comprehension becomes more challenging when its immediate language contexts are impoverished or not available

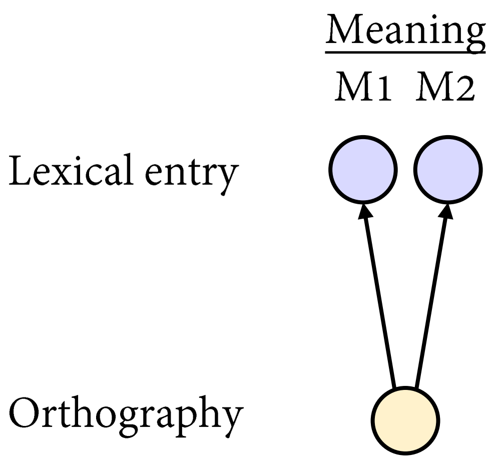
Who are studying lexical ambiguity?



Number of studies on the topic of lexical ambiguity since 1960 (Web of Science)

The processing of ambiguous words

- **Ambiguous vs. unambiguous words:** recognized ***faster*** and ***more accurately*** in lexical decision (e.g., Borowsky & Masson, 1996; Ferraro & Hansen, 2002; Hino & Lupker, 1996; but see Gernsbacher, 1984)
  - Multiple meanings are represented as ***individual lexical entries*** within a network (Klein & Murphy, 2001)
  - Simultaneous activation of the lexical entries results in ***greater inhibition*** to their competitors (Kellas et al., 1988)



The processing of ambiguous words

- *Relatedness* between a word's various meanings was underappreciated

Linguistic taxonomy of words	Examples	
• Homonymy: words with *unrelated* meanings	"bank"	financial institution side of river
• Polysemy: words with *related* senses	"paper"	white sheet academic article document
• Monosemy: words with a *single* sense	"wacky"	funny in an odd way

The processing of ambiguous words

- ***Relatedness*** between a word's various meanings was underappreciated

Linguistic taxonomy of words	**Effects on word recognition**
• Homonymy: words with ***unrelated*** meanings	Inhibition
• Polysemy: words with ***related*** senses	Facilitation
• Monosemy: words with a ***single*** sense	

(Rodd et al., 2002; Yap et al., 2011)

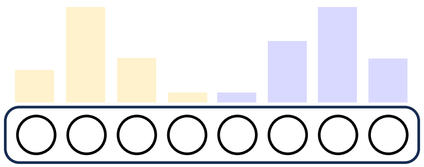
The processing of ambiguous words

- *Relatedness* between a word's various meanings was underappreciated

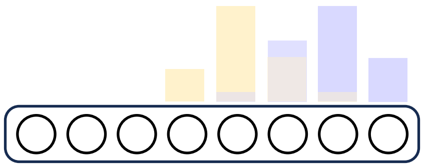
Linguistic taxonomy of words

Representation / Semantic activation

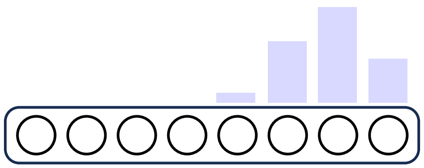
- **Homonymy:** words with ***unrelated*** meanings



- **Polysemy:** words with ***related*** senses



- **Monosemy:** words with a ***single*** sense



(Armstrong & Plaut, 2016; Rodd et al., 2002)

How to characterize a word's meanings?

- ***Number of meanings:*** number of dictionary meanings (dNoM)

  Critical issues

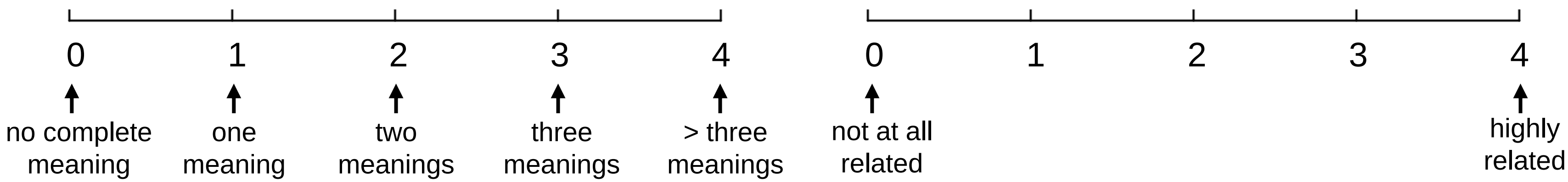
  - How different must two uses of a word be to qualify as separate senses (Hoffman et al., 2013; Hoffman & Woollams, 2015)?
  - Could measures based on dictionary definitions reflect native speakers' perception of the word's number of meanings (Gernsbacher, 1984)?

- ***Relatedness of meanings:*** distinction between homonymy, polysemy, and monosemy is a simplification of reality

homonymy polysemy monosemy

Alternative approaches to characterize a word’s meanings

- Norms for the ***perceived number of meanings (pNoM)*** and ***relatedness of meanings (pRoM)***



Norms for Chinese characters

pNoM: 4,363 characters
pRoM: 1,052 characters (with pNoM > 1.45)

Chen, H., Xu, X., & Wang, T. (2023). Assessing lexical ambiguity of simplified Chinese characters: Plurality and relatedness of character meanings. *Quarterly Journal of Experimental Psychology*.

- Corpus-based approaches ⟶ The present study

Ambiguity in Chinese characters

- Chinese is ***morpho-syllabic*** in nature, and there is usually a correspondence between a morpheme, a syllable, and a single character

- The mapping between characters and morphemes is ***not always one-to-one***
  - Forming a complex word based on one specific meaning of an ambiguous Chinese character makes it semantically concrete (Xu, 1994)

Example

	Meanings	**Word formations**	
花	flower	鲜花 “flower”	花园 “garden”
	to spend	花钱 “spend money”	花费 “cost”
	blurred	眼花 “blurred vision”	昏花 “dim-sighted”

→ Can we measure character ambiguity based on the variability of word formations?

Computed metric

- **Computed dissimilarity of meanings (cDoM)**
  - Probe the meanings of a character using its ***word formations***
  - Measure the ***dispersion*** of their vector representation in a ***distributional semantic space***

Distributional hypothesis

Words appearing in similar contexts have similar meanings

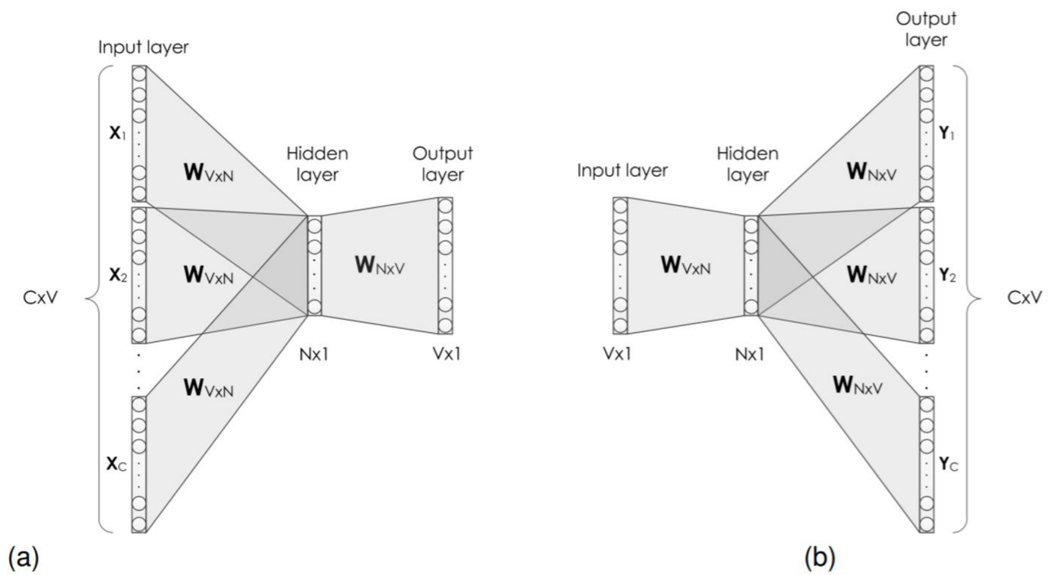


Illustration of the *word2vec* models: (a) CBOW, (b) skip-gram

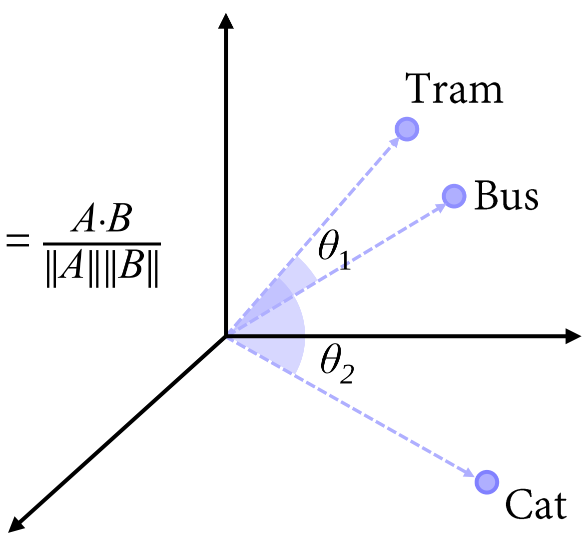
Computed metric

- **Computed dissimilarity of meanings (cDoM)**
  - Probe the meanings of a character using its ***word formations***
  - Measure the ***dispersion*** of their vector representation in a ***distributional semantic space***

Distributional hypothesis

Words appearing in similar contexts have similar meanings

$$sim(A,B) = cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$

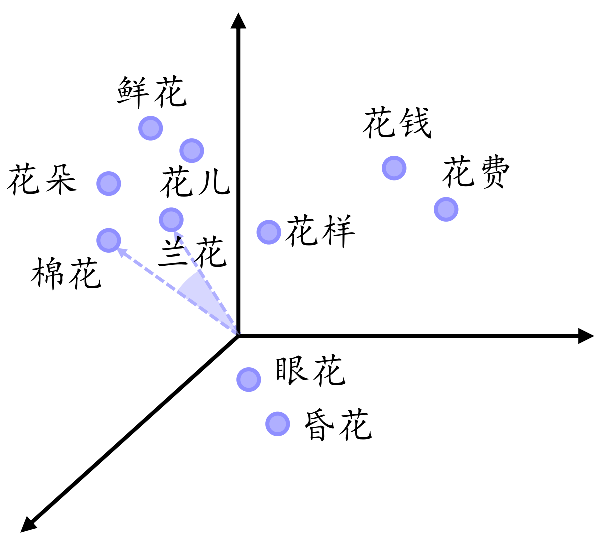


Computed metric

- **Computed dissimilarity of meanings (cDoM)**
  - Probe the meanings of a character using its ***word formations***
  - Measure the ***dispersion*** of their vector representation in a ***distributional semantic space***

Top 10 word formations based on frequency ⇒ Pairwise cosine similarity

花

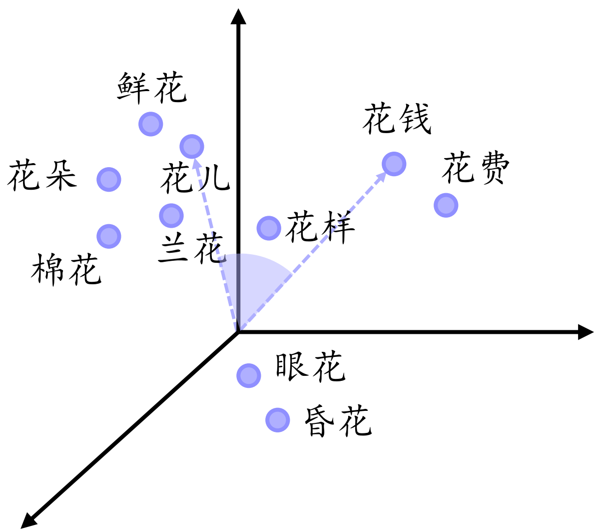


Computed metric

- **Computed dissimilarity of meanings (cDoM)**
  - Probe the meanings of a character using its ***word formations***
  - Measure the ***dispersion*** of their vector representation in a ***distributional semantic space***

Top 10 word formations based on frequency ⇒ Pairwise cosine similarity

花

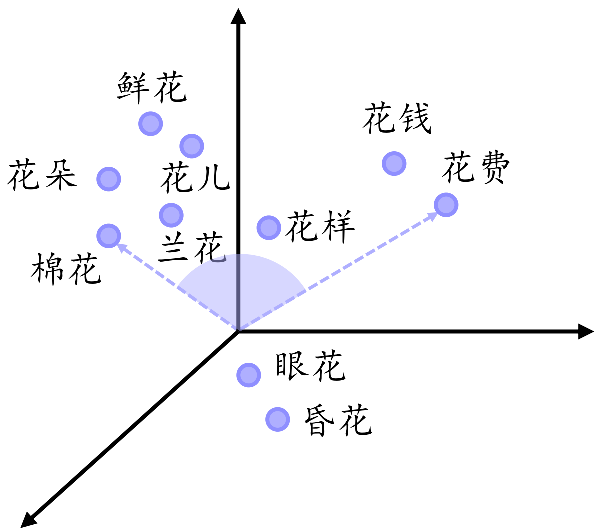


Computed metric

- **Computed dissimilarity of meanings (cDoM)**
  - Probe the meanings of a character using its ***word formations***
  - Measure the ***dispersion*** of their vector representation in a ***distributional semantic space***

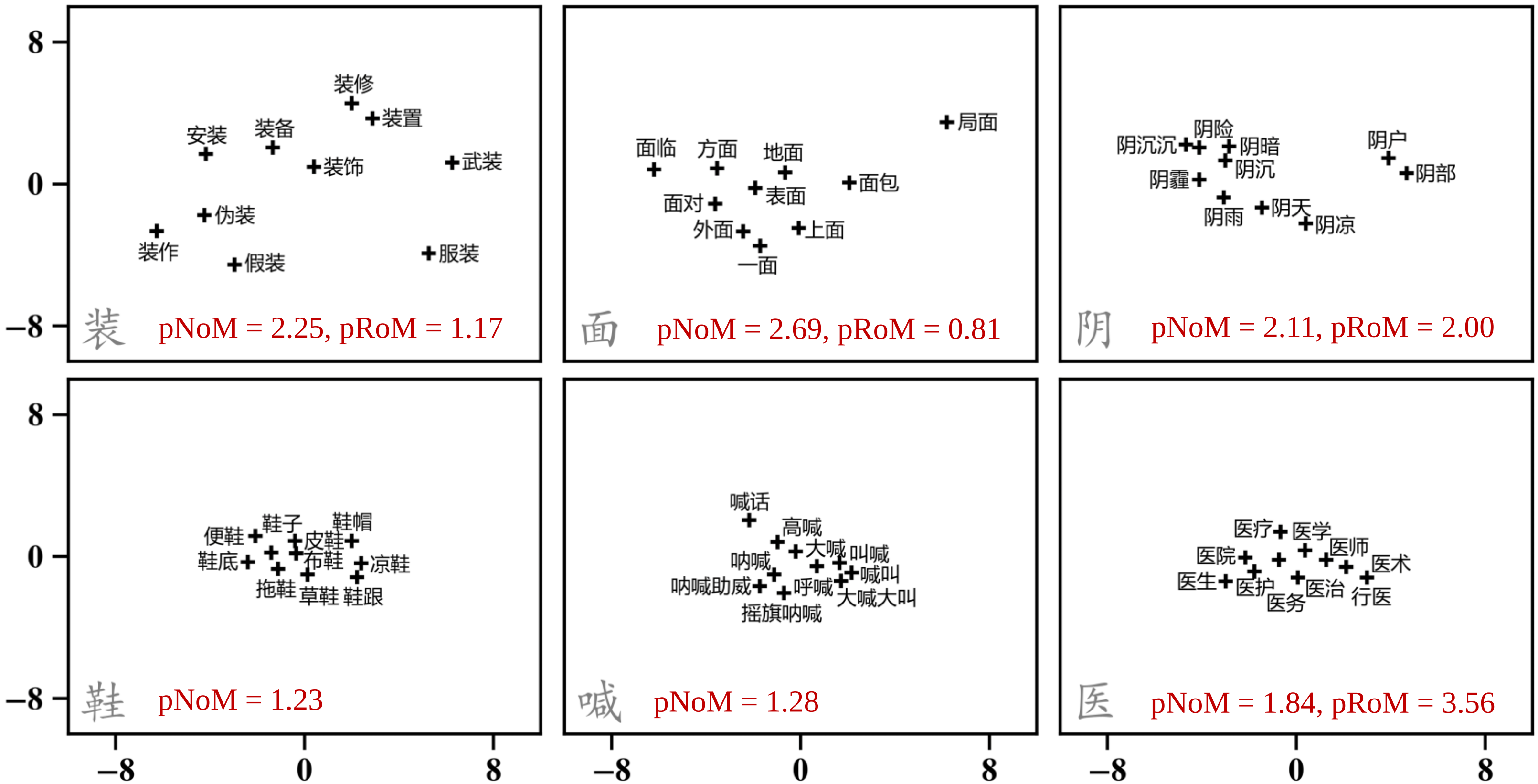
$cDoM = -\ log\ [min\ cos(\theta)]$ $\qquad cDoM \uparrow \Rightarrow$ dissimilarity between sense probes $\uparrow$

花



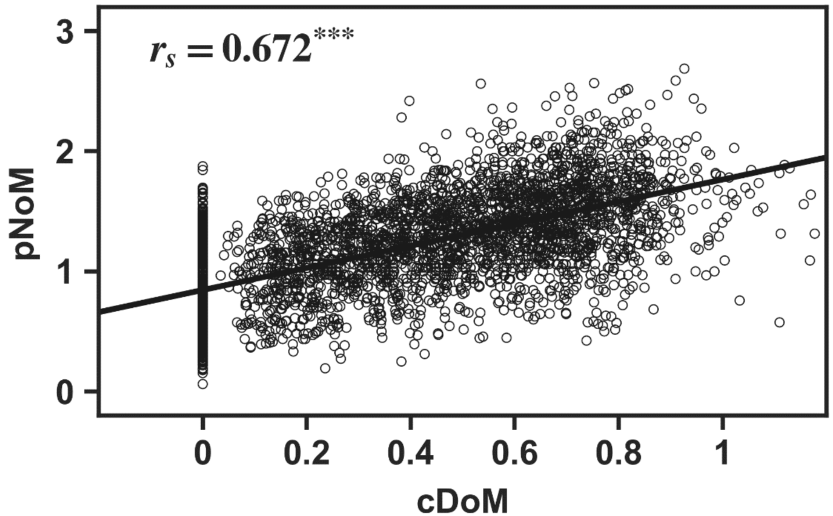
Visualization

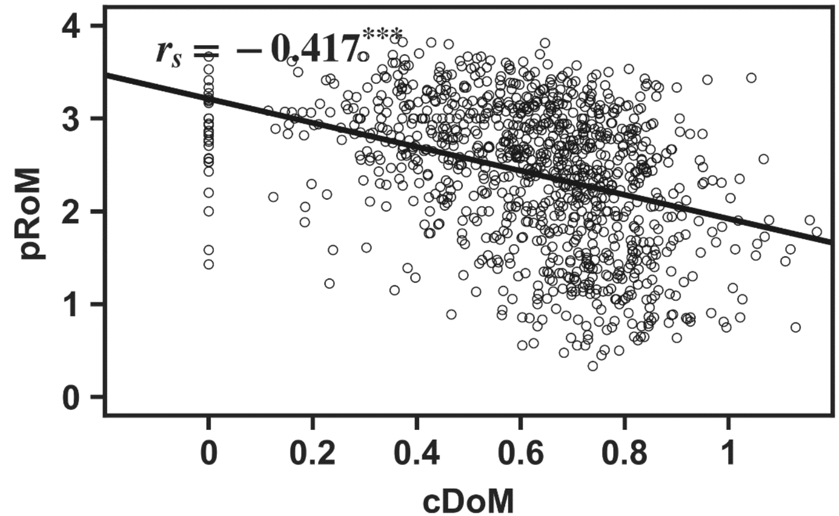
- Multidimensional scaling plots for the configuration of six example characters



Validity and efficacy of the computed metric

- Correlation with pNoM and pRoM





Validity and efficacy of the computed metric

- Correlation and partial correlation with ***number of word formation (NWF)***

	pNoM	pRoM	cDoM
logNWF	0.727^{***}	-0.123^{***}	0.792^{***}
logNWF (control: logFreq)	0.443^{***}	0.016	0.523^{***}

- Partial correlation with pNoM and pRoM

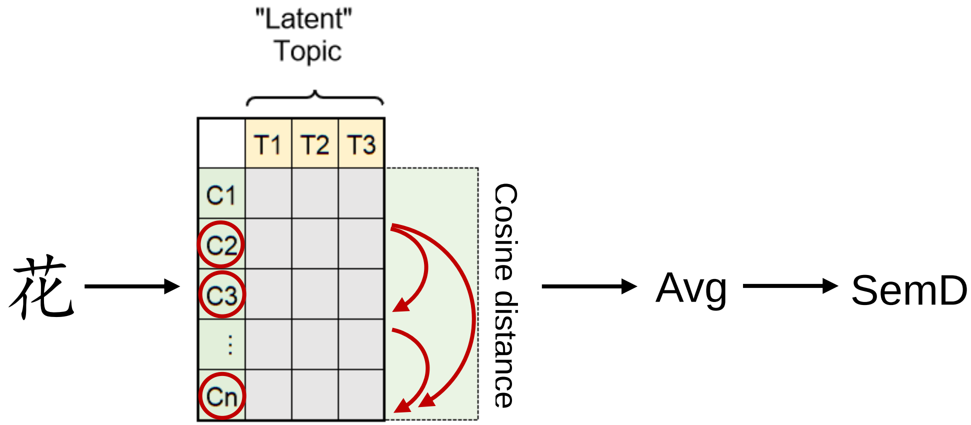
	pNoM	pRoM
cDoM (control: logNWF)	0.214^{***} ↓	-0.427^{***}

- *NWF* is the "common ground" between the pNoM and cDoM
- pRoM draws upon deeper ***conceptual analysis*** about the meanings associated with a character rather than statistical perception about the word formations of the character
- cDoM can capture the foundation of human perception about the degree to which a character would be regarded as polysemantic, and the essence of pRoM

Comparison to semantic diversity (SemD)

- **SemD** (Hoffman et al., 2013): semantic similarity of a word's different contexts

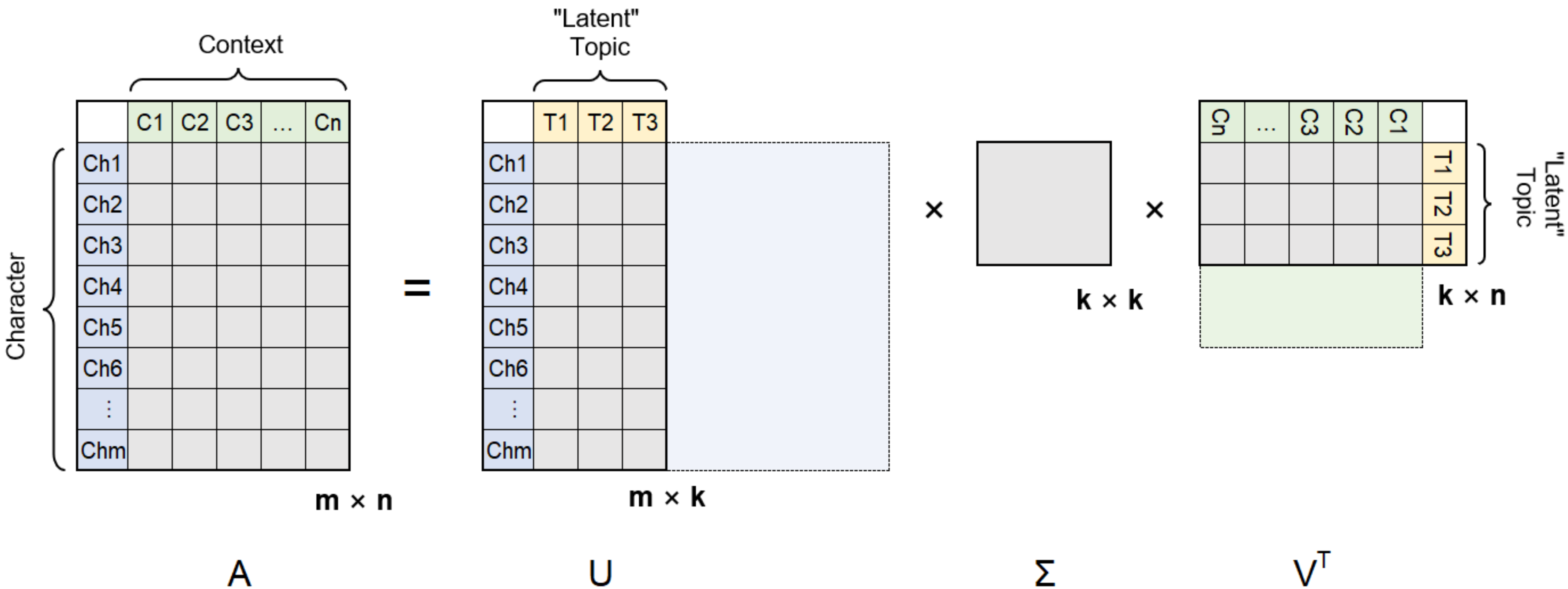
Character-based SemD



Comparison to semantic diversity (SemD)

- **SemD** (Hoffman et al., 2013): semantic similarity of a word's different contexts

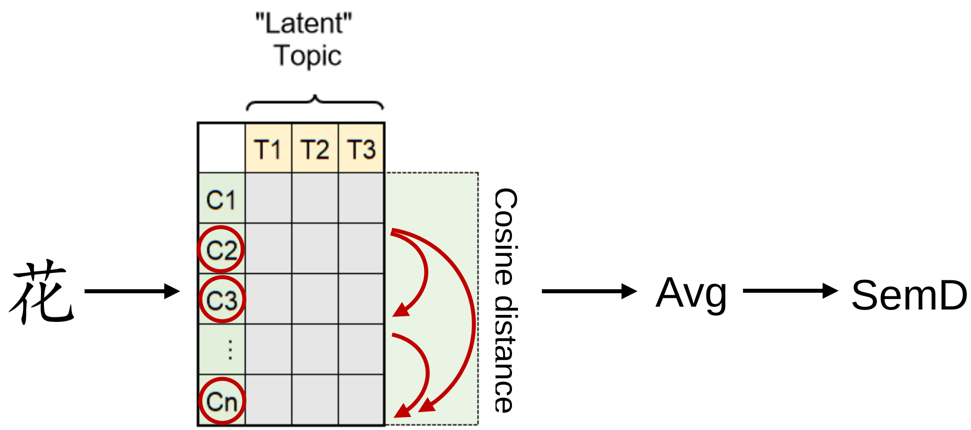
Character-based SemD



Comparison to semantic diversity (SemD)

- **SemD** (Hoffman et al., 2013): semantic similarity of a word's different contexts

Character-based SemD



- Correlation and partial correlation with pNoM and pRoM

	pNoM	pRoM
SemD	0.375***	-0.225***
SemD (Control: logNWF)	-0.007	-0.189***

Summary

- We can compute lexical ambiguity in Chinese characters based on the variability of their *word formations*
  - cDoM could inform about the ***number of the characters' meanings***
  - For characters that are definitively polysemantic, cDoM could also reflect people's conceptual knowledge about the ***relatedness between these meanings***
- cDoM reflected the *graded* nature of lexical ambiguity
- Gaps between computed metrics and human performance

Thanks !