



# Radiant suns, burning fires and brilliant flowers:

## The onomasiology and radical support of Chinese literary LIGHT ideophones

ICLC-15

15th International Cognitive Linguistics Conference

6-11 August 2019,

Kwansei Gakuin University, Nishinomiya,

Thomas Van Hoey  
National Taiwan University

國立台灣大學

# Roadmap

Goal: study the relation between the meanings and forms of Chinese ideophones in a diachronic manner

Introduction

Study 1: Sound symbolic clusters

Study 2: Manual case study of LIGHT ideophones

Study 3: Computational adaptation

Concluding thoughts

# Ideophones, a cross-linguistic concept

Ideophones are

- marked
- words
- that depict
- sensory imagery
- and which belong to an open lexical class

(Dingemanse 2011; 2012; 2019)

Japanese:

*koro koro* ころころ 'small thing rolling'

*fuwa fuwa* ふわふわ 'fluffy, puffy'

(Lu 2006:125ff.; 2011:325)

Basque:

*tipi tapa* 'walk in small steps'

*zirri zarra* 'drag clumsily'

(Ibarretxe-Antuñano 2019:152)

Ewe:

*kpakpa* 'quack (sound of duck)'

*dábòdábò* 'duck waddling'

(Westermann 1927; 1937

by Dingemanse 2017)

# Chinese ideophones

霧 茫茫

wù mángmáng

mist hazy.IDEO

"It's foggy."

狗 吠 汪汪

gǒu fèi wāngwāng

dog bark woofwoof.IDEO

"The dog is barking woofwoof."

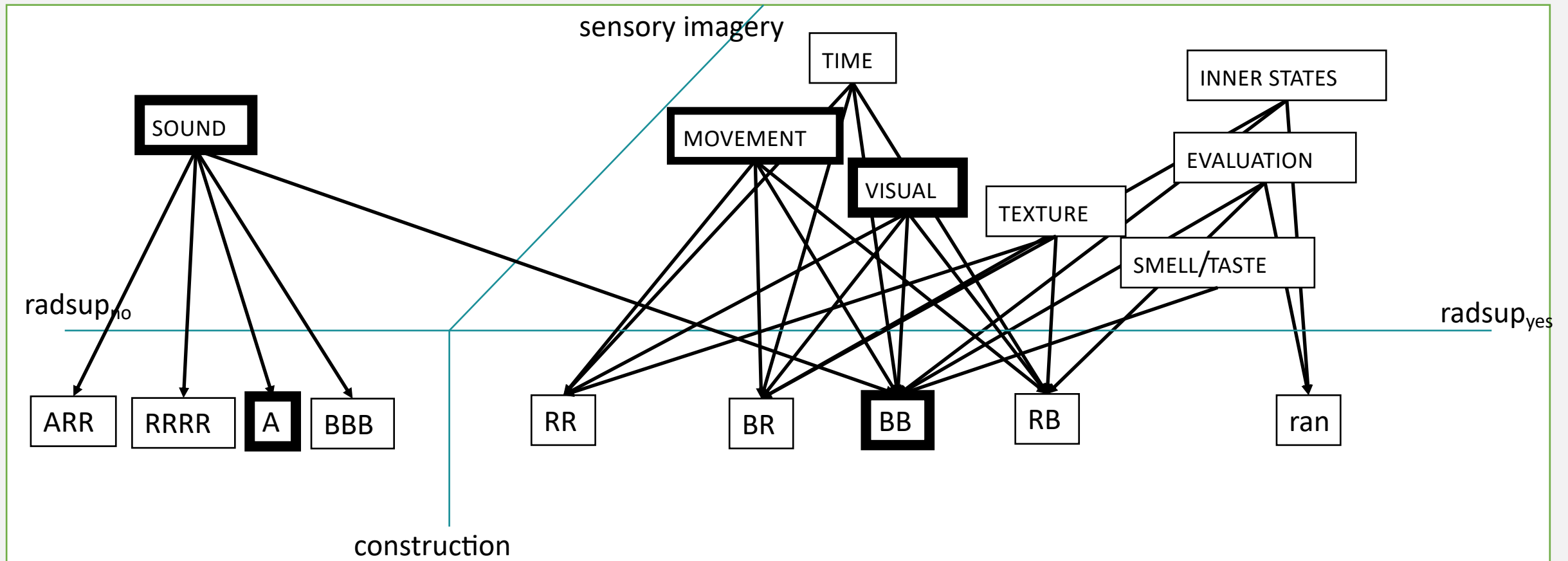


Instagram: @nickprometheus31



# Ideophones as language-particular constructions

Prototypical (Childs 1994; Akita 2009; Van Hoey 2019)



# Motivation

星-光                      熠熠  
xīng-guāng    yìyì  
start-light    twinkle.IDEO  
'twinkling star light'

Some people pronounced *zhézhé*  
but did know the meaning.

There can be a dissociation between phonological form,  
written form and meaning.



# Folk model of Chinese words

$\left[ \frac{\textit{writing}}{\textit{sound}} \mid \text{MEANING} \right]_{\Sigma}$

Folk model of Chinese based on Langacker's symbolic assemblies (1987; 1991; 2008)

Traditionally called 漢字的「形音義」 'the form-sound-meaning' of Chinese characters (Hsieh 2006)

# Folk model of Chinese words

$\left[ \frac{\textit{writing}}{\textit{sound}} \mid \text{MEANING} \right]_{\Sigma}$

$\left[ \frac{\text{熠熠}}{\textit{yìyì}} \mid \text{TWINKLING} \right]_{\Sigma}$

Folk model of Chinese based on Langacker's symbolic assemblies (1987; 1991; 2008)

Application to the previous ideophone

Traditionally called  
漢字的「形音義」 'the form-sound-meaning' of  
Chinese characters (Hsieh 2006)

# Research questions

Using the folk model as a guideline, this study aims to study the relations between these poles, especially how they evolved through time.

1. Phonological variation: are there any groupings that engage in a systematic/iconic relationship with the meaning (sound-symbolism)?

2. Semasiological variation: what does a given ideophone mean? (What does it collocate with?)

3. Onomasiological variation: what variants in the form are at play?

Study 1

Study 2

Study 3

# Study 1: Sound-symbolic clusters

(term paper)

# Study 1: Sound-symbolic clusters

For this study, which investigates the **SOUND pole vs. MEANING**, we ended up with **17 types** from the group that had an obstruent coda in Old Chinese (reconstruction Baxter & Sagart 2014), that were of the full reduplication / AA type (Van Hoey 2015; 2019)

Material came from

- Kroll's (2015) *A Student's Dictionary of Classical and Medieval Chinese*
- Ministry of Education's online dictionary (found on [zdic.com](http://zdic.com))

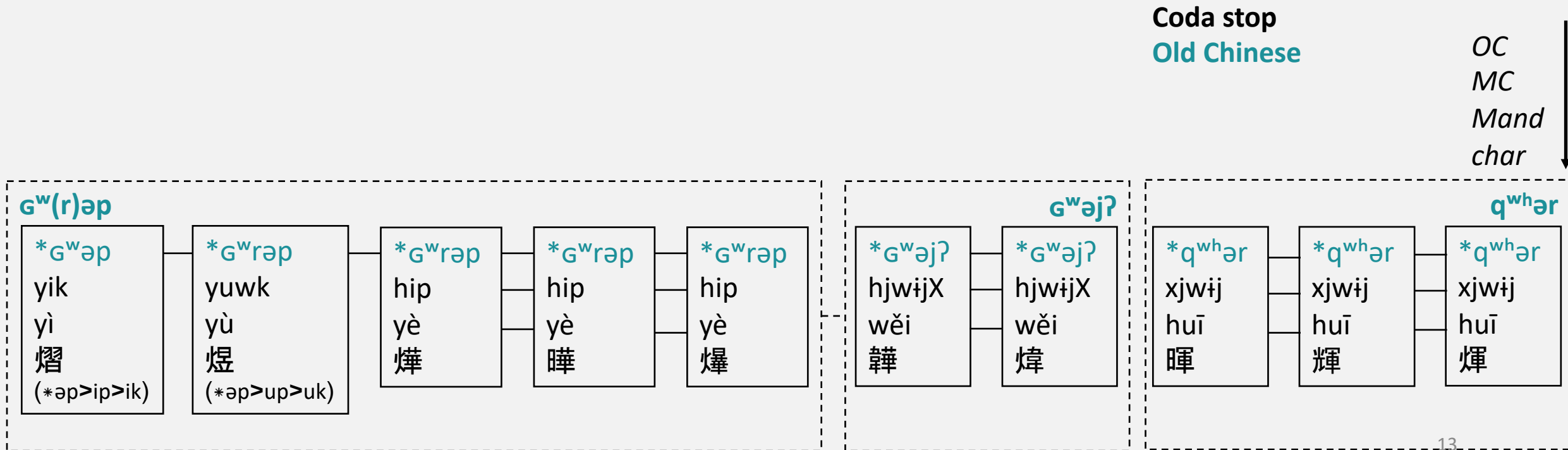
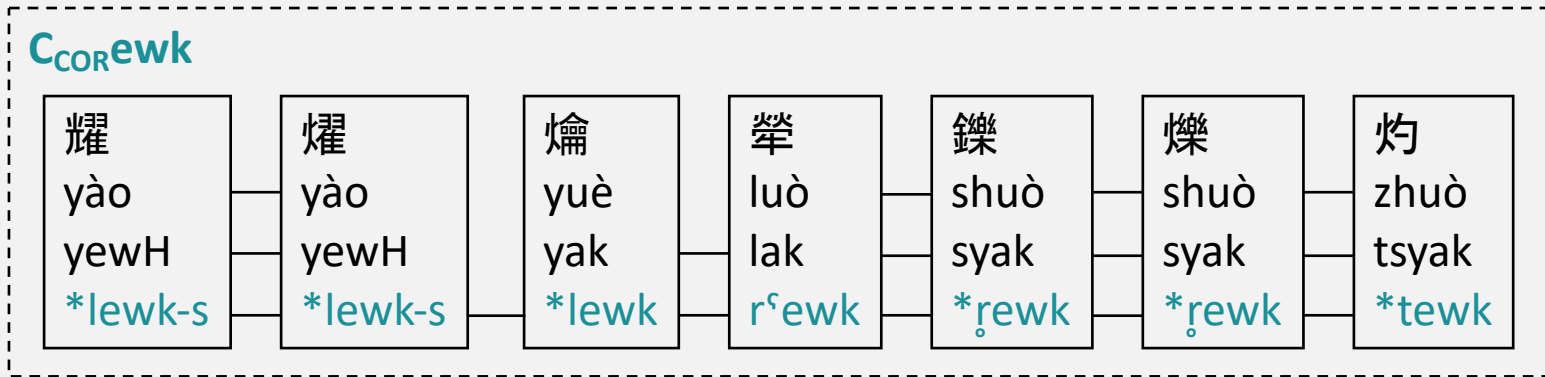
(These days I would recommend the Chinese Ideophone Database CHIDEOD, Van Hoey & Thompson 2019; in prep.)

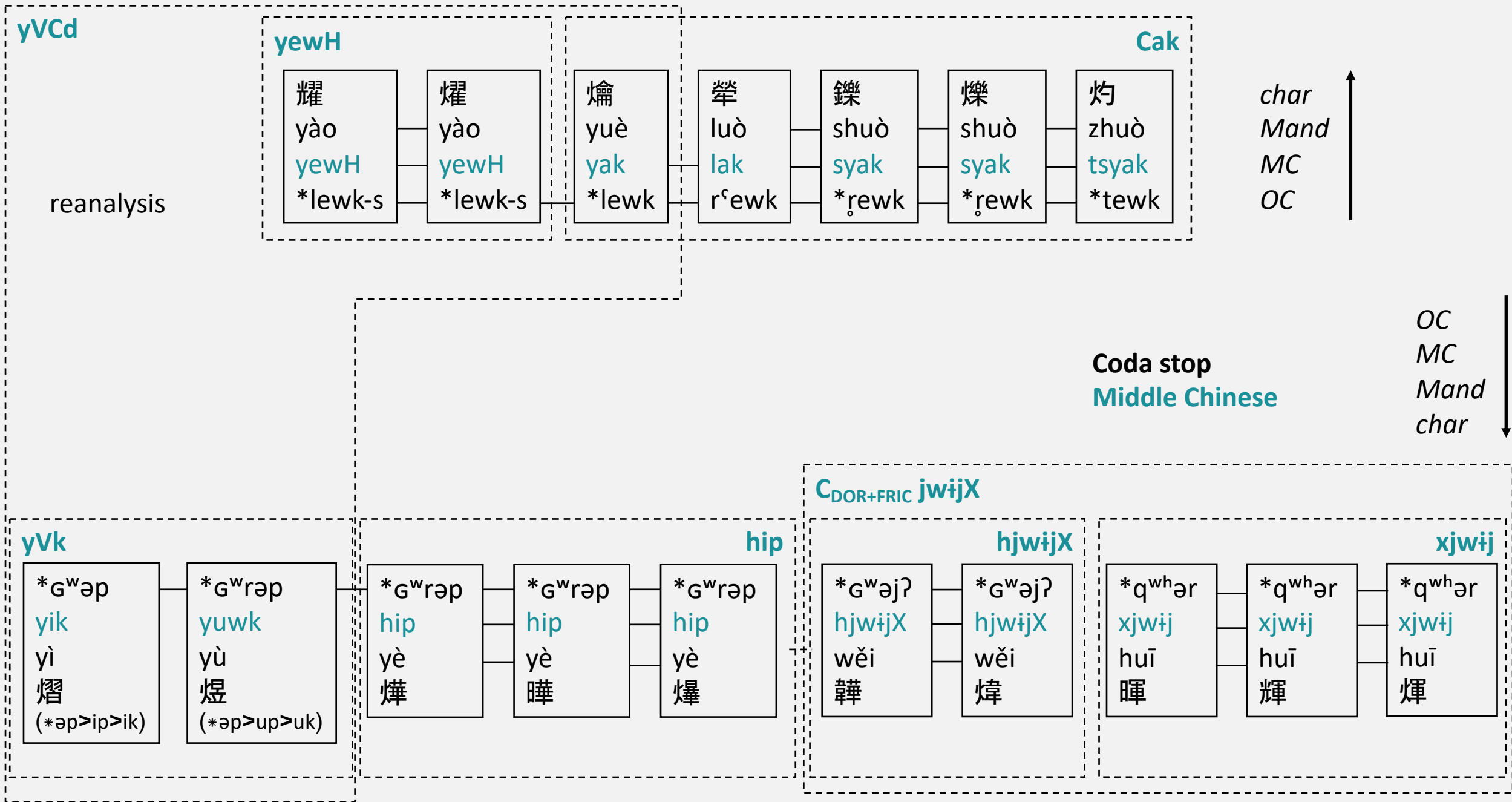
# Study 1: Sound-symbolic clusters

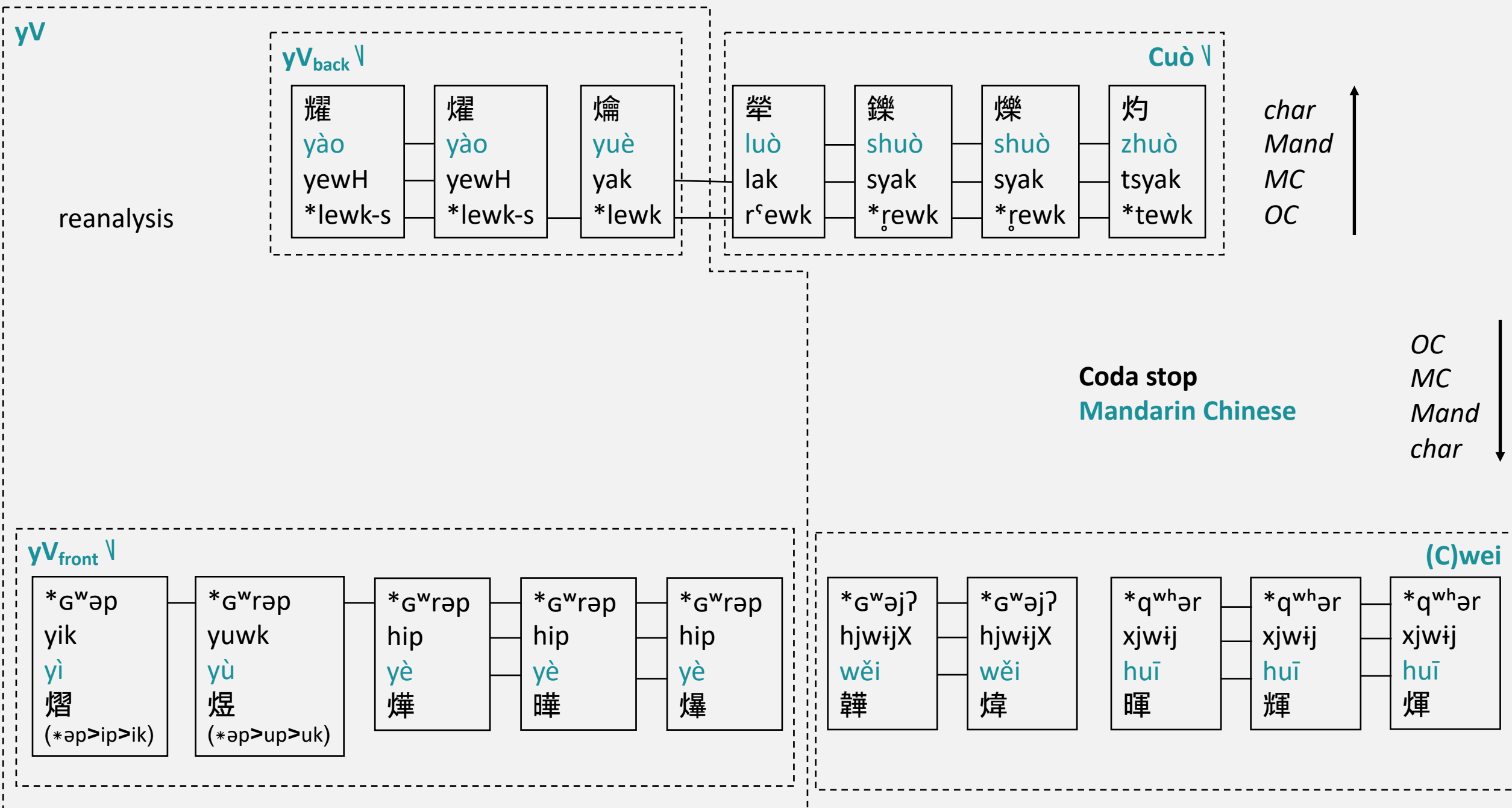
yì~yì	熠熠	làn~màn	爛熳	shuò~shuò	鑠鑠	yù~yì	煜熠
yuè~yuè	爚爚	wèi~wèi	煒煒	hù~hù	扈扈	yì~yù	熠煜
yào~yào	耀耀	wèi~yè	煒燁	shǎn~shǎn	閃閃	yè~yè	燁燁
yào~yào	耀耀	zhuò~zhuò	灼灼	zhēng~zhēng	錚錚	yè~yè	曄曄
yì~yào	熠耀	zhuò~shuò	灼爍	huǎng~huǎng	晃晃	càn~làn	燦爛
yì~yào	熠耀	hào~hào	皓皓	jīng~jīng	晶晶	càn~càn	燦燦
yù~yù	煜煜	jiǎo~jiǎo	皎皎	guāng~guāng	光光	làn~làn	爛爛
...	...	...	...	...	...	...	...

**Semantic domain of LIGHT**









yV

yV<sub>back</sub> ʌ

Cuò ʌ

reanalysis

char  
Mand  
MC  
OC

Coda stop  
Mandarin Chinese

OC  
MC  
Mand  
char

yV<sub>front</sub> ʌ

(C)wei

*G <sup>w</sup> əp yik yì 熠 (*əp>ip>ik)	*G <sup>w</sup> rəp yuwk yù 煜 (*əp>up>uk)	*G <sup>w</sup> rəp hip yè 燁	*G <sup>w</sup> rəp hip yè 曄	*G <sup>w</sup> rəp hip yè 燿
---	---	---------------------------------------	---------------------------------------	---------------------------------------

*G <sup>w</sup> əjʔ hjewijX wěi 韡	*G <sup>w</sup> əjʔ hjewijX wěi 燁	*q <sup>wh</sup> ər xjewij huī 暉	*q <sup>wh</sup> ər xjewij huī 輝	*q <sup>wh</sup> ər xjewij huī 燿
--	--	---	---	---

# Sound-symbolic clusters: conclusions

Real 'sound-symbolism' is hard to detect,

- especially through the ages
- especially with Chinese

Clusters with phonesthemes are possible

Clusters change over time

English *gl-* phonestheme

*glimmer*

*glitter*

*glisten*

...

Closed-class

(Kwon & Round 2014)

so not canonical ideophone

(Dingemanse 2019)

# Study 2: Manual analysis of semasiology and onomasiology

Presented at ICPEAL 17 – CLDC 9 (Van Hoey & Lu 2018)

# Study 2:

For same group of words looking at interplay between meanings and (written) forms

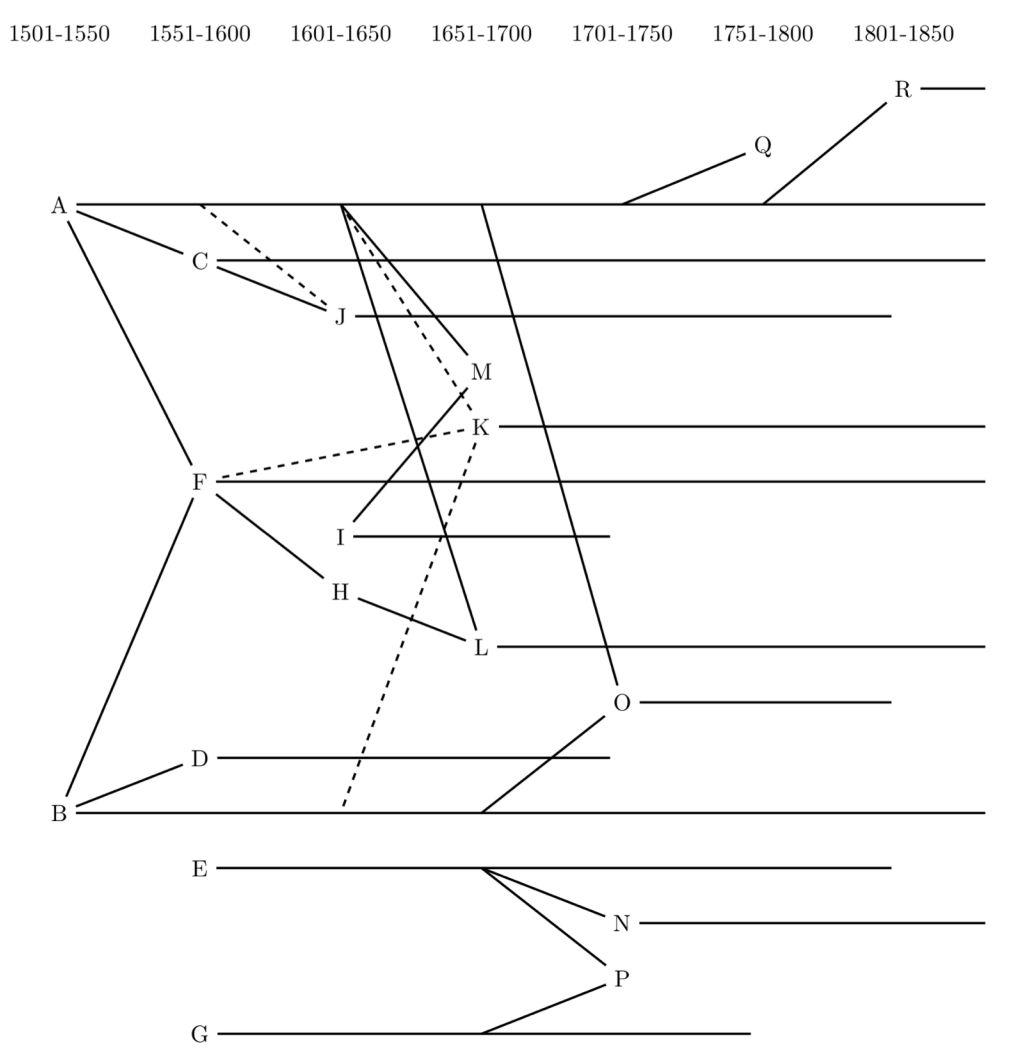
Follows the example of Geeraerts *vergrijpen*

Geeraerts (1997)

Discussion based on the four-tiered metaphor model from

Kövecses (2017)

# Diachronic prototype semantics



**'to do something wrong'**

**'to do something wrongly'**

- A: to use physical violence against (someone)**
- B: to oppose someone to whom one owes respect and obedience**
- C: to harm (someone) in a non-physical way
- D: to oppose an abstract principle
- E: to mis-take
- F: to do something forbidden
- G: to make a mistake
- H: to adulterate
- I: to do something inadvisable, unwise, improper
- J: to harm (something) in a non-physical way
- K: to steal
- L: to violate a woman's honour
- M: to eat or drink excessively
- N: to hurt while catching or seizing
- O: to rebel violently
- P: to catch the wrong person
- Q: to commit suicide
- R: to damage (something)

Geeraerts  
(1997: 47-62)

Figure 1: *Vergrijpen* (Adapted from Geeraerts 1997)

# Case study: *huīhuī*, *huīhuī*, and *huīhuī*

Item	Written	Phonological evolution	Meaning	Translation
huihui <sub>LIGHT</sub> 光 light	輝輝	huī < MC xjwɨj < OC *q <sup>wh</sup> ər	<ol style="list-style-type: none"> <li>顯赫貌。</li> <li>光耀貌。</li> <li>亮光。</li> <li>光澤，潤澤。</li> </ol>	<ol style="list-style-type: none"> <li>Illustrious</li> <li>Bright</li> <li>Light</li> <li>Glossy, sleek</li> </ol>
huihui <sub>SUN</sub> 日 sun	暉暉	huī < MC xjwɨj < OC *q <sup>wh</sup> ər	<ol style="list-style-type: none"> <li>形容日光灼熱。</li> <li>艷麗貌。</li> <li>清輝貌。</li> <li>象聲詞。</li> </ol>	<ol style="list-style-type: none"> <li>Bright sun</li> <li>Bright-coloured, beautiful</li> <li>Clear and bright</li> <li>ideophone</li> </ol>
huihui <sub>FIRE</sub> 火 fire	煇煇	huī < MC xjwɨj < OC *q <sup>wh</sup> ər	明亮貌。	bright

Based on definitions in the Hànyǔ Dà Cídiǎn 漢語大詞典, the meanings are very similar.

The phonology is also the same.

As semanticists, we want to know how these meanings evolved, extended, rather than just the dictionary ‘snapshot’.



# Methodology

## Step 0. Getting the data from corpus

Scripta Sinica (Hànjí 漢籍電子文獻資料庫):

manual copying + 'wrangling' into concordance

## Step 1. Identifying the collocates per item

huīhuī 輝輝 with LIGHT radical

"Spring clouds gently drifting, the sun blazing" 春雲澹澹日輝輝, in collection of 御定佩文齋廣群芳譜

"On a clear morning, the lantern burning brightly" 清晨輝輝燭, in collection of 御定佩文齋廣群芳譜

"Bright cinnabar" 輝輝丹, in collection of 御定佩文齋廣群芳譜

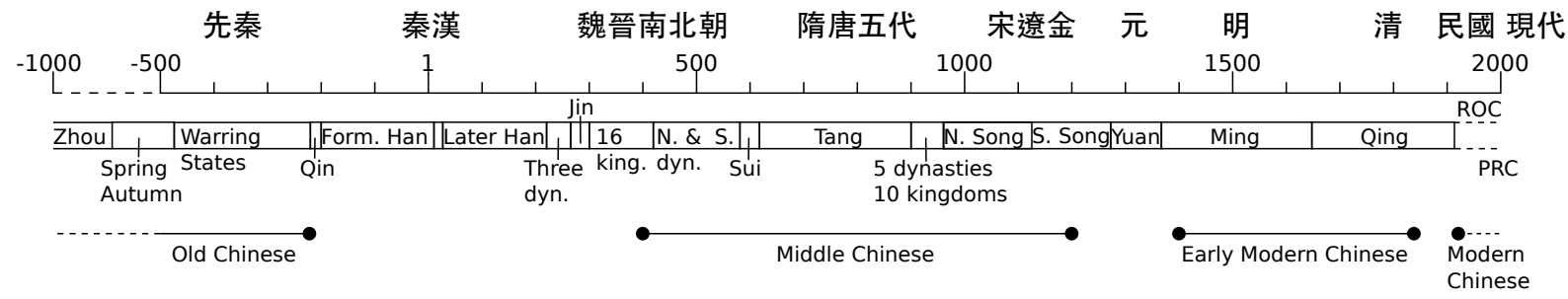
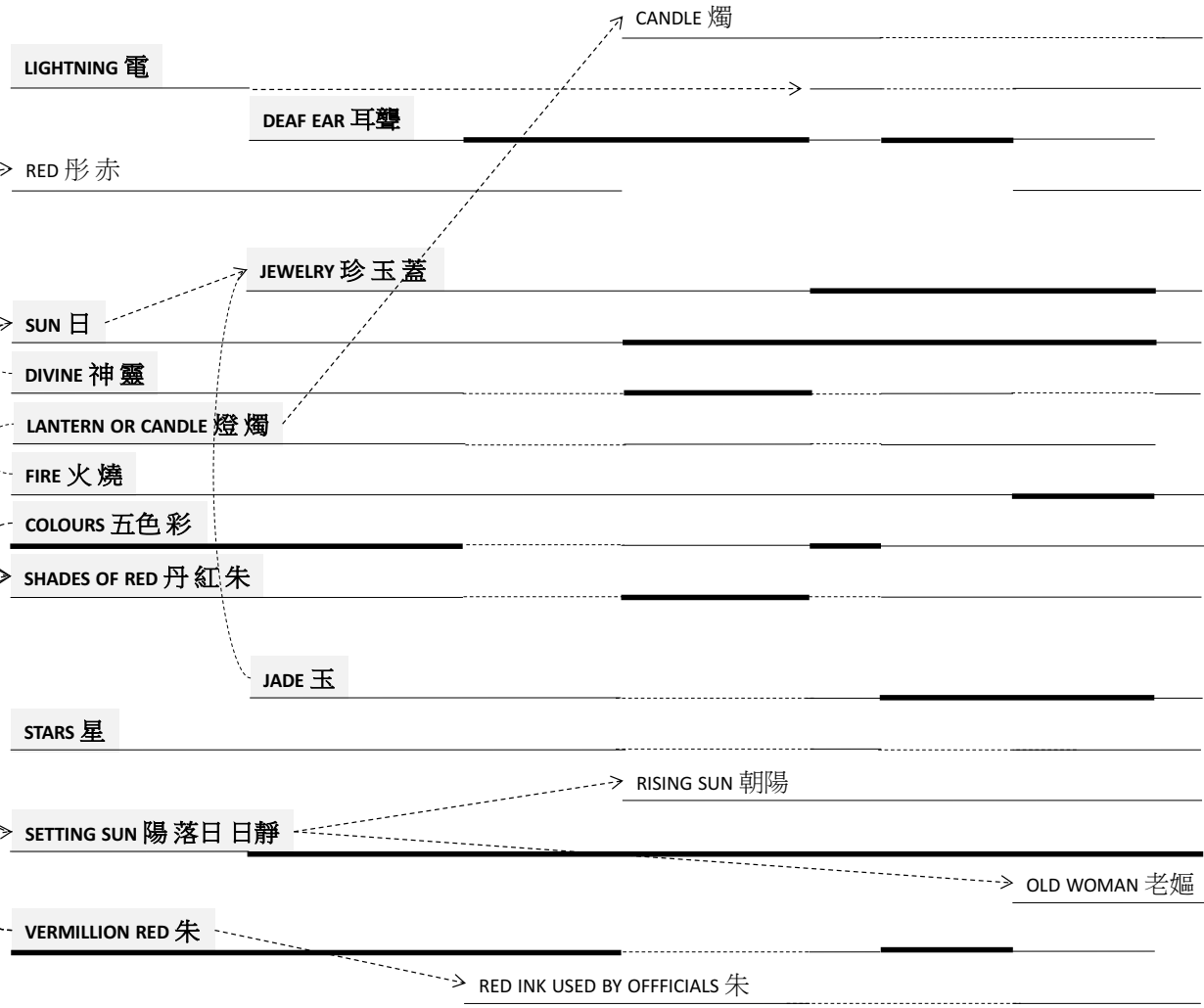
"Bright is the light, shining in the 5 colours" 輝輝有光曜五色, in 全後漢文

## Step 2. Count these collocates per period

輝輝  
FIRE

輝輝  
LIGHT

暉暉  
SUN



Vagueness / polysemy  
different shades of RED

(Referential) meaning extensions  
CANDLE

Metaphors  
OLD WOMAN

Token frequency effects

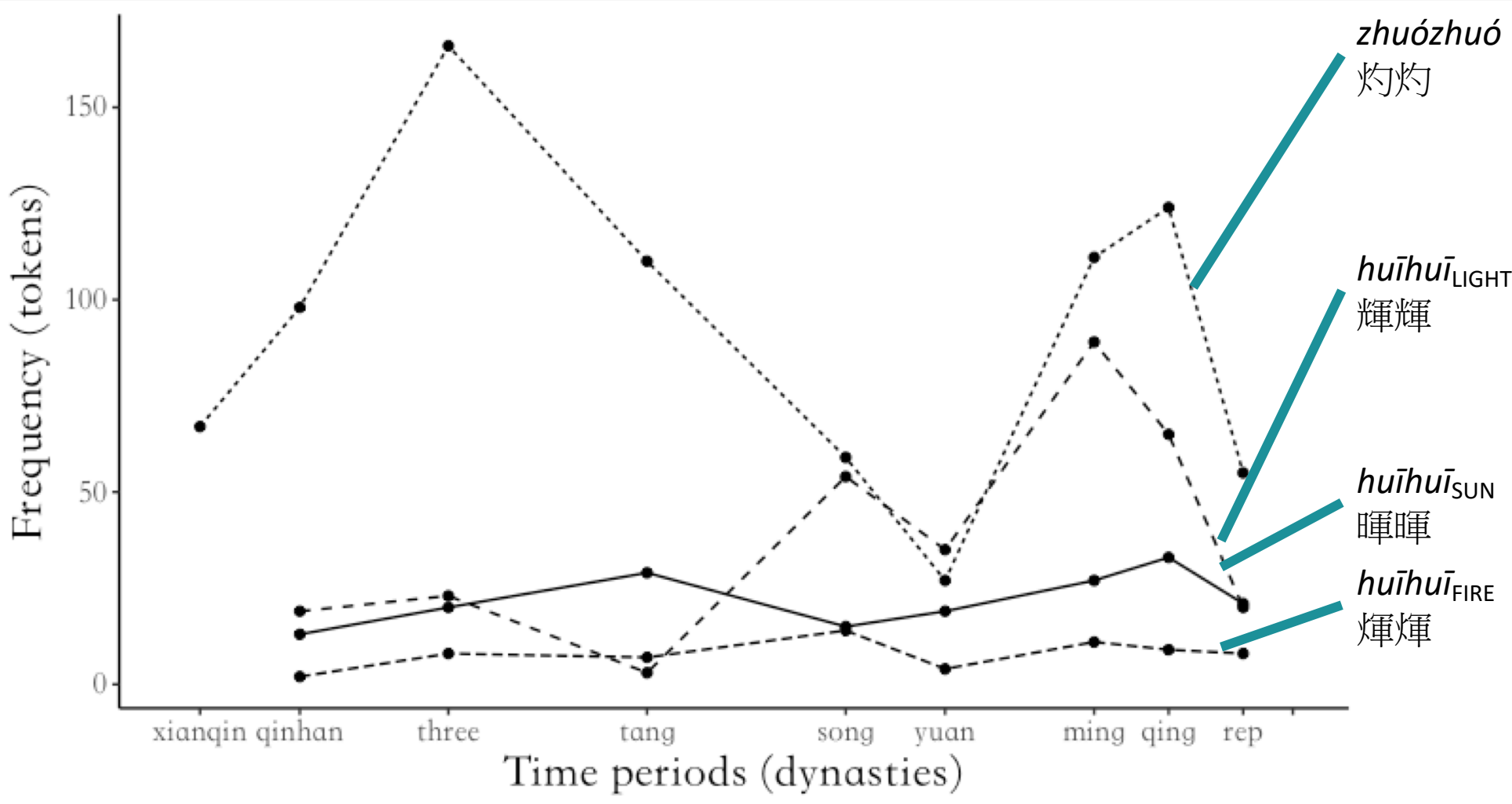
Type frequency effects

Prototypical structure

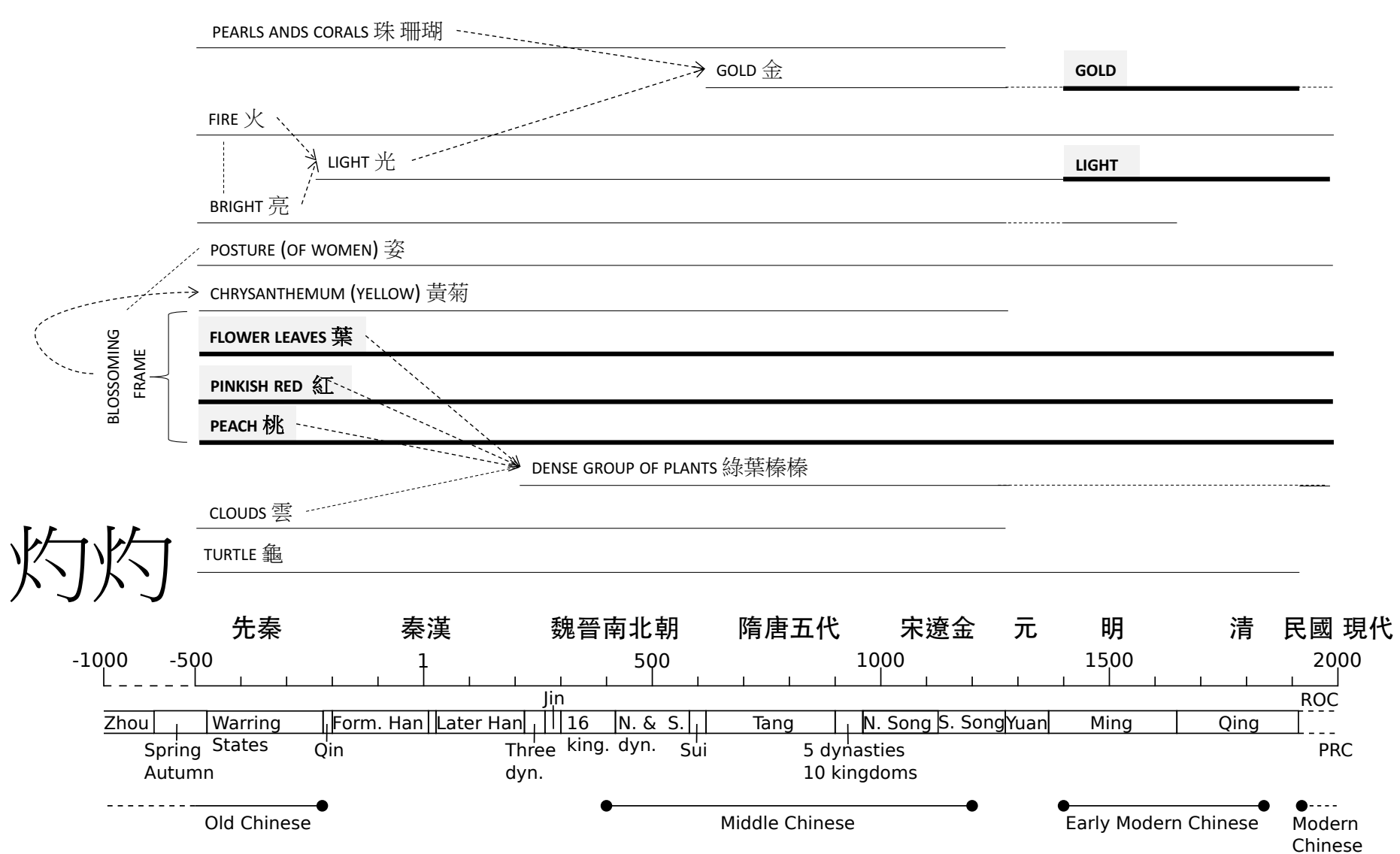
# Three *huīhuī*'s vs. *zhuózhuó*

- There is a big **type frequency** difference: *huīhuī* 輝輝 with LIGHT radical has a much higher **type frequency** in terms of different referential collocates.
- Still, they each are similar in meanings, e.g. different shades of red per ideophone.
- This difference in **type frequency** inspired an inquiry into **token frequency**.
- The ideophone with the highest token frequency in my data was *zhuózhuó* 灼灼.

# Three *huīhuī*'s vs. *zhuózhuó* token frequency



# Case study 2: *zhuó*灼灼



# Horizontal and vertical results

## Horizontal

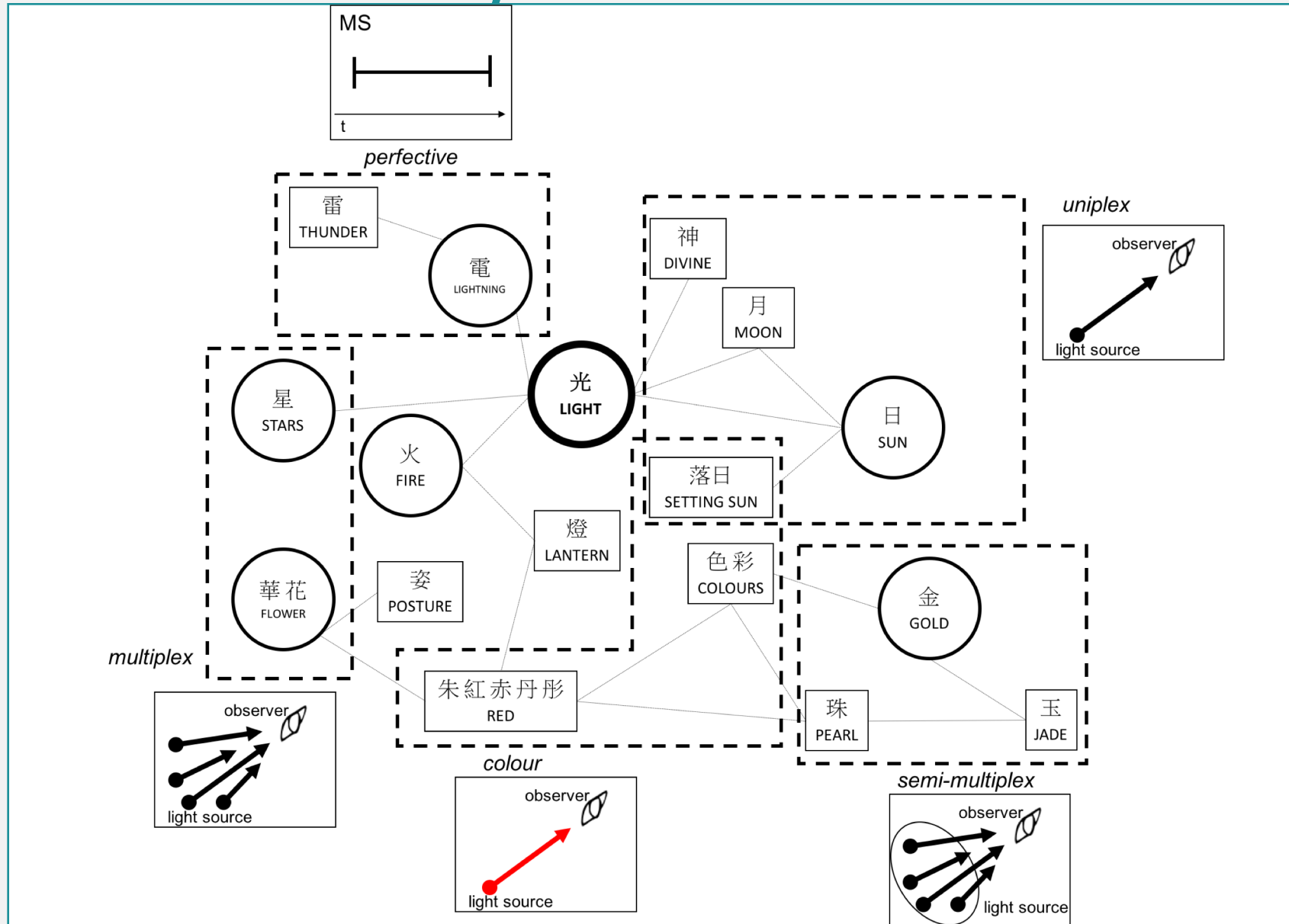
- Historical change with prototypes and frequency effects
- Mutual influence of the written form (even with a constant phonological form)

## Vertical

- **Mental spaces**      the lower level patterns, 'real data' as we go
- **Frames**              entrenched constructions, slightly bigger
- **Domains | ICMs**      more entrenched collocations
- **Image schemas**      licensed by embodiment:  
our bodies know about physics (optics)

# Domains / ICMs: collections of frames

(Kövecses 2017)



# Study 3: computational adaptation

Today's core study



# The challenge of scaling

**Main question: How can I look at more than the 17 ideophones in the sample?**

Until now, the manual methods (study 2) have yielded a great amount of insight into the dynamic semantic structure of these ideophones.

But to include more target items, I needed computational methods.

**Operationalized question:**

**How can computational methods help us better understand the semantic structure of ideophones?**

# The challenge of scaling

Methodological answer

## **Distributional semantics**

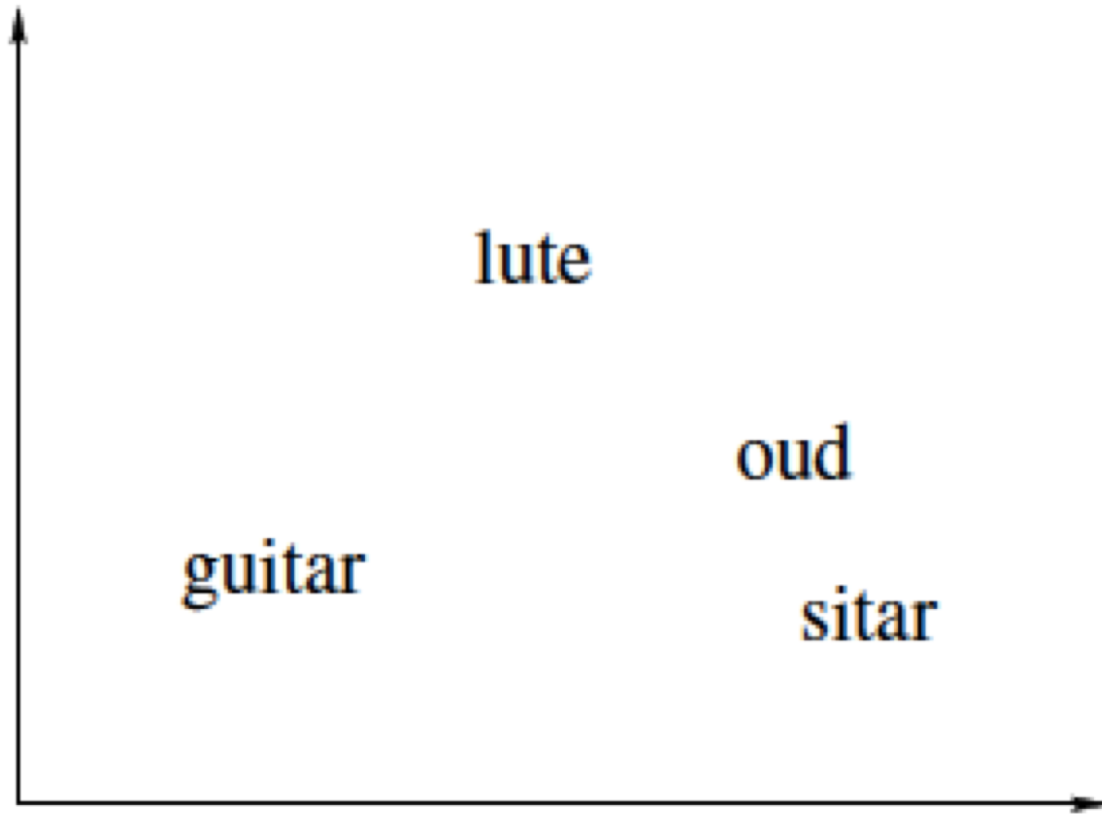
- Cognitive linguistics:

Heylen, Speelman & Geeraerts 2012; Wielfaert, Heylen & Speelman 2013; Heylen et al. 2015; Peirsman, Geeraerts & Speelman 2015, etc.

- word2vec-like approaches:

Mikolov, Yih & Zweig 2013; Mikolov et al. 2013; Goldberg & Levy 2014, etc.

# The challenge of scaling



Using conceptual metaphors:  
SIMILARITY IS PROXIMITY ;  
ENTITIES ARE LOCATIONS

Sahlgren (2006:18-19) shows how distributed semantics attempts to calculate the relative distance between different concepts

For instance: on the x-axis *oud* is mostly similar to *sitar* and then *lute*, but not to *guitar*

# Methodology

## Step 0. Gathering data

- Material: **Scripta Sinica corpus** 漢籍全文資料庫計畫
- Method: python (selenium library) to get data

## Step 1. Segmenting text

- Method: python (jieba library) + added 'dictionaries'
  - List of ideophones (CHIDEOD)
  - List of Premodern Chinese particles (Jonker, Esch & Mansvelt Beck 2011)

## Step 2. Choice of model and units

- word-based (as opposed to syntax-based or text-based)
- skip-grams and R script (cf. work done by Julia Silge through the R package tidytext)

# Methodology

## Step 3. **Frequencies and co-occurrence strength**

- calculating the pointwise mutual information PMI for every pair of words

## Step 4. **Similarity**

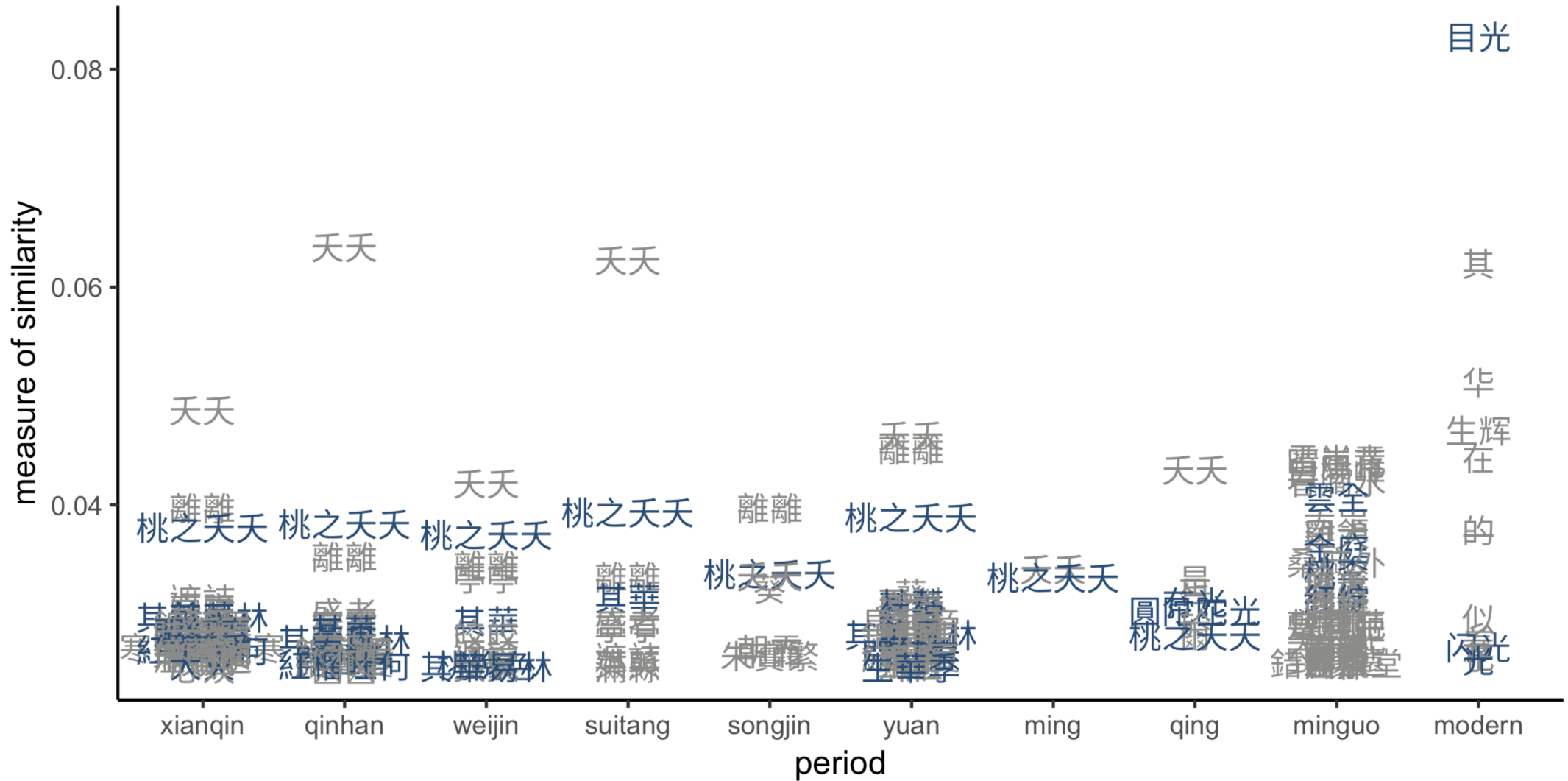
- reducing the dimensionality with singular value decomposition SVD

## Step 5. **Interpreting the results**

Now there is a long data frame ('table-like object in R'), with a score for each 'neighbour' ('calculated collocate')

This can be plotted.

# Distributional semantic nearest neighbours of zhuozhuo 灼灼



a in original a not in original

# Intermediate discussion of computational methodology

Computational methods are able to calculate similar results as with the manual analysis.

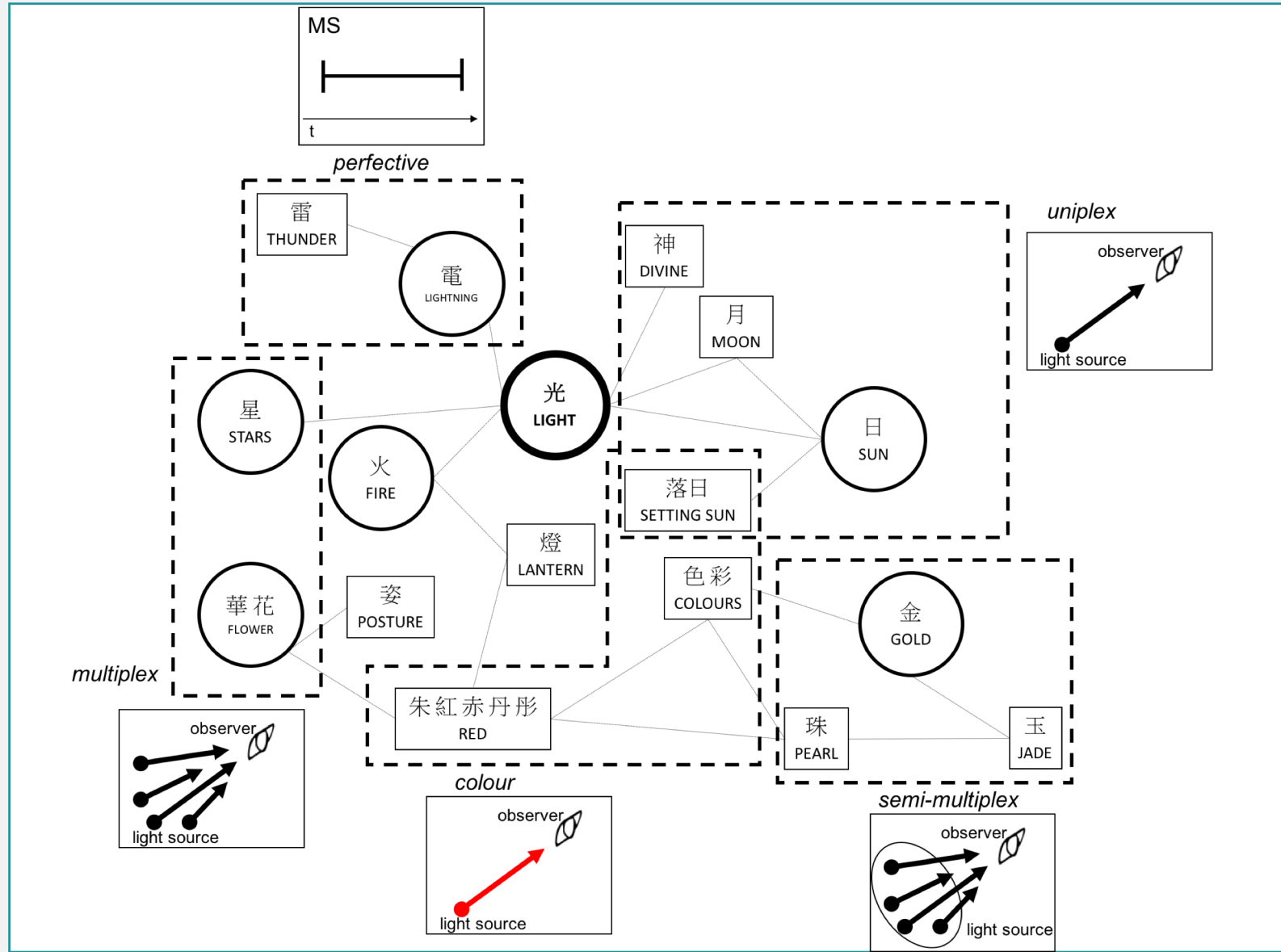
But:

- mostly higher-frequency items — if there are hapaxes or many similar scores, it can become messy
- + we now know **relative distances**

Thus, they can still aid in the analysis of higher levels of abstraction (vertical abstraction, cf. Kövecses 2017), but with relative distances!

mental spaces ← frames ← domains ← image schemas

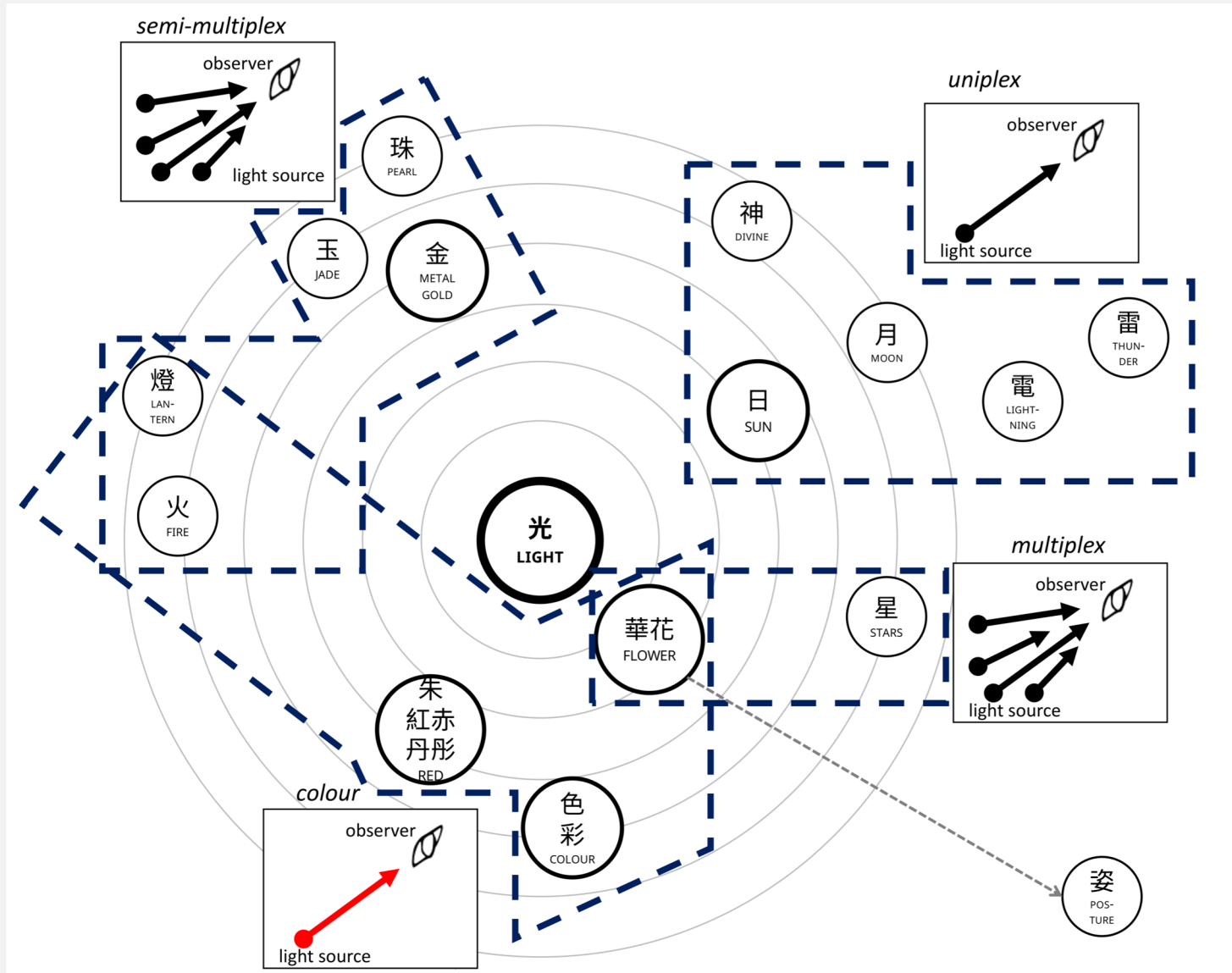
# Domains / ICMs: collections of frames



**Study 2:**  
Manual analysis of 17  
ideophones



# Domains / ICMs: collections of frames



Study 3 (current):  
Manual analysis of some  
40 ideophones

The distances from the  
core of LIGHT reflect  
conceptual distance.

This is a manually  
redrawn interpretation  
of the data.

# How do the three *huīhuī*'s fit in with this computational method?

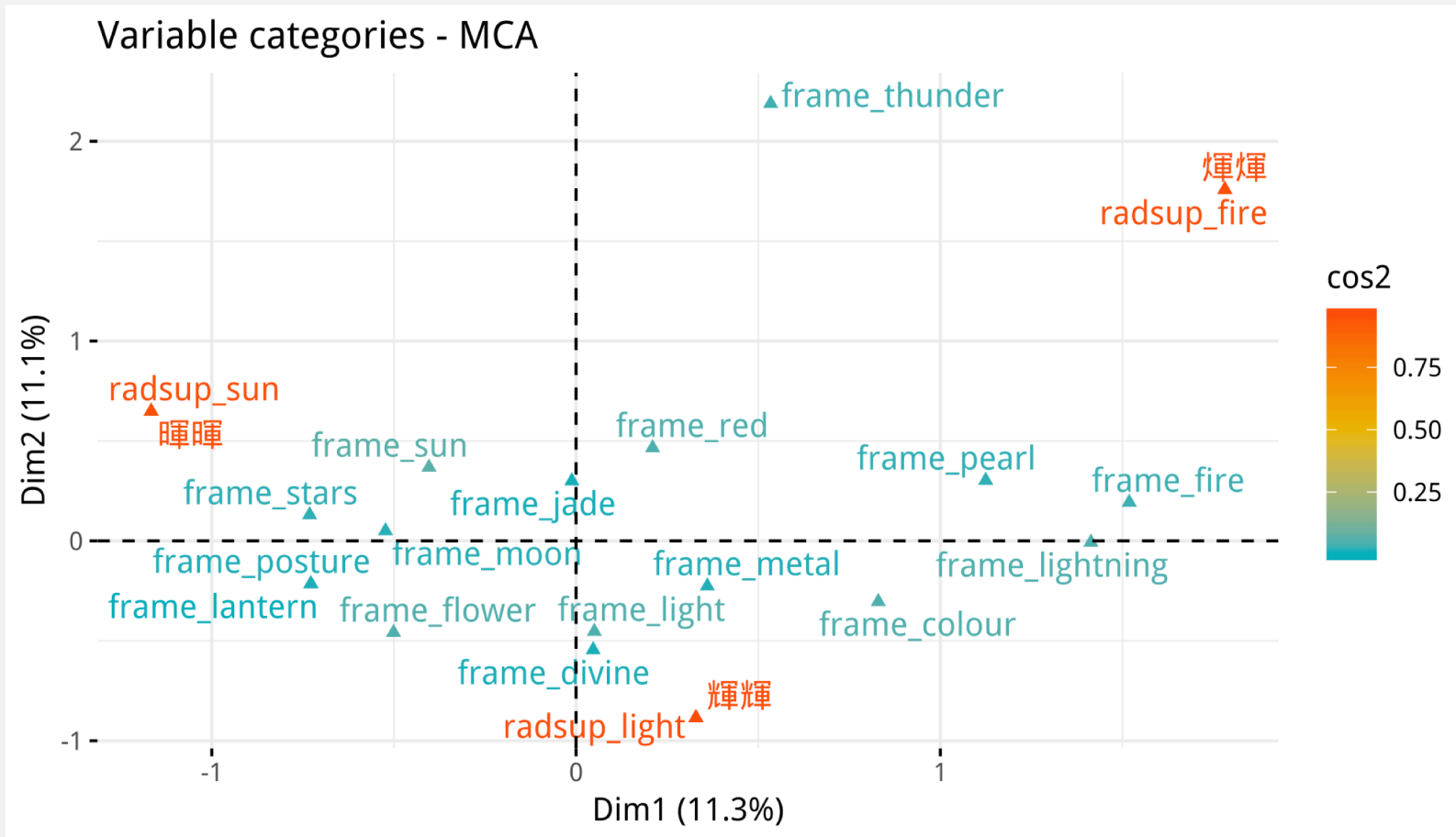
The preceding diagram leads to the following question, in relation to the case study of *huīhuī*:

**If we collapse the variable time, which frames co-occur with the three *huīhuīs*?**

In other words:

**Given that there are three different radicals (部首/偏旁 /functional components), how much do certain radicals *attract* certain frames:**

# How do the three *huīhuī*'s fit in with this computational method?

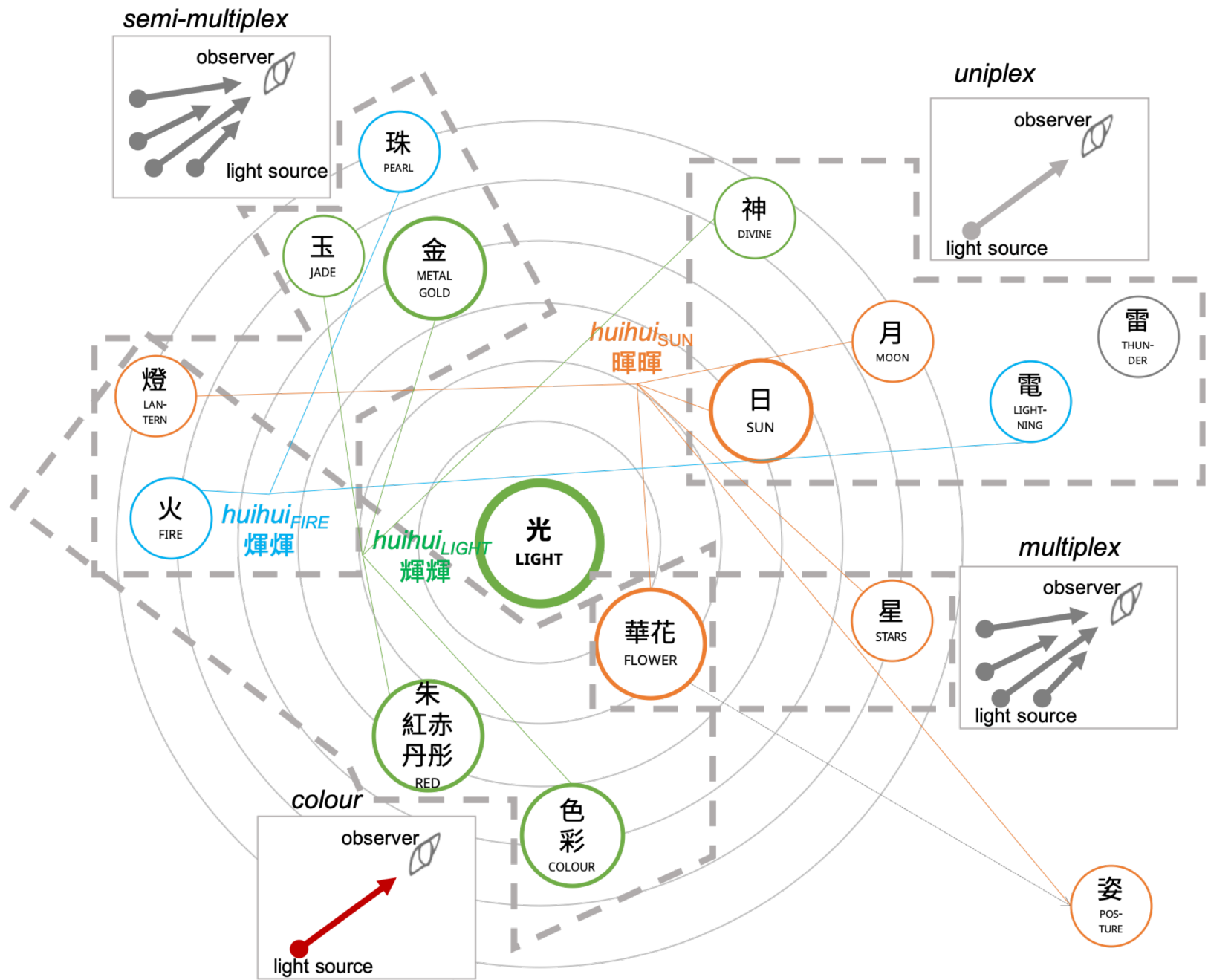


Multiple Correspondence Analysis  
(Glynn 2014; Levshina 2015)

R packages:  
FactoMineR, factoextra,  
mca, ca

Adjusted frequencies  
Dim1: 43.5 %  
Dim 2: 39.7 %  
Dim 1 + 2 = 83.2 %

# Alternative presentation



# Concluding thoughts

# Validity of computational approach

- Is good for exploratory analysis
- Can handle a lot more data more quickly
- Still needs manual check
- Also allows for many different perspectives
  - Collapsing dimensions like time
  - Or tracing the relative distances per time block  
(cf. Gries's studies on Behavioural Profiles, e.g. Gries 2006; Gries & Divjak 2009; Jansegers & Gries 2017)

# Ideophones are rampant with variation

There are many specific ideophones

(cf. case made by Akita 2013 for onomatopoeia)

But there are also ideophones that are

- highly vague / polysemous (1 form → many meaning)
- near-homophonic (1 meaning → many written forms, 1 phonological form)
- near-synonymic (1 meaning → many written & phonological form)

So there are both cases of semasiological and onomasiological variation.

# Future work

- More case studies that study the interplay between the different poles of the Chinese folk model of language
- In broader work on iconicity, these studies also look at cross-modality, but mostly focus on the interplay between the written modality and the sound modality.
- Currently mostly corpus-related, but possibilities for experimental work are becoming clearer.



Thank you!  
ご清聴ありがとうございました！

# References

- Akita, Kimi. 2009. *A grammar of sound-symbolic words in Japanese: theoretical approaches to iconic and lexical properties of mimetics* (日本語音象徴語文法：擬音・擬態語の類像的・語彙的特性への理論的アプローチ). Kobe: Kobe University PhD dissertation.
- Akita, Kimi. 2013. Constraints on the semantic extension of onomatopoeia. *Public Journal of Semiotics* 13(1). 21–37.
- Baxter, William Hubbard & Laurent Sagart. 2014. *Old Chinese: a new reconstruction*. Oxford ; New York: Oxford University Press.
- Childs, G. Tucker. 1994. African ideophones. In Leanne Hinton, Johanna Nichols & John J. Ohala (eds.), *Sound symbolism*, 178–204. Cambridge [England]: Cambridge University Press.
- Dingemanse, Mark. 2011. *The meaning and use of ideophones in Siwu*. Nijmegen: Radboud University Nijmegen dissertation.
- Dingemanse, Mark. 2012. Advances in the cross-linguistic study of ideophones. *Language and Linguistics Compass* 6(10). 654–672.
- Dingemanse, Mark. 2017. Beyond bouba and kiki. *CLS-MPI Iconicity Focus Group Workshop: Types of Iconicity in Language Use, Development and Processing*. Nijmegen: Max Planck Institute for Psycholinguistics.
- Dingemanse, Mark. 2019. “Ideophone” as a comparative concept. In Kimi Akita & Prashant Pardeshi (eds.), *Ideophones, mimetics and expressives* (Iconicity in Language and Literature, ILL 16), 13–33. Amsterdam : Philadelphia: John Benjamins Pub. Co.
- Geeraerts, Dirk. 1997. *Diachronic prototype semantics: a contribution to historical lexicology* (Oxford Studies in Lexicography and Lexicology). Oxford ; New York: Clarendon Press ; Oxford University Press.
- Glynn, Dylan. 2014. Correspondence analysis: Exploring data and identifying patterns. In Dylan Glynn & Justyna A. Robinson (eds.), *Corpus methods for semantics: Quantitative studies in polysemy and synonymy* (Human Cognitive Processing volume 43), 307–341. Amsterdam ; Philadelphia: John Benjamins Publishing Company.
- Goldberg, Yoav & Omer Levy. 2014. word2vec explained: Deriving Mikolov et al.’s negative-sampling word-embedding method. *arXiv:1402.3722 [cs, stat]*. <http://arxiv.org/abs/1402.3722>.

# References

- Gries, Stefan Th. 2006. Corpus-based methods and cognitive semantics: The many senses of *to run*. In Stefan Th. Gries & Anatol Stefanowitsch (eds.), *Corpora in cognitive linguistics: corpus-based approaches to syntax and lexis* (Trends in Linguistics. Studies and Monographs 172), 57–99. Berlin ; New York: Mouton de Gruyter.
- Gries, Stefan Th. & Dagmar Divjak. 2009. Behavioral profiles: A corpus-based approach to cognitive semantic analysis. In Vyvyan Evans & Stephanie Pourcel (eds.), *New directions in cognitive linguistics* (Human Cognitive Processing v. 24), 57–75. Amsterdam ; Philadelphia: John Benjamins Pub. Co.
- Hàndiǎn 漢典. 2004. Hàndiǎn 漢典 [Chinese dictionary]. <http://www.zdic.net/> (24 May, 2018).
- Heylen, Kris, Dirk Speelman & Dirk Geeraerts. 2012. Looking at word meaning: An interactive visualization of semantic vector spaces for Dutch synsets. *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH* 16–24.
- Heylen, Kris, Thomas Wielfaert, Dirk Speelman & Dirk Geeraerts. 2015. Monitoring polysemy: Word space models as a tool for large-scale lexical semantic analysis. *Lingua* 157. 153–172. doi:[10.1016/j.lingua.2014.12.001](https://doi.org/10.1016/j.lingua.2014.12.001).
- Hsieh, Shu-Kai. 2006. *Hanzi, concept and computation: A preliminary survey of Chinese characters as a knowledge resource in NLP*. Tübingen: Universität Tübingen PhD dissertation.
- Ibarretxe-Antuñano, Iraide. 2019. “Ideophone” as a comparative concept. In Kimi Akita & Prashant Pardeshi (eds.), *Ideophones, mimetics and expressives* (Iconicity in Language and Literature, ILL 16), 137–166. Amsterdam : Philadelphia: John Benjamins Pub. Co.
- Jansegers, Marlies & Stefan Th. Gries. 2017. Towards a dynamic behavioral profile: A diachronic study of polysemous sentir in Spanish. *Corpus Linguistics and Linguistic Theory* 0(0). doi:[10.1515/cllt-2016-0080](https://doi.org/10.1515/cllt-2016-0080).
- Jonker, Dirk R., Daan J.J. van Esch & Burchard J. Mansvelt Beck. 2011. *Wenyan partikels*. Leiden: Stichting Shilin.
- Kövecses, Zoltán. 2017. Levels of metaphor. *Cognitive Linguistics* 28(2). doi:[10.1515/cog-2016-0052](https://doi.org/10.1515/cog-2016-0052). <http://www.degruyter.com/view/j/cogl.2017.28.issue-2/cog-2016-0052/cog-2016-0052.xml> (14 October, 2017).

# References

- Kroll, Paul W. 2015. *A student's dictionary of Classical and Medieval Chinese* (Handbook of Oriental Studies: Section 4 China 30). Leiden: Brill.
- Kwon, Nahyun & Erich R. Round. 2015. Phonaesthemes in morphological theory. *Morphology* 25(1). 1–27. doi:[10.1007/s11525-014-9250-z](https://doi.org/10.1007/s11525-014-9250-z).
- Levshina, Natalia. 2015. *How to do linguistics with R: data exploration and statistical analysis*. Amsterdam ; Philadelphia: John Benjamins Publishing Company.
- Lu Chiarung 呂佳蓉. 2006. *Giongo, gitaigo no hiyuteki kakuchō no shosō: ninchi gengogaku to ruikeiron no kanten kara 擬音語・擬態語の比喩的拡張の諸相——認知言語学と類型論の観点から [Figurative extensions of mimetics: A Cognitive Linguistic and typological study]*. Kyōto: Kyōto University PhD dissertation.
- Mikolov, Tomas, Wen-Tau Yih & Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 746–751.
- Peirsman, Yves, Dirk Geeraerts & Dirk Speelman. 2015. The corpus-based identification of cross-lectal synonyms in pluricentric languages. *International Journal of Corpus Linguistics* 20(1). 54–80. doi:[10.1075/ijcl.20.1.03pei](https://doi.org/10.1075/ijcl.20.1.03pei).
- Sahlgren, Magnus. 2006. *The word-space model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Stockholm: Stockholm University PhD dissertation.
- Van Hoey, Thomas. 2015. *Ideophones in Middle Chinese: A typological study of a Tang dynasty poetic corpus*. Leuven: KU Leuven Master thesis.
- Van Hoey, Thomas. 2017. Defining Chinese ideophones: A family of constructions. *International Workshop on Mimetics (Ideophones, Exxpressives) III: Crucibles of Mimetics*. Nagoya: Nanzan University.
- Van Hoey, Thomas & Chiarung Lu. 2018. All that glitters is not gold: Prototypical semantic change in shiny Literary Chinese ideophones. Oral presentation presented at the ICPEAL 17 [International Conference on the Processing of East Asian Languages]-CLDC 9 [Conference on Language, Discourse, and Cognition]. 19-21 October 2018, Taipei: National Taiwan University.
- Van Hoey, Thomas & Arthur Lewis Thompson. 2019. Bridging phonology, meaning, and written form across time: Introducing a database of Chinese literary ideophones. Oral presentation presented at the ILL 12 [International Symposium on Iconicity in Language and Literature], Lund: Lund University. <https://konferens.ht.lu.se/en/ill-12/>.
- Wielfaert, Thomas, Kris Heylen & Dirk Speelman. 2013. Interactive visualization of semantic vector spaces for lexicological analysis. *TALN-RÉCITAL 2013* 154–166.