

國立臺灣大學文學院語言學研究所
博士論文
Graduate Institute of Linguistics
College of Liberal Arts
National Taiwan University
Doctoral Dissertation



漢語擬聲（態）詞的原型與顯著特徵：

以認知與語料庫語言學方法探討

Prototypicality and salience of Chinese ideophones:
A cognitive and corpus linguistics approach

司馬智

Thomas Van Hoey

指導教授：呂佳蓉博士

Advisor: Chiarung Lu, Ph.D.

中華民國 109 年 8 月

August, 2020



國立臺灣大學博士學位論文
口試委員會審定書

漢語擬聲（態）詞的原型與顯著特徵：以認知與語料庫語言學方法探討

Prototypicality and salience of Chinese ideophones: A cognitive and corpus linguistics approach

本論文係司馬智君（D04142001）在國立臺灣大學語言學研究所完成之博士學位論文，於民國 109 年 5 月 25 日承下列考試委員審查通過及口試及格，特此證明

口試委員：

呂德慕

(簽名)

(指導教授)

鍾曉芳

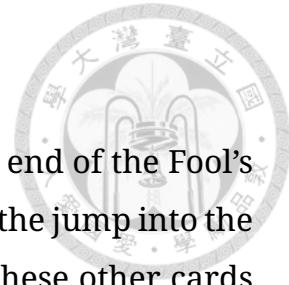
高庭明

許淑君

蔡宜仁



Acknowledgments



Finishing this dissertation is very much like coming to the end of the Fool’s Journey as it is known in tarot. The Fool needs to dare take the jump into the unknown and start his journey. On his way he meets all these other cards and archetypes, from the motherly Empress to the institutional Hierophant, to the isolating Hermit, to transformational Death, to a glimmer of hope in the Star, the glory of basking in the Sun, and finally the Universe (or the World). It takes the Fool quite a while to reach the Universe card, but when he does, he realizes that there is still so much to learn, and the journey starts again.

So too I daringly jumped into this PhD adventure and came to the Graduate Institute of Linguistics at National Taiwan University, where I met a great number of people whom I am grateful to, for allowing me to develop my ideas and research on Chinese ideophones. First and foremost I would like to thank my advisor Chiarung Lu for supporting me all of these years, and even before, resulting in my receiving of the Research Fellowship for Outstanding International Doctoral Students. I fondly treasure our time spent at conferences in Japan and Hong Kong, as well as discussions between the four classroom walls of 304. Thank you for teaching me about many of the unwritten rules of Eastern societies. まあ一～我覺得，想感謝老師！

Next there is of course my thesis committee, whose helpful comments made this dissertation better. Thank you Shu-Kai “Iakuhs” Hsieh for teaching me R and for pointing out philosophical questions that have no answer. I also greatly appreciate the use of the servers for leveling up my coding skills. Thank you Siaw-Fong Chung for the very thorough comments and suggestions of my manuscript. And lastly, thank you I-Ni Tsai and Zhao-Ming Gao for the inspiring discussion during the defense. I would also like to extend my thanks to the comments received during the proposal defense, from Chinfa Lien and Lili Chang.

Of course, I have the deepest thanks to the other professors at our institute, who have always been kind to me and interested in how I was doing

and my research. Thank you to my 蛤妮 Li-May Sung, who introduced me to the wonderful world of Austronesian languages and who sparked my interest in Isbukun Bunun. I hope the frog can be found. Thank you to: Lily I-Wen Su, for earnest discussions on constructions, such as the qǐlái 起來 construction; Wen-Yu Chiang, for diverting my attention away from poetic usage of ideophones to LIGHT ideophones; Chia-Lin Charlene Lee for the supportive chats in the hallway which always made me happy; Chenhao “Chuck” Chiu for the shared love of Kpop (BLACKPINK!); and Janice Fon, for making our joint conference of ICPEAL-CLDC in 2018 work out, as well as taking up the role of editor for the CLR edited volume that came out of CLDC in 2016. Additionally, I want to express my deepest gratitude to our institutional assistant Meiling “Merlin” Liu. Her help has been instrumental in dealing with administrative affairs, and has facilitated my research. I am equally grateful to 白小姐 for making cleanliness and order a priority in our facilities, and 劉姊 for the help with the application of reimbursements.

I am fortunate to have met a number of my personal scholarly heroes at different conferences around the world. Most influential for this dissertation, apart from the aforementioned, are Mark Dingemanse, whose insights and support I greatly appreciate; Kimi Akita, who invited me twice to Nagoya and has provided valuable feedback; Dirk Geeraerts, for introducing me to the world of lexical semantics with a strong focus on prototypes and salience; Willy Vande Walle, for instilling a lifelong passion for ideophones ever since taking his Chinese literature classes; and finally Arthur Lewis Thompson: thank you for evolving from what seemed like an improbable scholarly friendship into one that transcends the Taiwan Strait and the South China Sea. I am glad to have been able to work with you on CHIDEOD and look forward to future projects together. Oh, and I’ll gladly go with you on coffee runs! Lastly, I also feel content that the “ideophone community” is so vibrant with young scholars, such as Ian Joo and Bonnie McLean, both of whom have provided me with inspiration at crucial junctions.

Other scholars I had the pleasure of meeting include (but are not limited to): Barbara Meisterernst, Keiko Murasugi, Janis Nuckolls, Hilde De

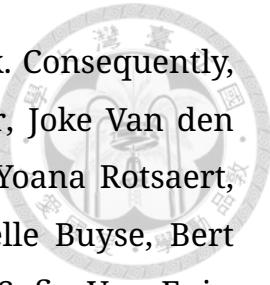
Weerdt, Fresco Sam-Sin, Youngah Do, Mutsumi Imai, Rik De Busser, Kathleen Ahrens, Chu-Ren Huang, Weiwei Zhang, etc. I would like to thank Tara Brabazon for her informative vlogs that provided the occasional support and insight. Special thanks are also due to Ann Heylen, who became a dear friend in the latter half of my candidature. And of course I would like to thank Chihkai Lin and Fuhui Hsieh for hiring me to teach English in the past year at Datung University. My students were awesome. Lastly, I want to thank Hanaivaz Takistaulan, the voice of Isbukun Bunun, for taking an interest in this daingaz hulbu. Uninang, mihumisang!

My candidature would have been a lot bleaker without my friends. First, there is the 303 crew, who I met with on a daily basis. Although the cast has changed a few times throughout the years, I will cherish our times and talks together: Yu-yun “Taco” Chang, Chester “Chestie my bestie” Hsieh, Sally Chen, Tzu-Min Li, Craig Yang, April Yeh, A-Sheng “Mergen” Chang, Hsiao-Han Wu, Miffy Lin, Chiung Yu, Ivy Chen, Freya Yeh, Brian Ke, Richard Lian, Jessica-Sharon Lin, Yolanda Chen, Jessy Chen, Ting Zeng, and of course, Lix-ing Yang and Yin-Ching Chang.

Second, I want to express my deep-felt appreciation for the other PhD members of my lab (“Lu Laoshi’s Lexicological Lab – try saying that 10 times in a row!): Alex Wei-Hao Chen, Thanh Hoa 清華 Nguyễn, Rui-Liang “Costco-Ikea” Xu, and Hui-Chun “Adrw Adrw” Chuang. Discussions during our weekly meetings have made all of us better. Thanks for your willingness to be introduced to R during the R-bootcamps.

Third, I would like to thank the other students that accompanied me on my “Fool’s Journey”: the members of my cohort, Po-Heng Chen, Iju Hsu and Hayato Saito – it is nice knowing that we are in the same boat together; Simon Shih, Thomas Schlatter, Christian Schmidt, Po-Ya Amberla Wang, Aarin Sirima, Rita Tsao, Angel Chung, Bonnie Liu, Claire Chang, Abner Chen, David Chen, Sam Fisher and Hui-Yu Chien also deserve to be thanked for interesting conversations over the years.

Fourth, there is big group of friends who came to visit me or I went to visit. They are of course also to be thanked as well, for showing that the



world is bigger and smaller than you would otherwise think. Consequently, I want to thank (in no particular order): Katrin Pletscher, Joke Van den Borre, Anneleen Paulissen, Lai Man Lung, Silke De Vos, Yoana Rotsaert, Tessa Willems, Ruben Van Dijck, Cedric Van Dijck, Isabelle Buyse, Bert Collin, Lili Vanden Wijngaert, Ewoud De Sadeleer, Ann-Sofie Van Enis, Matthijs Van Damme, Quoc Anh Tran, Kayi Tsang, Matthew Leung, Tsz Ka Wong, Mona Yip, Yuya Ishii, Mai Takeshita, Jan Matoušek, Helena Formánková, Ane Hamar, and Celia González. I fondly recall the good times we had.

Fifth, I was lucky enough to have long-distance support from numerous friends while studying abroad: Margaux Wuyts, Céline Debourse, Tahnee De Langhe, Elisabeth Erreygers, Ellen Demuynck, Dominique Van Bourgonie, Ikumi Yamamoto. I want to thank you all for checking in once in a while. I have missed you and cannot wait to catch up.

Lastly, I thank my lucky stars to have made numerous friends in Taiwan as well. So, last but not least, I would like to express deepest gratitude to Guillermo 聖威力 San Vicente, who has become one of my best friends from year 1. Also my closest group of friends, containing Greg Vondiziano, Chia-Ho Lai and of course CJ Young. Here's to our fridays! I like to think that I wouldn't have started here if I hadn't attended the summer school at NTU in 2014, where I met some wonderful people: Snow Kuo, Franzi Wang, Margaret "ntu", Antonio Sánchez among many others. A number of people were with me from the start of my candidature, like Po-Wei Li; a number I met near the end, like Andrew Black, Erin Hale and Ryan Kilpatrick Ho; and a number I met along the way, like Lucio Lu, Andy Chi, Luke Feng and Johnson Jiang.

Finally, I wish to thank my parents, Danielle Venckeleer and Marc Van Hoey. Without their everlasting support throughout the years, this journey could have gone quite differently. I have missed you, but am grateful to live in an age in which online video and audio communication is readily available. I also want to thank my family, notably Martine Venckeleer, Vincent Wuyts and Sophie Wuyts, as well as Jenny Van Hoey. Hearing from you ev-

ery so often lifted my spirits. I regret that I have had to say goodbye to three of my grandparents while I was here. I will miss them.

On a more pleasant note, I have felt very welcome in the family of Hsuan Yang, whose mother, aunt and grandparents have opened their homes and invited me in for traditional holidays. I will cherish those moments forever.

I have also felt at home-away-from-home in Neihu, where I have enjoyed spending many weekends with the Young family. Thanks to TJ, but of course most, if not all, thanks to CJ. Every since you came into my life I have felt less alone. Thank you for listening to me telling the same stories over and over, for your interest in my culture, for mastering Dutch through peak Belgian television and music, for being there for me when I needed you, and for being my own personal coffee barista – the gods know that when coffee goes in, research comes out.

I want to give a big shout out to my mom, CJ and Arthur, for helping me check typos in the manuscript. Of course, any errors that remain, are my own.

Finally finally, I am dedicating this dissertation to my grandfather Theo “Thoth” Venckeleer. I hope I made you proud.





摘要

本論文旨在研究漢語擬聲(態)詞的原型與顯著特徵：以統一一些曾經受各自研究的語言現象，如語言學重疊、聯綿詞、象聲詞，漢語含有所謂的擬聲(態)詞(ideophone)。然而，所謂的擬聲(態)詞詞彙(ideophonic lexicon)的結構並沒有同質性，而有原型性。本論文採取歷時的觀點及共時語言學的觀點，顧及到資料的多模態性，特別是書面資料，以補充擬聲(態)詞(ideophone)的類型語言學研究，如補充漢語的資料、歷時觀點、書面語的擬聲(態)詞的用法。

本研究之貢獻包含下列幾點：第一，本研究呈現出擬聲(態)詞的資料庫，名為 Chinese Ideophone Database (CHIDEOD)，v. 0.9.3 的類型頻率為 4948 筆，包含上古、中古、現代漢語的資料，以不同的格式存取得如.rds、.xlsx、.csv、R 套件與網路應用程序。第二，本論文利用 CHIDEOD 資料庫，以四個個案研究研究擬聲(態)詞詞彙的異樣性。

(一) 第一個案例研究，如何畫漢語擬聲(態)詞的邊線，以多重對應分析(multiple correspondence analysis)來摸索本類別的結構，結果證實擬聲詞與單詞素的關係最強；擬態詞與疊字的關係最強，但也更明確地表示不同變數之間的關係。後續研究顯示、語料庫資料也包含所呈現的關係。漢語擬聲(態)詞詞彙沒有一個原型中心，反而有兩個：聲音與非聲音。

(二) 第二個案例研究「光」的漢語擬聲(態)詞，以歷時原型語義學(diachronic prototype semantics)本研究呈現該語義場的心理空間(mental spaces)、框架(frames)、語義領域(domains)與其意象基模(image schemas)。「光」語義場的意思形成成群組，其特徵包含多義、互相有關、動態、包含原型中心的群組。

(三) 第三個案例研究先建立資料的向量空間模型(semantic vector space model)，接著摸索表「光」語義場的擬聲(態)詞，以三個顯著性觀點分析：語義顯著性(semasiological salience)、命名顯著性(onomasiological salience)與

結構顯著性 (structural salience)。

(四) 第四個案例研究採取構式搭配分析法 (collostructional analysis)，呈現擬聲 (態) 詞與構式之間的吸引、排斥。本研究亦論及 ABB 構式的擬聲 (態) 詞，表示 ABB 構式是更抽象「搭配擬聲 (態) 詞」構式的實例。

總而言之，四個案例研究表示漢語的擬聲 (態) 詞詞彙並沒統一性，依方法與抽象度來看卻有許多顯著性的特徵。因此，本論文提供資料與方法，以便仔細地考漢語擬聲 (態) 詞的各種特點。

關鍵詞：

擬聲詞、擬態詞、象聲詞、擬繪詞、聯綿詞、原型理論、詞彙顯著性、中文、認知語言學



Abstract



This dissertation explores prototypicality and salience effects of the variation within the Chinese ideophonic lexicon. Chinese is demonstrated to have ideophones, by unifying previously separately studied phenomena such as reduplication, binomes, and onomatopoeia. However, the “ideophonic lexicon” is not homogeneous; rather, it is prototypically structured. This is demonstrated from a synchronic and diachronic perspective, as well as across different modalities, with special attention devoted to the written modality. Thus, this dissertation aims to address the lacunae within the literature on ideophones, in which Chinese is often underrepresented, diachronic perspectives are scarce, and the ideophonic usage of writing is often neglected.

My original contributions to knowledge include (1) the creation of an open-source database of Chinese ideophones and (2) four methodological perspectives that show how the variation of and within this category is structured. The Chinese Ideophone Database (version 0.9.3) collects 4948 unique onomatopoeia and ideophones (mimetics) of modern Mandarin, as well as Middle Chinese and Old Chinese. It follows a framework that can be reused and updated in future research, and is accessible in different formats (.rds, .xlsx, .csv, R package and online app interface). Based on this database and corpus evidence, the variation of the ideophonic lexicon in Chinese is studied, in four case studies, each with its own methodological lens.

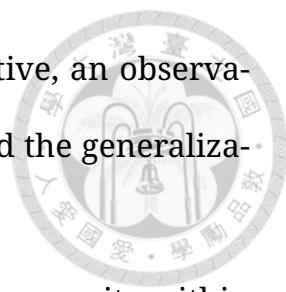
The first case study delineates the boundary of (Mandarin) Chinese ideo-

phones as a category and investigates how it is structured. Using Multiple Correspondence Analysis, the interactions between morphological patterns, orthographic motivation and depiction of sensory domain were calculated. These confirm that sound-depicting ideophones (onomatopoeia) correlate mostly with single morphemes, and that the depiction of movement and sound mostly correlates with full reduplication. However, the analysis also shows how strong other correlations between different values of the parameters are. A follow-up application of Multiple Correspondence Analysis finds that these correlations are also found with corpus data. Important in both applications, however, is the fuzzy overlap between correlations, which strongly suggests that the ideophonic lexicon in Chinese has a dual prototypical core.

The second case study investigates the diachronic prototype semantics of Chinese ideophones in the semantic field of LIGHT. Through manual study, the mental spaces, frames, domains and image schemas for a sample are followed. The meanings form interrelated polysemous clusters, which are dynamic throughout time, with clear prototypical cores that semantically extend over time and can be transient.

The third case study studies lexical variational salience within the field of LIGHT ideophones from three perspectives by constructing a semantic vector space based on a historical corpus. These perspectives are semasiological salience, onomasiological salience, and structural salience. These three types of salience show that within the semantic field of LIGHT, Chinese ideophones are not a homogeneous block. Instead they have different features

and elements that stand out, depending on one's perspective, an observation that can be extended to other types of ideophones and the generalizations that can be made about ideophones as a category.



The fourth case study continues the probing of the heterogeneity within the Chinese ideophonic lexicon, by adopting collostructional analysis to study ideophones used in Mandarin Chinese constructions. The association measures obtained through this method show to what degree individual ideophonic items are attracted or repulsed by certain constructions, and that some items also depend on these constructions to even occur. Furthermore, the well-known ABB construction is addressed and is argued to be an instance of a more schematic construction COLLOCATE-IDEOPHONE.

The case studies reveal that the Chinese ideophonic lexicon is not homogeneous, and that many different elements of salience can be found, depending on the perspective and the granularity of the analysis. They constitute an important addition to previous research by nuancing certain intuitive truths about the nature of ideophones.

Keywords:

ideophones, onomatopoeia, mimetics, prototype theory, lexical salience, Chinese, Cognitive Linguistics



Contents



Acknowledgments

i

摘要

vii

Abstract

ix

Table of Contents

xiii

List of Figures

xix

List of Tables

xxvii

Abbreviations and symbols

xxxiii

Abbreviations in glosses xxxiii

Other conventions xxxiii

Other abbreviations xxxiv

Transcription systems xxxiv

Preface

1

1 Introduction

5

1.1 Aims of this dissertation 10

1.2 Scope 17

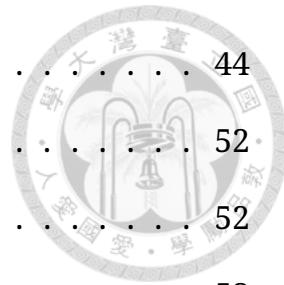
1.3 Semiotic folk model for Chinese and variation 20

1.4 Prototypicality and salience 23

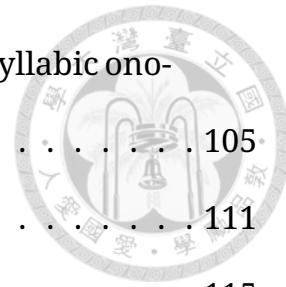
2 Background

31

2.1 Ideophones in the West 36

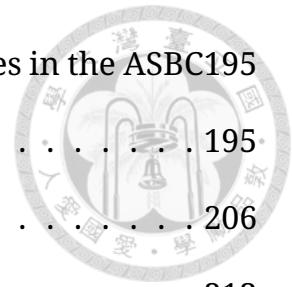


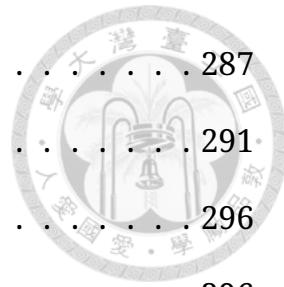
2.2 Ideophones in the East	44
2.3 Chinese approaches to expressive words	52
2.3.1 The scope of onomatopoeia	52
2.3.2 Binomes and reduplicative patterns	53
2.3.3 Character-based onomatopoeia	60
2.4 The need for Chinese ideophones as a category	66
3 Data sources	71
3.1 Dictionaries	71
3.2 CHIDEOD – the Chinese Ideophone Database	78
3.2.1 Data variables	82
3.2.2 Descriptive variables	84
3.2.2.1 Orthography	84
3.2.2.2 Phonology	87
3.2.2.3 Semantics	91
3.2.3 Analytical variables	91
3.2.3.1 Morphological template	91
3.2.3.2 Radical support and radical repetition	95
3.2.3.3 Interjection	97
3.2.3.4 Sensory imagery	99
3.2.4 Frequency variables	100
3.2.5 Other variables	104
3.2.6 Tutorial	104
3.2.6.1 Using the online app version of CHIDEOD . . .	104



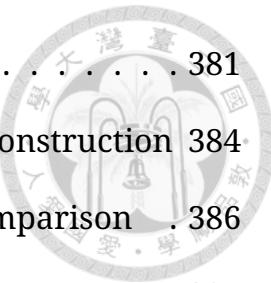
3.2.6.2	Tonal distribution of mono- and disyllabic onomatopoeia in Mandarin	105
3.2.7	Future applications of CHIDEOD	111
3.2.8	CHIDEOD in this dissertation	115
3.3	Corpora	116
3.3.1	Scripta Sinica	119
3.3.2	DIACHIC	122
3.3.3	ASBC 4.0	128
4	Defining ideophones in Chinese	131
4.1	The prototypicality of Japanese mimetics	132
4.2	The canonical ideophone	138
4.2.1	Ideophones are marked	140
4.2.2	Ideophones are words	144
4.2.3	Ideophones depict rather than describe	151
4.2.4	The meanings of ideophones pertain to sensory imagery	158
4.2.5	Ideophones belong to an open lexical class	166
4.2.6	Dingemanse's (2019) criteria in a Canonical Typological framework	168
4.2.7	Non-canonical Chinese ideophones	173
4.3	Finding the prototype with Multiple Correspondence Analysis	176
4.4	Case study 1: Ideophones in CHIDEOD	181
4.4.1	Data and feature selection	181
4.4.2	The MCA of CHIDEOD	185
4.4.3	Interim summary	193

4.5	Case study 2: Analyzing the structure of ideophones in the ASBC195	195
4.5.1	Data and feature selection	195
4.5.2	The MCA of ASBC and CHIDEOD	206
4.5.3	Interim summary	213
4.6	Conclusion	214
5	Diachronic prototype semantics	223
5.1	Introduction	223
5.1.1	Phonesthemes: one meaning for multiple forms	224
5.1.2	LIGHT syllables through time	234
5.1.3	Data of the current study	242
5.2	Methodology	243
5.2.1	Unifying three Cognitive Linguistics definitional frame- works	244
5.2.2	Diachronic prototype semantics	249
5.3	Mental spaces and Frames: corpus-based case studies	256
5.3.1	Divergent networks	257
5.3.2	Conflation of forms and semantic transfer	260
5.3.3	Frequency effects	267
5.3.4	Transient prototypicality	274
5.3.5	Types of extensions	277
5.4	Domains/ICMs and Image Schemas	280
5.5	Conclusion	282
6	Variational salience and LIGHT ideophones	287





6.1	Introduction	287
6.2	Distributional relational semantics	291
6.3	Methodology	296
6.3.1	Step 1: Segmentation	296
6.3.2	Step 2: Context models and units	297
6.3.3	Step 3: Frequencies and co-occurrence strength	303
6.3.4	Step 4: Similarity	304
6.3.5	From DIACHIC to semantic vectors for ideophones	309
6.4	Semasiological salience	311
6.5	Onomasiological salience	319
6.6	Structural salience	329
6.7	Conclusion	341
7	Constructions	345
7.1	Introduction	345
7.1.1	Zhū and Paul's adjectives	345
7.1.2	Meng's ideophonic constructions	350
7.2	Collostructional analysis	356
7.3	Collostructional analyses of ideophone constructions	362
7.3.1	Predicative constructions	365
7.3.1.1	IDEOPHONE _{PRED} DE construction	365
7.3.1.2	BARE IDEOPHONE _{PRED} construction	370
7.3.2	Adverbial constructions	374
7.3.2.1	IDEOPHONE DE _{ADV} construction	375
7.3.2.2	BARE IDEOPHONE _{ADV} construction	379

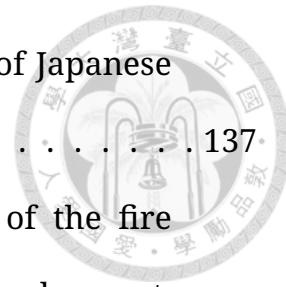


7.3.2.3 DE _{COMP} IDEOPHONE construction	381
7.3.3 Attributive constructions: IDEOPHONE DE _{ATT} construction	384
7.3.4 Collostructional analysis as a method for comparison .	386
7.4 ABB constructions and ideophones	391
7.4.1 From ABB to COLLOCATE + IDEOPHONE	391
7.4.2 Cues and the COLLOCATE + IDEOPHONE construction . .	402
7.5 Conclusion	413
8 Conclusions	417
8.1 Summary	418
8.2 Limitations and future extendability	424
References	431
Appendix 1: Data and code used in this dissertation	470
Appendix 2: Typological data concerning terminology	473
Appendix 3: Recategorizing semantic radicals	477
Appendix 4: Reproduction of selected figures	483

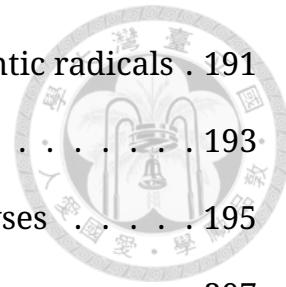
List of Figures



1.1 A: Basic semiotic model of Cognitive Grammar; B: Basic semiotic model for mimetics in Japanese (Lu 2006), and by extension Chinese	21
1.2 Taxonomy for lexicological salience (based on Geeraerts 2000)	26
2.1 Map of languages for which ideophones have been described (non-exhaustive)	44
2.2 "Western" terminology for ideophones	47
2.3 Japanese terminology for ideophones	48
2.4 Chinese terminology for ideophones	50
3.1 The variables of CHIDEOD (0.9)	81
3.2 Chinese words analyzed according to word level, character level and below-character level. Illustrated with <i>líng~líng 鈴~鈴</i>	85
3.3 App version of CHIDEOD	105
3.4 Visual representation of the periodization in the Scripta Sinica and the terms used in this dissertation.	120
3.5 The number of words per branch per dynasty in DIACHIC	128
3.6 The number of articles per year in ASBC 4.0	129
4.1 The ABAB-construction as the prototype in Japanese mimetics (Lu 2006:97)	134
4.2 Extension across the senses for the ABAB-construction (Lu 2006:100) treatment of prototypes	135



4.3 The internal structure of the prototype category of Japanese mimetics (Akita 2009:135)	137
4.4 Left panel: the gesticulations of the losing up of the fire cracker which was accompanied by the first loud report, <i>pōng</i> . Right panel: the gesticulations of the second, and louder, muffled bang, <i>pà</i> , which was accompanied by a flash. (Sam-Sin 2008:23–24)	154
4.5 Etic grid for the sensory imagery cross-linguistically depicted by ideophones	161
4.6 Semantic map for Old and Middle Chinese ideophones	162
4.7 Lattice representation of the Canonical Typological criteria in Dingemanse (2019)'s definition. M = marked, W = words, D = depiction, SI = sensory imagery, O = open class; JSL = Japanese Sign Language	172
4.8 Eigenvalues for the MCA analysis	186
4.9 MCA plot of the barycenters of the CHIDEOD data with the supplementary values (the ideophones themselves) showing. A landscape version of this figure is provided in Appendix 4 (Figure 8.1).	187
4.10 MCA plot of the barycenters of the CHIDEOD data	188
4.11 Confidence ellipses around the centroid of sensory modality .	189
4.12 Confidence ellipse around the exemplars of sensory modality .	190
4.13 Confidence ellipse around the exemplars of morphological template	190



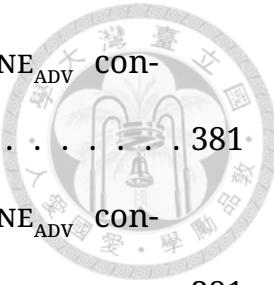
4.14 Confidence ellipse around the exemplars of semantic radicals	191
4.15 Biplot of the simple CA	193
4.16 Diagram based on previous correspondence analyses	195
4.17 Eigenvalues for the MCA analysis after correction	207
4.18 MCA plot of the barycenters of the CHIDEOD-ASBC data with the supplementary values (the ideophones themselves) show- ing. A landscape version of this figure is provided in Appendix 4 (Figure 8.2).	208
4.19 MCA plot of the barycenters of the CHIDEOD-ASBC data	208
4.20 Correlation between variables and principal dimensions	209
4.21 Confidence ellipse around the exemplars of morphological template	209
4.22 Confidence ellipse around the exemplars of sensory imagery .	210
4.23 Confidence ellipse around the exemplars of semantic radicals .	211
4.24 Confidence ellipse around the exemplars of modality	211
4.25 Confidence ellipse around the exemplars of topical class	212
4.26 Confidence ellipse around the exemplars of frequency class .	212
4.27 Diagram based on previous correspondence analyses	214
5.1 The image schema representation of <i>polang</i> (adapted from Nuckolls et al. 2017:163–168)	244
5.2 The ICM of <i>korokoro</i> (adapted from Lu 2006:133)	245
5.3 Frame semantic representation of <i>suta~suta</i> walking briskly, adapted from Kiyama & Akita (2015)	247
5.4 Levels of Metaphor (adapted from Kövecses 2013)	249



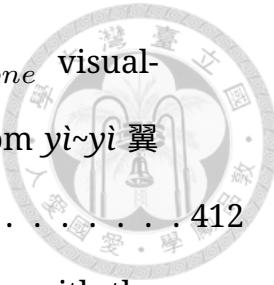
5.5	<i>Vergrijpen</i> (Adapted from Geeraerts 1997)	254
5.6	(ref:yueyue	259
5.7	<i>yàoyào</i> 燿耀 and <i>yàoyào</i> 耀耀	262
5.8	<i>shuò~shuò</i> 爍爍 and <i>shuò~shuò</i> 鑠鑠	264
5.9	<i>yì~yì</i> 熠熠 and <i>yù~yù</i> 煙煙	266
5.10	<i>huī~huī</i> 輳暉, <i>huī~huī</i> 輳暉 and <i>huī~huī</i> 暉暉	269
5.11	Token frequencies of <i>huī~huī</i> 輳暉, <i>huī~huī</i> 輳暉 and <i>huī~huī</i> 暉暉 暉 vs. <i>zhuó~zhuó</i> 灼灼	270
5.12	<i>zhuó~zhuó</i> 灼灼	273
5.13	<i>yè~yè</i> 爝暉, <i>yè~yè</i> 煙暉 and <i>yè~yè</i> 暉暉	275
5.14	Frames and Domains / ICMs	280
5.15	Image Schema of LIGHT ideophones vs. the standard folk model	282
6.1	(1) A 1-dimensional word space, (2) A 2-dimensional word space (adapted from Sahlgren 2006:18)	293
6.2	Word math	305
6.3	The number of words per dynasty in DIACHIC	310
6.4	The semasiological analysis of <i>zhuó~zhuó</i> 灼灼 (see Chapter 5) .	313
6.5	Semasiological salient collocates for <i>zhuó~zhuó</i> 灼灼	314
6.6	Comparing nearest neighbors for <i>càn~càn</i> 燦爍, <i>càn~làn</i> 燦爛 and <i>làn~làn</i> 煙燄	317
6.7	The nearest neighboring ideophones for <i>guāng</i> 光 LIGHT	321
6.8	The nearest neighboring ideophones for <i>huǒ</i> 火 FIRE	322
6.9	The nearest neighboring ideophones for <i>yuè</i> 月 MOON	323



6.10 The nearest neighboring ideophones for <i>xīng</i> 星 STARS	324
6.11 The nearest neighboring ideophones for FLOWERS: <i>huā</i> 花 and <i>huá</i> 華	326
6.12 Distribution of frames for LIGHT ideophones	332
6.13 The original frames and domains (ICMs)	333
6.14 The revised frames and domains (ICMs)	333
6.15 Distribution of frames for LIGHT ideophones per period	336
6.16 Structural salience: association measures between semantic radical and frame – <i>xianqin</i> to <i>suitang</i>	338
6.17 Structural salience: association measures between semantic radical and frame – <i>songjin</i> to <i>qing</i>	339
7.1 The iconic LIH-GFH mapping model for mimetic syntax in Japanese (Akita 2009:247)	355
7.2 The iconic “LIH”-GFH mapping model for mimetic syntax in Middle Chinese (Van Hoey 2015:84)	355
7.3 Association measures for ideophones in the IDEOPHONE _{PRED} DE construction	369
7.4 $\Delta P_{construction \rightarrow ideophone}$ in the BARE IDEOPHONE _{PRED} con- struction	373
7.5 $\Delta P_{construction \rightarrow ideophone}$ in the BARE IDEOPHONE _{PRED} con- struction	374
7.6 $\Delta P_{construction \rightarrow ideophone}$ in the IDEOPHONE DE _{ADV} construction	377
7.7 $\Delta P_{ideophone \rightarrow construction}$ in the IDEOPHONE DE _{ADV} construction	378



7.8 $\Delta P_{construction \rightarrow ideophone}$ in the BARE IDEOPHONE _{ADV} construction	381
7.9 $\Delta P_{ideophone \rightarrow construction}$ in the BARE IDEOPHONE _{ADV} construction	381
7.10 $\Delta P_{construction \rightarrow ideophone}$ in the DE _{COMP} IDEOPHONE construction	383
7.11 $\Delta P_{construction \rightarrow ideophone}$ in the IDEOPHONE DE _{ATT} construction	386
7.12 $\Delta P_{ideophone \rightarrow construction}$ in the IDEOPHONE DE _{ATT} construction	386
7.13 All constructions for <i>pái~huái</i> 徘徊 ‘waver, hesitate’	389
7.14 All constructions for <i>páng~huáng</i> 徨徨 ‘waver, hesitate’	389
7.15 Three types of ABB compositional structures	397
7.16 Simple schematic network of the COLLOCATE + IDEOPHONE construction	400
7.17 Revised schematic network of the COLLOCATE + IDEOPHONE constructions	405
7.18 $\Delta P_{ideophone \rightarrow collocate}$ and $\Delta P_{collocate \rightarrow ideophone}$ visualized from the perspective of <i>hēi</i> 黑 and from <i>qī~qī</i> 漆漆	407
7.19 $\Delta P_{ideophone \rightarrow collocate}$ and $\Delta P_{collocate \rightarrow ideophone}$ visualized from the perspective of <i>bái</i> 白 and from <i>ǎi~ǎi</i> 壴𡿵	409
7.20 $\Delta P_{ideophone \rightarrow collocate}$ and $\Delta P_{collocate \rightarrow ideophone}$ visualized from the perspective of <i>huó</i> 活 and from <i>shēng~shēng</i> 生生	411



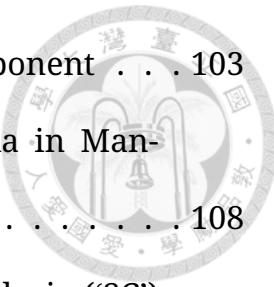
7.21 $\Delta P_{ideophone \rightarrow collocate}$ and $\Delta P_{collocate \rightarrow ideophone}$ visualized from the perspective of <i>xiǎo-xīn</i> 小心 and from <i>yì~yì</i> 翼翼	412
8.1 MCA plot of the barycenters of the CHIDEOD data with the supplementary values (the ideophones themselves) showing. Reproduction of Figure 4.9.	483
8.2 MCA plot of the barycenters of the CHIDEOD-ASBC data with the supplementary values (the ideophones themselves) showing. Reproduction of Figure 4.18.	484



List of Tables



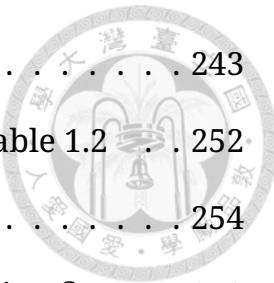
1.1	Themes in the research questions	15
1.2	Four types of prototypicality effects	24
2.1	Overview of what is to be considered <i>liánmiáncí</i>	58
2.2	Frequencies of onomatopoeic and ideophonic patterns for Mandarin, Cantonese and Hakka in Mok (2001:45-46)	61
3.1	An overview of the main sources of data used in this dissertation	71
3.2	Coverage and absolute frequency of items in CHIDEOD	84
3.3	Orthographic variables on the word and character level for <i>líng~líng</i> 鈴鈴.	85
3.4	Semantic radical and phonetic component variables for <i>líng~líng</i> 鈴鈴.	86
3.5	Examples of phonological variables for Standard Chinese ('SC') for <i>guān~guān</i> 關關 'onom. cry of the osprey'	87
3.6	Examples of historical phonological variables for <i>guān~guān</i> 關關 'onom. cry of the osprey'	89
3.7	The semantic variable #definitions for <i>guān~guān</i> 關關 'onom. cry of the osprey'	91
3.8	Morphological templates of ideophones included in CHIDEOD .	93
3.9	Examples of radical support (Van Hoey 2018a:250)	95
3.10	The top radicals in partially reduplicated items in CHIDEOD . .	97
3.11	The sensory imagery values allocated to items in CHIDEOD . .	100
3.12	Frequency variables concerning characters	103
3.13	Frequency variables concerning the semantic radical	103



3.14 Frequency variables concerning the phonetic component	103
3.15 Tonal distribution for monosyllabic onomatopoeia in Mandarin ('SC') based on CHIDEOD	108
3.16 Tonal patterns for disyllabic onomatopoeia in Mandarin ('SC') based on CHIDEOD	109
3.17 Tonal distribution for monosyllabic onomatopoeia in Mandarin from CHIDEOD and ASBC 4.0	110
3.18 Tonal distribution for disyllabic onomatopoeia in Mandarin from CHIDEOD and ASBC 4.0	110
3.19 An overview of the corpora used in this dissertation	118
3.20 The number of words per class in ASBC 4.0	130
4.1 The coverage of mimetic morphophonological templates (adapted from Akita 2009:110)	136
4.2 Modality norms for <i>yellow</i> and <i>harsh</i> (Adapted from Winter 2019:143)	163
4.3 Reported magnitude of some well-documented ideophone inventories (Dingemanse 2018:15)	166
4.4 Distribution of top radicals participating in the MCA of CHIDEOD184	
4.5 Distribution of morphological templates participating in the MCA of CHIDEOD	185
4.6 Distribution of sensory imagery participating in the MCA of CHIDEOD	186
4.7 Regression analysis of the variables and the MCA model	192



4.8 DP, DP _{norm} and corrected frequency for <i>wāng~wāng</i> 汪汪 ‘woof-woof’	199
4.9 Distribution of morphological templates participating in the MCA of ASBC	201
4.10 Distribution of sensory imagery participating in the MCA of ASBC	201
4.11 Distribution of radical support (binary variable) participating in the MCA of ASBC	202
4.12 Distribution of ASBC tags participating (Huang, Hsieh & Chen 2017) in the MCA of ASBC	203
4.13 Distribution of ASBC modalities participating in the MCA of ASBC	206
4.14 Distribution of ASBC tags participating (Huang, Hsieh & Chen 2017) in the MCA of ASBC	206
4.15 Regression analysis of the variables and the MCA model for ASBC	213
5.1 The <i>gl</i> -phonestheme according to Magnus (2001)	225
5.2 The <i>gl</i> -phonestheme according to Sadowski (2001)	226
5.3 /m/ phonestheme for items present in Baxter & Sagart (2015) .	233
5.4 Sample of 35 ideophones expressing LIGHT	234
5.5 <i>Hànyǔ dà cídiǎn</i> definitions for some LIGHT ideophones	236
5.6 Kroll (2015) definitions for some LIGHT ideophones	236
5.7 Syllable types with their Mandarin, Middle Chinese and Old Chinese reconstructions	237
5.8 Reconstructions to Old and Middle Chinese	239



5.9 Ideophone types used as material in this study	243
5.10 Four types of prototypicality effects (as shown in Table 1.2)	252
5.11 the different meanings of <i>vergrijpen</i> in Figure 5.5	254
5.12 Types of semantic extensions present in the preceding figures	278
6.1 Overview of the four positions on saliency	288
6.2 Bigrams	302
6.3 Trigrams	302
6.4 Contingency table that takes all skip-grams in (86b) into account	303
6.5 Top collocates for <i>zhuó~zhuó</i> 灼灼 per period in the DIACHIC	312
6.6 Operationalization of structural salience	330
7.1 Crosstabulation of <i>accident</i> and the [N waiting to happen] construction, adapted from Stefanowitsch & Gries (2003:219)	358
7.2 Crosstabulation of items and constructions	358
7.3 Constructions identified by Meng (2012) and their abbreviations	363
7.4 Types within the dataset as compared to Premodern and Modern data within CHIDEOD	364
7.5 Tokens within the dataset as compared to Premodern and Modern data within CHIDEOD	365
7.6 The contingency table for <i>jiàn~jiàn</i> 漸漸 in the IDEOPHONE _{PRED} DE construction	368
7.7 The contingency table for <i>máng-rán</i> 茫然 in the IDEOPHONE _{PRED} construction	371
7.8 The contingency table for <i>shēn~shēn</i> 深深 in the IDEOPHONE DE _{ADV} construction	376



7.9 The contingency table for <i>fēn~fēn</i> 紛 紛 in the BARE IDEO-PHONE _{ADV} construction	380
7.10 Collocates in the DE _{COMP} IDEOPHONE construction (n > 3)	383
7.11 The contingency table for <i>càn~làn</i> 燥爛 in the IDEOPHONE DE _{ATT} construction	385
7.12 <i>Hànyǔ dà cídiǎn</i> definitions for <i>pái~huái</i> 徘徊	387
7.13 <i>Hànyǔ dà cídiǎn</i> definitions for <i>páng~huáng</i> 徨徨	388
7.14 Affixes in Chinese (adapted from Packard 2000:174)	393
7.15 Examples of ABB (adapted from Wang (2014:352))	394
7.16 Examples of ABB (adapted from Wang 2014:353)	395
7.17 Patterns and frequencies in the COLLOCATE + IDEOPHONE construction	404
7.18 The contingency table for <i>hēi-qī~qī</i> 黑漆漆 in the COLLOCATE + IDEOPHONE construction	406
8.1 R packages used in this thesis	470
8.3 Coverage and absolute frequency of items in CHIDEOD	473
8.4 Distribution of the radicals participating in the MCA of CHIDEOD	477



Abbreviations and symbols



Abbreviations in glosses

1	first person	INCH	inchoative
2	second person	LE	Chinese particle <i>le</i> 了
3	third person	LNK	linker, often Chinese particle <i>de</i> 的
ADV	adverb(ial)	LOC	localizer
BA	Chinese particle <i>bǎ</i> 把	NEG	negative
CAUS	causative	NMLZ	nominalizer
CLF	classifier	NOM	nominative
COMP	complementizer	NPST	non-past
CONJ	conjunction	PASS	passive
COP	copula	PFV	perfective
DAT	dative	PL	plural
DE	Chinese particle <i>de</i> 的, 地, or 得	PN	personal name
DEM	demonstrative	PST	past
DET	determiner	QILAI	Chinese complement <i>qǐlái</i> 起來
DUR	durative	QUOT	quotative
EMPH	emphatic	RAN	Chinese suffix <i>rán</i> 然 'be thus'
EXCLAM	exclamative	RED	reduplication
EXIST	existential	SFP	sentence final particle
EXP	experiential	SG	singular
GEN	genitive	SHI	Chinese copula <i>shì</i> 是
IDEO	ideophone	TOP	topic
IDIOM	idiom, idiomatic expression	~	reduplicative pattern
		-	morpheme boundary
		=	clitic

Other conventions

SMALL CAPS	highlighted terms, sensory domains, conceptual metaphors and metonymies
<...>	graphemic representation; or phonological, semantic, or orthographic pole
/.../	phonemic representation
[...]	phonetic representation
# ...	variable used in CHIDEOD
', ...'	values of a variable in CHIDEOD
$\Delta P_{\text{ideophone} \rightarrow \text{collocate}}$	cue validity with ideophone as cue and collocate as response

Other abbreviations

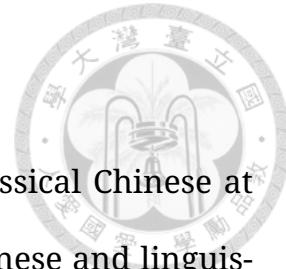
ASBC	Academia Sinica Balanced Corpus of Chinese
CA	Correspondence Analysis
CHIDEOD	Chinese Ideophone Database
CLD	Chinese Lexical Database
DIACHIC	Diachronic Chinese Ideophone Corpus (based on Scripta Sinica)
ICM	Idealized Cognitive Model
MCA	Multiple Correspondence Analysis
MC	Middle Chinese
OC	Old Chinese



Transcription systems

Standard Chinese	<i>Hànyǔ pīnyīn</i> 漢語拼音
Middle Chinese	Middle Chinese transcription (Baxter 1992)
Old Chinese	Middle Chinese transcription in (Baxter & Sagart 2014)
Japanese	JSL romanization (Jorden 1987)

Preface



Eleven years ago, I attended an introductory class to Classical Chinese at University of Leuven (Belgium), where I would study Chinese and linguistics before coming to Taiwan. The quote that marks the beginning of this introductory chapter belonged to the text that was read in class that day, a philosophical story about the ‘happiness of fish’ (yú zhī lè 魚之樂).

Zhuangzi and Huizi were strolling along the dam of the Hao River when Zhuangzi said, “See how the minnows come out and dart around where they please! That’s what fish really enjoy!”

Huizi said, “You’re not a fish—how do you know what fish enjoy?”

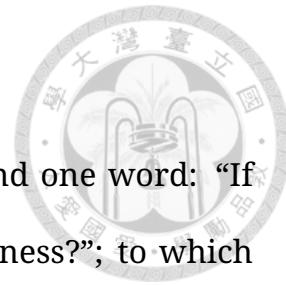
Zhuangzi said, “You’re not I, so how do you know I don’t know what fish enjoy?”

Huizi said, “I’m not you, so I certainly don’t know what you know. On the other hand, you’re certainly not a fish—so that still proves you don’t know what fish enjoy!”

Zhuangzi said, “Let’s go back to your original question, please. You asked me *how* I know what fish enjoy—so you already knew it when you asked the question. I know it by standing here beside the Hao.” (translation Watson 2003:111)

莊子與惠子遊於濠梁之上。莊子曰：「儻魚出遊 從容，是魚樂也。」惠子曰：「子非魚，安知魚之樂？」莊子曰：「子非我，安知我不知魚之樂？」惠子曰：「我非子，固不知子矣；子固非魚也，子之不知魚之樂全矣。」莊子曰：「請循其本。子曰『汝安知魚樂』云者，既已知吾知之

而問我，我知之濠上也。」



The main rhetorical point of this story revolves around one word: “If you are not a fish, *how* (*ān* 安) do you know their happiness?”; to which the answer is: “Since you are asking *how*, it means you already know that I know their happiness.”

Yet I found myself more interested in another word of the text, *cōng~róng* 從容. These two little characters turned out to be quite hard to define yet easy to imagine: fish swimming at ease, calmly, leisurely roaming through the water. This was my first conscious exposure to Chinese, let alone a classical variant, but the succinctness and wit of the text, and intrigue of this word stuck with me, and half a year later I enrolled and began my “Journey to the East”.

For my MA thesis I investigated ideophones, or psychomimetic words as I had called them up until that point, in the *300 Tang poems* (*Táng shī sān bǎi shǒu* 唐詩三百首). One finding was the overwhelming amount of VISUAL ideophones, used to evoke boundless landscapes, mountain ranges so high the eye could not see them, or rivers stretching beyond the horizon. Yet, I wondered, were these words still present in daily usage? After all, it is often claimed that Modern Mandarin Chinese only makes use of SOUND-depicting onomatopoeia. However, as I have come to experience during my candidature at National Taiwan University, depiction is used on a daily basis, and it is not limited to sound alone. I remember going for a haircut and explaining to my hair stylist that I wanted him to cut off more hairs on the side than on the top of my head, but that this transition should be

gradual, i.e., *jiàn~jiàn=de* 漸漸的. As I got my haircut, pop music was playing in the background. My ears caught a few verses from G.E.M.'s *guāng nián·zhī wài* 光年之外, the soundtrack for the movie Passengers:



The universe is *boundless* and freezing; 宇宙 磅礴而冷漠

Our love is infinitesimal yet *scintillating*; 我們的愛微小卻 閃爍

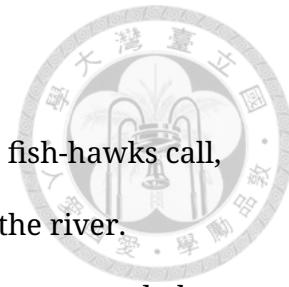
Jolting yet making me forget about ourselves. 頽簸卻如此忘我

These marked words stood out as they were included in my ideophone inventory. Afterwards, my hair stylist asked if I wanted my hair *blown upwards and outwards*, i.e., *chuī-péng~péng=de* 吹蓬蓬的¹, to which I of course said yes. It has been an interesting experience to pay attention to these words in the so-called ideophonic lexicon, because they appear quite often. Yet I started to wonder how this lexicon was structured: could we find out the prototype, and what other variational phenomena were present, e.g., salience. Intrigued by my previous studies on different aspects of Chinese ideophones, I decided to make these two questions the underlying propulsion of my dissertation. And now I have written that dissertation about the prototypicality and salience of ideophones in Chinese, a group of marked words which depict sensory imagery and which belong to an open lexical class, a feat that would not have been possible without my experiences here in Taipei.

¹Or is it *péng~péng=de* 澎澎的?



1 Introduction



*Kron, kron the fish-hawks call,
on the islet in the river,
delicate, demure, young lady,
for the lord a good mate she.*

關關雎鳩，在河之洲。

窈窕淑女，君子好逑。

Shijing 詩經

In the last three decades or so, iconicity – defined as the “(perceived) resemblance-based mapping of form and meaning” (Dingemanse 2013; Dingemanse, Perlman & Perniss 2020) or more succinctly “form miming meaning” (Näny & Fischer 1999) – has progressively gained a prominent status as a research topic within the different domains of Cognitive Science. It has been studied from various perspectives in a number of linguistic subfields, such as the iconic ordering of clausal constituents in relation to their temporal order in syntax and pragmatics (e.g. Haiman 1985; Simone 1995; Radden & Panther 2004); iconic relations between sounds and meaning, i.e., phonesthemes and sound symbolism in phonetics, phonology and morphology (Hinton, Nichols & Ohala 1994; Perniss & Vigliocco 2014; Lockwood 2017; Nielsen & Dingemanse 2020); iconicity across different modalities such as in sign language (Taub 2001; Occhino et al. 2017) or metaphor (Hiraga 2005); and in the lexicon (Voeltz & Kilian-Hatz 2001; Dingemanse 2012; 2018; Haiman 2018). This dissertation is concerned with the latter: ONOMATOPOEIA and IDEOPHONES OR MIMETICS.

First and foremost, they constitute a group of words that are typologically widespread (Dingemanse 2018). Some well-known English examples of the phenomenon include *woof~woof*² for the ‘barking of a dog’, *bang* for the ‘sound of a bomb explosion’, and *zig~zag* for ‘running alternately to left and right’. Similar groups of words have been identified in many languages, with onomatopoeia being a near-universal group of words. However, as a large body of research on Japanese mimetics shows, they can also express other modalities, e.g. the tactile quality in *nuru~nuru*ぬるぬる ‘slimy, slippery’, or the internal feeling in *kuyo~kuyo*くよくよ ‘worrying’. Despite the fact that the Sinitic languages possess a large inventory of ideophones, as we will call these words, they are generally not named in overview articles. Here and there, small exceptions can be found, such as Dingemanse (2018), who refers to Bodomo’s (2006) comparative study between Cantonese and Dagaare. There are a number of reasons for this unfortunate glossing over of a part of the literature, briefly outlined below and revisited in subsequent chapters. The examples in (1) provide a small sample of the scope of the words under investigation. Each of the examples refers to the article number in the Academia Sinica Balanced Corpus of Chinese 4.0, to be introduced in Section 3.3.3. These are abbreviated with “ASBC”, followed by its article number in that corpus. In other words, “ASBC (nº 100622)” in (1) is to be read as “article 100622 in the Sinica Balanced Corpus of Chinese 4.0”.

²Reduplication is marked with a tilde <~>, following the Leipzig Glossing Rules (Bickel, Comrie & Haspelmath 2008).



(1) ASBC (nº 100622)

她 哇的一聲 大 哭起來，
tā wā=de-yì-shēng dà kū-qǐlái,
3SG waa.IDEO=LNK-one-sound big cry-INCH
“Waaaa, she began to wail.”

(2) ASBC (nº 202869)

公車 咻——的 過 站 不 停，
gōngchē xiū=de guò zhàn bù tíng
bus whoosh=LNK pass stop NEG stop
“Whoosh, the bus rushed past the bus stop.”

(3) ASBC (nº 202246)

還 有 燕子 俯身 飛 向 屋簷的 咻咻聲。
hái yǒu yànzi fǔ-shēn fēi xiàng wūyán=de xiūxiū-shēng
also EXIST swallow bend-body fly to eaves=LNK chirp.IDEO-sound
“There’s also the chirping sounds of swallows flying to eaves [of the house].”

(4) ASBC (nº 100871)

咕嚕 咕嚕的 一飲而盡，
gū~lū gū~lū=de yī-yǐn-ér-jìn
glug.IDEO glug.IDEO=LNK one-drink-conj-finish
“[He] chugged it down in one go.”



(5) ASBC (nº 101485)

天色 灰濛濛的，
tiān-sè huī-méng~méng=de
sky-color grey-dim.IDEO=LNK
“The sky is dim grey.”

Traditional Chinese studies have mainly limited their scope of this group of words to a morphological point-of-view, which studies them in three main ways that often overlap, depending on the scope a certain scholar takes. First, the interest may lie in the reduplicative patterns (3-5) (e.g. Sun 1999) – effectively leaving out a group of words that share related meanings or functions but are not reduplicated (1). Second, morphological interests also touch upon the traditional study of so-called *liánmiáncí* 連綿詞 ‘alliterative words’ (e.g. Xú 2000; Xú 2013; Li 2013)³, which tend to include more than just ideophones. For instance, the names for many plants and insects are formed with similar sounding syllables, although this similarity may have become opaque over time, e.g., *géjiè* 蛤蚧 ‘gecko’, which in Mandarin *gé~jiè* comes from Middle Chinese *kop~keajH* and is reconstructed in Old Chinese as **kʰop~kʰrep*. Third, at the most abstract level, words or phrases are classified and studied using the Latin alphabet to represent that structure. For example, *gūlǔ* in (4) would be ‘AB’; *wā* in (1) or *xiū* in (2) would be ‘A’; *xiū~xiū* in (3) would be ‘AA’; *gū~lǔ gū~lǔ* in (4) consequently ‘AB AB’ (or two times

³We follow the recent conventions of journals such as *Cahiers de Linguistique Asie Orientale*: Chinese authors publishing in Chinese get a transliterated reference that includes tone marks, unless they have well-established names without, e.g., Chao Yuen Ren; Chinese authors publishing in English typically do not have tone marks on the name in the reference.

AB, depending on how separate you consider the sounds to be); and *wù-méng~méng* in (5) the ‘ABB’ pattern (see Mok 2001; Lu 2006; Chang 2009).

From a semantic standpoint, however, most research has tended to focus on ideophones that depict SOUND, namely onomatopoeia,⁴ e.g., Lǐ (2007). Because onomatopoeias constitute the most iconic, if not also largest group of ideophones cross-linguistically, they definitely attracted the most scholarly interest. This is a consequence of adhering to basic linguistic tenets such as the “arbitrariness of the sign” (Saussure 1916/ 2005) and the limited amount of sensory modalities expressed by ideophones in Standard Average European languages (Dingemanse 2018). However, as briefly surveyed at the beginning of this section, iconicity has been slowly but steadily reclaiming its place from the margins to which it was banished by mainstream linguistics (Joseph 1997; Dingemanse 2018).

Within the studies of Sinitic languages, the past 20 years have seen an increase in studies devoted to onomatopoeia and ideophones as well. For varieties of Mandarin we find among others Mok (2001); Lu (2006); Sam-Sin (2008); and Meng (2012). For other varieties there are T’sou (1978) (Cantonese); Mok (2001) (also Hakka and Cantonese); Bodomo (2006) (Cantonese); Wu (2014) (Southern Sinitic), Thompson (2018); (2019a) (Teochew among others). Studies of collections can be found in Zhào (2005); Mǎn (2009); Van Hoey (2015) etc. Thus it can be seen that ideophones have been carving out their spot within Chinese linguistics. One of the goals of this dissertation is to bring these studies to the typological forefront, so they

⁴It is worth noting that the Japanese term *onomatope* オノマトペ is a cover term for all these ideophones depicting all sensory modalities, see Chapter 2.

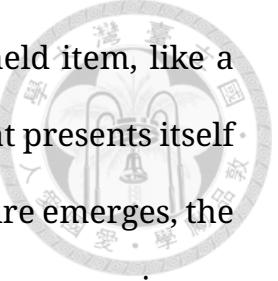
can aid in the theorisation on ideophonic phenomena across languages.



1.1 Aims of this dissertation

Cross-linguistically, it has become relatively well-accepted that “the ideophone” as a concept is prototypically structured (Childs 1994; Dingemanse 2019), *pace* (Heath 2019). See Section 1.4 in this chapter for an introduction to prototype theory and its use in this chapter. On the language-particular level, however, the mimetic lexicon (Akita 2009; 2016; Ruiz Martínez 2019) or ideophonic lexicon (Samarin 1970a), is often contrasted to the prosaic lexicon (Nuckolls et al. 2016; Kwon & Masuda 2019). Such statements suggest that these are two monolithic blocks, and that a hard boundary can be found, despite nuanced studies showing fuzziness and mutual interactions, e.g., Akita (2009); and Dingemanse (2017). Of course, we know that this is far from true for prosaic items, and most studies use these monolithic terms as a starting point to study structural unevenness within the ideophonic lexicon. We also have a quite good understanding of what this structural heterogeneity looks like for the Japanese mimetic lexicon, with prototypical constructions and other more salient form-meaning mappings (cf. Chapter 4). But despite the growing interest in Chinese onomatopoeia and ideophones, mentioned in the previous section, there are still many questions that remain unanswered.

This dissertation aims to lay bare the structural heterogeneity of “the ideophonic lexicon” in Chinese languages, by focusing on contemporary Mandarin Chinese, Middle Chinese and Old Chinese. A helpful metaphor



is to think about the ideophonic lexicon as a small hand-held item, like a Rubik's cube, with the different colored sides as a puzzle that presents itself to us. We can twist and turn the sides, until a complete picture emerges, the point being exactly that ideophones are a multi-faceted phenomenon in any language. To make things more challenging, our ideophonic Rubik's cube is not opaque, but consists of transparent stained glass. We can peer inside, and see how the different components take up different sizes yet work together to form this object. So too ideophones in Chinese will have larger subgroups or clusters, that will differ when we look at them from another approach. We can take apart our cube, and inspect the machinery of it, and see that within items themselves, there are parts that stick unevenly, which will turn out to be true for ideophone items. And then we can reassemble the cube, and find that some parts work closely with other parts, while others are better kept apart. This will resemble how ideophone items interact with constructions.

To leave the metaphor behind, the main question behind this dissertation aims to uncover the variational phenomena of Chinese ideophones, that is, the prototypicality and salience (see Section 1.4) inherent in the Chinese ideophonic lexicon. This question can be explored from different angles, as shown in (6).

(6) Research questions:

Given that there is a language-particular category of CHINESE IDEOPHONES, what variational phenomena can be observed:

- a. What is the scope and structure of this category? (Chapters 2 to 4)

- b. How did the prototypical variation of ideophones change over time? (Chapters 5 and 6)
- c. How did other instances of variational salience evolve over time? (Chapter 6)
- d. How do ideophones interact with the constructions they appear in? (Chapter 7)

Let us first address the assumption made above (6): “Given that there is a language-particular category of CHINESE IDEOPHONES”. This ontological assumption rests on arguments from two main perspectives, namely cross-linguistic typology and advances in Chinese linguistics. In typology, a popular definition of IDEOPHONE in recent years has been formulated by Dingemanse: “*Ideophones are marked words that depict sensory imagery* (2011a; 2012), *belonging to an open lexical class* (2019)”. As will become clear in the following chapters, the definition captures the essence of ideophones cross-linguistically but is formulated with an amount of vagueness which allows for language-particular interpretations of the features mentioned. From the typological perspective, then, the challenge is identifying items that could fit within this definition. Conversely, from the Chinese linguistics perspective, we can see that there are a number of studies that identify onomatopoeia and ideophones based on mostly formal grounds (e.g. Mok 2001) or criteria relying largely on meaning (Zhào 2008). Chapter 2 will delve deeper into these two perspectives and propose that Chinese ideophones covers a group of words comprising of onomatopoeia, binomes, and other reduplicated patterns.

The next step, as shown in Chapter 3, consists of a data collection of such words. In other words, a database is needed. We introduce the Chinese Ideophone Database (CHIDEOD), a novel open-source database which centralizes data on Chinese ideophones based on previous research, material and textual study. Furthermore, other usage-based sources like corpora will be necessary when analyzing variational phenomena of Chinese ideophones.

This allows us to tackle the first question (6a) (Chapter 4): first Chinese ideophones will be theoretically delineated as a category. Subsequently, the structuring of the items within the category will be investigated, based on data from CHIDEOD and items of CHIDEOD which occur in a synchronic corpus of Mandarin Chinese. We will make use of Multiple Correspondence Analysis, an exploratory statistical technique that can chart correlations between different values. This technique is able to display the fuzziness of clusters as well, making it an adequate tool for approaches that aim to investigate variation and prototype structures.

Question (6b) asks about variation of and between items from a diachronic perspective. This question will be first dealt with in Chapter 5 and revisited in Chapter 6. As a case study we will investigate ideophones belonging to the semantic field of LIGHT. The choice for this particular semantic domain is motivated by the scholarly interest in the *gl-* phones-theme in the past century, see Sadowski (2001). Because this phonestheme is argued to be a depiction of LIGHT, this semantic field seems an obvious choice for inquiry in the context of Chinese ideophones.

In Chapter 5, a sample of these will be traced throughout time, with a

manual analysis of how their referential meanings evolved through time.

This will result in dynamic semantic networks which will be proposed for these items. Furthermore, we will be able to observe phenomena like prototypicality, token frequency effects, type frequency effects and competition between different variants.

Chapter 6 makes use of more complex statistical methods to expand on question (6b) by recategorizing prototypicality as a type of variational salience within the lexicon. In other words, this chapter also aims to answer question (6c). Using a technique from the family of distributional relational semantics, semantic vector spaces will be calculated for all ideophones, in which the same semantic field of LIGHT ideophones will be highlighted. Prototypicality is best seen as a form of variational salience (Geeraerts 2006a). More precisely, semasiological salience, which studies how the meanings of one label hang together. The salience lies in prototypicality effects. However, turning that relation around, from meaning to label, allows one to study onomasiological salience. We will observe how different ideophones have different entrenchment values. Lastly, the relations between concepts can also be studied. This structural salience will enable us to study the interplay between collocates and semantic radicals.

The last question (6d), approached in Chapter 7, asks about ideophones in relation to constructions. This issue will be approached with synchronic corpus data and collostructional analysis. The association measures presented here will show that there is considerable difference between ideo-phonic items and their occurrence in constructions. This has important

consequences for generalizations about the usage of ideophones, or even the characterization about this behavior in terms of tendencies.

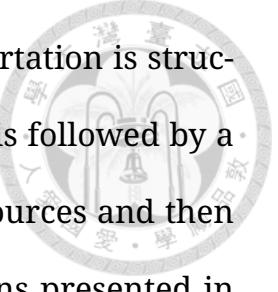
Another way of looking at these research questions is thematically presented in 1.1. After the introduction to the field in Chapter 2 and the presentation of different data sources in Chapter 3, it can be seen that there are three main dualistic oppositions that characterize the queries of interest: temporal scope and thematic focus, by virtue of a better term. Data scope concerns the magnitude of the data investigated: as a whole category (macro), as one semantic field (micro), or somewhere in between those two (meso). With temporal scope we mean the binary opposition between synchronic and diachronic approaches to the data. Thematic focus refers to the theme of the chapter: are we looking at instances of prototypicality or of the broader phenomenon called salience.

Table 1.1: Themes in the research questions

research question	temporal level	thematic scope	methodology
a (Chapter 4)	macro	synchronic	prototypicality Multiple Correspondence Analysis
b (Chapter 5)	micro	diachronic	prototypicality Diachronic Prototype Semantics

research question	temporal level	thematic scope	methodology
c (Chapter 6)	micro	diachronic	Distributional Relational Semantics
d (Chapter 7)	meso	synchronic	Collostructional Analysis

Viewing the research questions in terms of these themes allows us to stress that this dissertation does not aim to provide an exhaustive study of Chinese ideophones, for that is impossible. Instead, what this thesis offers is a set of methodologies and case studies to approach the research questions. Question a, which asks about scope and structure of the entire category (macro level), will be approached by exploring prototypicality through a Multiple Correspondence Analysis based on synchronic data. Question b is interested in the diachronic evolution of prototypical variation and warrants a case study of a semantic field, ideophones depicting LIGHT (micro level), analyzed with Diachronic Prototype Semantics. Question c takes the same basic premise as question b, but here prototypicality is seen as a special case of salience. Using the computational approach of Distributional Relation Semantics it becomes possible to explore this aspect. Finally, question d investigates the interplay between items and constructions. Based on synchronic data and Collostructional Analysis we can investigate notions of salience in this matter as well.



As a consequence of these research questions, the dissertation is structured in eight chapters. The current introductory chapter is followed by a sketch of the state of the field, a chapter devoted to data sources and then the four case studies that respond to the research questions presented in this section, and lastly the conclusion.

1.2 Scope

What we understand to be within the scope of Chinese ideophones will be explored in depth in Chapters 2-4. However, that is a conceptual question different from the scope of the dissertation. Our strong usage-based position follows from wanting to study the semantics of ideophones from both synchronic and diachronic perspectives. The data that will be used, therefore, mostly consists of corpus data, which reflects non-elicited language use (Tummers, Heylen & Geeraerts 2005). The great advantage, especially for Chinese, is that we can study cross-modal iconicity and motivation in terms of phonology and orthography vis-à-vis meanings (see Section 1.3).

However, there are also two nuances that need to be sketched in relation to this data source. First, the data is limited mainly to Mandarin Chinese, for which the term ‘Standard Chinese’ will be used interchangeably within the remainder of this work. It does not cover other Sinitic languages like Southern Min, Hakka, Cantonese, Wu etc. It can be conceived that these languages have similar ideophone systems and that some findings will be extendable to those language varieties as well. However, we would need more in-depth research with the methods proposed in this dissertation in order to investi-

gate this issue. Unfortunately, the non-standard status of these languages, or even their dialects, has resulted in their being understudied. As a result, resources are scarce and not well-known or hard to access. It is my hope that Chinese Ideophone Database (CHIDEOD), which will be presented in Section 3.2, can provide a framework to collect onomatopoeia and ideophones of these other varieties as well. Second, given that ideophones are mostly studied in spoken (synchronic) settings, and often co-occur with gesture (Dingemanse 2019), this is a phenomenological side of their nature that will be downplayed in this dissertation, i.e., the data used are mainly written in nature, see Chapter 3. The reasons are twofold: for diachronic sources, spoken data are unobtainable; for synchronic sources they do not exist yet⁵. Consequently, this thesis makes use of textual sources, but provides a foundation to revisit the findings in the future, when such sources will exist. Furthermore, it is not unlikely that ideophones used in spoken language vary considerably when compared to those used in written sources. In fact, that is one of the findings that will follow from the Multiple Correspondence Analysis in Section 4.5. On the other hand, a strict theoretical distinction between spoken and written language (of Mandarin Chinese) does not do justice to the data, as has been argued before (Féng 2010; Zhang 2017; Eifring 2019; see also Iwasaki 2015). For example, Zhang (2017) argues that there are more dimensions that need to be taken into account than just spoken or written modality, when classifying texts. Such statements also count for

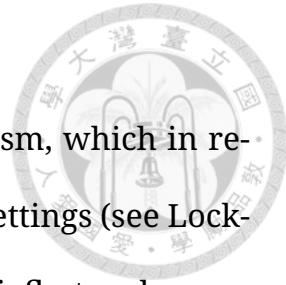
⁵This is not to say that there are no spoken Chinese corpora – there are. However, informal exploratory searches showed that the occurrence of ideophones is on the low side. This shows that on the one hand ideophones are perhaps not used as frequently, or that the corpora are not large enough to do the study of them justice.

ideophonic items and usage.

Related to this issue is the literature on sound symbolism, which in recent years has been studied many times in experimental settings (see Lockwood 2017 among others). This dissertation will only briefly touch upon the issue of sound symbolism, not because we oppose it, but because it is a phenomenon that is conceptually different from ideophones. That is to say, ideophones *can* display sound symbolism – or sound iconicity as recently proposed (Hoshi et al. 2019) – but not always, and it becomes rarer for NON-SOUND ideophones. But what about phonesthemes? As will be argued in Section 5.1.1, it is not claimed that there are no submorphemic networks present in ideophones, in terms of motivation and systematicity (Dingemanse et al. 2015), there most certainly are. However, they do not appear to be as iconic as often assumed. Future research is expected to take up this issue and convincingly provide a place for phonesthemes within a theory of Chinese ideophones.

The current dissertation is also more quantitative than qualitative in nature, a natural consequence of using corpus-based approaches (Tummers, Heylen & Geeraerts 2005; Geeraerts 2010a). The case studies, will often refer to dictionary usage, but then go beyond this usage by relying on corpus data and making use of the statistical technique that takes center stage in each chapter, respectively Multiple Correspondence Analysis (MCA) in Chapter 4, Diachronic Prototype Semantics in Chapter 5, Semantic Vector Spaces in Chapter 6, and Collostructional Analysis in Chapter 7.

Lastly, in terms of scope, the studies presented here are not meant to



be exhaustive; they are meant to provide methods of approaching ideophones from a usage-based and Cognitive Linguistic perspective. In doing so, I aim to strike balances between synchronic (Chapter 4, 7) and diachronic (Chapter 5-6) approaches (see also the preceding section); and paradigmatic (Chapter 3-4) and syntagmatic (Chapter 5-7) perspectives.

1.3 Semiotic folk model for Chinese and variation

It is not uncommon to come across folk statements that suggest that “the Chinese language equals the written language”. That is to say, rather than the usual linguistic fundamental notion that language primarily revolves around spoken mappings between form and meaning (Langacker 1987a), of which the majority is arbitrary (Saussure 1916/ 2005), traditional Chinese approaches include characters very much in the conception of the word (Packard 2001; 2016). Consequently, it is possible to apprehend this traditional notion of the unity of shape, sound and meaning (*Hànzì de xíng yīn yì* 漢字的「形音義」 (Hsieh 2006) as a folk model.

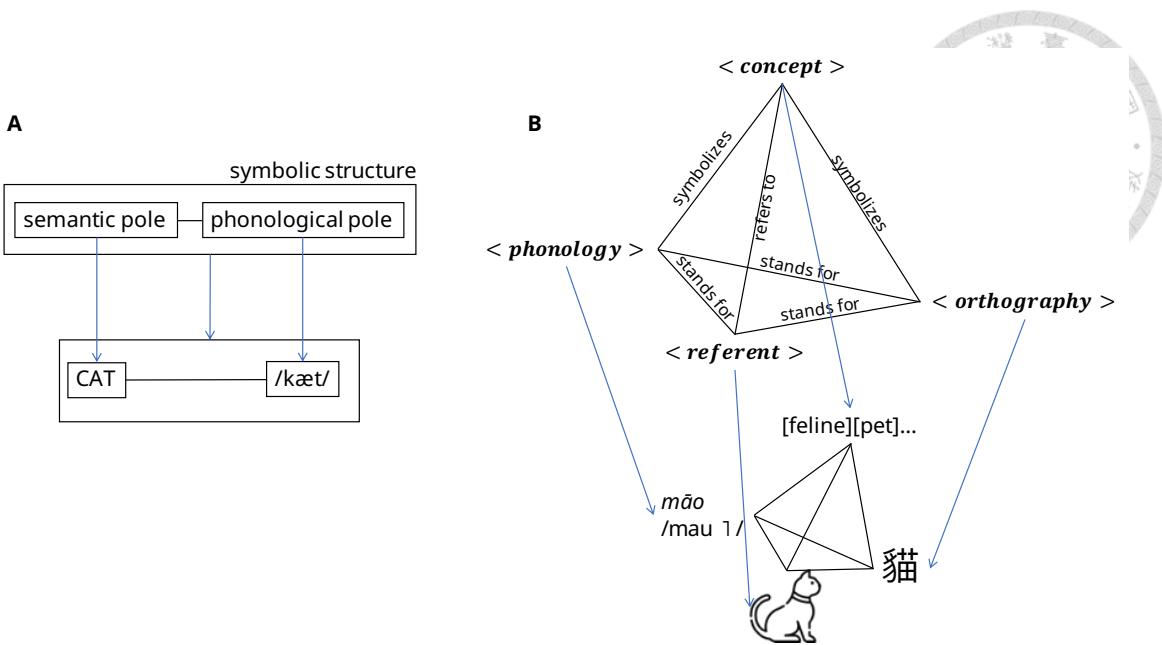


Figure 1.1: A: Basic semiotic model of Cognitive Grammar; B: Basic semiotic model for mimetics in Japanese (Lu 2006), and by extension Chinese

In other words, in Langacker's Cognitive Grammar, (1987a; 1991; 2008b) symbolic structures contain a mapping between a semantic pole and a phonological pole, as illustrated in Figure 1.1.A. For him, the semantic pole can be thought of as meanings. The phonological pole does not merely include sounds, but also gestures and orthography. It thus is a general formal structural pole, although he nuances this by saying that most attention will be devoted to phonology (Langacker 2008b:15). So for English *cat*, we get the representation of a semantic pole [CAT] that is connected to a phonological pole [/kæt/], and forms a symbolic structure: [[CAT] / [/kæt/]].

However, Chinese speakers with their cultural specific semiotics, have three poles: <referent / semantics>, <phonology> and <orthography>, as is shown in Figure 1.1.B, for *māo* 貓 'cat', which form a symbolic structure⁶. Lu (2006) primarily adapted Ogden and Richards's (1923) well-known semiotic model to Japanese mimetics, but this can easily be further ex-

⁶(cat icon: Pixel perfect on www.flaticon.com)

tended to Chinese ideophones (see Van Hoey 2018a). It does not entirely contest this model, but explicitly differentiates between <phonology> and <orthography> on the formal side, and <semantics> on the meaning side. A useful shorthand presentation is shown in example 7.

(7)

$$\frac{\text{sound}}{\text{writing}} \mid \text{meaning}$$

In the rest of the thesis I will often refer to this model as the cognitive SEMIOTIC FOLK MODEL of Chinese. Its advantages are that different variational phenomena on one pole vis-à-vis other poles can be observed and studied. For instance, in a word like *mōu* 哟 ‘moo, sound of a cow’, there appears to be no variation – in Chinese the concept ‘moo, sound of a cow’ is expressed by /mōu/ and written as 哟. But a word depicting ‘vastness, boundless’ has onomasiological variation in terms of <phonology>: speakers can choose to use words like *guǎngdà* 廣大, *guǎngkuò* 廣闊, or an ideophone, e.g., *máng~máng*, which depicts a scene where something is boundless and unclear, difficult to observe. If the latter is chosen, then there is another variational choice that needs to be made, namely on the <orthographic> pole. Throughout much of Chinese history, *máng~máng* was either written as 茫茫 or as 芒芒 (Van Hoey & Lu 2019a). Now, the relative preferences for one form over another are issues that will be explored in Chapter 6, when salience phenomena within LIGHT ideophones are explored. However, the model deserves a first introduction here, as it is present in subsequent chapters. It also sets the stage for the Cognitive Linguistics orientation that is

present in this dissertation.

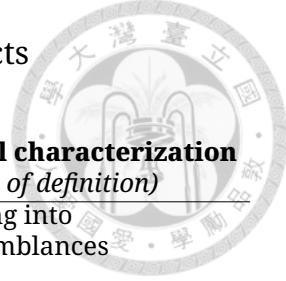


1.4 Prototypicality and salience

The Cognitive Linguistics orientation can also be deduced from two key-words in this dissertation's title: prototypicality and salience. Of these, PROTOTYPICALITY is more well-known. It originated in the mid-1970s as an alternative to traditional conception of definition, which relies on necessary and sufficient conditions that cover all representatives of a given category defined by these conditions. For instance, if one defines BIRDS as 'being able to fly', then logically all birds should be able to fly. However, we know this is not the case, e.g., the ostrich or penguin cannot fly. The research on prototypicality, originally commenced in psycholinguistics by Rosch and colleagues (see overviews in Rosch 1978; 1988; Mervis & Rosch 1981) revealed that for the semantic category of BIRD indeed not all members are equally typical, i.e., equally well described by the same necessary and sufficient conditions. Because if that were the case, there would be no difference in representativity between members. Rosch & Mervis (1975) showed that for Americans, the robin is the most typical bird, followed by the sparrow, bluejay, bluebird, canary, blackbird, dove, lark, swallow, and parakeet. This could be opposed to the least typical bird-like members in their experiment: chicken, turkey, ostrich, titmouse, emu, penguin, and bat. Of course, such kind of classifications were not limited to BIRD, but were also found for \term{furniture, FRUIT, VEHICLE etc.} (Rosch & Mervis 1975).

After the introduction of the notion of PROTOTYPE within linguistics in the

Table 1.2: Four types of prototypicality effects



	Extensional characterization (on the level of exemplars)	Intensional characterization (on the level of definition)
Non-equality <i>(salience effects, core/periphery)</i>	(a) differences of typicality and membership salience	(b) clustering into family resemblances
Non-discreteness <i>(demarcation problems, flexibility)</i>	(c) fuzziness at the edges, membership uncertainty	(d) absence of necessary and sufficient definitions

early 1980s (Geeraerts 2010b:183–192) it became clear that this term had been used to refer to a number of different related, yet distinct phenomena. For starters, prototypicality could refer to the degree of clear category membership, such as the case of the birds mentioned in above (Rosch 1978:36). It could also refer to their family resemblance (Rosch & Mervis 1975:574–575). For instance, a chicken and turkey resemble each other more than they do a robin, making is plausible that they are judged with the same representativeness in the experiments. Furthermore, it is unclear what is the clear dividing line between chickens and turkeys, or chickens and robins. In other words, the category boundaries are fuzzy (Mervis & Rosch 1981:109). Lastly, not all conditions can cover all members of the category, as we saw above with the condition of ‘flying’. The realization that ‘prototypicality’ is actually prototypically structured as well (Posner 1986) allowed Geeraerts (2010b:189) to organize these four approaches in a contingency table, shown in Table 1.2, which will be revisited in Section 5.2.2.

It is important to note that these four interpretations or features of ‘prototype’ are not mutually exclusive. For instance, we find features (a), (b), and (d) readily in the category of BIRD: not all birds are equally birdy, they

cluster in different resembling groups, and there is no single definition that covers all of them. Feature (c), however, is can not be found for BIRD: we know what a bird is and what is not. It is no wonder that bats were the least typical for the ‘bird’ stimuli in Rosch & Mervis (1975).

Turning now to ideophones, it has often been argued that they are prototypically structured categories (Childs 1994:195) both within languages as well as across languages. In Childs’s interpretation, feature (a) differences of typicality is highlighted. However, as the literature has shown and as will be shown in this dissertation, the four features are present in the prototypicality notion of IDEOPHONE. In Chapter 4 the category in Chinese will show that not all items are equally representative (a), but that they cluster together in different groups (b). We will be able to identify a boundary, but it will be a fuzzy one, only being able to state that some items fall inside the category while others should be excluded. And as will be shown from next chapter onwards, recent definitions such as the one by Dingemanse (2012; 2019) rely on a number of important features that together aim to capture the “canonical ideophone”. Chapter 5 retains the notion of prototypicality, but mostly focuses on feature (b) in Table 1.2. This detailed study of ideophones in one semantic field will show that non-equality and heterogeneity can be identified in a number of ways, especially if the interplay between the three poles of the semiotic folk model is borne in mind (Section 1.3). Summarizing, these two chapters show how the notion of prototypicality is important for our understanding of Chinese ideophones.

However, now that prototypicality has been introduced, it is time to turn

to the notion of SALIENCE. In Chapter 6 we take the analytical turn of interpreting prototypicality as a kind of variational salience, following Geeraerts's (2000; revisited in Geeraerts 2006b; 2010b; 2017) development of a taxonomy for salience phenomena in lexical semantics, presented in Figure 1.2. In this interpretation (for others, see Schmid & Günther 2016), salience refers to the differences in structural weight, as can be observed in features (a) and (b) of Table 1.2.

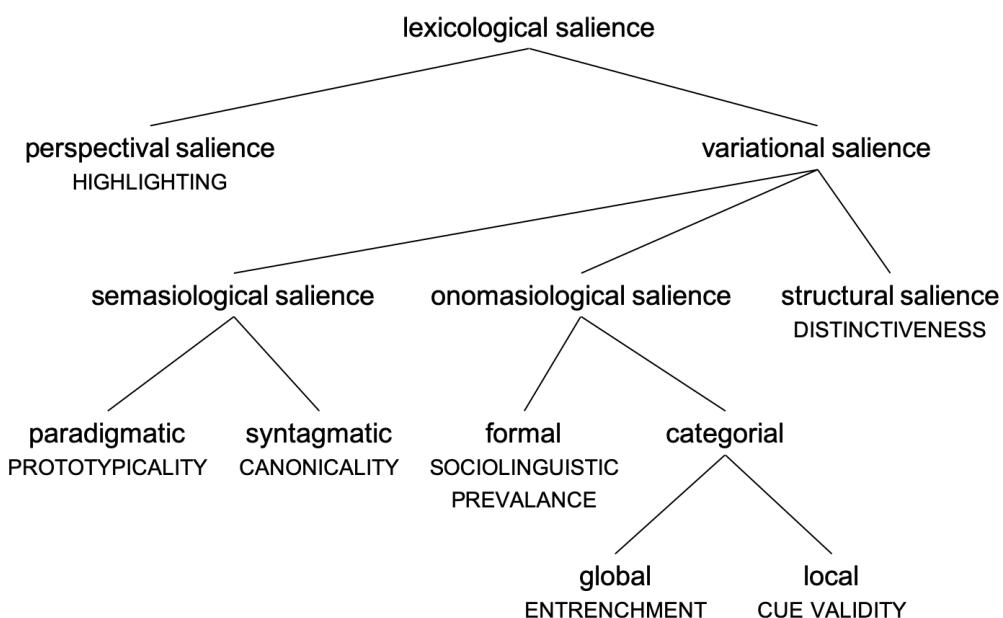


Figure 1.2: Taxonomy for lexicological salience (based on Geeraerts 2000)

As can be seen in Figure 1.2 the first split in lexicological salience is into perspectival salience and variational salience. Perspectival salience, also called highlighting by Geeraerts, refers to the kind of differences in construal that are part and parcel in Langacker's Cognitive Grammar (1987a; 1991; 2008a). For instance, *to give* and *to receive* both encode a transmission of an item on a path that goes from giver to receiver, but the perspective is different. Consequently, when using *to give*, the giver will be the most

salient actor; for *to receive* it is the receiver.

When one of these alternatives is more frequently used than the other alternative, we can say that there is a form of variational salience at play: one of the variants is more salient. In this case, since our choice depends on how we want to express the Giving frame by choosing from the two alternatives *to give* and *to receive*, it will be a case of onomasiological salience. In other words, onomasiological salience departs from the meaning side (here, ‘giving’) and investigates the labels.

This can be contrasted to going from label to meaning, the semasiological perspective. It is here that we encounter ‘prototypicality’ in Geeraerts’s taxonomy, referring to the typical semantic exploration of an item or category. Answering the question of what a *a bird* is by finding definitional features, is a form of doing semasiology. And as we now have seen, these features are not all equally important or salient, e.g., ‘being able to fly’ is quite important to categorize something as a bird, but ‘having wings’ may be even more important.

The third main group within variational salience is structural salience, or distinctiveness. This instantiation of salience explores the different structural features that set different variants apart. Let us say that we are investigating the different (folk) names of birds and have taken a special interest in pelicans. We can find the Dalmatian pelican, brown pelican, Australian pelican, great white pelican, pink-backed pelican and spot-billed pelican. Structurally, we would then be interested if color (brown, great white, pink-backed, spot-billed) is a more salient coding mechanism for the name than



geographic location (Australian).

In terms of ideophones, these three kinds of variational salience will be explored in Chapter 6 by making use of computational methods. Of the three, semasiological salience is the perspective that has most often been explored in relation to ideophones, asking ‘what does this ideophone mean’, and coming up with a number of senses. Good examples of such research are the three frameworks discussed in Section 5.2.1. Onomasiological salience or structural salience have, to our knowledge, not received the attention they need, and this dissertation hopes to provide a foundation for these perspectives as well.

In Chapter 7 we explore salience in a slightly different manner, namely in terms of cue validity, as it was called in Figure 1.2. Geeraerts (2000) illustrates this with two example sentences, shown here in (8-9). His argument is that, while Example (9) definitely occurs, i.e., the ‘across’ meaning with the building as its landmark, it is far more natural to use the variational alternative of *tegenover* ‘facing, across’ in this local case.

- (8) Martha woon-t **tegen-over** het=museum.

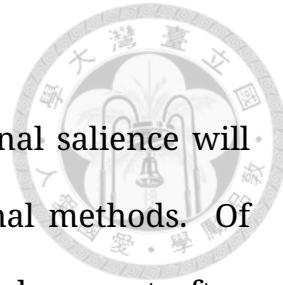
Martha live-s against-across DEF.ART=museum

“Martha lives across of the museum.”

- (9) Martha woon-t **over** het=museum.

Martha live-s across DEF.ART=museum

“Martha lives across the museum.”

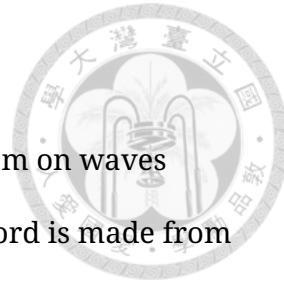


So in terms of ideophones, we can then ask the following: given constructional constraints, what are the best items that fit a constructional slot? In other words, which items are most attracted by a construction? But also, which items are repulsed the most? And, which items depend most on a construction to appear? Chapter 7 uses Collostructional Analysis, with the statistic ΔP to investigate these kinds of salience. While we will go over these methods and statistics in detail in Chapter 7, it is of interest to sketch why this statistic is chosen. Rather than relying on raw frequencies as can be obtained through simple corpus tally, ΔP takes contingency (as well as directionality) in consideration. For the learning of the category BIRD, this means that “while [features like] eyes and wings are equally frequently experienced features in the exemplars, it is wings which are distinctive in differentiating birds from other animals. Wings are important features to learning the category of birds because they are reliably associated with class membership, eyes are neither. Raw frequency of occurrence is less important than the contingency between cue and interpretation.” (Ellis & Ferreira-Junior 2009:194). For ideophones and constructional slots we thus want to know what kind of constructions attract which ideophones and vice versa.

However, we can only get to this point after developing the literature survey and the introduction of the data sources, presented in the next two chapters.



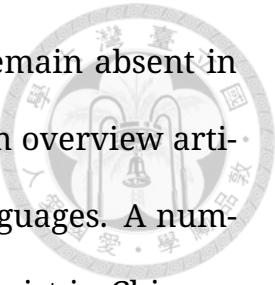
2 Background



We call the foam on waves
sukien: that word is made from
two words of the Old Speech,
suk, feather, and *inien*, the sea.
Feather of the sea is foam. But
you cannot charm the foam
from calling it *sukien*; you must
use its own true name in the Old
Speech, which is *essa*. [...] Thus,
that which gives us the power to
work magic sets the limits of
that power. A mage can control
only what is near him, what he
can name exactly and wholly.

Ursula K. Le Guin

What is an ideophone? And why do we need this term for the study of Chinese? In this section, we will provide the foundation of this dissertation by examining the state of the field. These theoretical preliminaries concerning ideophones first include the cross-linguistic perspective, by succinctly tracing about 150 years of research devoted to ideophones. This history adopts in effect a Western linguistics perspective. However, the very rich tradition of Japanese linguistics will be shown to have developed its own vocabulary to investigate mimetics. In the last two decades, Japanese linguistic studies of ideophones have found their way into the cross-linguistic literature. Cu-



riously, however, Chinese studies of ideophones largely remain absent in this (English-speaking) research, being barely mentioned in overview articles which discuss ideophonic phenomena in different languages. A number of scholars even simply state that ideophones hardly exist in Chinese, being only or mostly confined to onomatopoeia, or even go beyond that, stating that the language on the whole seems iconic⁷, because the language is tonal. Such a generalization takes things too far, although clear statistical preferences for COLLOQUIAL ideophones and the high level tone have been observed, see Mok (2001); Arthur Lewis Thompson (2019a), as well as Section 3.2.6.2. Three quotes can be used to illustrate such stances on onomatopoeia and ideophones in Chinese. The first consists of Sun & Shi's (2004) study on ideophones in rGyalrong Tshobdun:

狀貌词 (ideophone, expressive) 是一种以“音”直接表“义”，模拟或描写声音、形状、颜色、性质、动作的特殊语词，说话者藉以表达感官经验或主观态度，他们常具明晰性 (iconicity)，与一般词汇以抽象语言符号间接转达语义有所不同。所有语言都有一定数量的状貌词。有些语言（如汉语、英语）状貌词用途较有限，未能引起研究者的关注。有些语言，如美洲印第安语 (Mithun 1982; Nuckolls 1996)，非洲语言 (Childs 1994)，及日、韩、南亚、苗瑶、藏缅等亚洲语言则有特别发达的状貌词。

Ideophones are a special type of words that depict all kinds of sounds, shapes, colors, qualities, and actions in a direct matching of sound and meaning to convey sensory experiences and

⁷Imai Mutsumi (p.c.)

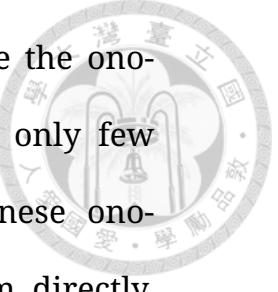


attitudes. They often display iconicity, unlike the normal lexicon which consists of abstract symbols that express meaning. All languages have a certain number of ideophones. **The ideophones of some languages, such as Chinese and English, have a rather limited usage, and thus did not attract the attention of scholars.** Some languages, e.g. Native American languages (Mithun 1982; Nuckolls 1996), African languages (Childs 1994), as well as Japanese, Korean, South[-East] Asian languages, Hmong-Mien languages, Tibeto-Burman languages etc. have an exceptionally well-developed [system] of ideophones.

Sun & Shi (2004:1); translation and emphasis mine

Arguably, their lumping of Chinese ideophones together with English is mistaken. It is well-known that the ideophone inventory in West-European languages is of a smaller size (Leskien 1902; Wälchli 2015). But to claim this for Chinese disregards the growing body of literature that is referenced in this chapter.

In a similar vein, Luo, Masui & Ptaszynski (2014), who introduce a system of automatic machine translation for translating Japanese mimetics to Chinese, identify a number of problems regarding onomatopoeia. They use a list of 174 onomatopoeia recorded in the *Xiàndài Hànyǔ cídiǎn* (现代汉语词典) (Lǚ 2005). Consequently, it is claimed by Luo et al. that there are major lacunae in Chinese lexicographic studies. This is erroneous, as there are a number of specialized onomatopoeia dictionaries they could have used instead (see Section 3.2.1). They state:



The words equivalent to Japanese onomatopoeia are the onomatopoeia in Chinese. However, there have been only few Chinese onomatopoeia when compared with Japanese onomatopoeia, thus it is not possible to translate them directly. Furthermore, onomatopoeia have plural meaning [sic] in Chinese. If there is no clear context, Chinese readers often cannot guess the true meaning of onomatopoeia. For example, the Chinese onomatopoeia *wang wang* (汪 汪) has two meanings. One is the sound of dog barking, and the other one is a name of a pop-idol. Saying only “wang wang” in a dialogue or writing will not be understood in most cases. Furthermore, the words equivalent to Japanese mimetic expressions existed in Chinese in the past, but they are rarely used today.

Luo, Masui & Ptaszynski (2014:371)

To reiterate, Luo, Masui & Ptaszynski (2014) state that there are fewer onomatopoeia in Chinese. However, they use Japanese onomatopoeia to mean a concept that covers the whole spectrum of sensory imagery, yet their examples of Chinese show that they treat onomatopoeia for Chinese to only refer to sound. Limiting oneself to only depictions of sound is not inherently problematic, as certain recent scholarship (Yu 2015) has shown, but when comparing two languages, one must take extra care in treating any given concept similarly and consistently across both languages, i.e. avoid to fall in the fallacy mistaking cross-linguistic concepts with language-particular category fallacies (Haspelmath 2010; 2018; Van der

Auwera & Sahoo 2015; Croft 2016). What is meant by “onomatopoeia have plural meaning in Chinese” is unclear, but it can be presumed that they mean they are vague or polysemous (cf. Akita 2013a). Unfortunately their (homophonous) example of *wāng~wāng* 汪汪 has not aged well, since the pop-idol’s star seems to have waned in recent years. In any case, even if we go with their example, in most cases the context and co-text will disambiguate whether one refers to the barking of a dog or the pop-idol. Lastly, Chinese onomatopoeia are virtually presumed to have died out in current usage. This is reminiscent of the so-called Iconicity Treadmill Hypothesis (Flaksman 2017), which tries to explain how items become de-iconicized over time. Yet, what are we to make of such a statement? Are they really rarely used today, or are there just other semantic preferences and ways of stating things that have developed? Or are there modality effects at play, so that, for instance, spoken language has different ways of using ideophones from written language? Or genre effects or socio-linguistic variables? Or are we dealing with a case of “where have all the ideophones gone”, as Childs observed for Zulu in urban settings (Childs 1996)? We do not have a full answer to all of these questions, but hope that the subsequent chapters will convince the reader that ideophones are still in use.

The last quote in this series demonstrating the literature gap of cross-linguistic studies and Chinese studies comes from Wu (2014):

Thus, we know that having ideophones is indeed a universal or near-universal feature of language (Dingemanse 2012:655). Ideophones’ distinguished features like expressiveness, iconicity and

structural markedness are noticeable for many researchers. But for those who are accustomed to the traditional research method, ideophones are hard to classify or even recognize, let alone to research. **Chinese linguistics is such a field.**

Wu (2014:8); emphasis mine

This is very telling, and things are getting better, but if ideophones are to be considered a prevalent class of words in Chinese languages, then researchers would do well to pay additional attention to them. Maybe they have been paying attention to other related phenomena, or have been studying them under different names. The current chapter attempts to illuminate some aspects of this issue – it is not that there is no research, but the studies are found to be sinocentric, or analyzed through other subfields, such as reduplication studies. Let us first turn to a brief history of ideophone studies.

2.1 Ideophones in the West

The history of ideophone studies has been documented quite well, especially from a Standard Average European perspective. For instance, Dingemanse (2011a:57–74) provides an overview from the earliest descriptions of ideophones in African languages up until their current study in the 21st century:

Although ideophones touch on many things, an attempt is made here not to duplicate the voluminous literatures on sound-symbolism (some excellent overviews are Bühler 1934 ch. 13;

Jakobson & Waugh 1979; Hinton, Nichols & Ohala 1994; Magnus 2001:12–33), interjections (e.g. Ameka 1992; Kockelman 2003) and “expressive” language in its various senses – including affect, emotional lanugage and swear words (Foolen 1997; Potts 2007), phonological expressiveness (e.g. Fudge 1970), expressive morphology (e.g. Zwicky & Pullum 1987) and poetic expressiveness (Tsur 1992). Though I [i.e. Dingemanse] have done my best to cover the relevant literature on ideophones, my discussion may be somewhat biased towards research published in Western academia (in German, English, and French) and on African languages. For pre-1950 sources this bias is justified by the fact that African linguistics was the first linguistic tradition in which ideophones were recognised as a significant linguistic device. I do note with particular regret, however, that I miss out on rich bodies of literature in Japanese, Korean and Vietnamese.

Dingemanse (2011a:57)

Dingemanse (2018) also surveys the history, mentioning studies on Sanskrit ideophones and some Western observations of Japanese mimetics in the 17th century. Later, in the 1850s, some Africanists provided seminal studies of ideophones in African languages, namely Vidal on Yoruba, Koelle on Kanuri, and Schlegel on Ewe. However, these words are regarded as “playthings , not the tools of language” which “constitute a very small proportion of our dictionary” (Müller 1861:346). Notwithstanding, in-depth language-specific studies are performed in which ideophones of all sorts are

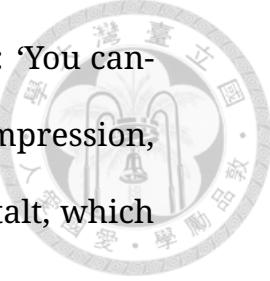
mentioned: McLaren (1886) on Nguni, Junod (1896) on Ronga, Whitehead (1899) on Bobangi, Meinhof (1906) on the Bantu languages, Westermann (1905; 1907) on Ewe, Aston (1894) on Japanese, Grammont (1901) on Indo-European languages, Leskien (1902) on Lithuanian, Winkler-Breslau (1907) on Caucasian languages, and Urtel (1919) on Basque. Of these, Dingemanse (2018) highlights contributions by Junod and Westermann who stated that:

On dira peut-être: « C'est là une manière enfantine de parler; il ne vaut pas la peine de s'y arrêter. » Bien au contraire! L'esprit infiniment mobile, primesautier de la race se reflète dans ce parler pittoresque. Il réussit à rendre par ces mots-là des nuances qu'un langage plus posé ne saurait exprimer.

One might say: 'This is an [sic] childish way of speaking; it is not worth the trouble.' Quite the contrary! The versatile and spontaneous mind of the people is reflected in this picturesque talk. It enables these words to render nuances which a more restrained language could not express.

Junod (1896:196f.), translation in Dingemanse (2018:7)

Ewe has two dialectally separated words for duck (...), *kpakpa* after its quacking and *qabodabo*. When I asked a local whose dialect does not have the latter why it is that other people would say *qabodabo*, his answer was, "Well, because...", and he used his upper body to imitate the waddle of a duck. Ideophones describe a process or object as a whole, not focusing on a single aspect but highlighting primarily its living, moving features. Asking for the



meaning of an ideophone often leads to the objection: 'You cannot just describe it, you have to see it.' It is the total impression, the whole Gestalt, or the movement of the whole Gestalt, which is important.

Westermann (1937:159), translation in Dingemanse (2018:7)

Later, another study on Bantu languages performed by Doke saw the coinage of the term ideophone. According to him, this group of words, or part-of-speech, could be defined as:

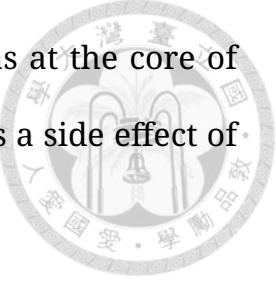
A vivid representation of an idea in sound. A word, often onomatopoetic, which describes a predicate, qualificative or adverb in respect to manner, colour, sound, smell, action, state or intensity.

Doke (1935:118)

By then, ideophones had gained some standing as a study field. Some scholars, like Newman, accepted ideophones as a class for some languages (such as Hausa), but not for others (Newman 1968). And again, more research followed: Hoffkann (1952) on ancient Indian languages, Emeneau (1969) on languages in India, Uhlenbeck (1952) on Indonesian (Javanese), Carr (1966) on Malay, Durand (1961) on Vietnamese, Banker (1964) on Bah-nar, Henderson (1965) on Khasi, Watson (1966) on Pacoh, and explosion in Japanese studies as well (Akita 2009).

In the 1970s, Dingemanse (2018) states, three main figures appeared: Kunene, Samarin, and Diffloth. Kunene (1965)] highlighted the performance-like quality of ideophones. While linguists like Vidal

(1852) had argued that the intensifier meaning ('very') was at the core of ideophones, for Kunene this intensification simply arose as a side effect of the depictive quality of ideophones:



The ideophone attempts to bring before the listener, for first-hand perception, actions or states (...) It is an attempt to make the audience see for themselves what happened — or will happen.

Kunene (1965:35)

Samarin (1970a) then proposed a first cross-linguistic generalization. Based on ideophones in African languages, Azerbaijani, Malagasy, Mon-Khmer, Korean, Tamil, Thai, Yokuts and Waiwai, he found that in these languages, similar classes of words could be identified that "(1) they display a great deal of play with sounds, that (2) they are predominantly reduplicative, that (3) their phonology is in some respects different from that of all other words, and finally, that (4) they have very specific meanings sometimes difficult to define. (Samarin 1970a:160)".

Diffloth (1972), lastly, added to this that they must be a characteristic of natural languages, if they appear so widespread. However, "they are conspicuously undeveloped and poorly structured in the languages of Europe (Diffloth 1972:440)".

A number of naysayers, like Newmeyer, argued that "the number of pictorial, imitative, or onomatopoetic nonderived words in any language is vanishingly small" (1992:758) – not based on data, but on his gut feeling (or so). On the other hand, ideophones started gaining attention from theoretical linguistics (mostly phonology), providing real data to test hypothe-



ses. Dingemanse mentions explorations in: the nature of phonological features (Kim-Renaud 1978; Mester & Itô 1989), non-concatenative morphology (McCarthy 1983; McNally 1991), aesthetic and expressive functions of language (Samarin 1970b; Jakobson & Waugh 1979), lexical discreteness and the nature of words (Diffloth 1976; Mithun 1982), gradients and iconicity in prosody (Liberman 1975; Bolinger 1985), psychological reality of lexical iconicity (Fischer-Jørgensen 1978; Fordyce 1988), and expressive vs. prosaic levels of structure (Diffloth 1979 (misquoted as 1980 in by Dingemanse); Zwicky & Pullum 1987). Continuing along this trend, sound symbolism came to bloom as its own (related) field (Jakobson & Waugh 1979; Waugh 1992; Hinton, Nichols & Ohala 1994). With the eye on cross-linguistic comparison, Kulemeka (1995) argued there are two ‘traditions’ of ideophones: the African stream which focused on the word class status of ideophones, and the Asian stream that focused on the iconic patterns of ideophones.

Ideophones gradually gained more widespread interest as a study field. First and foremost, a conference on ideophones led to the volume *Ideophones* (Voeltz & Kilian-Hatz 2001) on ideophones all around the world, but with a strong focus on Africa. Next, a major branch of studies elicits them through linguistic anthropology, notably by Nuckolls (1996; 1999; 2010) on the linguistic and cultural ecology of ideophones in Pastaza Quechua, Noss (1989; 1999; 2001) on ideophones and verbal art in Gbaya, Webster (2008) and the aesthetics and semiotics of iconicity in Navajo poems, or Dingemanse’s multimodal studies of ideophones in Siwu (Dingemanse 2011a; 2013; 2016; Lockwood & Dingemanse 2015a). Third,

interdisciplinary studies have been increasingly carried out in the last decade (see Vigliocco & Kita 2006; Akita 2009; 2015; Armoskaite & Koskinen 2017; Ibarretxe-Antuñano 2017 for overviews). Such research has moved the needle regarding the nature of ideophones towards the need for a new definition. Currently, the definition that cross-linguistically finds the highest degree of acceptance in the (mostly English-speaking) scholarly world of ideophone researchers, is presented in (10).

- (10) Ideophones are marked words that depict sensory imagery (Dingemanse 2011a; 2012), and which belong to an open lexical class (Dingemanse 2019).

As he specifies (Dingemanse 2011a:25–29), ideophones are formally MARKED in that they stand out from other words, be it through phonology, morphology, intonation or other means. They are WORDS in the sense of being “conventionalized minimal free forms with specifiable meanings”. They DEPICT rather than DESCRIBE their referents – ideophones are performances rather than commentative terms. For instance, the English *zig~zag* depicts ‘the way of alternately running leftward and rightward’. That is to say, a definition can usually be found, but it is not the same as using the ideophone itself. Lastly, SENSORY IMAGERY is “understood here as perceptual knowledge that derives from sensory perception of the environment and the body” – embodiment. Some languages also depict ‘internal perceptions’, see Section 2.2. This definition forms an important starting point for the discussion of the boundaries of the Chinese ideophonic lexicon. It will be revisited in detail in Chapter 4.

Dingemanse's (2018) overview of the historiography of ideophone studies thus convincingly achieves what he set out to do, and manages to stress the importance of ideophones as linguistic 'MARGINALIA', i.e., study objects that have been relegated to the margins of research, while actually being quite common across languages. They are contrasted with 'RARA', which are typologically rare phenomena, e.g., clicks as common phonemes. He also shows that the terminology may be different, but that cross-linguistically the phenomenon of ideophones, mimetics, expressives, *impressifs* (French), onomatopoeia, etc. can be compared across languages. In the map below, we have visualized the terminology used by Western scholars for ideophones in different language-specific studies, based on Voeltz & Kilian-Hatz (2001); Dingemanse (2011a; 2018); and Kwon (2015). An overview of the language sample is presented in Appendix 2.

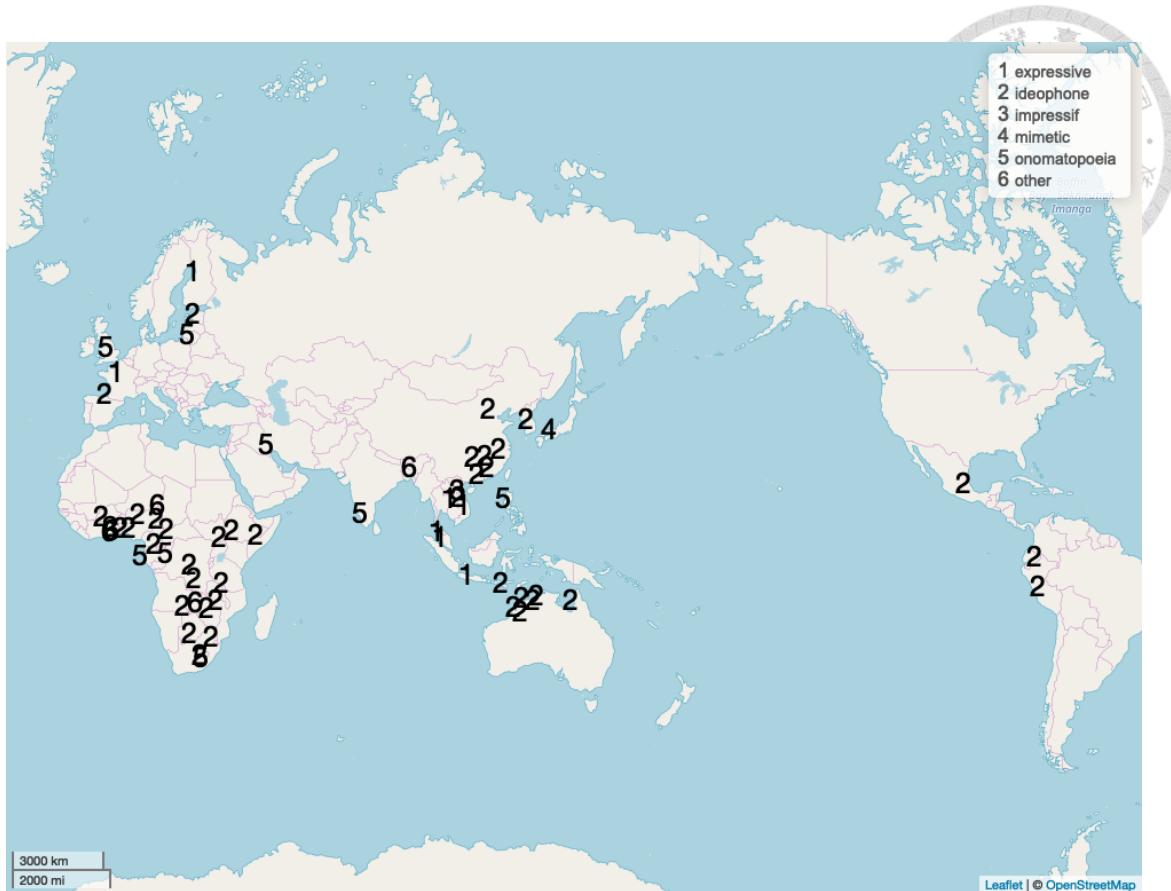


Figure 2.1: Map of languages for which ideophones have been described (non-exhaustive)

This visualization shows how widespread the phenomenon is. Furthermore, Africa and Oceania are mostly dominated by the term IDEOPHONE, while EXPRESSIVE occurs mostly in South-East Asia. MIMETIC is a term favored by Japanese linguists writing in English. Of these three terms, IDEOPHONE seems to be gaining the most coverage in recent years, especially when compared to other traditional terms such as ONOMATOPOEIA, which in its narrow sense only takes the modality of sound into its scope (see below).

2.2 Ideophones in the East

As stated above, the growing body of literature in Western languages on ideophones has achieved a certain extent of agreement on the terms that



can be used and the scope of data they pertain to. However, Japanese research on mimetics has also made great advances in the past three decades or so (see Kita 1993; Hamano 1998; Lu 2006; Akita 2009; Hiraga et al. 2015; Kizu & Cross 2017; Iwasaki, Sells & Akita 2017). While the Japanese language-internal studies retain the greatest volume, some studies have made their way into the English-speaking scholarly world. This has led to such cross-over studies as Dingemanse & Akita (2016) or Dingemanse et al. (2016), where Dutch speakers were tested for their attentiveness to real Japanese mimetic stimuli and performing in a statistically significant way.

This body of research has also not gone unnoticed by Chinese linguists. Zhao (2008) mentions that the comparative studies between Chinese and other languages, such as Japanese, make up an important array of work that investigates onomatopoeia and mimetics. It is then no surprise that the Japanese terminology has influenced the Chinese vocabulary used to describe these phenomena. The diagrams presented below in Figures 2.2-2.4 display the different terminological apparatus that can be found in across different linguistic communities.

We will first illustrate this for the Western conceptions of ideophones, followed by Japanese terms, and then address their Chinese parallels, in order to disambiguate the field. The *tertium comparationis* will consist of Dingemanse's (2012) implicational hierarchy of semantic domains that ideophones can occupy cross-linguistically.⁸ These semantic domains are SOUND, MOVEMENT, VISUAL PATTERNS, OTHER SENSORY PERCEPTIONS, and INNER FEEL-

⁸This may not be the hierarchy that best fits Chinese data, see Van Hoey (2016b) as well as Section 3.2.3.4.

INGS AND COGNITIVE STATES, illustrated with Chinese examples in (11). As Dingemanse argues, SOUND is extremely common across languages ('onomatopoeia' in their narrow interpretation), while INNER FEELINGS AND COGNITIVE STATES are rarer (Dingemanse 2012), but do occur more frequently in Japanese, Korean and Chinese.

- (11) a. SOUND *miāo* 喵 'mew, miaow'
- b. MOVEMENT *yōu~yōu* 悠悠 'move swiftly, quickly; smoothly slipping by, rapidly rushing'
- c. VISUAL PATTERNS *zhēng~róng* 嶙嶜 'craggy, precipitous masses of rock; loftily lifted'
- d. OTHER SENSORY PERCEPTIONS *bì~bì* 范范 'deeply fragrant'
- e. INNER FEELINGS AND COGNITIVE STATES *xǐ~xǐ* 恰恰 'glad and gay, happy and light-hearted, smug and satisfied'

As mentioned before, in Section 2.1 and the map in Figure 2.1, the current dominant term for the words under investigation in mostly Western approaches is IDEOPHONE, with a close contender being *mimetic*. Other terms, such as *expressive*, *impressif*, *echo-word*, *emphatic*, have also been used in the past (Childs 1994). On the diagram, presented in Figure 2.2, we can see that the top category covers the whole hierarchy. However, there also exists a body of literature that singles out ONOMATOPOEIA versus non-onomatopoeia, often referred to with IDEOPHONE in this kind of research. In this dissertation, the term ideophone generally understood on line A in the diagram. As a consequence, in a dichotomy as can be found on line B,

onomatopoeia⁹ are best described as SOUND IDEOPHONES versus NON-SOUND IDEOPHONES.

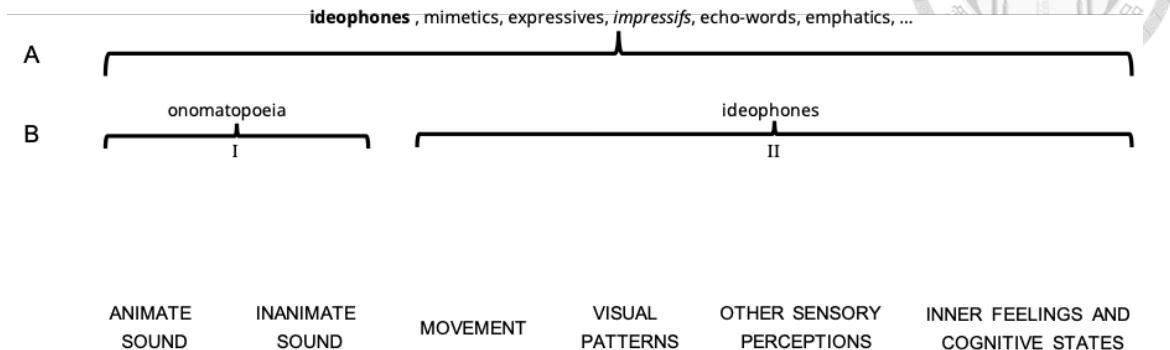


Figure 2.2: "Western" terminology for ideophones

For Japanese, the terminological apparatus is more complicated, as is diagrammed in Figure 2.3. A first grouping is simply to make no distinction between different sensory modalities depicted by these words (Line A in Figure 2.3). The Japanese use ‘onomatopoeia’ or ‘onomatopoe(t)ic words’ (*onomatope* オノマトペ) as a cover term for all sound-symbolic words (Akita 2009:9). This practice is still widespread. For example, Sasamoto (2019) uses *onomatopoeia*¹⁰ to refer to the broad range of Japanese mimetics.

⁹The term ‘onomatopoeia’ stems from the Greek *ónoma* ὄνομα ‘name’ and *poiéō* ποιέω ‘to make, to produce’.

¹⁰She states: “In Japanese linguistics, the term *mimetics* is generally used rather than *onomatopoeia*. Another term that is often used is *ideophone*, which, according to Akita and Dingemanse (2019), supersedes what is generally covered as onomatopoeia and includes a wide range of words that denote imagery from different sensory domains such as motion, texture, states, and sounds. While such distinctions could be useful in a descriptive study of onomatopoeia, it is not within the scope of this study to make such a distinction and, as such, the term *onomatopoeia* will be used to cover them all. Note that the use of the term *onomatopoeia* in this study is for convenience and is not to be taken as having any particular commitment to the debate on terminology” (Sasamoto 2019:1–2). It is not fully clear why she does not want to have a stake in this debate or how her study is not descriptive, as she provides a qualitative description of ideophones embedded in Relevance Theory, based on Sasamoto & Jackson (2016).

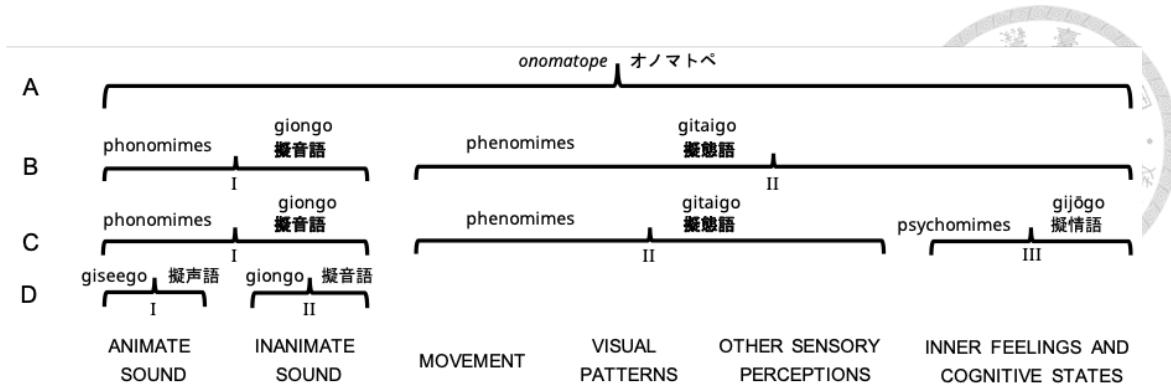


Figure 2.3: Japanese terminology for ideophones

Next, a special status is accorded to ideophones that depict SOUND (Groups B.I, C.I, D.I-D.II in Figure 2.3). Some Japanese studies further differentiate between ANIMATE SOUND and INANIMATE SOUND (Line D in Figure 2.3), calling them *gi-see-go* 擬声語 ‘mimic-voice-word’ and *gi-on-go* 擬音語 ‘mimic-sound-word’ respectively. However, the latter term is now the preferred term; a distinction between the two is not necessarily maintained according to the animacy of words (Akita 2009:11). Akita (2009) calls this group of sound-depicting ideophones PHONOMIMES. Non-phonomimes (Group B.II in Figure 2.3), in contrast, are usually called *gi-tai-go* 擬態語 ‘mimic-state-word’. It seems that most Japanese studies of mimetics adopt this semantic classification into phonomimes and non-phonomimes (Akita 2009:11; see also e.g. Lu 2006; Iwasaki, Sells & Akita 2017).

However, a final distinction that has gained some traction in the last years is allocating a special term to mimetics that depict INNER FEELINGS AND COGNITIVE STATES, namely *gi-joo-go* 擬情語 ‘mimic-feeling-word’. Akita (2009:11) differentiates this group of mimetics, because they “behave differently from the rest of non-phonomimes with respect to some grammatical phenomena and cross-linguistic distribution”. He thus discerns three

groups (see Line C in Figure 2.3): PHONOMIMES, PHENOMIMES (*gitaigo*) and PSYCHOMIMES (*gijoogo*), illustrated in (12).



(12) Examples of PHONOMIMES, PHENOMIMES and PSYCHOMIMES, adapted from Akita (2009:11–13)

- a. ネコが にやあにやあ 鳴きながら 出て きた。
neko-ga nyaanyaa naki-nagara de-te ki-ta
cat-NOM IDEO cry-while exit-CONJ come-PST
'A cat came out crying *meow-meow*.'
- b. 太陽が ぎらぎら 輝いて いる。
taiyoo-ga giragira kagayai-te i-ru
sun-NOM IDEO shine-CONJ be-NPST
'The sun is shining *glaringly*.'
- c. マイは 失恋に クヨクヨ 悩んで いた。
mai-wa situren-ni kuyokuyo nayan-de i-ta
PN-TOP lost.love-DAT IDEO worry-CONJ be-PST
'Mai was *worrying* about [her] lost love.'

Chinese approaches have been inspired by the distinctions made by Japanese studies of mimetics, as can be seen in Figure 2.4. For example, as a cover term (Line A in 2.4) it is not unusual to use *xiàng-shēng-cí* 象聲詞 ‘resemble-sound-word’ or *nǐ-shēng-cí* 擬聲詞 ‘mimic-sound-word’, although these may also be interpreted in the same way as *onomatopoeia* in Western linguistics. That is to say, they either are cover terms (e.g. Zhao 2008; Li 2018), or comprise only the sound-depicting group (e.g. Mok 2001;

Li 2007). For the latter group of scholars, if they do discuss the whole range of mimetics, the more common term for non-phonomimes is *nǐ-tài-cí* 擬態詞 ‘mimic-state-word’, *zhuàngtài-cí* 狀態詞 ‘state-word’ (Sun & Shi 2004; Kwok 2012), although *zhuàng-cí* 狀詞 ‘state-word’ also occurs (Li 2006). Mok (2001), who treats the issue in English from a phonological perspective for Mandarin, Cantonese, and Hakka, explicitly discerns ‘onomatopoeia’ from the rest (‘ideophones’).

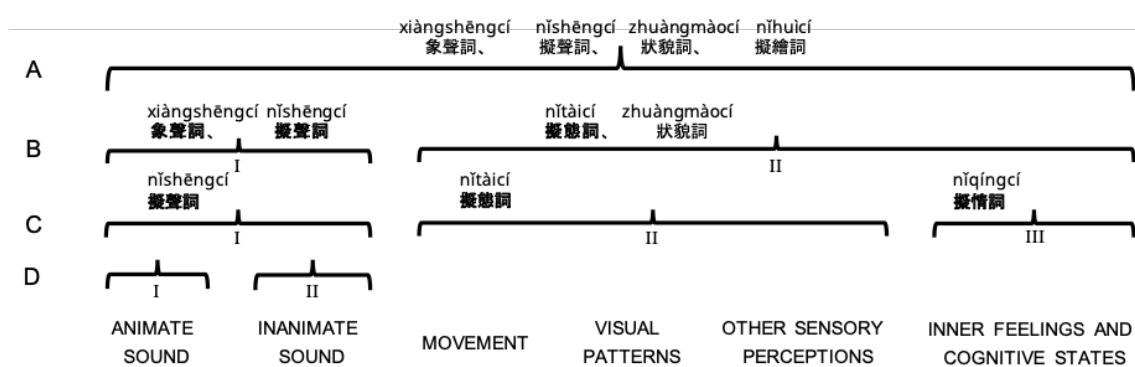
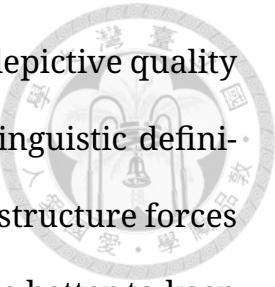


Figure 2.4: Chinese terminology for ideophones

The trichotomy proposed by Akita (2009) has also been discussed in other studies of Chinese ideophones. Meng (2012:21), for instance, sees their usefulness, but proposes her own three-fold classification to supplement Akita’s analysis to better analyze mimetics in the Beijing dialect: O-IDEOS (onomatopoeic words), A-IDEOS (ideophonic adjectives), and M-IDEOS (manner-mimicking words of mixed identities).

Van Hoey (2015) has also used Akita’s trichotomy in Chinese transliteration (13) as a starting point to investigate ideophones in the *Táng shī sān bǎi shǒu* 唐詩三百首 ‘300 Tang poems’. However, it became clear that the terminology in Chinese is ambiguous and vague: as a cover term for the whole range of IDEOPHONES Van Hoey (2015) ultimately suggested *hùi-yì-cí* 繪



意詞 ‘depict-meaning-word’; partly because it captures the depictive quality that is so important in Dingemanse’s (2011a; 2012) cross-linguistic definition; partly because the ‘mimic-SOUND/STATE/FEELING-word’ structure forces one to choose (a range of) sensory modalities. It thus seems better to keep one term as a cover term (IDEOPHONES or MIMETICS in English), but still keep a special term for those that depict sound, namely ONOMATOPOEIA (or NON-SOUND IDEOPHONES).

- (13) a. *nǐ-shēng-cí* 擬聲詞 ‘mimic-sound-word’
b. *nǐ-tài-cí* 擬態詞 ‘mimic-state-word’
c. *nǐ-qíng-cí* 擬情詞 ‘mimic-feeling-word’

While for Chinese the cover term of *nǐ-shēng-cí* has been adopted the most in treatments of ideophones – with mostly applications to phonomimes – the need for a cover term that unambiguously comprises phonomimes and non-phonomimes has also been voiced (Féng 2016). Féng proposes the term *nǐ-huì-cí* 擬繪詞 ‘mimic-depict-word’ as a solution to blend the sound-mimicking nature (*nǐ-shēng-fǎ* 擬聲法 ‘mimic-sound-way’) of one group of words and the depictive nature (*hùi-jǐng-fǎ* 繪景法 ‘depict-(land)scape-way’) of the other group. These two different natures were noticed by the well-known Chinese linguist Wáng Lì 王力 (1984:385; 1985), as Féng states. To summarize then, it is not claimed that previous terminology is wrong. However, it is useful to have an unambiguous cover term for all ideophones, while recognizing the special position sound-depicting ideophones hold. That is, onomatopoeia are often treated as a class on

their own, with special modes of (imagic) iconicity. Our terminological apparatus should reflect this, but currently the field has yet to agree on one term.



2.3 Chinese approaches to expressive words

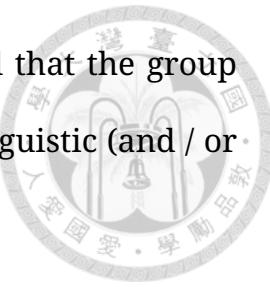
Until now, this chapter has traced the cross-linguistic literature, some Japanese perspectives on the terminology and the Chinese perspectives that came into contact with these two. However, such discussions using this particular terminology are still quite recent, and Chinese data is generally not included in cross-linguistic studies. This makes one wonder, why were Chinese ideophones (mimetics, expressives) largely left out of that debate, while they obviously occur in great numbers? The answer points in the direction of three language-internal subfields, each taking a different perspective on subsets of the same data. These three are: narrow vs. broad scopes of such onomatopoeia, binomes (*liánmiáncí* 聯綿詞) and reduplicative patterns, and character-based onomatopoeia.

2.3.1 The scope of onomatopoeia

The first of three subfields – the scope of onomatopoeia – has already been mentioned. That is to say, the terminology is ambiguous¹¹. Looking at the meaning of the different terms (see Section 2.2), most linguists have shown interest in investigating phonomimes, but a smaller number have used similar vocabulary to also include non-phonomimes. Since the former group

¹¹The terminology is ambiguous from the perspective of meaning of terms (semasiology) as well as that of naming, the terms (onomasiology).

has had more studies devoted to it, it is not unanticipated that the group of non-phonomimes may have gone unnoticed by cross-linguistic (and / or Western) studies.



2.3.2 Binomes and reduplicative patterns

A big part of reduplication literature is devoted to the Chinese phenomenon of *liánmiáncí* 聯綿詞 ‘binome’, also known as BINOMES (Kroll 2015) or TWIN-WORDS (Barnes 2007:115–130). Typical for these *liánmiáncí* is that they consist of two marked syllables that constitute one morpheme (Li 2013). These syllables can be marked by full reduplication or partial reduplication (or no reduplication).

This second subfield takes a morphological approach, rather than a form-function perspective present in the studies mentioned so far. More precisely, it is concerned with reduplication phenomena in Chinese – and it is through this perspective that Chinese has made its way into cross-linguistic studies (see Li 2015; Li & Ponsford 2018 for a recent approach).

An extensive overview of the history of the terminology is given by Xú (2013), and will be briefly sketched here. Xú states that the term itself is quite old, first appearing as *liánmiánzì* 聯綿字 ‘binome characters’, rather than *liánmiáncí* 聯綿詞 ‘binome’¹², in Zhāng Yǒu’s 張有 *Fùgǔ biān* 《復古編》‘*Returning to the Ancients*’, which collected 58 of these binomes. Later, during the Ming dynasty (1368–1644), works like *Gǔyīn pián zì* 古音駢字 ‘*Ancient*

¹²This is an important note, because up until today, distinguishing between ‘character’ (zì 字) vs. ‘word’ (cí 詞) remains an issue in Chinese linguistics (see Packard 1998; 2000)



sounds and parallel characters' by Yáng Shèn 楊慎 or *Pián yǎ* 駢雅 'Parallel elegance' by Zhū Mōuwěi 朱謀偉 continued along the same lines. The traditional approach reached its peak during the Qing dynasty (1644-1912) and Modern period, represented by Fāng Yǐzhì's 方以智 article *Liányǔ* 諺語 'Connecting words', Chéng Jìshèng's 程際盛 *Piánzì fēnjiān* 駢字分箋 'Analytical commentary on parallel characters', and Wáng Niànsūn's 王念孫 article *Liányǔ* 諺語 'Connecting words'; the latter by Wáng Guówěi's 王國維 *Liánmiánzì pǔ* 聯綿字譜 'Record on binome characters', and Fú Dìngyī's 符定一 *Liánmiánzì diǎn* 聯綿字典 'Dictionary of binome characters' among others. Wáng Niànsūn explains the phenomenon as follows:

凡連語之字，皆上下同義，不可分訓。

In general, characters of connecting words all consist of juxtaposed with similar meanings, that cannot be analyzed and explained separately.

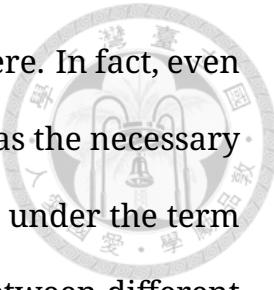
Wáng Niànsūn 王念孫 《讀書雜誌·漢書第十六·連語》 in Xú (2013:1–2), translation mine

These traditional authors took the character as the basic unit, rather than the morpheme, and thus included many cases that nowadays are analyzed as compounds rather than single morphemes. Xú (2013) lists ten categories, illustrated through the works of Wáng Guówěi and Fú Dìngyī, and exemplified here in (14).

- (14) a. *tóng-yì-lián.hé* 同義聯合 ‘same-meaning-connected’: *yǒngyuè* 蹦躍 ‘leap’
- b. *lèi-yì-lián.hé* 類義聯合 ‘category-different-connected’: *guǎng-dà* 廣大 ‘broad-big’
- c. *fǎn-yì-lián.hé* 反義聯合 ‘opposite-meaning-connected’: *xiān-hòu* 先後 ‘before-after’
- d. *piān-zhèng-shì* 偏正式 ‘tend-formal-style’: *jiā-yán* 嘉言 ‘commend-words > wise words’
- e. *zhǔ-wèi-shì* 主謂式 ‘subject-predicate-style’: *dì-zhèn* 地震 ‘earthquake > earthquake’
- f. *shù-bīn-shì* 述賓式 ‘modify-object-style’: *yǐn-qì* 飲泣 ‘drink-cry > swallow one’s tears’
- g. *shù-bǔ-shì* 述補式 ‘modify-complement-style’: *tián-mǎn* 填滿 ‘fill-complete > fill out [a form]’
- h. *fù.jiā-shì* 附加式 ‘add-style’: *péi-rán* 裴然 ‘long.and.flowing-RAN’
- i. *shuāng-yīn-xū-cí* 雙音虛詞 ‘double-sound-empty-phrase’: *hū-zāi* 乎哉 ‘!!-!!’
- j. *cí-zǔ* 詞組 ‘phrase-group’: *bù-zú* 不足 ‘not-(be.)enough’

In a way these classifications can be called precursors of modern flavors of construction grammar – the constructions consisting of two characters, which in these cases coincide with two syllables each. However, a cursory glance at the examples quickly reveals that much more than single mor-

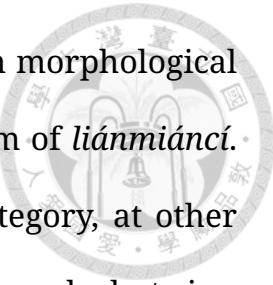




phemes consisting of two syllables (*liánmiáncí*) are listed here. In fact, even with this criterion of “two characters and one morpheme” as the necessary and sufficient condition for delineating the group of words under the term *liánmiáncí* ‘binome’, there has been some disagreement between different scholars as to the scope of words to be investigated. That is to say, Chinese linguists generally agree that reduplication is an important feature of these words, and thus divide the different kinds of characters composed of two characters into the following logical possibilities, presented in (15).

- (15) a. **full reduplication** (*dié-zì 疊字* ‘reduplicated-characters’ / *dié-yīn 疊音* ‘reduplicated-sounds’), e.g., *jīn~jīn* 津津 ‘flowing out (of water)’
- b. **alliteration (partial reduplication)** (*shuāng-shēng 雙聲* ‘double-sounds’), e.g., *sī~xū* 斯須 ‘a little while, a moment’
- c. **rhyme (partial reduplication)** (*dié-yùn 疊韻* ‘reduplicated-rhymes’), e.g., *pái-huái 徘徊* ‘pace up and down’
- d. **ablaut (partial reduplication)** (*shuāng-shēng jiān dié-yùn 雙聲兼疊韻* ‘double-sounds CONJ reduplicated-rhymes’), e.g., *líng~lóng* 玲瓏 ‘delicate (things), clever and nimble (girls)’
- e. **non-reduplication** (*fēi-shuāng-shēng dié-yùn 非雙聲疊韻* ‘NEG double-sounds reduplicated-rhymes’), e.g., *zhóu-lǐ 姉娌* ‘sisters-in-law’

The range of lexical items to be considered under the term of *liánmiáncí* has differed depending per scholar and period, as is diagrammed in Table



2.1. There is a general consensus that partial reduplication morphological mechanisms like alliteration and rhyme fall under the term of *liánmiáncí*. Sometimes, ablaut changes are included as a separate category, at other times they may not be recognized as a separate group of words, but simply included in alliteration or rhyme, depending on the item in question. It is also clear that the membership status of full reduplication is not always secure, let alone that of non-reduplication. That being said, Xú (2013) does include all of these formal properties as possible candidates for *liánmiáncí*, as long as they only contain one morpheme¹³. In other words, Xú (2013) regards them as NON-COMPOSITIONAL. Other special features he accords to *liánmiáncí* are their VIVIDNESS, their VARIATION IN THE WRITTEN FORM – which will be explored in subsequent chapters of this dissertation – and, once again, their marked forms (ALLITERATION, RHYME, PARARHYME (ablaut change), FULL REDUPLICATION and NON-REDUPLICATION).

¹³Xú (2013) further mentions the origins and mechanisms of *liánmiáncí*: interjections (*dòngqíng=de gāntàn* 動情的感歎 ‘sighing of emotions’); onomatopoeia (*ní-shēng-cí* 擬聲詞 ‘mimic-sound-words’); reduplication (*shēng.yīn chóng.dié* 聲音重疊 ‘sound reduplication’), resulting in nouns, verbs or adjectives; fossilization (*tóng-yì jìn-yì dān-yīn-cí=de lián-yòng* 同義近義單音詞的聯用 ‘connections of single-syllable words that are (near-)synonyms’); partial reduplication (*dān-yīn-cí=de yǎn-yīn* 單音詞的衍音 ‘sounds generated from single-syllable words’), which can either be BASE-REDUPPLICANT or REDUPPLICANT-BASE, in other words, leftward or rightward; loanwords (*wài-lái-cí=de yīn-yì* 外來詞的音譯 ‘loan translations’); and phonesthemes (*fúfùyīn shēngmǔ=de fēnlì* 複輔音聲母的分立 ‘discrimination of consonant cluster initials’). As can be seen in this overview, this grouping of *liánmiáncí* does not entirely consist of homogeneous categories. The focus here is mostly on the function of the *liánmiáncí*, while the latter four highlight the different mechanisms. This is probably due to a polysemy of the Chinese term *láiyuán* 來源: the former group answers the question “what is the source, the referent?” but the latter answers to the question about the constructional formation, “how were these formed?” (*zěnme láide?* 怎麼來的?).



Table 2.1: Overview of what is to be considered *liánmiáncí*

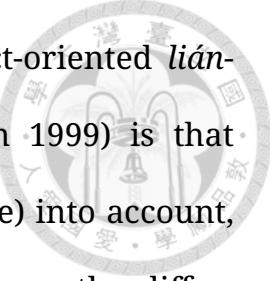
FULL REDUPLICATION	ALLITERATION	RHYME	ABLAUT	NON- REDUPLICATION	representatives
✓	✓	✓	not differentiated	✓	Wáng Guówěi ¹⁴ , Fú Dìngyī, Zhōu Fǎgāo <i>Cíhǎi</i> 辭海
✓	✓	✓	not differentiated	✗	Wáng Lì, Yáng Bójūn, Zhōu Dàpú
✗	✓	✓	✓	✗	Jiǎng Lǐhóng, Rèn Míngshàn
✗	✓	✓	not differentiated	'different kind'	Zhōu Bǐngjūn, <i>Hànyǔ dà cídiǎn</i> 漢語大詞典
✗	✓	✓	✗	✗	Approach at Northeast Normal University 東北師範大學
✓	✓	✓	✓	✓	Dǒng Wéiguāng
'full'	'partial'	'partial'	'partial'	✗	Chéng Xiāngqīng
✓	✓	✓	✓	✓	Xú Zhènbāng

¹⁴These are the date spans of these scholars: Wáng Guówěi 王國維 (1877-1927), Fú Dìngyī 符定一 (1877-1958), Zhōu Fǎgāo 周法高 (1915-1994), Wáng Lì 王力 (1900-1986), Yáng Bójūn 楊伯峻 (1909-1992), Zhōu Dàpú 周大璞 (1909-1993), Jiǎng Lǐhóng 蔣裡鴻 (1916-1995), Rèn Míngshàn 任銘善 (1912-1967), Zhōu Bǐngjūn 周秉鈞 (1916-1993), Dǒng Wéiguāng 董為光 (1946-), and Chéng Xiāngqīng 程湘清 (1937).

It is important to note that Xú (2013) is not the only representative of this subfield of reduplication. A crucial contribution (Sun 1999; 2008) studied reduplication in Old Chinese. He does not adopt a product-focused perspective, instead of with the by-now familiar terms of ALLITERATION, RHYME, PARARHYME (ablaut change), FULL REDUPLICATION and NON-REDUPLICATION. Instead, Sun (1999) looks at the process of the marked form. He discerns two big groups of processes: directional reduplication and non-directional reduplication. They each comprise two patterns: respectively the progressive pattern (16a)¹⁵ and retrogressive pattern (16b), and fission reduplication (16c) and total reduplication (16d). In (16) the Mandarin Chinese phonology is given with reconstructed Middle Chinese (MC) and Old Chinese (OC), followed by the characters and a gloss and explanation.

- (16) a. *cōng~róng* < MC dzjowng~yowng < OC *[dz]oŋ~*[G]ron 從容 ‘at leisure, casually’,
 which follows BASE~REDUPLICATE
- b. *zhǎn~zhuǎn* < MC trjenX~trjwenX < OC *tre[n]?~tron? 輾轉 ‘toss and turn’,
 which follows REDUPLICATE~BASE
- c. *jīng* < MC tsjeng < OC *tsenj 精 ‘astute, smart’
 becomes *jí~líng* < MC tsik~leng < OC *[ts]ik~[r]in 卽零
- d. *mò~mò* < MC mak~mak < OC *m^fak~m^fak 莫莫 ‘quiet’

¹⁵ Sun (1999) uses a slightly different reconstruction system, namely that of his advisor Pulleyblank. Here are the corresponding transcriptions: For *cōng~róng*: Eearly Middle Chinese {tsʰuàwŋ~juàwŋ} < OC {tsʰaŋ~laŋ} (1999:60); for *zhǎn~zhuǎn*: EMC trian’~trwian’ < OC trən?~trwən? (1999:99–102); for *jīng* > *jí~líng*: *tsiajŋ* > * *tsiak liajŋ* (1999:133), note that this is a Middle Chinese occurrence, as example 16c shows.



What is typical for these two approaches, the product-oriented *lián-miáncí* and process-oriented reduplication patterns (Sun 1999) is that they take two syllables (and thus two characters in Chinese) into account, but do so from a phonological point-of-view. They compare the different consonants and nuclei in the two syllables, to see how they may be marked. This stands in stark contrast with the last subfield of “traditional” Chinese perspectives, which is based on the writing system rather than the phonology, and which forms the point of departure for studies like Mok (2001).

2.3.3 Character-based onomatopoeia

One of the most popular ways of analyzing onomatopoeia (and ideophones) in Chinese is by looking at the number of same and different characters that make up a given items. That is to say, items are characterized in what could be called “ABB-notation”. A host of recent comparative studies employ this line of thinking, such as Yu (2004), Wang (2014), Tran (2017), Kǒng (2017), and Le (2017). A thorough study that also takes this as a starting point by Mok (2001) contains the following patterns for her two groups of onomatopoeia (phonomimes) and ideophones (non-phonomimes), shown in Table 2.2.

Table 2.2 shows the 30 different syllabic/character patterns that Mok (2001) identified. In her notation, $\langle A \rangle$, $\langle B \rangle$, and $\langle C \rangle$ stand for sound-symbolic elements in a word, while $\langle X \rangle$ and $\langle Y \rangle$ indicate non-sound-symbolic elements, usually nouns, verbs or adjectives, or a related

Table 2.2: Frequencies of onomatopoeic and ideophonic patterns for Mandarin, Cantonese and Hakka in Mok (2001:45-46)

pattern	mand_onom	mand_ideo	canto_onom	canto_ideo	hakka_onom	hakka_ideo
A	79		69		47	
AA	38		14		10	
AA'	15		16			
AA"	78					
AB	163		2		2	
AAA	33		3		16	
AAA'			2			
AA"A"	37					
AAB			1			
AAX	3		93	25	82	13
AA'X	1					
AA"X	2					
ABX	2					
ABB	87					
XAA	3	450	5	207		191
AAAA	2		4		1	
AA'AA'	2		7		4	
AAA'A'	14		29		20	9
AAA"AA"	7		1		3	
AAA"A"			2		1	5
AABB	16		6		13	16
ABAB	6		3		4	
ABA'B	1					
ABA'B'	6					
AA"BB"	4		1			
AA"BB"/ABA'B'	27		20		8	
AA"BB"/ABCB'	20					
AA"BC	5					
AA"BC/ABCB'	13					
AA"BC/ABA'B'	3					
ABCB		6				
ABCB'		21				
AA'XY				4		
AAAAA				2		
XA'A/XAA'				9		6
XYA'A/XYAA'				1		
XA"A/XAA"				11		
XYA"A/XYAA"				7		
XYAA				1		
XAB				2		
AA'BB'						2

phrase. A single apostrophe <’> means that the duplicate syllable alliterates, while a double apostrophe <”> denotes rhyme. This type of analysis, which has been very influential in Chinese linguistic studies of onomatopoeia, is product-oriented: it takes into account what syllables are the same (or partially the same), but does not generalize over the processes

that were used to arrive there, like e.g. Sun (1999) does. Parallel notations can be found in Japanese (and Korean) mimetic studies, where an item like *pika~pika* ピカピカ ‘glitter, sparkle’ is analyzed as ABAB.

There are many types for onomatopoeia in Mok (2001) (Table 2.2). It appears that the ones with highest type frequency in Mandarin are AB (n=163), ABB (n=87), A (n=79), and variants of AA (n=38 for full reduplication, n=15 for alliteration, and n=78 for rhyme), see (17). Compare this to Cantonese onomatopoeia where the dominant types are AAX (n=93) and A (n=69), shown in (18). Hakka onomatopoeia apparently have similar frequent patterns as to Cantonese ones: AAX (n=82) and A (n=47), as can be observed in (19). It is of significance that Mandarin only has a few AAX onomatopoeia (n=3), according to Mok’s study. Other patterns, with much lower type frequencies than the aforementioned ones are also included in Mok’s summary. Note that Mok provides the data in IPA transcription, without tone marking or characters. Therefore, we have followed this convention here.

(17) Examples of Mandarin onomatopoeia patterns in Mok (2001)

AB /pu tɕi / ‘fizz’

ABB /pu tɕi tɕi / ‘fizz’

A /kɑ / ‘quack; sound of laughter; sound of squeezing’

AA /kɑ kɑ / ‘quack’

AA’ /ti ta / ‘tick tock; sound of dripping water’

AA” /kɑ ta / ‘crash; sound of opening and closing automatic doors’

- (18) Examples of Cantonese onomatopoeia patterns in Mok (2001)

AAX	/tap tap səŋ /	'sound of licking or tasting something'
A	/fɪt /	'fizz'



- (19) Examples of Hakka onomatopoeia patterns in Mok (2001)

AAX	/pʰo pʰo kun /	'sound of running water'
A	/tɕia /	'chirp'

However, looking at Mok's (2001) ideophones for the three languages, the picture is radically different, see examples (20-22). Clearly, the overwhelmingly dominant pattern is XAA (Mandarin n=450, Cantonese n=207, Hakka n=191). For Cantonese and Hakka, the next pattern is AAX (Cantonese n=25, Hakka n=13), but for Mandarin, this type apparently does not exist. This gap for Mandarin indicates that it is special, when compared to the other two Sinitic variants, or that it has not been recorded, or that perhaps there is something wrong with the classification scheme.

- (20) Examples of Mandarin ideophone patterns in Mok (2001)

XAA	/ɕian pʰu pʰu /	'fragrant'
-----	-----------------	------------

- (21) Examples of Cantonese ideophone patterns in Mok (2001)

XAA	/hək mɔ mɔ /	'dark'
AAX	/tsat tsat tʰiu /	'jumping up and down'

- (22) Examples of Hakka ideophone patterns in Mok (2001)

XAA	/ko tsʰan̩ tsʰan̩ /	'very tall'
AAX	/lit lit son /	'turning round and round'

Mok (2001) distinguishes quasi-ideophones from ideophones proper.

That means that the examples above all illustrate real ideophones, while quasi-ideophones, on the other hand, may look similar, especially with respect to reduplicated elements, but do not contain any sound-symbolic elements whatsoever (Mok 2001:7–9). Examples of Cantonese quasi-ideophones are presented in (23). While Mok (2001) rightfully points out that this conceptual distinction should be made on the basis of sound symbolism, from the perspective of construction grammar there hardly seems any distinction. And while she maintains that she will not be analyzing quasi-ideophones in her work, it is curious that her data does seem to contain them. For instance, is *mɔ mɔ* in *hek mɔ mɔ*, seen in (21), not just a reduplication of what in Mandarin is *mò 黑* ‘dark’? The point here is not that /m/ is not a sound symbolic phonestheme (see Section 5.1.1) but rather, that other, prosaic, words can be coerced into an ideophonic construction, a process referred to as ideophonization (Dingemanse 2017). The issue of ideophonization will only be briefly touched upon in this dissertation. While this process warrants further investigation, pioneering studies exist, e.g., how reduplication of prosaic words can be coerced into ideophonic constructions (Liú 2009).

- (23) Examples of Cantonese quasi-ideophones in Mok (2001:7–9)

<u>Quasi ideophone</u>		<u>Original form</u>
[pʊk6 lei1 lei1]	‘hide and seek’	[pʊk6] ‘lie down’ + [lei1 lei1] ‘hide’ (V+V)
[ŋan6 tsam2 tsam2]	‘winking’	[ŋan6] ‘eye’ + [tsam2] ‘wink’ (N+V)
[ŋan3 sep1 sep1]	‘tearful eyes’	[ŋan3] ‘eye’ + [sep1 sep1] ‘wet’ (N+A)
[sei2 pan2 pan2]	‘stubborn’	[sei2] ‘dead’ + [pan2] ‘board’ (AN compound)

Another useful distinction made by Mok (2001) in her data is the distinction between sound-symbolic elements and non-sound-symbolic head morphemes or suffixes (Mok 2001:44), marked by X in the examples above, e.g., ‘XAA’ and ‘AAX’. These are notational variants of the well-studied ABB and BBA constructions (see Bodomo 2006; Sew 2008; Lai 2015 for Cantonese; Chang 2009 for Southern Min; and Cáo 1995; Mok 2001; Zhāng 2005; Yáo 2006; Lǐ 2008; Sūn 2012; and Wang 2010; 2014 for Mandarin). It is, in our view, more fruitful to treat these items as constructions of a non-ideophonic collocate and an ideophone, i.e., the AA and BB part of these items are seen as ideophones, and the B and A as collocates. As a result, from /tap tap səŋ /in (18), the /səŋ /part, presumably cognate to Mandarin *shēng* 響 ‘sound’, should be left out of the ideophone proper. However, when studying how this ideophone /tap tap /is used, one may find that a collocate /səŋ /indeed is the most common. Or another example is /cian pʰu pʰu /in (20), where /pʰu pʰu /is the ideophone and /cian /, in Pīnyīn *xiāng* 香 fragrance, fragrant is

the collocate.

This distinction between collocate and ideophone is maintained throughout the larger part of this dissertation. The main reason for doing so is that there exists an introspective bias, based on dictionary data, to include only a few collocate-ideophone combinations. However, within a usage-based paradigm, we will find that a probabilistic model of different combinations reflects the usage of ideophones better. This issue will be revisited in depth in Chapter 7.



2.4 The need for Chinese ideophones as a category

Given the overview thus far, the reader may be wondering if there really is a need for a Chinese ideophone category, and how that category is to be structured. Let us review the points made in this chapter. A number of quotes made it clear that there is an almost denial of the abundance of onomatopoeia in (Mandarin) Chinese. This was countered by Wu (2014), from whom it can be inferred that as a near-universal of languages, an absence of ideophones in Chinese would be typologically peculiar, especially since they do occur in Sinitic languages. A natural consequence of defining ideophones as marked words that depict sensory imagery, belonging to an open lexical class (Dingemanse 2019) is that a number of previously separately studied phenomena of Chinese can be taken under this umbrella term. First and foremost, onomatopoeias make a good candidate, as they are well-known for depicting sounds in marked ways. Through comparison with Japanese, it was also made clear that other depictive items sometimes have pendants

in Chinese. These can also be incorporated under the term ‘ideophone’. Another group of words that are morphologically marked are binomes. A large number of binomes, though not all of them, can be seen as ideophones as well, because they share that depictive quality.

This dissertation is of course not the first time that these things are being argued for. For instance, Mok (2001) already includes onomatopoeia and ideophones. However, ideophones in Mok (2001) are interpreted as non-sound depicting sound-symbolic items. Mok thus follows the Line B in Figure 2.4. This dissertation, on the other hand, falls more in line with current typological research, where the term ideophone is seen as an overarching category, comprising (possibly) multiple sub-categories, i.e., Line A in Figure 2.4. If these fields that seemingly existed parallel to one another, such as onomatopoeia research, binome studies, or sound-symbolic investigations, are converged, the scope of the terminological apparatus is confronted with similar cases that it would not even consider in the first place. For example, an ideophone like *xīlīhuālā* can be interpreted in a number of ways, as shown in (24). For some perspectives, it can be considered a valid word, for others it falls outside the scope of analysis. Is it not better then to see it as an instance of the Chinese ideophone category, namely one that is a depiction of VISION, with a possibly synesthetic component to it? It is marked, but as a product it is unclear if we should categorize it with ABCD, AA”BC or ABCB’, or even with the more fine-grained character-based analysis. From a process point-of-view, which will be adopted in the Chinese Ideophone Database (see Section 3.2), no Base can be found in this item, and

we thus categorize it as a Reduplicant-Reduplicant-Reduplicant-Reduplicant ('RRRR')¹⁶ template.

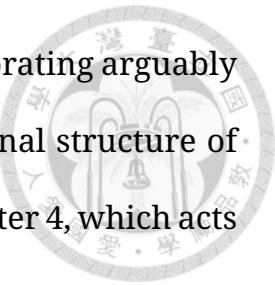


- (24) a. *xīlīhuālā* 希哩嘩啦 ‘streaming and sluicing, heavy rainfall’
- b. In “ABB-notation”: strictly speaking ABCD, but Mok gives AA”BC / ABCB’ (2001:227)
- c. In reduplication literature: $(AB)_{\text{high V}} (CD)_{\text{low V}}$, or $(A B_{\text{liquid}})_{\text{high V}} (C D_{\text{liquid}})_{\text{low V}}$ for a slightly more finegrained analysis.
- d. According to binome literature: impossible to analyze, falls out of the scope.
- e. Under a narrow scope of ‘onomatopoeia’ (only phonomimes): falls out of the scope.
- f. Under a broad scope of ‘onomatopoeia’ (including non-phonomimes): acceptable.

To reiterate then, an overarching category of ‘ideophones’ is necessary. Dingemanse’s cross-linguistic concept is a good starting point, because it contains a formal and a functional component. That way it converges a number of phenomena that were studied separately. Furthermore, it allows us to collect data from different sources, i.e., dictionaries and previous studies. Together with corpora, these data can provide a more comprehensive foundation for the study of different instances of the Chinese ideophonic lexicon (see Chapter 3). It should, however, be borne in mind that this lexi-

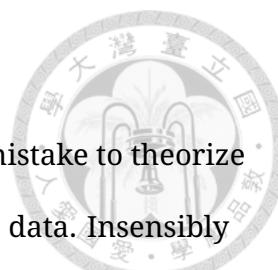
¹⁶This is also not wholly satisfactory, but it will facilitate the analyses that are performed in this piece of research.

con is not homogeneous – a natural consequence of incorporating arguably different linguistic phenomena. The boundaries and internal structure of the category of Chinese ideophones will be revisited in Chapter 4, which acts as a sequel to the current chapter.





3 Data sources



It is a capital mistake to theorize
before one has data. Insensibly
one begins to twist facts to suit
theories, instead of theories to
suit facts.

Sherlock Holmes

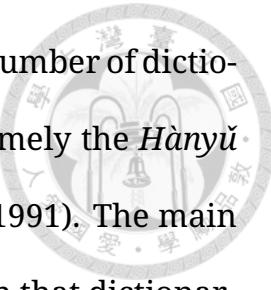
After surveying the state of ideophone studies and related field in Chapter 2, it is time to discuss the data sources used in the remainder of this dissertation. They fall into three categories: dictionaries (Section 3.1), a new database for Chinese ideophones (Section 3.2), and a selection of corpora (Section 3.3). The latter two sources are presented in an overview in Table 3.1.

Table 3.1: An overview of the main sources of data used in this dissertation

data source	abbreviation
Chinese Ideophone Database	CHIDEOD
Scripta Sinica corpus	Scripta Sinica
Academia Sinica Balanced Corpus of Modern Chinese	ASBC, or Sinica Corpus
Diachronic Chinese Ideophone Corpus	DIACHIC

3.1 Dictionaries

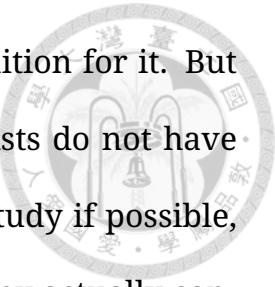
A straightforward datasource for a lexical semantic study is the consultation of lexicographic material, i.e., dictionaries. However, dictionaries are



not directly used as a separate resource here. We include a number of dictionaries in the Chinese Ideophone Database (Section 3.2), namely the *Hànyǔ dà cídiǎn* 漢語大詞典, Kroll (2015), Wáng (1987), and Gōng (1991). The main reason for not listing them directly is that it has been shown that dictionaries can provide rich information that is, however, incomplete. That is, practical concerns of (finite) paper dictionaries, or conceptual choices made by lexicographers all shape the dictionary, and tie into this issue. Some helpful overviews can be found in Geeraerts (2001); (2003); and Adamska-Sałaciak (2015), among others.

Despite these concerns, in historical studies, dictionaries can contribute important data about the introspective insights of the lexicographer, as Geeraerts (1997:172) claims for the synonym dictionaries he consults to study the meaning of *vernielen* ‘to destroy’ and *vernietigen* ‘to destroy’. The synonym dictionaries reflect the contemporary understanding of these items, and as such form an important source of data. Chinese definitely has a long history of those kind of dictionaries, as is surveyed in e.g. Xue (1982) and updated in Xue (2003).

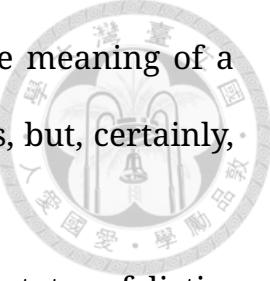
The earliest dictionary referred to by most semantic studies regarding Chinese words is the *Shuōwén jiězì* 說文解字 ‘Explaining Graphs and Analyzing Characters’. Notwithstanding the criticisms such as those voiced by Chi (2014), it still remains a beloved starting point, with many studies taking the synonym definitions provided there as the “real etymological meaning”, e.g., Chen et al. (2019). This is of course not true; rather, it merely reflects what Xǔ Shèn 許慎 (58-148), the editor of the *Shuōwén jiězì*, considered to be



the etymology of a lemma, or a fitting contemporary definition for it. But that is not to say that these early dictionaries and word lists do not have any value; on the contrary, they should be included in a study if possible, yet should not be accorded more importance than what they actually contribute. A semantic study should then also look at language as it is used in context, and verify how the two compare. As for its legacy, the *Shuōwén jiězì* initiated the dictionary type of character dictionaries (*zì shū* 字書), and has successors in the *Zìhuì* 字彙 (1615) and the even more famous *Kāngxī Zìdiǎn* 康熙字典. Even now it is still a popular model for the arrangement of modern Chinese dictionaries, such as the *Hànyǔ dà zìdiǎn* 漢語大字典 (Great Compendium of Chinese Characters) (Xǔ 1995).

Although these dictionaries offer a wealth of information, they are nonetheless character dictionaries, and that is the biggest drawback. As we have seen in Chapter 2, many Chinese ideophones are not confined to a character which conforms to the one-syllable-one-meaning principle, as Sun (1999) has called it, one syllable also equating to one character. Maybe the *Hànyǔ dà cídiǎn* 漢語大詞典 can be of help. After all, it uses a notion of *cí* 詞 as the main unit for the lemmata, instead of individual characters, , see Section 4.2.2. This source is also included in the Chinese Ideophone Database, as well as other specialized onomatopoeia dictionaries in the Chinese Ideophone Database. But even then the definitions provided for a lemma are classified by the editors into different meanings based on, and illustrated by authoritative quotations. The crux of the problem for semantic studies is there: how can the user trust that the right lexicographic

choices have been made? If he or she wants to know the meaning of a word, they should get a guiding direction from dictionaries, but, certainly, no dictionary would want to claim that it is exhaustive.



Added to this issue is the discussion about the theoretical status of dictionaries, especially as opposed to encyclopedias. An early voice in this debate (Haiman 1980) concluded that all meaning is, in fact, encyclopedic in nature, and saw dictionaries as a reduced entity underneath the larger category of encyclopedias. More debate ensued: Frawley (1981) vehemently disagreed with Haiman's (1980) conceptualization, which was then somewhat hastily countered by Haiman (1982). Yet, with the arrival of Cognitive Linguistic approaches, the voices arguing in favor of the encyclopedic approach grew louder, as is surveyed in detail by Peeters (2000). That is to say, while we subscribe to the idea that our semantic knowledge is very much encyclopedic in nature, dictionaries as a data source for lexical semantic investigation are helpful yet limited in their aid.

We see such discontent with traditional dictionaries in the writings of a number of ideophone lexicographers. Writing about Zulu in particular and Bantu in general, de Schryver & Dje (2009:38) state that “ideophones are a lexicographer’s worst nightmare”, because in Zulu they constitute a small category, for which the metalexicographic description is highly relevant. They need to be defined not too broad neither too specific in order to capture the depiction, as illustrated in example (25). An adequate dictionary definition thus should strive to be balanced.

(25) Dictionary definitions for the Zulu ideophone *mpo* (de Schryver & Dje

a. (*of being erect*)

Wavuka wama mpo uGovu wathi ... *Govu woke up, stood straight up, and said: ...*

b. (*of extreme action*)

Sekubanda, mpo, eNingizimu. *It is now cold, extremely cold, in the South.;*

Kuthe ilanga selithe, mpo, ibandla lihlezi phansi nje emthunzini. *When it is extremely hot, the group of men simply sits on the ground in the shade.*

c. (*of being full*)

Ligcwaliseni isaka lithi mpo *Fill up the bag to the brim.*

But does such a definition with examples truly capture the meaning of the ideophone? A minimalist perspective in lexicology would say that the referents shown in example (25) are not of major importance, i.e., the '*of being erect*' is not essential to fundamentally understand the meaning of a given ideophone. Natural Semantic Metalanguage (Wierzbicka 1972; Goddard & Wierzbicka 1994; 2002; 2014a), a neostructuralist framework within the development of lexical semantics (Geeraerts 2010b) that takes this perspective, defines terms based on atomic universal semantic primitives. Currently, there are 65 such primitives¹⁷ (Goddard 2018). It is argued

¹⁷The currently 65 semantic primitives in Natural Semantic Metalanguage I~ME, YOU, SOMEONE, SOMETHING~THING, PEOPLE, BODY, KINDS, PARTS, THIS, THE SAME, OTHER~ELSE, ONE, TWO, MUCH~MANY, LITTLE~FEW, SOME, ALL, GOOD, BAD, BIG, SMALL, THINK, KNOW, WANT, DON'T WANT, FEEL, SEE, HEAR, SAY, WORDS, TRUE, DO, HAPPEN, MOVE, BE (SOMEWHERE), THERE IS, BE (SOMEONE/SOMETHING), (IS) MINE, LIVE, DIE, WHEN~TIME, NOW, BEFORE, AFTER, A LONG TIME, A

that through a script using these primitives, the core meaning of a given term can be explicated. For instance, the Japanese mimetic *doki~doki* ドキ・ドキ ‘pitter-patter’ has been defined in (26) according to the principles of Natural Semantic Metalanguage (Hasada 1994).

(26) Meaning of DOKI DOKI

X feels something

sometimes a person thinks something like this:

I know now: something will happen to me

something is happening to my heart because of this

because of this, this person feels something

this person feels like someone who thinks this:

‘I hear the sound of my heart beating’ [doki-doki]

this person feels this for some time X feels like this

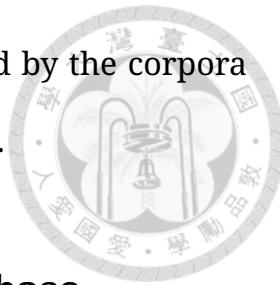
Hasada (1994:111–114)

While a certain amount of psychological insight is obtained through such a definition, it does not inform us how to use it and in what context. Furthermore, such definitions have been critiqued for their lack of cognitive plausibility (see in particular Geeraerts 2010b:127–137 for an overview), especially in terms of chunking. That is to say, when a situation is characterized as *doki~doki*, do we follow this script every time, as suggested by this definition; or do we gradually associate the kind of situation that can be depicted with *doki~doki*? The Multimedia Encyclopedia of Japanese

SHORT TIME, FOR SOME TIME, MOMENT, WHERE~PLACE, HERE, ABOVE, BELOW, FAR, NEAR, SIDE, INSIDE, TOUCH, NOT, MAYBE, CAN, BECAUSE, IF, VERY, MORE, LIKE~AS, see Goddard & Wierzbicka (2014a).

Mimetics (Akita 2012a; Akita 2016) indicates that *doki~doki* mostly collocates with *mune* 胸 ‘chest, heart’. But even this may not be enough information for a maximalist perspective. Since it is well-known that ideophones often co-occur with gesture, special foregrounding, as well as other means of markedness (cf. Section 4.2.1), a maximalist perspective would require converging all possible evidence, in order to capture the meaning of an ideophone in a usage-based sense. There are a number of projects that strive to do this, such as Akita’s Multimedia Encyclopedia of Japanese Mimetics and Nuckolls’s Anti-dictionary, which will be briefly discussed in Section 3.2, or analysis of ideophones based on Image Schemas (Nuckolls 1996), Idealized Cognitive Models (Lu 2006), or Frame Semantics (Akita 2012b), which will be discussed in Section 5.2.1. Such definitional frameworks all require that referents be present to observe how ideophones are used in context, rather than positing a script that arguably captures the prototypical usage of a given ideophone. And rather than the concise dictionary model, they take the encyclopedic model in lexical semantics as their point of departure (Peeters 2000). In other words, lexicographic sources are not bad per se, but they need to be supplemented with real data. The way this is operationalized in this dissertation is by collecting data on ideophones from paper dictionaries (lexicographic sources) and previous studies into one large-scale database. Based on these data, a large number of other parameters (descriptive, analytical, etc.) are investigated, so as to go beyond a mere amalgam of dictionaries. The database can then be connected to corpora, providing the usage-based data that can be probed with the corpus linguistics methodology.

ical apparatus. Let us now turn to that database, followed by the corpora that will provide the basis for investigations of their usage.



3.2 CHIDEOD – the Chinese Ideophone Database

This section introduces the Chinese Ideophone Database¹⁸ (CHIDEOD), a newly constructed database aimed at supporting research on Chinese ideophones, from both a diachronic and synchronic perspective. It is constructed under the CC BY-NC-SA 4.0 license.

Currently there is no digital means for examining Chinese ideophones from either semantic, phonological, historical, and orthographic perspectives. Until now researchers interested in Chinese ideophones have had to rely on a scattered and limited number of largely un-digitized resources, only three of which are exclusive to ideophones, such as dictionaries (Gōng 1991; Wáng 1987) or detailed studies of onomatopoeia (Lǐ 2007). A major issue is that detailed phonological studies exist, but they only provide IPA transcriptions of the data, making it difficult to replicate and reuse those studies orthographically with native speakers (e.g., Meng 2012; Mok 2001). In order to systematically investigate the nature of semantic, morphophonological, and historical properties of Chinese, the Chinese Ideophone Database (CHIDEOD) was devised, an open-source framework for studying ideophones across a large number of parameters, including data from Mandarin as well as Middle and Old Chinese. This makes it

¹⁸Van Hoey & Thompson. In press. This section represents an adapted version of that article. The database was first presented at the 12th International Symposium on Iconicity in Language and Literature (ILL 12), see Van Hoey & Thompson (2019).

possible to share data in a reusable fashion (see Forkel et al. 2018), as well as provide transparency into the sources of a given item on which an analysis can be based.

The database is mostly conceived in conjunction with the R programming language (R Core Team 2019), and is stored in a repository of the Open Science Framework¹⁹. In this repository, one can find the most recent versions of the database, in three formats: Comma Separated Values (.csv), Excel (.xlsx) and R Data Structure (.rds). For the R-agnostic researcher who wishes to make use of CHIDEOD, I have also developed an online app version²⁰ that allows you select the variables you need and gives a search functionality. Data from the app version can be exported as an Excel file, a CSV file or copied to the clipboard for further analysis. A tutorial will be provided in Section 3.2.6.1.

Similar database projects have been undertaken for other languages, e.g., the Word Profiler for the Balanced Corpus of Contemporary Written Japanese or BCCWJ (NINJAL 2016), Multimedia Encyclopedia of Japanese Mimetics or MEJaM (Akita 2012a; Akita 2016), and Quechua Realwords, an audiovisual corpus of Quechua expressive ideophones (Nuckolls 2016a–; Nuckolls et al. 2017; Nuckolls & Swanson 2019). They each have their own research question in mind. The BCCWJ corpus is mostly geared toward generating frequency information and provides no semantic information such as definitions or whether an ideophone is literary rather than colloquial and orthographic preferences are not specified. MEJaM is an audiovisual collec-

¹⁹ Available at <https://osf.io/kpwgf/>.

²⁰ Available at https://simazhi.shinyapps.io/chideod_appversion/.

tion of ideophones to illustrate the visual depiction of Japanese mimetics. Quechua Realwords provides detailed semantic definitions and examples of each ideophone used in natural speech. However, what Quechua Realwords makes up for in semantic description, it currently lacks in morphophonological detail such as information about reduplication or whether a form is phonologically marked.

For Chinese ideophone data, we wanted to provide an alternative to these three projects: a database that contains detailed information about morphology, dictionary definitions, orthography etc. CHIDEOD is not a finished product, but a place to store data from previous studies for which a number of variables have been entered, e.g., the traditional and simplified character variants. The database is modeled after the Chinese Lexical Database (Sun et al. 2018). As of yet (version 0.9.3), CHIDEOD comprises 4948 unique onomatopoetic and ideophonic items in traditional characters. The dimensions of the whole database are 14925 rows and 59 columns, i.e., values and variables.

Below, the variables are introduced. CHIDEOD has five major types of variables: data variables, descriptive variables, analytical variables, frequency variables and other variables. A summary is provided in Figure 3.1. Note that bold capital **N** stands for four variables, as will become clear in the discussion concerning Table 3.3.



DATA VARIABLES	DESCRIPTIVE VARIABLES	ANALYTICAL VARIABLES
	PHONOLOGY	
data_source	pinyin_tone pinyin_tonenumber pinyin_without_tone ipa_toneletter ipa_tonenumber middle_chinese_baxter middle_chinese_ipa old_chinese_ipa	morphological_template
language_stage		radical_support sensory_imagery interjection
	SEMANTICS	
	definitions	OTHER VARIABLES
		note
	ORTHOGRAPHY	
	traditional traditionalN simplified simplifiedN	FREQUENCY VARIABLES
	character_semantic_radicalN	characterN_freq characterN_family_size
		characterN_semantic_radical_freq characterN_semantic_radical_family_size
	character_phonetic_componentN	characterN_phonetic_component_freq characterN_phonetic_family_size
	orthographic_variants	

N = character 1-4 in template of 4 slots

$$\text{e.g. } traditionalN = \begin{Bmatrix} traditional1, \\ traditional2, \\ traditional3, \\ traditional4 \end{Bmatrix}$$

Figure 3.1: The variables of CHIDEOD (0.9)



3.2.1 Data variables

There are three mandatory variables when entering data into CHIDEOD: #data_source²¹, #language_stage, and the ideophone itself. The #language_stage variable broadly denotes different historical stages of Chinese such as Old or Middle Chinese. In this section we discuss the variables #data_source and #language_stage. At present, data in CHIDEOD comes from dictionaries as well as studies on onomatopoeia and ideophones. For dictionaries, we have consulted *Xiàngshēngcí lì shì* 象声词例释 ‘Examples and explanations of onomatopoeia’ (Wáng 1987); *Xiàngshēngcí cídiǎn* 象声词词典 ‘Dictionary of onomatopoeia’ (Gōng 1991); *A student’s dictionary of Classical and Medieval Chinese* (Kroll 2015); and the *Hànyǔ dà cídiǎn* 漢語大詞典 ‘Comprehensive Chinese Dictionary’ (Luó 1993). Wáng (1987) and Gōng (1991) are onomatopoeia dictionaries that record Mandarin. We have included all of their entry words in CHIDEOD. While Kroll as a dictionary has its problems, e.g., typographical errors in the Middle Chinese reconstructions, see O’Neill (2016) as well as the revised edition (Kroll 2017), it is a good resource for English translations of Classical and Medieval Chinese. We have included in CHIDEOD words that were reported as either binome (‘bn.’) or onomatopoeia (‘onom.’) in Kroll (2015) together with their English definitions. Binomes, or *liánmiáncí* 聯綿詞, are monomorphemic words created through full, partial or fission reduplication (Li 2013; Kroll 2015 xi-xiii). Binomes for names of insects and animals as well as tools and

²¹When discussed in the main text, variables of CHIDEOD are marked with <#>, and values with apostrophes. For example, for the variable #data_source a possible value is ‘Shijing’, which refers to the *Shījīng* 詩經.

loanwords were not included, e.g., *gé~jiè* 蛤蚧 ‘gecko’. To avoid Kroll’s erroneous Middle Chinese reconstructions, we took our Middle Chinese reconstructions from Baxter & Sagart (see Section 3.2.2.2) only, and not from Kroll.

In terms of studies, CHIDEOD currently includes Lǐ (2007), a comprehensive study on onomatopoeia in Mandarin, as well as the ideophones occurring in the *Táng shī sān bǎi shǒu* 唐詩三百首, see Van Hoey (2015), and *Shījīng* 詩經, discussed before by Smith (2015) and Van Hoey (2016a). The variable #language_stage broadly denotes different periods in the historical development of Sinitic as a whole. Currently, we have data on Standard Chinese or Mandarin ('SC'), as well as Middle Chinese ('MC') and Old Chinese ('OC'). Future versions of the open-source database are planned to include data on other Sinitic languages like Cantonese, Southern Min, Hakka etc. as well. Thus, items that are (to be) included have to be accompanied by a value for #data_source and #language_stage.

Table 3.2. shows how #data_source and #language_stage interact which each on the whole as variables. The values ‘Shijing’, ‘Tang’, ‘Gong’, ‘Li’, and ‘Wang’, do not span multiple #language_stages. ‘Kroll’ spans the language stages of Old Chinese and Middle Chinese. Lastly, the *Hànyǔ dà cídiǎn* is a comprehensive dictionary that encompasses data from different language stages (Old Chinese, Middle Chinese and Standard Chinese), with chronologically ordered definitions which reflect the semantic development (Xue 2003). The comprehensive nature of the *Hànyǔ dà cídiǎn* warrants a cautious treatment of this #data_source as an authority on any given

#language_stage, see Section 3.2.7.

Table 3.2: Coverage and absolute frequency of items in CHIDEOD

data_source	n	language stage		
		Old Chinese	Middle Chinese	Mandarin
HYDCD	6541	YES	YES	YES
Kroll	3279	YES	YES	-
Gong	1329	-	-	YES
Shijing	1205	YES	-	-
Li	1174	-	-	YES
Wang	939	-	-	YES
Tang	458	-	YES	-

3.2.2 Descriptive variables

There are three groups of descriptive variables: orthography, phonology, and semantics. Descriptive variables will be illustrated using the ideophones *guān~guān* 關關 ‘onom. cry of the osprey’ and *líng~líng* 鈴鈴 ‘onom. small clapper bells on carriages’.

3.2.2.1 Orthography Descriptive variables concerning orthography contain information for ideophone entries on the word level, the separate character level and the below-character level, visualized in Figure 3.2. In total, there are 18 orthographic variables in CHIDEOD.

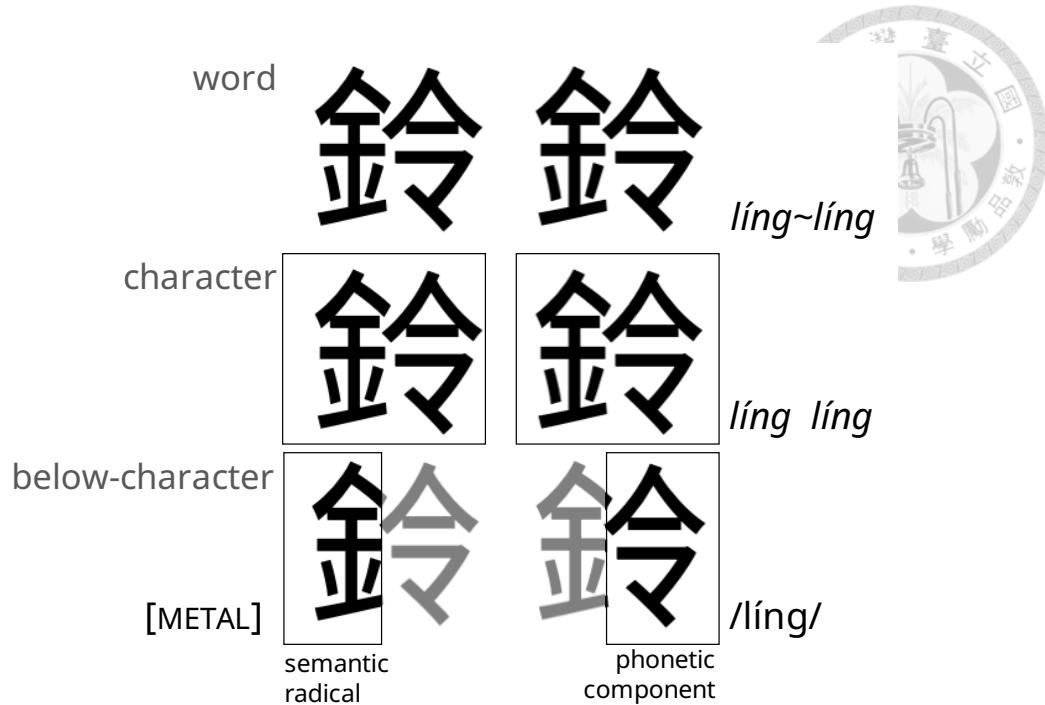


Figure 3.2: Chinese words analyzed according to word level, character level and below-character level. Illustrated with *líng~líng* 铃铛.

On the word level, CHIDEOD provides entries in #traditional and #simplified characters, e.g., ‘關關’ and ‘关关’ for *guān~guān*. On the character level, these entries can be spread out across a template of four character slots, as illustrated in Table 3.3. Since *líng~líng* 铃铛 is only disyllabic, the slots for the third and fourth characters are empty and marked with ‘NA’.

Table 3.3: Orthographic variables on the word and character level for *líng~líng* 铃铛.

variable	value	variable	value
traditional	鈴鈴	simplified	铃铛
traditional1	鈴	simplified1	铃
traditional2	鈴	simplified2	铃
traditional3	NA	simplified3	NA
traditional4	NA	simplified4	NA

Next is the below-character level. Most characters belong to the so-called ‘semantic-phonetic compound type’ (*xíngshēng* 形聲) (Norman 1988). According to data in the Chinese Lexical Database (Sun et al. 2018), 72% of items are *xíngshēng* characters. As shown in the below-character level in Figure 3.2, each character in the word *líng~líng* 鈴鈴 consists of left and right components. In this case, the left component ‘金/钅’ (‘metal’) is the semantic radical, which loosely indicates the semantic domain of a character, in this case METAL. The right component ‘令’ (pronunciation: *líng* /linj 1/) is the phonetic component which is homophonic to the *xíngshēng* character in and of itself. These are also generated across the template of four characters, as illustrated in Table 3.4.

Table 3.4: Semantic radical and phonetic component variables for *líng~líng* 鈴鈴.

variable	value	variable	value
character1_semantic_radical	钅	character1_phonetic_component	令
character2_semantic_radical	钅	character2_phonetic_component	令
character3_semantic_radical	NA	character3_phonetic_component	NA
character4_semantic_radical	NA	character4_phonetic_component	NA

The last relevant orthographic variable is #orthographic_variants, a well-known phenomenon for Chinese ideophones, see Xú (2000), Xú (2013), Li (2013), and Kroll (2017). For example, there are three written variants for *qiāng~qiāng* ‘onom. of rumbling and rattling of chariots; with assured ease; in orderly array’, e.g., 踏踏, 踣躕, and 鐙鎗.

3.2.2.2 Phonology There are eight descriptive variables pertaining to phonology in CHIDEOD. These reflect Mandarin phonology, as well as Middle and Old Chinese reconstructions when available. For Standard Chinese, we included the Pīnyīn transcription system, and provide variants with tone accents, tone numbers, and without tones. We also include transcriptions in the International Phonetic Alphabet (IPA) with tone letters and with tone numerals. For a historical overview of these transcription systems, as well as other alternatives, see Wilkinson (2015:58–61). The variables contained in CHIDEOD are illustrated with *guān~guān* 關關 ‘onom. cry of the osprey’ in Table 3.5. As a convention, syllables in CHIDEOD are split with a tilde (~), a convention in line with the Leipzig Glossing Rules (Bickel, Comrie & Haspelmath 2008).

Table 3.5: Examples of phonological variables for Standard Chinese ('SC') for *guān~guān* 關關 ‘onom. cry of the osprey’

variable	value
pinyin_tone	guān~guān
pinyin_tonenumber	guan1~guan1
pinyin_without_tone	guan~guan
ipa_toneletter	kwan˥~kwan˥
ipa_tonenumber	kwan55~kwan55

The main goal of these variables is further data-driven research into the ways ideophones *stretch* the phonological system, a phenomenon observed for e.g. Pastaza Quichua ideophones (Nuckolls et al. 2016). A number of

such phenomena include amongst other higher occurrence of palatalized voiceless stops, /kj/, /pj/, /tj/; expressively aspirated stops, e.g. /p^h/, /b^h/, /t^h/, /k^h/; and a high occurrence of the vowel /o/. In Chinese, we can see this happening to e.g. *xiū* 呻 (see example 2 in Chapter 1) which according to the rules of Standard Chinese phonology should be IPA /ɛjou˥/ but is actually closer to [ʃu^w˥], at least among my Taiwan Mandarin speaking peers.

Another phenomenon that can be tested through the variables mentioned in this section, is asymmetric tonal distribution. Previous research (e.g. Mok 2001; Thompson 2018) has shown that there does seem to be a certain skewedness towards high level tones for ideophonic words, especially the ones that depict SOUND (onomatopoeia). This is not entirely unsurprising, because they generally employ imagic iconicity to map between the phonological form of the ideophone and the sound it aims to imitate (Dingemanse 2012). This can include a multitude of sound symbolic (Hinton, Nichols & Ohala 1994; Dingemanse et al. 2016; Lockwood 2017), or in a recent proposal “sound iconic” (Hoshi et al. 2019:2), features, such as those that are well-known from the *maluma-takete* / *bouba-kiki* effect (Köhler 1929; Ramachandran & Hubbard 2001). LaPolla (1994) and Chan (1996) show early investigations of these phenomena for Chinese and while it falls outside of the scope of the current study, the studies that followed these can use data in CHIDEOD for further exploration and theorization of sound symbolism.

The next phonological descriptive variable pertains to Middle Chinese and Old Chinese reconstructions from Baxter and Sagart

(Baxter 1992; Sagart 1999; Baxter & Sagart 2014; 2015). These are `#middle_chinese_baxter`, `#middle_chinese_ipa` and `#old_chinese_ipa`, each is shown with *guān~guān* 關關 in Table 3.6. For Old Chinese, Baxter & Sagart use IPA. For Middle Chinese, however, Baxter first employed a notation that was easier to type in the early 1990s, but we also provide the IPA through a conversion using List’s (2018) Python library ‘sinopy’.

Table 3.6: Examples of historical phonological variables for *guān~guān* 關關 ‘onom. cry of the osprey’

variable	value
<code>middle_chinese_baxter</code>	kwaen~kwaen
<code>middle_chinese_ipa</code>	kwæn ¹ ~kwæn ¹
<code>old_chinese_ipa</code>	[k]fro[n]~[k]fro[n]

However, this does not entail that it is the only reconstruction system. For Middle Chinese, a number of frameworks exist that generally agree on the phonological system, represented by important scholars like Bernhard Karlgren (1889-1978), Tung T'ungho 董同龢²² (1911-1963), André-Georges Haudricourt (1911-1996), Li Fang-Kuei 李方桂 (1902-1987), Edwin G. Pulleyblank (1922-2013), Wáng Lì 王力 (1900-1986), William Baxter (1949), Laurent Sagart (1951), and Axel Schuessler (1940).

For Old Chinese, the situation is somewhat different. Other than acceptance in the field and ease of access (Baxter & Sagart 2015), the Baxter-Sagart system was chosen because it has won the Leonard Bloomfield Book Award

²²This particular stream is used in the courses of Chinese phonology at National Taiwan University.

in 2015, awarded by the Linguistic Society of America. Despite this achievement, it has received mixed reviews, as Jacques (2017) surveys: Goldstein (2015) and Ma Kun (2017) are the most positive; Starostin (2015) and Hill (2017) are quite sympathetic to the theory, but others like Schuessler (2015), Harbsmeier (2016), and especially Ho (2016) are hostile. Ho (2016) is particularly displeased with the six vowels that are posited by Baxter & Sagart (2014), see Hill (2012) for an overview of the development of this system. The criticism in Ho (2016), however, has been refuted by List et al.'s (2017) usage of network methodology in relation to the rhymes of the *Shījīng* 詩經. Baxter & Sagart (2017) then also responded to Schuessler (2015), arguing that Schuessler distinguishes between a Proto-Chinese and Old-Chinese on epistemological grounds rather than chronological ones. Furthermore, Baxter & Sagart claim that any reconstruction will remain hypothetical up to a certain degree, and that as long as it is data-driven rather than theory-driven, we will get closer to a robust reconstruction of Old Chinese. And this data-driven methodology seems to work, as for instance List, Hill & Foster (2019) has recently applied it to an analysis of rhymes in the *Shījīng* 詩經.

3.2.2.3 Semantics The only semantic descriptive variable in CHIDEOD, #definitions, is concerned with the meanings reported for each ideophone.

This variable is correlated to #data_source. As Table 3.7 illustrates, the item *guān~guān* 關關 has been entered via the data sources ‘Shijing’, ‘HYDCD’, and ‘Kroll’. The first of these is a corpus, so the value is ‘NA’ because no definitions were provided in the #data_source. The latter two are dictionaries, so they do have values.

Table 3.7: The semantic variable #definitions for *guān~guān* 關關 ‘onom. cry of the osprey’

data_source	definitions
Shijing	NA
HYDCD	1. 鳥類雌雄相和的鳴聲 [...examples...] 2. 和諧安適貌。 [...examples...] 3. 車行聲。 [...examples]
Kroll	onom. cry of the osprey

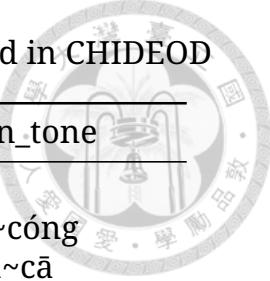
3.2.3 Analytical variables

The current version of CHIDEOD includes four analytical variables, each a result of analysis based on descriptive variables: #morphological_template, #radical_support, #interjection and #sensory_imagery.

3.2.3.1 Morphological template When discussing morphological templates in CHIDEOD, we take the theoretical standpoint of “reduplication as a process”, as it is typical of typological research (Hurch 2005), rather

than simply a linear string of Chinese characters, see Section 2.3.3. In treating reduplication as a process, we attempt to identify a Base and Reduplicant, where the Reduplicant is a reduplicative product of the Base. There are several instances of partial reduplication in CHIDEOD morphological templates. Partial reduplication (Sun 1999; Hurch 2005) encompasses the patterns RB, BR, and RR in Table 3.8. Although synchronic proof for a particular partial reduplicative process is not always available, there are certain diagnostics to categorize them as such. For example, *páng~huáng* 徨徨 ‘walk back and forth; nervously pacing’ is analyzed as the #morphological_template of Reduplicant-Base (‘RB’). This is because *páng* 後 on its own has no transparently identifiable or independent meaning. It is therefore a ‘meaningless bound form’ (see Sun 1999:120–122). The coding of *páng~huáng* 後徨 as Reduplicant-Base variable is supported by CHIDEOD, which contains *huáng~huáng* 徜徨 ‘jumpy, concerned’, but not *páng* 後 or *páng~páng* 後後. While there is no instance in CHIDEOD of *huáng* 徜 as a monosyllabic word, the fact that *huáng* 徜 can be fully reduplicated while *páng* 後 cannot implies that *páng* 後 is semantically dependent on *huáng* 徜 to be meaningful in *páng~huáng* 後徨, whereas *huáng* 徜 is not dependent on any syllables other than itself. With these factors in mind, *huáng* 徜 is justifiably the Base and *páng* 後 the Reduplicant. Another instance of Reduplicant in an RB or BR template is when one of the syllables is written with a character which, though has meaning on its own, seemingly has no relation (compositional or otherwise) to the overall disyllabic ideophone meaning. For example, *hàn* 汗 ‘sweat’ in the ideophone *hàn~màn* 汗漫

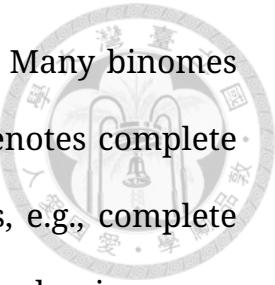
Table 3.8: Morphological templates of ideophones included in CHIDEOD



morphological_template	type_freq	examples	pinyin_tone
A	283	咚	dōng
BB	1460	濛濛	cóng~cóng
BBB	66	嚙嚙嚙	cā~cā~cā
BBBB	14	咕咕咕咕	gū~gū~gū~gū
BR	353	嘈啐	cáo~cuì
RB	109	汗漫	hàn~màn
RR	1557	演漾	yǎn~yàng
RRR	25	吉丁當	jí~dīng~dāng
RRRR	582	滴滴答答	dī~dī~dá~dá
ARR	210	忒楞楞	tè~léng~léng
ARR	210	忒楞楞	tēi~léng~léng
RRA	21	當當丁	dāng~dāng~dīng
RAN	288	茫然	máng~rán
YAN	5	苔焉	tā~yān
RU	11	恬如	tián~rú
ER	3	赫爾	hè~ěr

borderless and boundless is a Reduplicant while *màn* 漫 ‘overflowing, brimming over’ is the Base. Moreover, *màn* 漫 is found in other words (e.g., *màn~cháng* 漫長 ‘long, endless’) whose meanings resemble ‘borderless and boundless’ whereas *hàn* 汗 is not. Therefore *màn* 漫 is the Base and *hàn* 汗 is the Reduplicant in *hàn~màn* 汗漫 borderless and boundless.

Table 3.8 shows the type frequency for all morphological templates in CHIDEOD. They are coded as follows: A is a single syllable; BB stands for Base-Base (full reduplication, i.e., *huáng~huáng* 狩狩); BR and RB for Base-Reduplicant and Reduplicant-Base respectively. Type frequency of RR, Reduplicant-Reduplicant ($n = 1557$), is higher than BB Base-Base ($n = 1460$). This is because for many items we were not able to allocate Base status to a syllable, e.g., *yǎn~yàng* 演漾 ‘rolling (waves)’ whose individual syllables mean ‘perform, act’ and ‘pool; ripples’ respectively, yet their



onsets and nuclei suggest (partial) reduplication at work. Many binomes from Kroll (2015) fall into the Base-Base category. ARR denotes complete ideophones rather than collocate-ideophone constructions, e.g., complete ideophone *gū~lū~lū* 咕 噜 噜 where all syllables are ideophonic versus collocate-ideophone *liàng-jīng~jīng* 亮 晶 晶 ‘bright-sparkling’ where only *jīng~jīng* 晶 晶 is the ideophonic component and *liàng* 亮 is the adjective ‘bright’, otherwise known as the ABB construction (Wang 2014; Huang, Jin & Shi 2016; Hsieh 2017). CHIDEOD stores the ideophonic component of collocate-ideophone constructions only, i.e., *jīng~jīng* 晶 晶 (as type BB), while the co-occurrence of ideophonic component with collocates is usage-related and warrants separate investigation. In Section 7.4 this will be investigated. Finally, the morphological templates of RAN, YAN, RU, and ER are ideophones that are suffixed; Van Hoey (2015) referred to these forms as compositional ideophones. A seminal study by (Künstler 1967) describes their function in Old Chinese, but their exact relation to ideophones has yet to be studied. Lastly, we want to stress that the analysis of morphological template is not fixed: future versions can improve the current distribution. The database in general, i.e., ignoring/negating the variable #language_stage shows that disyllabic items (BB, RB, BR, RR) have the highest type frequency. However, the existence of other types strongly suggests that reduplication is not a necessary nor sufficient feature for ideophonicity in Chinese, as has been shown for other languages as well (Dingemanse 2015). The issue is revisited in Section 3.2.7.

Table 3.9: Examples of radical support (Van Hoey 2018a:250)



variable name	淋淋	淋漓
pinyin_tone	lín~lín	lín~lí
character1_semantic_radical	氵	氵
character2_semantic_radical	氵	氵
radical_support	氵	氵

3.2.3.2 Radical support and radical repetition Ideophones are usually analyzed on the morphological or phonological level in terms of markedness (Childs 1988; Akita 2009; Haiman 2011; Dingemanse 2012; Nuckolls 2016b; Thompson 2018; Arthur Lewis Thompson 2019b). However, since it is possible to break down Chinese characters into semantic radical and phonetic components, as illustrated in Figure 3.2 and Table 3.4, there may be markedness from an orthographic point-of-view as well. For meteorological expressions, Van Hoey (2018a) used the term ‘radical support’ whenever the same semantic radical was repeated across different characters within an ideophone, e.g., *lín~lín* 淋淋 ‘streaming, soaking wet’ and *lín~lí* 淋漓 dripping wet both have the semantic radical <氵> meaning WATER. Consequently, their #radical_support in CHIDEOD is ‘氵’. This is visualized in Table 3.9.

The ‘support’ part of the term ‘radical support’ stems from studies on linguistic motivation (Radden & Panther 2004) and systematicity (Dingemanse et al. 2015). The main idea is that when a semantic radical like <氵> ‘water’ is used to effectively express WATER, we can speak of motivation. A well-known example where this is not the case is *shā* 沙 ‘sand’, in which <氵> is incongruent for the meaning. The current version of CHIDEOD, however, treats RADICAL SUPPORT more as orthographic repetition of semantic

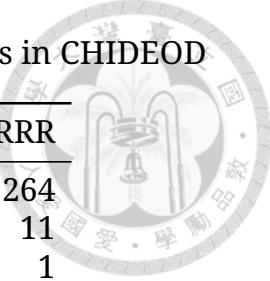


radicals. Therefore, it has been suggested that a term RADICAL REPETITION better reflects the current state of analysis. However, we will continue to keep using ‘radical support’, but bear this nuance in mind. Radical support is a systematic tendency rather than a necessary or sufficient condition for markedness. However, literary ideophones tend to exhibit more various radicals in their radical support than colloquial ideophones, while colloquial ideophones are more likely to display radical support with ‘mouth’ than literary ideophones. CHIDEOD allows us to see the type frequency of radicals for partially reduplicated²³ items (#morphological_pattern is A, BR, RB, RR, RRR, RRRR). Table 3.10 shows that □ ‘mouth’ is most frequent. This is not surprising given that □ ‘mouth’ is a common indicator for onomatopoeic characters (Lǐ 2007:109). Examples of this radical include *dōng* 咚 ‘boom, thud’ and *gū~lū~gū~lū* 咕噜咕噜 ‘glug, glug, drinking’. However, other radicals appear to cover a range of sensory meanings.

The majority of the high frequent semantic radicals seem to occur in ideophones belonging to the visual sensory domain, e.g., *hàn~màn* 汗漫 ‘borderless and boundless’ has 氵 WATER, or *cuó~é* 嶙峨 ‘jutting and jagged, gnaw-toothed, looming’ has 山 MOUNTAIN. A list of the semantic radicals displayed in Table 3.10 includes WATER (氵), GRASS (艹), MOUNTAIN (山), FOOT (足), PERSON (亼), JADE (玉), HEART1 (忄), HAND (扌), SILK (糸), WOMAN (女), HEART2 (心), STONE (石), WALKING (辵), METAL, GOLD (金), CART (車), TREE (木), FIRE (火), and DOOR (門). It may be of interest to explore whether these semantic radicals differ significantly from the prosaic lexicon. This could be accomplished in

²³It is only logical that for a fully reduplicated ideophone (morphological template ‘BB’) like *lín lín* 淋淋 radical support occurs; the two characters that make it up are the same, after all!

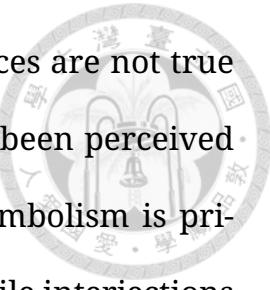
Table 3.10: The top radicals in partially reduplicated items in CHIDEOD



radical_support	A	BR	RB	RR	RRR	RRRR
口	128	8	2	213	6	264
氵	NA	48	7	71	NA	11
艹	NA	11	2	27	NA	1
山	NA	21	4	22	NA	NA
足	1	3	2	22	NA	3
亼	NA	2	2	19	NA	NA
王	NA	3	1	15	NA	5
宀	2	8	6	14	NA	1
扌	NA	5	NA	13	NA	5
糸	1	4	2	12	NA	NA
女	NA	5	3	10	NA	NA
心	NA	1	NA	10	NA	NA
石	4	4	4	9	NA	NA
辵	NA	4	2	9	NA	NA
金	2	NA	NA	9	NA	6
車	NA	2	1	7	NA	NA
木	NA	2	1	5	NA	1
火	NA	3	1	5	NA	NA
門	1	NA	NA	5	NA	NA

a future study by comparing them to the Chinese Lexical Database (Sun et al. 2018). What is sure, however, is that ideophones not only are written with the MOUTH radical, but that in fact a whole range of senses (Dingemanse 2012; Akita & Dingemanse 2019) is included. The importance of radical support, then, is an extra cross-modal motivation for a group of ideophones. In the Chinese Ideophone Database, about 55% of the unique items display a form of radical support. This percentage will be used in Chapter 4 when the variable of #radical_support is recoded as a binary value.

3.2.3.3 Interjection In their authoritative volume *Sound Symbolism*, Hinton, Nichols & Ohala (1994) point out that certain instances of sound symbolism pertaining to emotional responses (e.g., oh!, ouch!, ew!, ah!)



should be classified as interjections, seeing as these instances are not true symbols, but rather signs of a stimulus or referent having been perceived by a speaker (Hinton, Nichols & Ohala 1994:2). Sound symbolism is primarily iconic due to its imitative depiction of a referent, while interjections are primarily indexical because they signal an internal state or response to a stimulus rather than a depiction of that stimulus itself, see also Clark (1997) and Dingemanse (2017). Similar observations for Chinese have since been made by Xíng (2004) and Yáng (2006), but the difference between interjections and onomatopoeia was discussed far earlier (see Zhào 2008 for overview). For instance, Zhū & Lǚ (1951) categorize what they call “mimetic words” into three groups: interjections, response words, and onomatopoeia. As the aforementioned studies imply, there are important semantic distinctions to be aware of when differentiating ideophones and interjections, i.e., the symbol versus the sign or the depictive versus the indexical. On the other hand, the difference is not as clear-cut as one would hope. For instance, *wū~hū* 嘿呼 ‘sigh of sadness, alas’ is indeed indexical of a sad or mournful reaction. But it can also be used to depict the act of crying, much like the English *boohoo*. For these reasons, we have opted for keeping the interjections in the database, but marking them with a separate variable, *interjection*, which has the values ‘interjection’ and ‘notinterjection’. The value ‘interjection’ was assigned if the definition was described as *tàncí* 嘆詞 ‘interjection’ in the *Hànyǔ dà cídiǎn* (#definitions of ‘HYDCD’) or ‘interjection’ in Kroll (2015) (#definitions of ‘Kroll’). While this is not a foolproof interjection identification process, it serves as a point of

departure for future researchers interested in Chinese interjection data.



3.2.3.4 Sensory imagery The last analytical variable is #sensory_imagery. This variable is concerned with the sensory domain expressed by a given ideophone. The cross-linguistic implicational hierarchy proposed by Dingemanse (2012:663) includes the domains SOUND < MOVEMENT < VISUAL PATTERNS < OTHER SENSORY PERCEPTIONS < INNER FEELINGS AND COGNITIVE STATES. These are ordered such that if a language contains haptic ideophones (belonging to OTHER SENSORY PERCEPTIONS) then that language should also contain ideophones belonging to the sensory domains of SOUND, MOVEMENT and VISUAL PATTERNS. This hierarchy has had considerable impact on empirical questions in subsequent ideophone research, but is not the only alternative (Lu 2006; Akita 2009; Akita, Zhang & Tamaoka 2020; Van Hoey 2015; Van Hoey 2016b; McLean 2019; Arthur Lewis Thompson 2019b; Thompson, Akita & Do). CHIDEOD follows the basic classification proposed by Van Hoey (2015), see Table 3.11. Items defined as ‘sound of’, ‘onom.’, *xiàngshēngcí* 象聲詞, or *nǐshēngcí* 擬聲詞 were marked as belonging to the sensory domain of SOUND. Other domains were coded according to the definitions and/or the examples listed therein, e.g., ‘drifting’ was interpreted as belonging to the sensory domain of MOVEMENT. VISUAL, on the other hand, has to do with luminescence or color. The difference between EVALUATION and INNER_FEELING is mainly one of perceived directionality of judgment by the conceptualizer (Langacker 2008b), see also Talmy’s discussion of ception (2000a). That is, evaluations are value judgments of external referents, while inner feelings mainly about cognitive states or

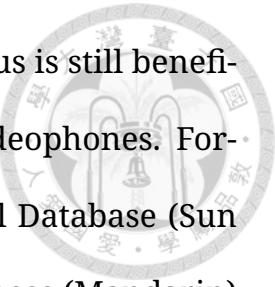
emotions, although it must be stressed that the current classification is still preliminary and needs to be supplemented by surveys or experiments, as discussed in Section 3.2.7. Finally, TIME denotes time duration rather than point in time.

Table 3.11: The sensory imagery values allocated to items in CHIDEOD

#sensory_imagery	example (#traditional)	#pinyin_tone
SOUND	嚶嚶	yīng~yīng
MOVEMENT	塵塵	méi~méi
VISUAL	熒熒	càn~làn
TEXTURE	萎萎	wěi~něi
SMELL	苾苾	bì~bì
TASTE	譚譚	tán~tán
TEMPERATURE	凜凜	lǐn~lǐn
EVALUATION	錄錄	lù~lù
INNER_FEELINGS	怡怡	yí~yí
TIME	慆慆	tāo~tāo

3.2.4 Frequency variables

As Sun et al. (2018) relate, for their Chinese Lexical Database, the goal was to aid in psycholinguistic experiments as well as computational studies. The goal of CHIDEOD is to provide a storage space that can aid in the analysis of ideophones. However, it is hoped that the inclusion of these measures in CHIDEOD can also help with future studies that wish to go the experimental



route, or want to compare these data to new findings. It thus is still beneficial to provide frequency measures for the characters in ideophones. Fortunately, these are readily available in the Chinese Lexical Database (Sun et al. 2018), based on simplified characters in Standard Chinese (Mandarin) from Mainland China. Frequency variables allow for further comparison with frequency information associated with words in the prosaic lexicon. CHIDEOD frequency variables form a set which contains an identifying variable, a frequency variable, a family size variable for the character-level, semantic radical, and phonetic component, against the four characters template shown in Table 3.3 and Table 3.4. We illustrate the frequency variables with *bā* 叻, as it occurs in *bā~cā* 叻𡇗 ‘sound of dropping something (like a mug)’. On the character level, we have the #traditional and #simplified orthographic representations, spread out over four-character slots, shown above in Table 3.3. For #simplified1 (“the first value in the character template of the simplified orthographic representation”) the value is ‘𠮩’ and for #simplified2 it is ‘𡇗’; the slots for the third and fourth characters are empty, i.e., they have the value ‘NA’. In terms of frequency, we find in the Chinese Lexical Database that 叻 occurs 7.4321 times per million characters: for the variable #character1_freq, the value here is ‘7.4321’. In the same database, 叻 is used in 6 prosaic words, resulting in the value ‘6’ for our variable #character1_family_size in CHIDEOD. Looking at the semantic radical of *bā* 叻, we find #character1_semantic_radical = ‘匚’. This semantic radical has a frequency (#character1_semantic_radical_freq) of ‘50016.56’ per million characters. Its #character1_semantic_family_size is ‘297’. This

number is quite high for prosaic words, but will undoubtedly be even higher if data from CHIDEOD would be combined with that of the Chinese Lexical Database, as tentatively explored in Table 3.10. Finally, for the phonetic component of our example *bā* 叻, we find that the value for the variable #character1_phonetic_component is ‘八’ (also pronounced *bā*). Its character frequency (#character1_phonetic_component_freq) is ‘43.3432’ per million characters its family size (#character1_phonetic_family_size) is ‘3’. A systematic overview of the frequency variables is given in Table 3.12, Table 3.13, and Table 3.14.



Table 3.12: Frequency variables concerning characters

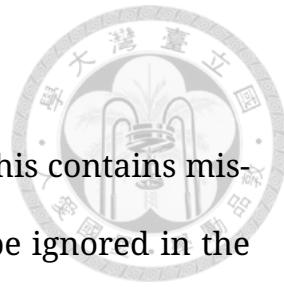
main variable	frequency variable	family size
#simplified1	#character1_freq	#character1_family_size
#simplified2	#character2_freq	#character2_family_size
#simplified3	#character3_freq	#character3_family_size
#simplified4	#character4_freq	#character4_family_size

Table 3.13: Frequency variables concerning the semantic radical

main variable	frequency variable	family size
#character1_semantic_radical	#character1_semantic_radical_freq	#character1_semantic_family_size
#character2_semantic_radical	#character2_semantic_radical_freq	#character2_semantic_family_size
#character3_semantic_radical	#character3_semantic_radical_freq	#character3_semantic_family_size
#character4_semantic_radical	#character4_semantic_radical_freq	#character4_semantic_family_size

Table 3.14: Frequency variables concerning the phonetic component

main variable	frequency variable	family size
#character1_phonetic_component	#character1_phonetic_component_freq	#character1_phonetic_family_size
#character2_phonetic_component	#character2_phonetic_component_freq	#character2_phonetic_family_size
#character3_phonetic_component	#character3_phonetic_component_freq	#character3_phonetic_family_size
#character4_phonetic_component	#character4_phonetic_component_freq	#character4_phonetic_family_size



3.2.5 Other variables

Currently there is only one other variable, called #note. This contains miscellaneous observations made during data entry. It will be ignored in the remainder of this dissertation.

3.2.6 Tutorial

In this section, we first guide the reader to the different formats in which CHIDEOD is available, and further illustrate the online app version. Next, we illustrate how CHIDEOD can be used by itself as well as in combination with a corpus to examine, e.g., tonal distributions in Standard Chinese.

3.2.6.1 Using the online app version of CHIDEOD We provide the whole CHIDEOD dataset in three formats in the OSF repository²⁴, namely excel (.xlsx), comma separated values (.csv) and R data serialized (.rds). The .rds version is shipped with the data package CHIDEOD to be used in R code²⁵. On the whole we highly recommend researchers to work with one of these full database formats in their academic analyses, using a software they are familiar with. For quick look-ups we have also deployed an online app.

Figure 3.3 displays the layout of the app. The pane on the left (marked with A in Figure 3.3) allows one to select the variables one is interested in. In the search bar (B), we have searched “líng~líng” with all the data variables selected by default. If one would prefer to search without tones, we sug-

²⁴ Available at <https://osf.io/kpwgf/>.

²⁵ Available at https://simazhi.shinyapps.io/chideod_appversion/.

gest selecting the #pinyin_without_tone or #pinyin_tonenum under the descriptive variables heading. After entering a search expression, the app displays all items for which that expression is found (C). The data can then be exported (D) by copying to the clipboard (“Copy”), to comma separated values (“CSV”) or to excel format (“Excel”). The tidy data format may lead to repeated values. If exported, users can choose to delete repeated values, in order to trim the data as users see fit.

The screenshot shows the CHIDEOD app interface. At the top, there are three tabs: CHIDEOD (selected), Table, and About. Below the tabs is a sidebar with three sections: Data variables, Descriptive variables, and Orthography. The Data variables section contains two checked checkboxes: Language stage and Data source. The Descriptive variables section contains one checked checkbox: Traditional Chinese. The Orthography section contains five unchecked checkboxes: Simplified Chinese, Traditional character 1, Traditional character 2, Traditional character 3, and Traditional character 4. To the right of the sidebar is a search bar with the placeholder "Search: ling-ling". Above the search bar are three buttons: Copy, CSV, and Excel. A red box labeled "D" highlights these buttons. Below the search bar is a table with columns: language_stage, data_source, traditional, and pinyin_tone. The table contains 10 rows of data, each with a unique ID (e.g., 68, 77, 310, 938, 1028, 1059, 1229, 1230, 1231, 1232), a source (e.g., SC, Wang, Li), a traditional name (e.g., 琅琅玲玲, 滴滴鈴鈴, 伶伶朗朗, 滴鈴鈴鈴, 鈴鐺鈴鈴, 玲玲瓏瓏, 叮聆玲, 叮鈴鈴, 叮鈴鈴, 叮鈴鈴), and a pinyin tone (e.g., cóng-cóng-líng-líng, dī-dī-líng-líng, líng-líng-lāng-lāng, dī-líng-líng-líng, líng-dāng-líng-líng, líng-líng-lóng-lóng, dīng-líng-líng, dīng-líng-líng, dīng-líng-líng). A red box labeled "C" highlights the table area. A red box labeled "B" highlights the search bar. At the bottom of the table, it says "Showing 1 to 10 of 45 entries (filtered from 14,936 total entries)". Below the table are navigation buttons: Previous, 1, 2, 3, 4, 5, Next.

language_stage	data_source	traditional	pinyin_tone
68	SC	Wang	cóng-cóng-líng-líng
77	SC	Wang	dī-dī-líng-líng
310	SC	Wang	líng-líng-lāng-lāng
938	SC	Li	dī-líng-líng-líng
1028	SC	Li	líng-dāng-líng-líng
1059	SC	Li	líng-líng-lóng-lóng
1229	SC	Wang	dīng-líng-líng
1230	SC	Wang	dīng-líng-líng
1231	SC	Gong	dīng-líng-líng
1232	SC	Li	dīng-líng-líng

Figure 3.3: App version of CHIDEOD

3.2.6.2 Tonal distribution of mono- and disyllabic onomatopoeia in Mandarin

In order to provide a practical demonstration of CHIDEOD, we explore the distribution of Mandarin tones across ideophones. Previous research has found that Mandarin onomatopoeia are mostly in the high level tone (IPA tone numeral: 55, tone letter: ˥). Thompson (2018) reports the following distribution for Mandarin tones: 57% for high level, 23% for rising, 5% for dipping or low, 14% for high falling. This distribution is remarkably different from that of the non-ideophonic lexicon: 8% for high level, 15% for rising, 32% for dipping or low, and 36% for high falling. The statistically

significant difference in distributions reported by Thompson (2018) highlights the role that high level tone plays in the markedness of Chinese ideo-phones. Contra to Thompson's (2018) findings, Mok (2001:67) uses wordlists and fieldwork to conclude that 97.6% of syllables in monosyllabic and disyllabic Mandarin onomatopoeia are high level tone. Meng (2012) agrees with Mok (2001) and goes on to say that monosyllabic onomatopoeia “almost always employ high level tone”, while disyllabic onomatopoeia “generally have a tonal melody [high level] – [high level]” (Meng 2012:26). Meng (2012) uses an array of resources, including dictionaries, corpus material, introspection, and consultation with native speakers. It is clear that the high level tone (tone 1) plays a crucial role in Mandarin onomatopoeia. However, there is still a large discrepancy between Thompson's reported 57% and Mok's 97.6%. With CHIDEOD, we can obtain a better picture of this distribution. We use R (R Core Team 2019), the tidyverse package (Wickham 2017) and the CHIDEOD dataset, which appears as ‘chideod’ in the code. First, we must subset CHIDEOD so that we have all the monosyllabic and disyllabic onomatopoeia from Mandarin, i.e., #sensory_imagery = ‘SOUND’ and #language_stage = Standard Chinese ‘SC’.

The following code chunk shows how to subset in two steps. Comments regarding each line of code follow the ‘#.’ The ‘%>%’ operator means “then do X”.

```
# Step 1  
  
chideod.modern.onomatopoeia <- # assign R object  
  chideod %>% # call CHIDEOD
```

```

filter(language_stage == "SC") %>% # standard chinese
filter(sensory_imagery == "SOUND") %>% # only onomatopoeia
select(traditional, # ideophone item in
       # traditional characters
       pinyin_tonenumbers # pinyin with tone number
) %>%
distinct() # unique because tidy data

# Step 2

chideod.subset <- chideod.modern.onomatopoeia %>% # take new R object
mutate(syllablenumbers = str_count(pinyin_tonenumbers,
                                     "\\\d")) %>% # count number of
# syllables based on
# digits in
# pinyin_tonenumbers
mutate(tonepattern = str_remove_all(pinyin_tonenumbers,
                                     "[[:alpha:] | ~")) %>%
filter(syllablenumbers == 1 |
       syllablenumbers == 2) # monosyllabic and disyllabic

```

After creating the ‘chideod.subset’, we need the numbers for the tonal distribution of mono- and disyllabic onomatopoeia. We count the combinations of the syllable types and the tonal patterns using the variables #syllablenumbers and #tonepattern. The code for this operation is shown

in the following code chunk.



```
chideod.subset %>%  
  
  count(syllablenumbers,    # count combinations of  
         tonepattern,        # mono/disyllabic and tonal pattern  
         name = "absolute",  
         sort = TRUE) %>%  
  
  group_by(syllablenumbers) %>%  
  
  mutate(relative.perc = # calculate relative percentage  
         absolute / sum(absolute) * 100) %>% # per syllable type  
  
  arrange(syllablenumbers,  
  
         desc(absolute)) # present in order
```

The results for the distribution of Mandarin tones across monosyllabic and disyllabic onomatopoeia are presented in respectively Table 3.15 and Table 3.16. At first glance, our simple percentages support those statistically verified in Thompson (2018), rather than Mok (2001) especially with respect to monosyllables. This could be to do with how the data was collected, i.e., using wordlists rather than usage-based data.

Table 3.15: Tonal distribution for monosyllabic onomatopoeia in Mandarin ('SC') based on CHIDEOD

High level (1)	Rising (2)	Dipping (3)	High falling (4)	Neutral (5)
137	28	8	21	2
69.9%	14.3%	4.1%	10.7%	1%

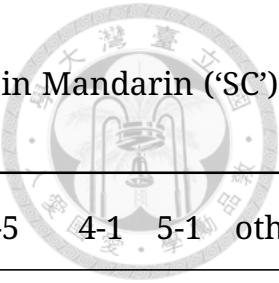


Table 3.16: Tonal patterns for disyllabic onomatopoeia in Mandarin ('SC') based on CHIDEOD

1-1	2-2	4-4	1-2	1-4	3-3	1-3	2-1	1-5	4-1	5-1	other
412	78	52	44	32	21	19	16	12	7	7	25
56.8%	10.7%	7.2%	6.1%	4.4%	2.9%	2.6%	2.2%	1.7%	1%	1%	3.4%

We can check whether the results in Table 3.15 and Table 3.16 differ based on a data source discrepancy between Thompson (2018) and Mok (2001) by including corpus data as well. To do so, we have collected all occurrences of the ideophones ($n = 1204$) in ‘chideod.subset’ that appear in ‘spoken’ or ‘written-to-be-spoken’ texts of the Academia Sinica Balanced Corpus (ASBC) 4.0 (CKIP group & Academia Sinica 2013). In doing so, we show that it is also possible to combine data from external corpora with CHIDEOD. First, the prepared dataset is fed into R, and then the previously chideod.subset is joined with it, see the next code chunk. After this, much of the same code from the previous code block is used to calculate the distributions of tonal patterns and onomatopoeia.

```
onom.in.asbc <- read_rds(here("data", "onomatopoeia_in_asbc.rds"))

# change file path accordingly

onom.in.asbc %>%
  # take onom.in.asbc object

  left_join(chideod.subset,          # join chideod.subset object
             by = c("onomatopoeia" = "traditional")) %>%
  count(syllablenumbers,           # count combinations of syllable
```

```

tonepattern,           # types and tonal patterns

name = "absolute",

sort = TRUE) %>%

group_by(syllablenumbers) %>%

mutate(relative.perc =           # calculate relative percentage

absolute / sum(absolute) * 100) %>% # per syllable type

arrange(syllablenumbers,
desc(absolute))           # present in order

```

Table 3.17: Tonal distribution for monosyllabic onomatopoeia in Mandarin from CHIDEOD and ASBC 4.0

High level (1)	Rising (2)	Dipping (3)	High falling (4)	Neutral (5)
38	9	4	3	2
67.9%	16.1%	7.1%	5.4%	3.6%

Table 3.18: Tonal distribution for disyllabic onomatopoeia in Mandarin from CHIDEOD and ASBC 4.0

1-1	1-4	1-5	1-2	3-3	4-4	5-1	5-5
20	2	2	1	1	1	1	1
68.9%	6.9%	6.9%	3.4%	3.4%	3.4%	3.4%	3.4%

Based on a sample of onomatopoeia from the spoken parts of the ASBC 4.0 corpus (see Section 3.3.3) ($n = 1204$), the results for monosyllabic and disyllabic onomatopoeia indicate that there is a serious skewedness towards high level tone at 68.9%. This result supports Thompson's (2018) statisti-

cally verified 57% over Mok's (2001) overly optimistic 97.6%. This may be due to how data were collected in Mok (2001), which was unfortunately not reported in a transparent manner. It may also be due to a prototypicality effect: if high level tone is implicitly understood as a marker for ideophones, then native speakers may exhibit a bias toward that tonal category during data elicitation. Prototypicality of the high level tone, however, deserves further exploration.

3.2.7 Future applications of CHIDEOD

CHIDEOD as a data repository which exists to serve and supplement future studies which analyze usage-based data. Usage-based data is obtained through four techniques: introspection, surveys, experiments, and corpus data (Tummers, Heylen & Geeraerts 2005). For introspection, such as Paul's (2006) discussion of Chinese adjective classes, our database provides a base indication of how frequent an ideophone has been recorded in previous work (although introspective studies notoriously lack frequency information). For survey-like tasks, such as Dingemanse and Majid's (2012) sorting-task showed that Siwu ideophones clustered and stratified according to different levels of granularity. In other settings, participants coded events from visual stimuli such as picture books (Slobin 2004) or video clips (Iwasaki 2017). In such cases CHIDEOD can act as a reliable centralized data repository that can either supply data for survey tasks or confer task outcomes with a pool of similar data points. As for experiments, a line of size sound symbolism research (Köhler 1929; LaPolla 1994; Ohala 1994;

Chan 1996; Ramachandran & Hubbard 2001; Lockwood & Dingemanse 2015a; 2015b) has shown that higher and more fronted vowels correlate with smaller and sharper concepts, while lower and more back vowels correlate with bigger and rounder shapes. It is not inconceivable that Chinese findings (e.g., LaPolla 1994; Chan 1996) can be expanded and supplemented with the centralization of iconic words in CHIDEOD. Our last point is in relation to corpus methods. Seeing as CHIDEOD contains information like type frequency, which correlates with schema productivity (Bybee & Hopper 2001a), such phenomena can already be studied with the database as it is. Moreover, as we briefly exemplified in Section 3.2.6.2, CHIDEOD can also be used in corpus-based explorations of token frequency, and distributional skewedness in language structure.

Aside from the four approaches detailed above, CHIDEOD holds much potential for lexical semantic research. For instance, the orthographic formal onomasiology of ideophones can be investigated through the variable #orthographic_variants. Previous research on three formal variants depicting ‘boundlessness’, namely *máng~máng* 茫茫, *máng~máng* 茫茫, and *cāng~máng* 蒼茫 (Van Hoey & Lu 2019a) shows how the different forms lead to different conceptualizations, and traces how their token frequencies evolved over time and across China. The usage of CHIDEOD can help scale such approaches.

One should, however, familiarize themselves with the data sources in CHIDEOD. For example, the *Hànyǔ dà cídiǎn* is a comprehensive dictionary whose entries are often vague in relation to #language_stage (Section 3.2.1).

Fortunately, each entry word is accompanied by an authoritative usage and context per definition. The definition for *páng~háng* 徨徨 ‘walk back and forth; nervously pacing’ in the *Hànyǔ dà cídiǎn* quotations range from the Zhōu dynasty up to the modern-day Standard Chinese. For lexicological studies, it is thus recommended that researchers pay special attention to the examples listed in *Hànyǔ dà cídiǎn* #definitions, see Allan (2012) for a discussion on the use of dictionary data.

Further applications of CHIDEOD data concern the co-occurrence of different variables, in order to describe the language-particular prototype of the Chinese ideophonic lexicon in a given language stage. After all, Japanese mimetics follow a prototypical structure (Akita 2009), as do ideophones in Siwu [Dingemanse et al. (2015)], and ‘ideophone’ as a cross-linguistic concept has been argued to be best understood as a prototype as well (Childs 1994; Dingemanse 2019). Thus, for ideophones in Modern Chinese, one needs to select the right language stage ('SC' for Standard Chinese) and then compute the co-occurrence of multiple variables, e.g., #morphological_template, #radical_support, and #sensory_imagery. A possible statistical approach is correspondence analysis (Glynn 2014), as we will use in Chapter 4, to identify a cluster of colloquial onomatopoeia (sound ideophones) and literary non-sound ideophones.

The multisensoriality, or synesthetic co-occurrences, of ideophonic items in Chinese has yet to be investigated. Nuckolls & Swanson (2019) identified sensory clusters for ideophones in Pastaza Quichua through linguistic surveys. Another recent method makes use of ratings for iconicity

as well as sensory imagery, allowing the researcher to identify semantic clusters on different levels of granularity, see Winter (2019). A first adaptation of this technique to Chinese (Chen et al. 2019) provides an avenue of future research for Chinese ideophones as well.

Ratings have also been used to investigate how iconic words are perceived by native speakers of English, Spanish, as well as Japanese, see Perlman et al. (2018), Perry, Perlman & Lupyan (2015), and Thompson, Akita & Do (in press). Ratings could prove useful for Chinese words whose iconic status is diachronically attested yet synchronically ambiguous. For example, partially reduplicated disyllabic words, like *hàn~màn* 汗漫 ‘borderless and boundless’, are treated as literary ideophones in this paper though their status as ideophones synchronically has not yet been empirically addressed. Though we established in Section 3.2.3.1 that *màn* 漫 is the Base in the morphological template of *hàn~màn* 汗漫, it remains to be seen whether *màn* 漫 would be considered iconic in non-reduplicative compound words such as *màn-cháng* 漫長 ‘long, endless’ or *màn-bù* 漫步 ‘stroll; roam’, especially because it has been claimed that /m-/ is a phonestheme denoting IMPAIRED VISION (Chan 1996). Ratings could be of use here, though see Thompson, Akita & Do (in press) for a discussion of methodological confounds. If ratings showed Bases like *màn* 漫 to be non-iconic outside their Base-Reduplicant context, then this may have to do with more frequent or more productive forms losing their ideophonic status, as proposed by Dingemanse (2017).

Lastly, CHIDEOD can be used for cross-linguistic comparisons with ideophone-rich Trans-Himalayan languages, such as the Rgyalrong (Jacques

2013) and Kiranti (Lahaussois 2017) families, to increase our understanding of the relation between Sino-Tibetan languages. In that case, one would subset the data to include only Old Chinese data (#language_stage = ‘OC’). Next, one would focus on the #old_chinese_ipa variable and forge comparisons with ideophones known from the field of Trans-Himalayan studies.

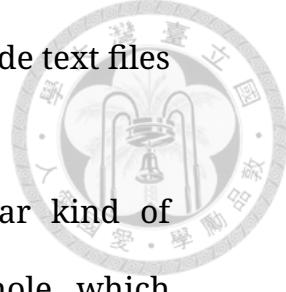
3.2.8 CHIDEOD in this dissertation

The current version of the Chinese Ideophone Database (CHIDEOD) collects 4948 unique onomatopoeia and ideophones (mimetics) of Mandarin, as well as Middle and Old Chinese, with future plans to include other Sinitic languages (Cantonese, Southern Min, etc.). It is a data repository that aims to help future research directions, be they introspective, surveys, experiments, or external corpus data. Although it is not feasible to gather ideophones as an exhaustive set, since they form an open lexical class (Dingemanse 2019), what can be done with CHIDEOD is begin devising a thorough and detailed picture of ideophones from Chinese languages based on various parameters. CHIDEOD can be used to further delineate “the ideophone category” according to historical stages of Chinese, by focusing on a combination of formal and functional approaches, as has been done for Japanese ideophones (Akita 2009). This is indeed the main manner in which the database will be used in this dissertation. In Chapter 4, three analytical variables of CHIDEOD, #(morphological_template), #(sensory_imagery) and #(radical_support) will be selected for participation in a statistical technique

called Multiple Correspondence Analysis (MCA). This will allow us to see how the values of these variables interact with each other and if a prototype can statistically be inferred. CHIDEOD will also be used for the selection of LIGHT ideophones that will be studied in Chapters 5 and 6. The benefit of this is once again a usage-based perspective: instead of relying solely on dictionary pointers of near-synonyms, with a database like CHIDEOD it becomes possible to include more data in a systematic and reproducible manner. Lastly, the database will also be used for the scope of the data selection in Chapter 7. This will be done in conjunction with corpora, such as the ASBC 4.0 corpus (see Section 3.3.3) and the Scripta Sinica corpus (Section 3.3.1). It is thus clear that this database, CHIDEOD, forms the backbone of this dissertation.

3.3 Corpora

Apart from CHIDEOD, the Chinese Ideophone Database, discussed extensively above, the current study draws its data from a number of different corpora. A good understanding of ‘corpus’ is provided by Gries & Berez (2017), who recognize that the term is polysemous and structured like a radial set. This means, that similarly to Lakoff’s introduction to the radially structured category of MOTHER (Lakoff 1987:74–76, 91–114), the category of CORPUS contains exemplars which are prototypical, i.e. share a number of widely accepted characteristics, but also exemplars which are situated further away from that salient core. The characteristics in question are argued to be that the corpus

- 
1. consists of one or more machine-readable Unicode text files [...];
 2. is meant to be representative for a particular kind of speaker, register, variety, or language as a whole, which means that the sampling scheme of the corpus represents the variability of the population it is meant to represent;
 3. is meant to be balanced, which means that the sizes of the subsamples (of speakers, registers, varieties) are proportional to the proportions of such speakers, registers, varieties, etc. in the population the corpus is meant to represent; and
 4. contains data from natural communicative settings, which means that at the time the language data in the corpus were produced, they were not produced solely for the purpose of being entered into a corpus, and/or that the production of the language data was as untainted by the collection of those data as possible.

Gries & Berez (2017:380)

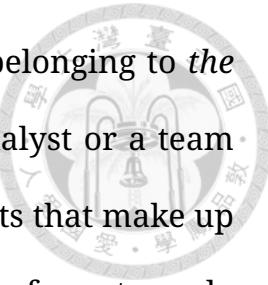
According to these features, Gries & Berez (2017) identify the British National Corpus (BNC) as prototypical. The corpora included in this dissertation generally conform to these four characteristics, although it is the third one that is violated most often. This has to do with a shift away from earlier corpus creation methodology, where balanced corpora were the gold standard, and when technology was not as advanced and the creation of

texts relied mostly on manual labor. As an example, the milestone Brown corpus (Francis & Kučera 1964) relied on punch cards, while these days it is computationally very easy to begin to create a corpus (de Marneffe & Potts 2017:427–428). In the last two decades, corpora have exploded in size, the idea being that if a corpus is sufficiently large, interesting phenomena can be found for study. As a consequence of this explosion, de Marneffe & Potts (2017:415) even define the corpus now as “any collection of language data [... , leaving] open the origin of this data, its size, its basic units, and the nature of the data that it encodes, which could come in any medium”. They also mention that by this definition, even dictionaries, specialized word lists, aggregated linguistic judgments etc., presumably even a database like CHIDEOD could be considered corpora. We do not want to go that far, and consider the four criteria identified by Gries & Berez (2017) as a better guideline.

The following corpora will be described in detail below: the Scripta Sinica, DIACHIC, and the Academia Sinica Balanced Corpus of Modern Chinese (ASBC 4.0). As a first introduction Table 3.19 shows them and judges them based on the four criteria identified by Gries & Berez (2017).

Table 3.19: An overview of the corpora used in this dissertation

corpus	text files	variety	balanced	natural
Scripta Sinica	yes	historical Chinese	no	yes
DIACHIC	yes	historical Chinese ideophones	no	yes
ASBC 4.0	yes	Standard Mandarin	yes	yes



Lastly, one other criterion, that is often mentioned as belonging to *the prototypical corpus* is annotation. This means that one analyst or a team of analysts, or a tagging scheme has marked up the elements that make up the corpus, in order to aid in later linguistic analysis. Voices for extremely detailed annotation as well as proponents for the opposite – raw texts – can be found (de Marneffe & Potts 2017:423). As a recent edited volume by Ide & Pustejovsky (2017) shows, there are seemingly endless annotation possibilities depending on the purpose of the research. As for annotation, the most common way is parts-of-speech. In the corpora in this dissertation, this can be found the ASBC 4.0. The historical corpus Scripta Sinica has not been segmented, and operates on literal string queries (as far as I understand the specifications). The DIACHIC has been segmented, as is explained in detail below.

3.3.1 Scripta Sinica

Since its inception in 1984, the Scripta Sinica²⁶ database (*Hànji quánwén zìliàokù* 漢籍全文資料庫), developed at Academia Sinica (2015), aimed to digitize all documents essential for traditional Sinological research²⁷. Later, the project turned into a full-text database for research into history and historical linguistics. Currently, more than 1,258 titles (718,132,225 characters of materials) pertaining to the traditional Chinese classics have been included in the database and have been categorized with meticulous care, according

²⁶ Available at <http://hanchi.ihp.sinica.edu.tw/ihp/hanji.htm>.

²⁷ The term *Hànji* 漢籍 seems to be a loan-translation from the Japanese term *kanseki* 漢籍, which refers to ‘old Chinese books’, i.e. those books used in Japanese traditional Sinology (cf. Wilkinson 2015:940)

to the introductory page.

The user-friendly web interface of the Scripta Sinica presents its data on a few different levels of granularity, the first of which is the periodization. This follows the traditional differentiation in groups of dynasties of the Chinese empire, most of the periods encompassing some 300 years. Figure 3.4 shows the label from Scripta Sinica (in Chinese) and the general names used for this period in this dissertation. In general, this classification has the benefit of following the traditional dynastic historiography. If one wants a more fine-grained analysis of a given linguistic phenomenon, it is possible to take one of the lower levels into consideration, such as the book title, and looking up when it (first) appeared. However, for analyses that are of a larger scope, such as those I present in this dissertation, the dynastic periodizations suffice. Furthermore, for some works the precise publication date is unknown, especially when they belong to earlier language stages. The reader is referred to Loewe's (1993) *magnum opus* for discussions on the periodization of Early Chinese works.

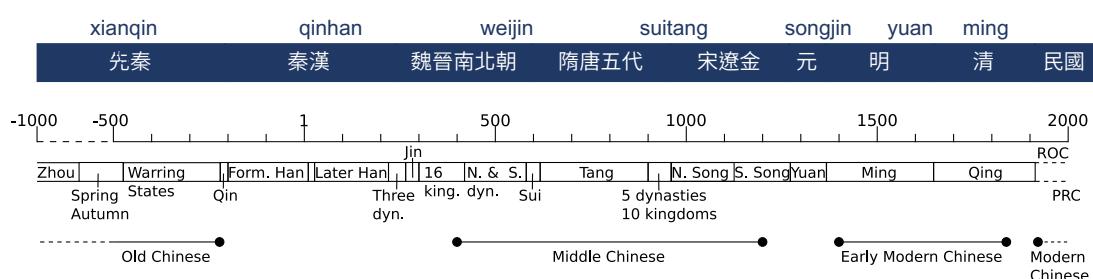
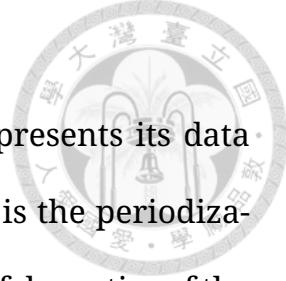


Figure 3.4: Visual representation of the periodization in the Scripta Sinica and the terms used in this dissertation.

For each period, the Scripta Sinica categorizes texts according to a five-fold scheme. The history of this classification originally was sixfold, but from the third century onwards, a simplified fourfold scheme came to be in





use, see Wilkinson (2015:936–940). This bibliographic classification scheme was known as the *sì bù* ‘四部’four branches. These included: the *jīngbù* 經部 ‘Classics’, generally referring to the Confucian canon in addition to works on textual criticism and philology; the *shǐbù* 史部 ‘Histories’, which are primarily works concerning history, but which also include geography and topography; the *zǐbù* 子部 ‘Masters’, which is a broad category including philosophical works other than the early Confucianists; and the *jíbù* 集部 ‘Literary Collections’, containing collectanea of individual authors or literary anthologies. As time unfolded, these four branches (*sì bù*) were kept in four different palace depositories, known as the *sì kù* 四庫, whence the famous *Sìkù quánshū* 四庫全書 ‘*Complete Library in Four Sections*’ project during the Qiānlóng 乾隆 Emperor’s reign (1711–1799) got its name: it follows the fourfold classification scheme. During the second half of the 19th century, however, a fifth branch was added, called *cóngshū* 叢書 ‘Collections’ to cope with the already huge number of collections (Wilkinson 2015:936–940). The Scripta Sinica also follows this fivefold classification scheme, with five branches (*bù* 部).

One level below the branches, we find so-called *lèi* 類 ‘categories’, which include text groupings like *zhèng shǐ* 正史 ‘official histories’ and *dìlì* 地理 ‘geography’ under the *shǐ* branch. The level below that there are the book titles (*shūmíng* 書名). After that there is detailed information like page number and so forth. It thus can be argued that the metadata of the Scripta Sinica is very detailed and can aid the researcher in studying language usage in different genres.

As mentioned before, the Scripta Sinica offers an intuitive user interface, where search terms can be entered and other variables such as the periodization can be checked for inclusion in the scope of the search. It is very easy to follow the development of a given term or conceptual metaphor, e.g., MIND metaphors in Pre-Qín Confucian texts (Hoon & Phua 2009). However, some practical comments about the large-scale extracting of data must be made. There currently are no direct plugin or API possibilities, so the user is forced to either use manual copy-pasting of the results, or resort to web crawling techniques. Because the website is dynamic, viz. the URL changes at every visit, this cannot be done systematically, unless one makes use of techniques that mimic the browser, such as the Selenium project²⁸, available through programming languages such as Python or R. We will return to this topic below when DIACHIC is introduced. As for usage in this dissertation, in Chapter 5 the Scripta Sinica will be used in the manual analysis of LIGHT ideophones.

3.3.2 DIACHIC

The next data source to be used in this dissertation is a subcorpus of the Scripta Sinica that I created, based on the distinct items in CHIDEOD. This subcorpus is called the DIAchronic CHinese Ideophone Corpus²⁹, DIACHIC. The need for such a corpus stems from the small amount of ideophones contained in the three historical tagged corpora developed at Academia Sinica, which are the Academia Sinica tagged corpus of Old Chinese

²⁸The development website of Selenium is available at <https://selenium.dev>.

²⁹Available here <https://osf.io/dc3uj/>.

(Academia Sinica 1990a), the Academia Sinica tagged corpus of Middle Chinese (Academia Sinica 1990b), and the Academia Sinica tagged corpus of Early Mandarin Chinese (Academia Sinica 1990c). While these corpora are very user-friendly, they do not provide enough data for the lexical semantic studies performed here. For some items, there can be as little as zero tokens. As an example, *bì~fèi* 霽沸 ‘spurt and bubbling’ does not appear in the Old Chinese corpus, but it does appear in the *Shijing*, as Ou (1992) 93 for instance shows. In itself there is nothing wrong with a (historical) corpus adopting a more balanced mindset – meaning that samples must be taken – but since this dissertation is focused on ideophones, it is necessary to include all data that can be found.

Thus, on the one hand these corpora are useful, because they follow a consistent tagging schema, but on the other hand the position taken here is that traditional studies on the parts-of-speech of Chinese have not been maximally inclusive of ideophones, so there are some issues that need to be resolved concerning the aforementioned tagging schema. The main tagging rule is that one character represents one word, which largely holds true for Pre-Modern Chinese (cf. Baxter & Sagart 1998). Some exceptions to this rule include the following five categories, as shown in (27).

(27) Patterns not conforming to the one-character-one-word principle in the historical corpora from Academia Sinica (Lee 2012:77)

- a. parallel compounds, e.g. *jūn-chén* 君臣 ‘the ruler and his minister’
- b. subordinating compounds, e.g. *tiān-xià* 天下 ‘the empire < all un-

der heaven'

- c. bisyllabic words (binomes), e.g. *jūnzi* 君子 ‘the gentleman’
- d. reduplications, e.g. *qī~qī* 蕤蕤 ‘luxuriant’
- e. proper nouns, e.g. *Zhōu gōng* 周公 ‘the Duke of Zhōu’

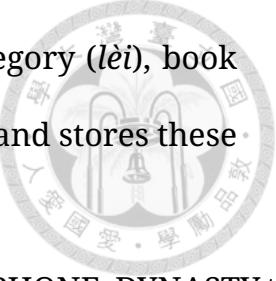


The problem with this otherwise valid tagging system is that ideophones do not necessarily come in a nice “reduplication” or “bisyllabic words” jacket; we have seen patterns up to four characters above (Section 3.2.3.1). This tagging scheme also confounds semantic notions with structural notions, for certain categories. For instance, example (27e) is more of a semantic notion; we need to understand who the Duke of Zhōu is, or at least that it is a person, so the segmentation should put these two characters together. In light of the present research, then, it is logical to adopt a more functional and practical approach to the segmentation (and tagging) of the subcorpus.

Making the most of CHIDEOD, we extracted all distinct ideophones in traditional Chinese characters and omitted single characters, in order to make sure that their non-ideophonic usage would not overgenerate data. This resulted in 4667 unique items that would need to be extracted from the Scripta Sinica corpus.

With Python 3.6 (van Rossum & Drake 2009) it was possible to iterate over each period of the Scripta Sinica to collect all tokens in one go. This means that for an item like *yìyì* 熠熠 ‘vividly bright’, the script loops over the periods *xianqin*, *qinhan*, *weijin*, *suitang*, *songjin*, *yuan*, *ming*, *qing* and

minguo (see Figure 3.4). Then, it takes the branch (*bù*), category (*lèi*), book title, up until the page number if filled out in the database, and stores these metadata with the corresponding paragraph into a .txt file.



In the following stage, all the different files (named IDEOPHONE_DYNASTY.txt for each item and dynasty) were combined into a structured corpus. The resulting structured corpus can be diagrammatically represented by the following directory tree:



```
##          levelName  
  
## 1  corpus  
  
## 2      |--xianqin  
  
## 3      |      |--經  
  
## 4      |      |--史  
  
## 5      |      |--子  
  
## 6      |      |      |--儒家  
  
## 7      |      |      |      |--荀子.txt  
  
## 8      |      |      |      °--...  
  
## 9      |      |      °--...  
  
## 10     |      |--集  
  
## 11     |      |--叢書  
  
## 12     |      °--...  
  
## 13     |--qinhan  
  
## 14     °--...
```

The resulting corpus contained about 5 GB of data in .txt files, which is quite large, even if spread out over different dynasties. Because the used method generated duplicate paragraphs, the duplicate rows per .txt file were removed. The resulting corpus ‘only’ had 1.33 GB of data, which is still quite large for a historical corpus.

Up until this point, the data was all unsegmented. However, to make it searchable, it needed to be segmented into words (in the sense of ‘separated by spaces’). During the pilot studies³⁰ I used the `jiebaR` package to achieve

³⁰This is similar to the data presented during my proposal exam.

this (Qin & Wu 2019). With the release of the `ckiptagger`³¹ developed at the CKIP group of Academia Sinica (Li, Fu & Ma 2019), there is a higher accuracy in segmenting Modern Chinese. However, at the moment of writing, there do not seem to be satisfactory segmentation tools or packages for Pre-Modern Chinese³². The `ckiptagger` library is an acceptable alternative to manually segmenting the 1.33 GB of .txt files, because there is the possibility to feed a list of words ('dictionary') into the segmentation function. In other words, the same ideophone list used to scrape the data was reused here, so as to make sure there are space boundaries around the items as they are being segmented. After all these steps, the resulting segmented DIAchronic CHinese Ideophone Corpus (DIACHIC) holds 1.1 GB of data, making it usable for the purposes of the salience research that will be presented in Chapter 6.

To better estimate the size of the data, the distribution of words per dynasty per branch (*bù 部*) is visualized in Figure 3.5. The distribution shows how the amount of data is not evenly dispersed, but that is not unusual, since in most cases the number of historical sources from more recent times will be more numerous than earlier ones. There is one period, however, the Yuán dynasty (1271–1368), for which there is less data than expected. Since it only lasted about 100 years, as opposed to the average of about 300 years, this actually is not surprising either. In any case, the number of words per period is certainly not small in this subcorpus of ideophones in Chinese. An-

³¹This library is available here: <https://github.com/ckiplab/ckiptagger>.

³²Although advancements are being made in this field. For instance, a recent dependency parser for Literary Chinese is presented in Yasuoka (2018), and available at <http://kanjizinbun.kyoto-u.ac.jp/%7Eyasuoka/kyodokenkyu/2018-12-01.html>.

other small comment should be made about the branches *cóngshū* 叢書 and *cóngshū 2* 叢書 2: this division was already present in the Scripta Sinica, and I do not know if this is the result of an input error or if they in fact are conceptually different. For this reason, both categories are kept.

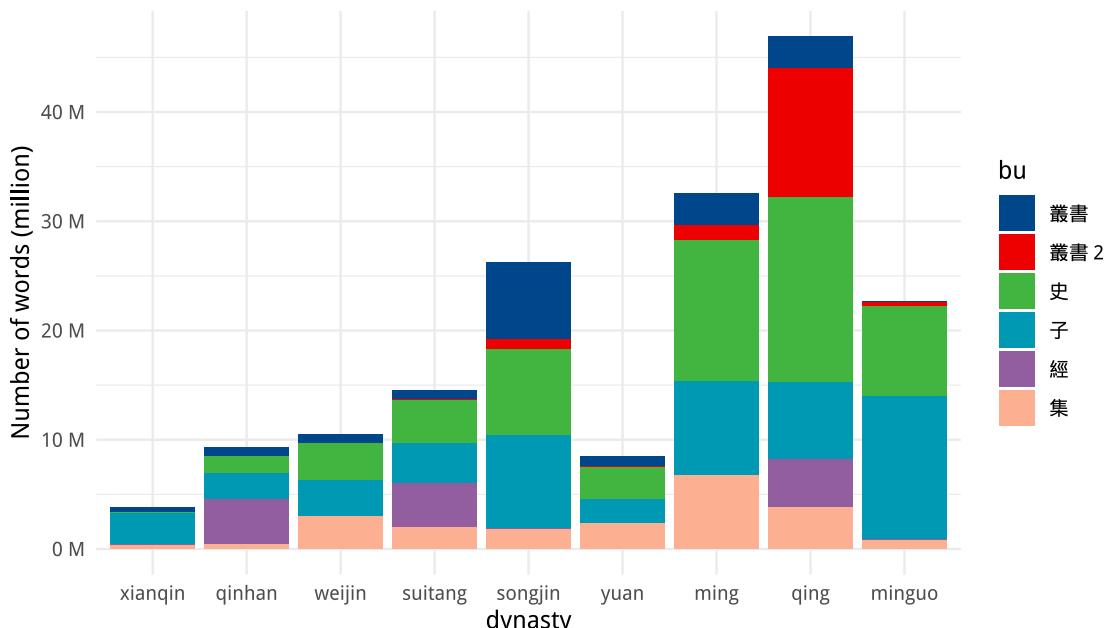


Figure 3.5: The number of words per branch per dynasty in DIACHIC

3.3.3 ASBC 4.0

For Mandarin Chinese, we include the Academia Sinica Balanced Corpus for Modern Chinese³³ (ASBC). Other alternatives would have been TenTen corpus or the Gigaword corpus. However, the ASBC was institutionally available and therefore easier to use with the methods outlined in this dissertation.

The first versions of the ASBC were small-scale and intended to draw feedback on issues concerning the balancing of categories, the part-of-speech tagging etc. (Huang & Chen 1992; Chen et al. 1993). Version 2.0 of the Sinica Corpus contained 5,345,871 characters, equivalent to 3.5

³³ Available here <https://ckip.iis.sinica.edu.tw/CKIP/engversion/20corpus.htm>.

million words (Chen et al. 1996). In 1997, version 3.0 was presented, which contained about 5 million words. The design of ASBC 3.0 as well as a few applications relating to mutual information, entropy etc. are discussed by Huang (2000). The most recent version of the ASBC is 4.0 (CKIP group & Academia Sinica 2013). It comprises about 10 million words. It was first completed in 2006, licensed in 2010 and accessible online in 2013. As can be seen in Figure 3.6, the bulk of the data stems from the early 1990s to the early 2000s, excluding 1835 files for which the date was not identifiable. While the current version of this corpus thus is made up of quite old data, for today's standards, the balanced nature of the corpus is still one of its advantages, as it facilitates a number of analyses.

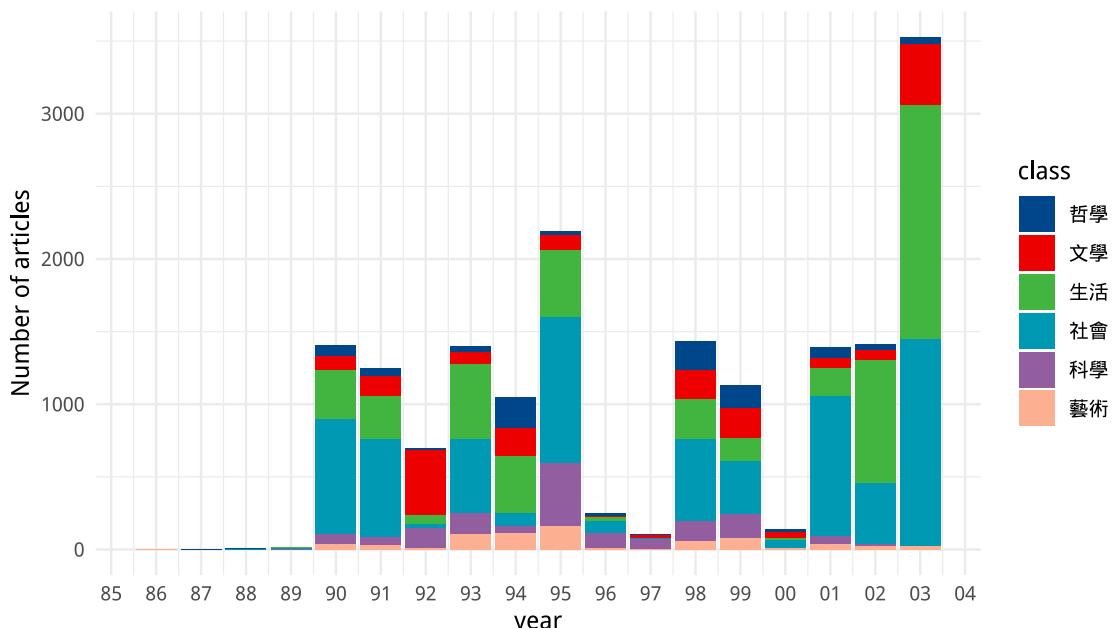
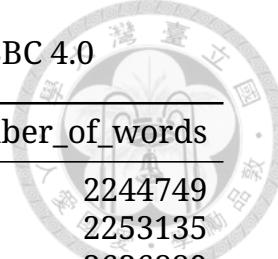


Figure 3.6: The number of articles per year in ASBC 4.0

It must be mentioned that there are a few discrepancies between the version accessible online³⁴ and the one that was institutionally available.

³⁴The url for the online version is <https://ckip.iis.sinica.edu.tw/CKIP/engversion/20corpus.htm>.

Table 3.20: The number of words per class in ASBC 4.0



class	class_eng	exp_number_of_words	actual_number_of_words
文學	literature	2244361	2244749
生活	life	2253102	2253135
社會	society	3636897	3636889
科學	science	1132298	1132397
哲學	philosophy	1129512	1129560
藝術	art	849160	849185

This latter version was originally in XML format, but I transformed it to a collection of .rds (R Data Structure) files. As for the discrepancies, the Table 3.20 shows the number of words per class (*zhǔtí* 主題 ‘topic’ on the online version) with the expected number of words and the ones I actually found in my data. These numbers do not appear to differ significantly, making it possible for us to use this version of corpus.

This corpus will be used quite extensively in this dissertation. In Chapter 4, it will be used in a Multiple Correspondence Analysis to investigate the prototypical structure of the Chinese ideophonic lexicon. In Chapter 7, constructions in which ideophones occur in will be probed with it.

4 Defining ideophones in Chinese



I can think of no better expression to characterize these similarities than "family resemblances"; for the various resemblances between members of a family: build, features, colour of eyes, gait, temperament, etc. etc. overlap and criss-cross in the same way.
—And I shall say: "games" form a family.

Ludwig Wittgenstein

In Chapter 2, we saw that the introduction of a category of "Chinese ideophones" can be beneficial for the converged analysis of related linguistic phenomena, such as research into onomatopoeia, binomes, and reduplication. However, that chapter could not answer what the boundaries of such a category are, nor how it is internally structured. This chapter³⁵ will investigate these two problems for Mandarin Chinese with the data sources introduced in Chapter 3. This will happen in four moves. First, a case study of a solution to the same issues in Japanese will be taken as a model. Second, the cross-linguistic state-of-the-art definition will be probed once more to identify the key features of "the canonical ideophone" and how it relates

³⁵Parts of this chapter were first presented at the International Workshop on Mimetics (Ideophones, Expressives) III: Crucibles of Mimetics, see Van Hoey (2019a).

to Chinese. From these two moves, it will follow that “the ideophone” as a cross-linguistic concept or as a language-particular category in Japanese cannot be explained through classical definitions based on necessary and sufficient conditions, but that membership to this group of words is prototypically structured, i.e., with some better representatives, and some worse cases near a fuzzy edge. There are a number of ways to investigate the prototypical structure of a category, but below we will adopt a statistical method, namely Multiple Correspondence Analysis, to explore the internal structure of the ideophonic lexicon of Mandarin. The third move is the application of this method to the available data in CHIDEOD (Chinese Ideophone Database). The fourth move rests on the linking between CHIDEOD and the ASBC 4.0 corpus. Instead of one single prototypical core, as is the case for Japanese ideophones, it will be shown that Chinese has a dual prototype structure, with considerable overlap between the two groups.

4.1 The prototypicality of Japanese mimetics

We start this chapter by summarizing how ideophones have been defined in Japanese. After all, Japanese linguistics have exerted considerable influence on Chinese work on onomatopoeia and ideophones (Zhào 2008), as mentioned in Chapter 2. Akita (2009:96–136) devotes a chapter to defining mimetics in Japanese from a prototype-theoretical point of view. Surveying the literature, he points out that Tamori & Schourup (1999) used “the categorization problem of mimetics” to refer to the problematic usage of native intuitions and other previous approaches to provide a sufficient def-

ition to the mimetic category (Akita 2009:98). What Tamori & Schourup (1999) suggest, then, consists of a set of formal features, unique to mimetics, such as an abundance of [p]-initial words, or suffixation in *-ri*, *-Q* (gemination), or *-N* (nasality), repetition and / or reduplication etc. Akita, however, does not regard such a featural definition of the mimetic class as a complete success; nevertheless other approaches, such as phonological or phonosemantic ones, are not entirely satisfactory either. Rather, his survey of previous literature leads him to the conclusion that a definition of mimetics in Japanese “cannot be formulated clearly with respect to both form and meaning” (Akita 2009:101), viz. it is mostly prototypical.

In a next step, Akita (2009) indicates that he is not the first one to point in this direction, e.g., Hamano (1998) has used iconicity as a fundamental notion in her treatment of Japanese ideophones, while Tamori & Schourup (1999) used mimeticity. Most interesting, however, is Lu (2006), who has argued that full reduplication patterns, called ‘ABAB’, *pi[^]kapika* ピカピカ ‘flashing’ and *to[^]NtoN* トントン ‘knocking’ are to be considered as a prototype for mimetics in Japanese (and possibly other languages). Let us explore this more in depth. Based on type frequency counts, Lu (2006) shows how this prototype can then be extended to other categories. This is visualized in Figure 4.1, where the ABAB-construction is the dominant morphological form and hence is marked with thick lines, while being connected to other morphological schemas, like ABN and ABQ. They all elaborate the real-word examples, drawn on a lower level. However, she also stipulates that there is a correlation between the senses depicted by mimetics and this construc-

tional paradigm. In Figure 4.2, the sensory prototype is considered to be SOUND AND MOVEMENT (*on ya doosa no renzoku hanpuku* 音や動作の連続・反復), which extends into STATE (*yootai no zizoku* 様態の持続), then into BODILY SENSES (*taikansei no zizoku* 体感性の持続), and then at last into PSYCHOLOGICAL STATES (*sinri zyootai no zizoku* 心理状態の持続). Crucial here is the schematization process, based on Langacker's (1987a; 1991) Cognitive Grammar, with constructions schematized (*tyuusyooka* 抽象化) from real language data, which in turn elaborate (*zireika* 事例化) the real data. For the current purposes of this chapter, the take-away message is that in Japanese the mimetic lexicon is prototypically structured in constructions, and that the senses associated with mimetics also display prototypical features.

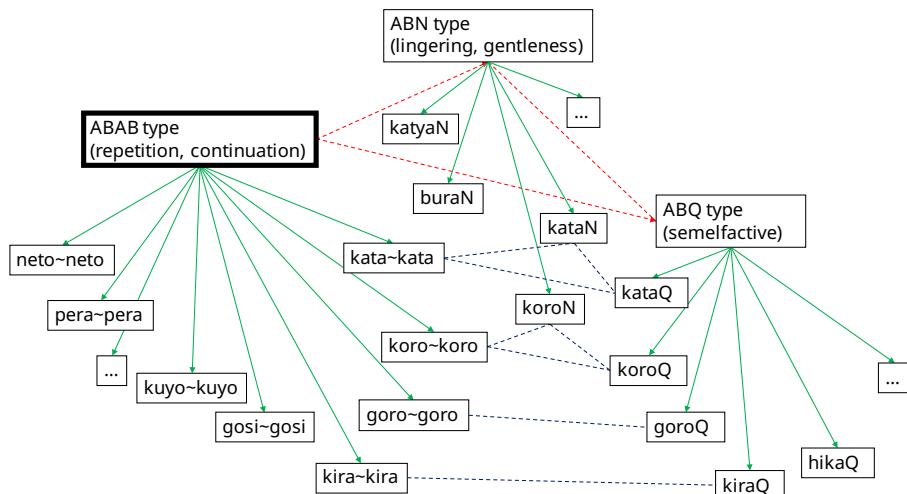


Figure 4.1: The ABAB-construction as the prototype in Japanese mimetics (Lu 2006:97)

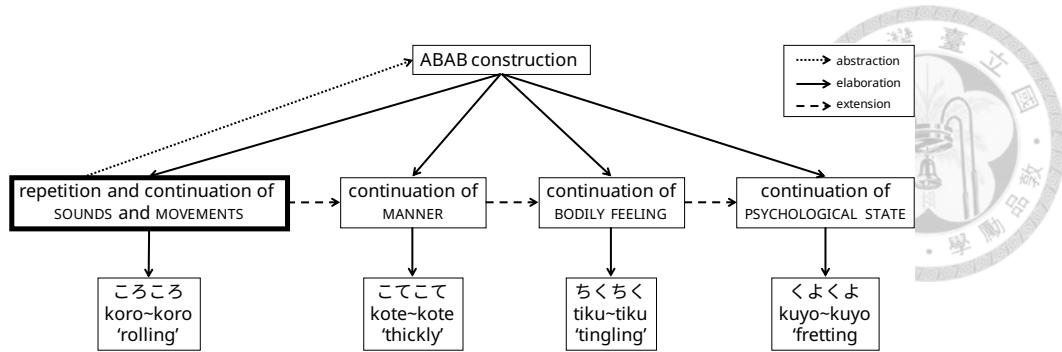


Figure 4.2: Extension across the senses for the ABAB-construction (Lu 2006:100) treatment of prototypes

How, then, does Akita (2009) go beyond this finding of prototypicality, as advanced by Lu (2006)? In a first step, he follows Tamori & Schourup's (1999) segmental criteria and aims to bring information about pitch accent into the equation (Akita 2009:113). This is included in order to discriminate mimetics from so-called 'quasi-mimetics'. For instance, examples (28a-28c) are derived from non-mimetic words. Of these, Akita shows that (28c) can have multiple pitch accent patterns: *iki~^iki* or *iki~i^ki*, which deviates from the prototype: CV[^]CV-CVCV (cf. 28d). Another category which is not deemed mimetic by Akita are referential reduplications (28e-28f).

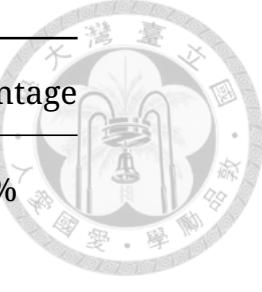
(28) Adapted from Akita (2009:104–106)

- a. *kona~gona* 粉々 ‘in pieces’ < *kona* 粉 ‘powder’
- b. *u^ki~uki* 浮き浮き ‘cheerful, lighthearted’ < *uku* 浮 ‘float’
- c. *iki(^)~i(^)ki* 生き生き ‘lively’ < *iku* 生 ‘live’ (non-modern verb)
- d. *si^to~sito* シトシト ‘wet’
- e. *mura~^mura* 村々 ‘villages’
- f. *ie~^ie* 家々 ‘houses’

Eventually, this leads Akita to investigate the distribution of pitch-information-rich constructions of ideophones: “it is perhaps true that the majority of mimetics have distinctly mimetic tones. Nevertheless, there are a certain number of words that should be located in a peripheral part of the mimetic category or on its boundary that is fuzzy. This fact leads us to the idea that Japanese mimetics form a prototype category with a fuzzy boundary” (Akita 2009:106). Based on a data set of dictionary items ($n = 1652$) he finds the following distribution of 15 morpho-phonological patterns (Table 4.1).

Table 4.1: The coverage of mimetic morphophonological templates (adapted from Akita 2009:110)

Root	Template	Number	Percentage
1 mora	CvQ(^)	50	3.03%
1 mora	CV(^N^)	29	1.76%
1 mora	CViQ	14	0.85%
1 mora	CV(^V^)	21	1.27%
1 mora	CV(^)V-CVV	46	2.78%
1 mora	CV(^)N-CVN	45	2.72%
1 mora	CV^i-CVi	9	0.54%
2 morae	CVCVQ^	213	12.89%
2 morae	CVCV(^N^)	101	6.11%
2 morae	CVCB^ri	130	7.87%
2 morae	CVCCV^ri	134	8.11%
2 morae	CV(^)CV-CVCV etc.	484	29.30%



Root	Template	Number	Percentage
Derivatives		332	20.10%
Fossilized		35	2.12%
no template		9	0.54%

Apart from iconicity³⁶, Akita thus uses the segmental criteria from Tamori & Schourup (1999) as well as his 15 morpho-phonological patterns to perform two experiments. The results indicate that indeed, the mimetic category in Japanese is internally structured with a prototypical core and a fuzzy boundary (Figure 4.3).

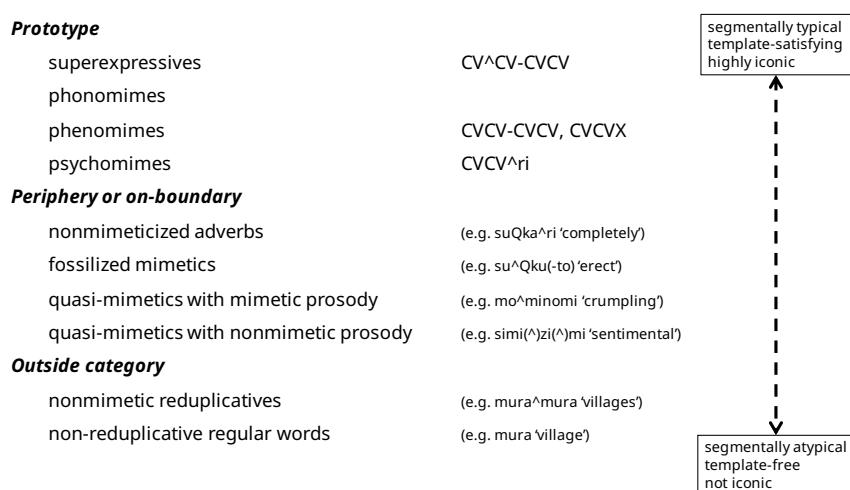
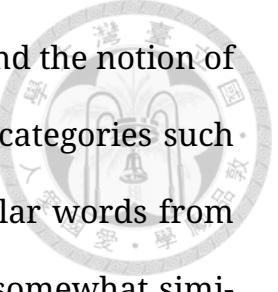


Figure 4.3: The internal structure of the prototype category of Japanese mimetics (Akita 2009:135)

Figure 4.3 is a slightly more fine-grained presentation of the different prototypical constructions that are associated with different sensory types in the mimetic lexicon. The steps Akita took to get at this summary are as follows: first he identified useful frameworks in the previous literature, such

³⁶This is not treated in the current version of this chapter, but concerns highly iconic words like コケコッコー (と) kokekoQko[^]o(-to) 'cock-a-doodle-doo', cf. [Akita (2009) 111; 116]



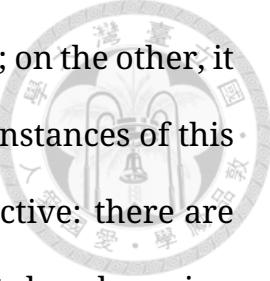
as the segmental criteria from Tamori & Schourup (1999) and the notion of prototypicality in Lu (2006). This allowed him to exclude categories such as non-mimetic reduplicatives and non-reduplicative regular words from the scope of the mimetic lexicon, as well as placing some somewhat similar forms on the periphery or boundary of the mimetic category; Next, he examined the distribution of the 15 different morphological templates in a data set, and based on type frequency arrived at the summary in Figure 4.3. Relevant to the structure of this chapter are exactly those steps. However, before we adapt this case study to Chinese ideophones, it is useful to zoom out and discuss the state of the art of cross-linguistic typology with regards to ideophones.

4.2 The canonical ideophone

As sketched in Chapter 2, the currently most widely adopted cross-linguistic definition for ideophones comes from Dingemanse (2011a; 2012; 2019), repeated in (29).

- (29) IDEOPHONE. A member of an open lexical class of marked words that depict sensory imagery.

Dingemanse is of the opinion, that “the term ‘ideophone’ is best seen as a comparative concept (Haspelmath 2010), like ‘adjective’, ‘future tense’ or ‘serial verb construction’” (Dingemanse 2019:14), but is also keenly aware of the two sides concealed in such an argument: on the one hand, ‘ideophones’ are to be considered language-general notions, that are not directly



defined in terms of their occurrence in particular languages; on the other, it leaves room for language-specific nuances in the concrete instances of this concept in particular languages. We agree with this perspective: there are many studies on the phenomenon in different languages that show how similar groups of words occur cross-linguistically (see Chapter 2), while each differing considerably as well. However, that does not absolve us from trying to chart these particular nuances in Chinese (and relate them to the cross-linguistic concept of ‘ideophone’). Let us first look at the five criteria which, according to Dingemanse (2019), constitute the key-properties of an ideophone, presented in (30). These are based on a thorough body of literature survey built up over the years (Dingemanse 2011a; 2012; 2019; Dingemanse 2016).

- (30) a. ideophones are MARKED, i.e. they have structural properties that make them stand out from other words
- b. they are WORDS, i.e., conventionalized lexical items that can be listed and defined
- c. they DEPICT, i.e., they represent scenes by means of structural resemblances between aspects of form and meaning
- d. their meanings lie in the broad domain of SENSORY IMAGERY, which covers perceptions of the external world as well as inner sensations and feelings
- e. ideophones form an OPEN LEXICAL CLASS, i.e., a set of lexical items open to new additions

In the next sections we will go over these one by one, discussing first the criteria as they have been represented in previous research, and then contemplate on how this works out in Chinese. It will become clear that no criterion by itself is really *the* defining (necessary or sufficient) criterion for ideophones, as the different criteria all subsume more than just ideophones.

4.2.1 Ideophones are marked

First, ideophones are MARKED (30a) because their properties make them stand out from other words – they indicate or signal a certain quality, different from the prosaic words around them (Dingemanse 2011b:25–26). Often in previous research, this markedness has been understood in terms of phonological or phonotactic peculiarities (e.g. Nuckolls et al. 2016), but special word forms, expressive morphology (Zwicky & Pullum 1987), relative syntactic independence or foregrounded prosody (Childs 1994) are possible ways in which this markedness criterion is manifested. A recent overview of this criterion in Japanese is offered by Dingemanse (2016:8–11), who uses a multimodal corpus of interviewees after the Tōhoku earthquake and tsunami disaster in 2011 in Japan. They state that intonational foregrounding – a markedly lower or, more frequently, higher pitch of an ideophone, preceded by an intonational pause – occurs often for Japanese ideophones. In (31), the Japanese ideophone *zabuun* ‘splash’ is uttered in a markedly higher pitch, after which the interviewee returns to the normal pitch range.

- (31) Intonational foregrounding (adapted from Dingemanse 2016:9)

ザブーンザブーンっていう 音は 私 聴こえてたの。
↑zabū:n-zabū:n↑-t:e-i-w oto-wa watafi kikoe-te-ta-no.
IDEO.splash.-QUOT-say-NPST sound-TOP I hear-CONJ-PST-SFP
“... I heard the sound like *splash-splash*.”



The same study also discusses phonational foregrounding – using different kinds of phonation such as breathy voice, growl, creaky voice, voicelessness, and whisper – as a way markedness can occur. They report phonational foregrounding of ideophones in the form of breathy voice, creaky voice, stiff voice, falsetto, voicelessness, or whisper in their corpus. For instance, *gu:t-to* ‘suddenly’ is produced as [gu:t:o] in this instance with stiff voice, i.e. with the glottal opening narrower than normal.

- (32) Phonational foregrounding (adapted from Dingemanse 2016:10)

もう 明らかに 流れが グット また こちの
mo: akiraka-ni nagare-ga **gur:t-to** **[gu:t:o]** mata kotš:i-no
just obviously flow-NOM IDEO.suddenly-QUOT again over.here-GEN
ほうに 広がって
ho:-ni hirogat-te
direction-DAT spread-CONJ
“... obviously, the flow spread far and wide over here again, and...”

Lastly, expressive morphological processes, such as vowel lengthening (*do:n*), partial multiplication (*dododon*), and stem repetition (*don-don-don*) occur as well. Examples embedded in a context are given in (33), where

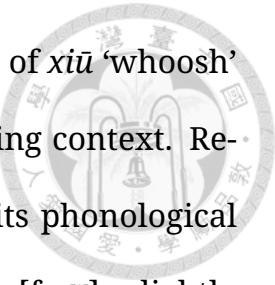
juuk:uri ‘slow’ is lengthened to [ju:k:uri], and *gat(-to)* ‘rattling’ is produced with partial reduplication. Dingemanse & Akita stress that while these three phenomena – intonational foregrounding, phonational foregrounding, and expressive morphology – are logically distinct, they do often occur together, and “contribute to the ‘performative foregrounding’ of ideophones (Nuckolls 1996)” (Dingemanse 2016:11).

(33) Expressive morphology (adapted from Dingemanse 2016:11)

瓦礫とかが	ゆーっくり	こう	動いて	いて
gareki-toka-ga	↑ju:k:uri↑	ko:	wgoi-te	i-te
debris-etc.-NOM	IDEO.slow	like.this	move-CONJ	be-CONJ
そして	また	ががががががががっと	引くのが	
sojite	mata	gagagagagagagat-to	cik-wi-no-ga	
and	again	IDEO.rattling-QUOT	draw-NPST-NMLZ-NOM	
朝までに	まあ	二、三回で	聞かないぐらい	
asa-made-ni	ma:	ni-san-kai-de	kika-na-i-gurai	
morning-until-DAT	well	2-3-time-in	suffice-NEG-NPST-degree	
あつたんじやないかな				
at-ta-n-za-na-i-kana				
be-PST-NMLZ-COP-NEG-NPST-SFP				

“...things like debris moved slowly and drew back with a rattling sound, which was [repeated] more than two or three times by the morning, I guess.”

Of course, markedness as it is used in these contexts can also be found



in Mandarin Chinese. For example in (34-35) the variations of *xiū* ‘whoosh’ are typically produced in a higher pitch than its surrounding context. Remarkable is that *xiū* is usually not produced as [ɛiou], as its phonological transcription (Hanyu pinyin) would suggest, but rather as [ʃw^w] – lightly stretching the phonological inventory of Mandarin. Furthermore, in (35) it occurs thrice, indicating that the same semelfactive event happens three times, viz. there are three planes that fly over at high speed. Lastly, there is also some markedness in the written form, which is presumably a hallmark of Chinese. For instance, it has been noticed that the *mouth*_Radical 口 often indicates onomatopoeia (cf. Li 2007), and this is also present in *xiū* 嘿.

- (34) 運動員 嘿的一聲 跑過來。
yùndòngyuán ↑xiū **[ʃw^w]**↑=de-yi-sheng pǎo-guò-lái
athlete IDEO.whoosh=LNK-one-sound run-cross-come
“The athlete *whooshed* in my direction.”

- (35) 飛機 嘿咻咻 = 地 飛過去。
fēijī ↑xiū~xiū~xiū **[ʃw^w]ʃw^w]ʃw^w]**↑=de fēi-guò-qù
plane IDEO.whoosh LNK-one-sound
“The (three) planes flew over, *whoosh*, *whoosh*, *whoosh*.”

Finally, it must be mentioned that Chinese ideophones display some phonological markedness without expressiveness: they are skewed to high tones, a bias not found in the prosaic lexicon (Thompson 2018).

4.2.2 Ideophones are words



Second, ideophones are WORDS, i.e. “conventionalised minimal free forms with specifiable meanings” (Dingemanse 2011a:26). Dingemanse refers to Haspelmath’s (2011) discussion on the problematic usage of the term word, viz. those cases where criteria for wordhood cannot differentiate words and smaller units, or words and larger units (phrases). Dingemanse resolves this issue by equating his usage of the term ‘word’ “for the more primitive and well-defined concept ‘root’ that Haspelmath (2011:70) proposes” (Dingemanse 2011a:27). What Dingemanse wants to avoid, is the fallacy that ideophones are just free expressive noises or spontaneous acts of mimicry; in contrast, they are conventionalized items. Valid evidence for him (and for us) lies in the fact that when native speakers are asked to define ideophones, they often are able to come up with coherent definitions.

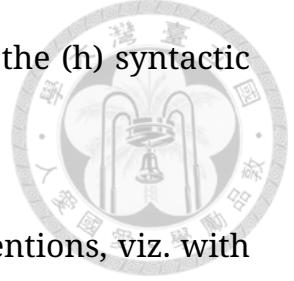
However, two addenda to this criterion must be made: the first from the perspective of Cognitive Linguistics, the second from that of Chinese in particular. From the point-of-view of Cognitive Linguistics, the notion of ‘word’ has received a lot of attention. For example, Taylor (2003) dedicates a chapter to differentiating words from affixes and clitics, by using a bundle of criteria such as (a) the independent utterance nature of a word, (b) stress patterns (in English), (c) phonological stability in different contexts, (d) their being “rather unselective” with regard to the kind of adjacent item, (e) movability in the sentence. As opposed to affixes and clitics, these criteria do seem to hold up, e.g., criterion (e) seems impossible for these two groups.

For Langacker as well, ‘words’ are not simply what lexicons are made of. Instead, the lexicon (in Cognitive Grammar) is defined as “the set of fixed expressions in a language” (Langacker 2008a:16). For him, these expressions are to be understood as symbolic assemblies: mappings between form and meaning. What this means, then, is that the lexicon is made up of well-entrenched expressions (which thus have attained so-called ‘unit’ status) or novel expressions, without being able to draw a clear boundary between these two extremes.

The same goes for ideophones: some will be readily available as *units* in the lexicon, e.g., *bang!* for the sound of an exploding bomb (Taylor 2004), while other ones will only vaguely be apprehended as being part of the group of ideophones. Therefore, the WORDS criterion in Dingemanse’s definition is useful for recognizing the conventionalization of ideophones, with the two caveats that (a) their being listable and definable is by no means unique to ideophones, but just a property of (well-formed) expressions in general, and (b) the lexicon and syntax are less apart than traditional approaches to languages, cf. discussions on the so-called rule-list fallacy (Langacker 1987a:28–29) – it is not that ideophones are ‘in the lexicon’ as opposed to not being ‘in the syntax’.

Second, from the perspective of Chinese then, this WORD criterion for ideophones might be particularly difficult. Packard’s work is particularly revealing in this aspect (first in Packard 1998; later in Packard 2000). He distinguishes between (a) the orthographic word, (b) the semantic word, (c) the psycholinguistic word, (d) the phonological word, (e) the morphological

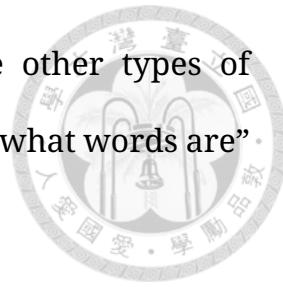
word, (f) the lexical word, (g) the sociological word, and the (h) syntactic word.

- 
- (a) The **ORTHOGRAPHIC WORD** is defined by writing conventions, viz. with spaces as the boundaries. Most computational approaches use this principle to perform word counts. This notion, however, is problematic in Chinese (and other languages, like Japanese) where graphemes aren't separated by empty spaces. For this reason, Packard does not discuss this conception of 'word' any further.
 - (b) The **SEMANTIC WORD** is defined using semantic criteria. As Packard (2000:9–10) explains, it is sometimes equated with the idea of a 'unitary concept' (Sapir 1921) or a 'basic expression' in formal semantics: a form with a semantic value that such expressions may combine to form complex expressions, but may not be further decomposed into subexpressions (Baxter & Sagart 1998). Semantic words, defined like this, are only minimally useful, "because reducing concepts to their semantic primitives is a notoriously difficult exercise", Packard argues, and I agree with him.
 - (c) The **PSYCHOLOGICAL WORD** is defined by Packard (2000) 13-14 as the "the operation of the language processor", which could be a cognitive compilation of different properties, such as: phonological or prosodic, semantic, morphological, and syntactic knowledge, with the relative proportions of such knowledge at any given point in processing time being dependent upon linguistic task demands or the state of the language

processor. Packard states that while this notion is plausible, it is not satisfactorily defined (in 2000) and thus temporarily cast aside.

- (d) The PHONOLOGICAL WORD can be defined through criteria such as the places where pauses in a sentence can possibly occur (Chao 1968:153–154). But, as Packard critiques, “‘word’ as defined by the phonological criterion of potential pause turns out to be of little use, since, like the orthographic and lexical definitions of ‘word’, this criterion turns out largely to be based upon other (i.e., syntactic, morphological or prosodic phonological) criteria. That is, the reason ‘pauses’ cannot go where a speaker feels it is inappropriate to place them is because their placement would violate the constituency of a syntactic, a morphological or a (otherwise defined) phonological word” (Packard 2000:10–11). To rectify the usefulness of the notion ‘phonological word’, Dai (1998) showed how phonological rules can identify words. Another approach characterizes the phonological word in prosodic terms (Duanmu 1998), with phonological tone and stress as the markers, or simply as a ‘prosodic word’ situated between the prosodic levels of ‘phonological phrase’ and ‘foot’ (cf. Feng 1998). While interesting analyses, as Baxter & Sagart (1998) show, there is often a mismatch between phonological words and other definitions of ‘word’ in Chinese. Thus we follow Packard in saying that “while phonological structure may indeed be sensitive to and correspond with word-sized entities as independently defined elsewhere, and phonology does provide another important piece of evidence that

converges on the construct *word*, nonetheless the other types of evidence correlate better with speakers' intuitions of what words are".
(Packard 2000:11).



- (e) The MORPHOLOGICAL WORD can be understood formally as the proper output of word-formation rules (Packard 2000:11–12). For Packard, this thus largely can be equated with the notion of 'syntactic word' (see below).
- (f) the LEXICAL WORD, as characterized by Packard (2000:8–9), refers to the 'listedness' of words in a 'lexicon', but, as he states, this "listedness criterion is neither sufficient nor necessary to define 'word', because it is common to have both 'listed' items that are not words (e.g., idiomatic phrases or 'listed syntactic objects' [...] and words that are not 'listed' (e.g., large numbers of complex words in languages such as Turkish or Italian that are productively constructed using members of affixation paradigms, and are not likely to be stored away as 'listemes')". Packard does not think this is a useful understanding of 'word' for Chinese, precisely because it would exclude all words created by rule, and by stating this falls prey to the rule-list fallacy (see above, and below).
- (g) The SOCIOLOGICAL WORD, attributed to Chao (1968:136), revolves around the terms used by native speakers when referring to linguistic units of a certain size. It is the type of unit the general, non-linguistic public has an everyday term for. In Chinese, Packard agrees, this term is *zì* 字, referring to either the Chinese written character or spoken

morpheme. In reality, these two separate concepts are conflated: “the *zì* as morpheme and the *zì* as written character are one and the same thing. This is due to the tacit assumption that the spoken *zì* (morpheme) can always be visually rendered with a written *zì* (character)” (Packard 2000:15). That being said, there is another term for ‘word’ which is distinct from ‘character’ in Chinese: *cí* 詞. Used predominantly by linguists, this term can be equated with the syntactic word in Chinese, as Packard states.

- (h) A SYNTACTIC WORD is “a form that can stand as an independent occupant of a syntactic form class slot, in other words, a syntactically free form, commonly designated in the literature as X^0 . This is probably the most common current linguistic characterization of the notion ‘word’ [...] Defining a syntactic word presumes that we can identify basic form class categories, and then use native speaker judgments to determine what entities are able to minimally occupy the category slots within utterances. This notion of syntactic word, as we shall see, will be one we crucially rely on in our description of Chinese words” (Packard 2000:12–13). It is clear that Packard (2000) favors X-bar theory in his comprehensive treatment of wordhood in Chinese. However, the formalisms of that theory, which assume complete compositionality and treat morphemes as building blocks, buys into the rule-list fallacy mentioned above. As an illustration, Packard characterizes ‘computer’ as follows: “[i]n the word *diànnǎo* 电 脑 electric brain ‘computer’, *the speaker perceives the meaning to be ‘elec-*

tric brain or ‘brain that is composed of electric circuits’ – an entirely reasonable semantic interpretation that also happens to conform to the general modification structure of Mandarin: the modifier virtually always precedes that which is modified” (Packard 2000:25, emphasis mine). Thus, these modification structures go to the lexicon, grab ‘electricity’ and ‘brain’, force them into the rule and get as output ‘electric brain > computer’. Unfortunately, this theoretical model of maximum compositionality simply doesn’t hold up: just because the elements ‘electricity’ and ‘brain’ are salient, does not mean that the compound of these two will compute the full constituency every time the compound is accessed. Counter proposals have been made for treating units with different levels of entrenchment differently, such as those in Cognitive Grammar by (Langacker 1987a; 1991; 2008b; Tuggy 1992). That being said, Packard does adequately describe what the notion of ‘syntactic word’ is about.

While the scope of this section is too short to comprehensively address these ‘word’ issues, it is worth pointing out that in recent years these different interpretations of ‘word’ have been successfully revisited, notably from the perspective of Construction Morphology (Booij 2005; 2007; 2010; 2017; 2018), which explores morphology through tenets used in Construction Grammar (e.g. Goldberg 1995; 2006; Croft 2001). For instance, Arcodia & Basciano (2018) take as their basic conceptualization of ‘word’ the ‘syntactic word’, but avoid the problems of the rule-list fallacy by filling out compositional slots (similar to Packard’s rules) and schemas, which are construed

bottom-up (diametrically opposed to Packard's rules). The last word about 'word' in Chinese has not been said, but it is a positive signal that the field of morphology is warming up to the notion that they can study morphology through constructions, mapping between form and function, or form and meaning (cf. contributions in Hoffmann & Trousdale 2013; Langacker 2005).

This lengthy discussion means that the WORD criterion can be used to characterize ideophones (in Chinese), since words are expressions with different levels of unit status in the grammar. However, this is obviously not sufficient to distinguish them from other prosaic words.

4.2.3 Ideophones depict rather than describe

Third, ideophones DEPICT (30c), rather than describe. That is, they represent their referents in a notable way. As Dingemanse (2011a:27) relates, instead of the 'normal' mode of representation, they have been called performative as opposed to discursive (Nuckolls 1996), dramatic as opposed to commen-tative (Kunene 1965; Fortune 1962), expressive as opposed to prosaic (Dif-flofth 1972), affecto-imagistic as opposed to analytic (Kita 1997), or mimetic as opposed to descriptive (Güldemann 2008). The next example (36) shows the difference between these two modes of representation. The ideophone (36a) depicts what is described in (36b). *Gbadara-gbadara* is uttered as a small performance.

(36) Siwu (Dingemanse 2011b:27)

- a. *gbadara-gbadara*

b. ‘be walking unevenly and out of balance’



For Dingemanse, this criterion is the crux of his PhD dissertation: “For that is my suggestion: ideophones are first and foremost depictive signs: words that enable others to experience what it is like to perceive the sensory imagery depicted” (Dingemanse 2011a:38). It is also the reason why they are sometimes erroneously viewed as iconic signs. If iconic signs are defined as “conventionalized linguistic signs that exhibit some form of iconicity” in which iconicity is equated with “a perceived resemblance between aspects of form and meaning” (Dingemanse 2019:18; based on Ahlner & Zlatev 2010; Dingemanse 2012; Clark 2016), then these notions do come very close to the definition of ideophones as posited by Dingemanse, which is why the two can be conflated. Yet, they are not the same. An advantage of Dingemanse’s choice to use ‘depiction’ as the criterion entails that the two can be separated, i.e., there are different communicative phenomena that fall under the term of depiction:

Because depiction is defined as a communicative act and not by reference to iconic signs or ideophones, important similarities become visible between ideophones, iconic gestures, direct quotations, bodily demonstrations, and enactments, all phenomena united by their fundamentally depictive nature (Güldemann 2008; Ferrara & Hodge 2018). While it pays off to be attentive to the semiotic kinship between these phenomena, there are also salient differences in terms of modality, gradience and conventionalisation (Okrent 2002). For ideophones, as we will

see below, the most important of these differences are captured by defining them as an open lexical class of marked words depicting sensory imagery.

Dingemanse (2019:19)

In the examples above (34-35) *xiū* is often accompanied by a gesture, namely a ‘hand palm cutting straight through the air in a forward motion’. A pioneering study on the relation between gesture and ideophones used in the Peking dialect (Sam-Sin 2008) shows how gesture supports the markedness of the two ideophones used. When asking about the meaning of *èrtíjiǎo* 二踢脚 ‘double-bang firecracker’, the informant used two ideophones, accompanied by gestures, to explain the meaning, as shown in example (37) and Figure 4.4. The interaction of gesture and ideophones unfortunately falls outside the scope of this dissertation, but deserves further study.

(37) *pōng* ‘bang’, *pà* ‘louder bang’ in Sam-Sin (2008:23–24)

二踢脚儿	啊？	嘛，	啪	就	那	种
èrtíjiǎor	a	pōongggggg	pàaa	jiù	nèi	zhǒng			
firecracker SFP IDEO.bang IDEO.louder.bang CONJ DEM kind									

“*Ertijiao?* [First] poongggggg, [then] paaa, that kind.”



Figure 4.4: Left panel: the gesticulations of the losing up of the fire cracker which was accompanied by the first loud report, *pōng*. Right panel: the gesticulations of the second, and louder, muffled bang, *pà*, which was accompanied by a flash. (Sam-Sin 2008:23–24)

A second way this depiction can play out is in the written form of Chinese characters. Ever since Xǔ Shèn's 許慎 traditional six-way classification of Chinese characters, illustrated in (38-43), known from the *Shuōwén jiězì* 說文解字 (2nd century CE) (Hsieh 2006:40–43), Chinese characters have been analyzed into components, which has been very useful, since the majority of Chinese characters follow the picto-phonetic principle (41). Often the semantic contributor in this type, the so-called RADICAL or FUNCTIONAL COMPONENT, is in fact a character ‘created’ through the picto-graphic principle (38). These radicals then work as indices, but in themselves are still often iconic – graphically depicting a (possible) semantic domain. See Van Hoey (2018a:248–252) for an overview of how this plays out in meteorological ideophones.



(38) The picto-graphic principle (*xiàng xíng* 象形 ‘form imitation’)

a. *mù* 木 ‘tree, wood’

b. *yǔ* 雨 ‘rain’

(39) The picto-logic principle (*zhǐ shì* 指事 ‘point at things’)

a. *běn* 本 ‘root (of tree)’

b. *mò* 末 ‘top (of tree)’

(40) The picto-synthetic principle (*huì yì* 會意 ‘assembling meaning’)

a. *lín* 林 ‘forest’

b. *sēn* 森 ‘woods’

(41) The picto-phonetic principle (*xíng shēng* 形聲 ‘form and sound’)

a. *gēn* 根 ‘root’ < a radical *mù* 木 ‘tree, wood’ and a sound component

gēn 艮 ‘seventh of the eight diagrams’

b. *jiāng* 江 ‘river’ < a radical *shuǐ* 氵 ‘water’ and a sound component

gōng 工 ‘work’

(42) The mutually interpretive symbolic principle (*zhuǎn zhù* 轉注 ‘turn and interpret’)

a. *lǎo* 老 ‘old’ > *kǎo* 考 ‘investigate’

(43) The phonetic loan principle (*jiǎ jiè* 假借 ‘false borrowing’)

- a. *yào* 要 ‘to want’, originally ‘waist’, so *yāo* 腰 ‘waist’ was introduced
- b. *běi* 北 ‘north’, originally ‘back (of body)’, so *bèi* 背 ‘back (of body)’ was introduced

A third way depiction can be found in Chinese is by investigating reduplicated words. The recent Chinese study (Li 2015) and its cross-linguistic follow-up study (Lǐ & Ponsford 2018) have revealed in more detail what kinds of extra meaning verbal reduplication can contribute to the semantics of a predicate. A small selection of their survey includes habitual, attenuative of extent, distributive object or subject, reflexive etc. They related these to five different domains of iconicity: identity, magnitude, discreteness, proximity, and sequentiality (Lǐ & Ponsford 2018:79).

A fourth way depiction can be found across languages is in the morphosyntactic behavior of reduplicated words, especially in relation to negation. That is to say, ideophones are cross-linguistically known to “display an antipathy towards negation and questioning (Diffloth 1972; Childs 1988; Kilian-Hatz 2006)” (Dingemanse 2017:363–364). While this is presumably a cross-linguistic tendency, rather than a rule set in stone, it has been noticed before that ideophones indeed are often best treated as “positive-polarity items” (Tolskaya 2011). This can also be found in Chinese. As an example, Paul (2006; 2015a) revisits the dichotomy between so-called “simplex adjectives” (*jiběn xíngshì* 基本形式) and “complex adjectives” (*fùzá xíngshì* 複雜形式), two terms established by Zhū (1956) (see also Chapter 7).

(44) Mandarin Chinese (Paul 2006:310-311)



- a. 他的 衣服 比 你的 更 乾淨。
 tā=de yīfú bǐ nǐ=de gèng gānjìng
 3.SG=LNK clothes COMP 2.SG=LNK even.more clean
 “His clothes are even cleaner than yours.”
- b. *他的 衣服 比 你的 更 乾乾淨淨的。
 *tā=de yīfú bǐ nǐ=de gèng gāngānjìngjìng=de
 3.SG=LNK clothes COMP 2.SG=LNK even.more clean.IDEO?=LNK
 “His clothes are even cleaner than yours.”
- c. 老 這麼 慢騰騰的 可 不 行。
 lǎo zhème mǎn-téngténg=de kě bù xíng
 always so slow-sluggish.IDEO=LNK can NEG possible
 “It’s impossible to be always so sluggish.”
- d. 他 不 胖。
 tā bù pàng
 3.SG NEG fat
 “He is not fat.”
- e. *他 不 胖胖的。
 *tā bù pàngpàng=de
 3.SG NEG fat.IDEO?=LNK
 “He is not fat.”



f. *他 非常 胖胖的。

*tā fēicháng pàngpàng=de

3.SG very fat.IDEO?=LNK

“He is very fat.”

As the examples in (44) show, the reduplicated forms resist comparative constructions (44b) and negations (44e). As for degree adverbs, some are accepted (e.g. *zhème* in 44c), while others are not (e.g. *fēicháng* in 44f). While we have adapted the glosses from Paul (2006) by marking the ideo-phone in relevant sentences, it is better to treat these constructions as IDEOPHONIZATIONS rather than pure ideophones. The only ‘real’ ideophone in this set of examples (44) is *téngténg* in (44c). Furthermore, it is important to point out that these constructions thus do occur in the formal linguistic literature, yet are hardly ever recognized as ideophones, and reluctantly as onomatopoeia. Summarizing, depiction in Chinese ideophones is expressed through concomitant gesture, through the writing system, and in grammar.

4.2.4 The meanings of ideophones pertain to sensory imagery

Fourth, what ideophones depict is SENSORY IMAGERY. Yet, what is sensory imagery? According to Langacker (1987a:110–113), there are three kinds of sensory imagery. First, there is his own usage of this term, later renamed CONCEPTUALISATION (Langacker 2008b:44), i.e. “our ability to construe a conceived situation in alternate ways – by means of alternate images – for purposes of thought or expression” (Langacker 1987a:110). As an example, he shows how the cognitive ability of sensory imagery / construal allows us to

contemplate the different configurations in (45).



(45) Langacker (1987:110)

- a. The clock is on the table.
- b. The clock is lying on the table.
- c. The clock is resting on the table.
- d. The table is supporting the clock.

Another way the term ‘sensory imagery’ is used, according to Langacker (1987a), is as an equivalent of METAPHOR or FIGURATIVE LANGUAGE. While he deems this also an important cognitive ability, he does not consider this the basic usage of ‘sensory imagery’ in his Cognitive Grammar. And it seems Dingemanse (2011b) did not intend the meaning of the term in this way either.

What then is sensory imagery? Langacker (1987a) distinguishes a third usage, used mostly in cognitive psychology, which is closest related to the senses. As he states: “If I close my eyes, I can nevertheless evoke a kind of visual sensation by imagining or visualizing a scene. Similarly, I can evoke a kind of auditory sensation even when surrounded by total silence, for instance by imagining the sound of a barking dog or recalling a certain passage from the performance of a symphony” (Langacker 1987a:110). This usage is perhaps closest to the way it has been used in previous ideophone research. As Dingemanse points out:

That ideophones evoke sensory imagery has been recognised commonly and from early on in ideophone research. For exam-

ple, Koelle (1854:283) notes that ‘they are eminently expressions of feelings (German, Gefühlsworte’; Westermann (1907:129) describes them as ‘means to recreate perceptions in sound’; Fortune (1962:5) notes that they refer to ‘colour, taste, smell, texture, postures, gaits, activities, and conditions of every kind’; Noss (1986:243) states that they ‘denote what is felt or what is observed through the senses’; Kita (1997:381) notes that ‘they can refer to perceptual events in different sensory modalities’; and Nuckolls (1995:146) observes that they communicate ‘salient sounds, rhythms, visual images, and psychophysical sensations that are drawn from perceptions of the environment and bodily experience’.

Dingemanse (2011a:28–29)

From the previous quote it is clear that the senses covered by SENSORY IMAGERY range from the classical folk model (vision, hearing, touch, taste and smell) to more modern scientific taxonomies that include intero-receptors and proprio-receptors (Dingemanse 2011a:29). That is to say, INNER FEELINGS and COGNITIVE STATES may also feature in the abstracted sensory domains that ideophones can depict. It is even stipulated, from a cross-linguistic perspective, that there might be a certain hierarchy for ideophones (Dingemanse 2012:663), shown in (46).

(46) SOUND < MOVEMENT < VISUAL PATTERNS < OTHER SENSORY PERCEPTIONS <
INNER FEELINGS AND COGNITIVE STATES

Dingemanse's hierarchy is implicational. That means that languages that have ideophones expressing e.g. OTHER SENSORY PERCEPTIONS will also have those in the semantic domains on the left, i.e. SOUND, MOVEMENT, and VISUAL PATTERNS, but not necessarily those on the right, i.e. INNER FEELINGS AND COGNITIVE STATES.

I have argued before (Van Hoey 2016b) that this hierarchy is presumably better conceived of as a semantic map, that splits the broad sensory domains into smaller domains, yet retains some conceptual grouping. Most importantly, I have argued that in the analysis of Chinese ideophones, it makes sense to also include TIME and EVALUATION into the classification. Figure 4.5 is the current proposal for the etic grid of such a cross-linguistic map. Figure 4.6 shows how Old and Middle Chinese would fit onto this map, based on data in Van Hoey (2015) and Van Hoey (2016a).

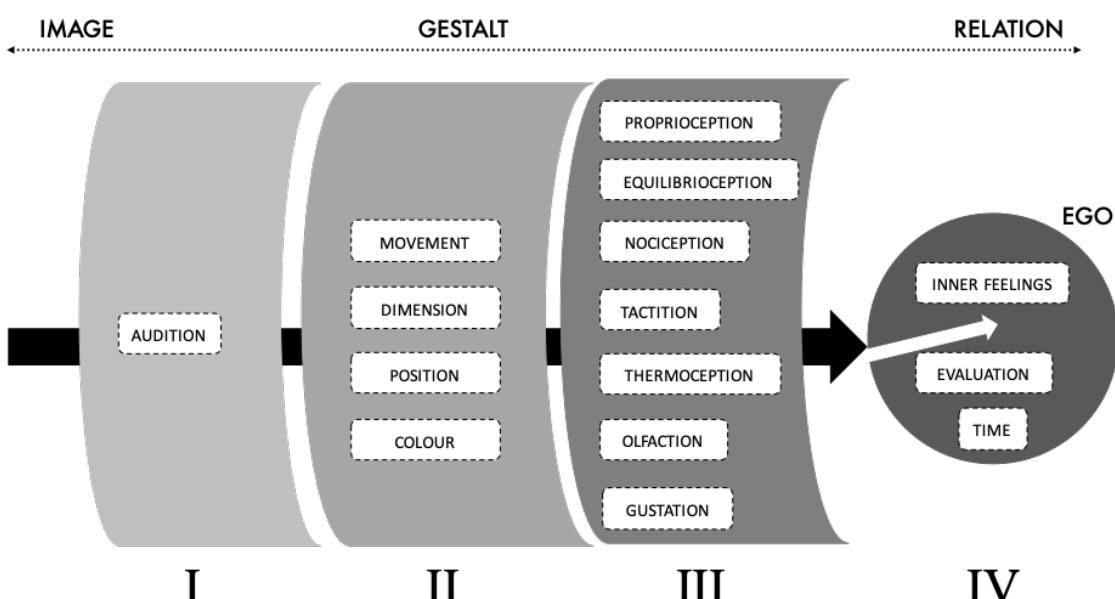


Figure 4.5: Etic grid for the sensory imagery cross-linguistically depicted by ideophones

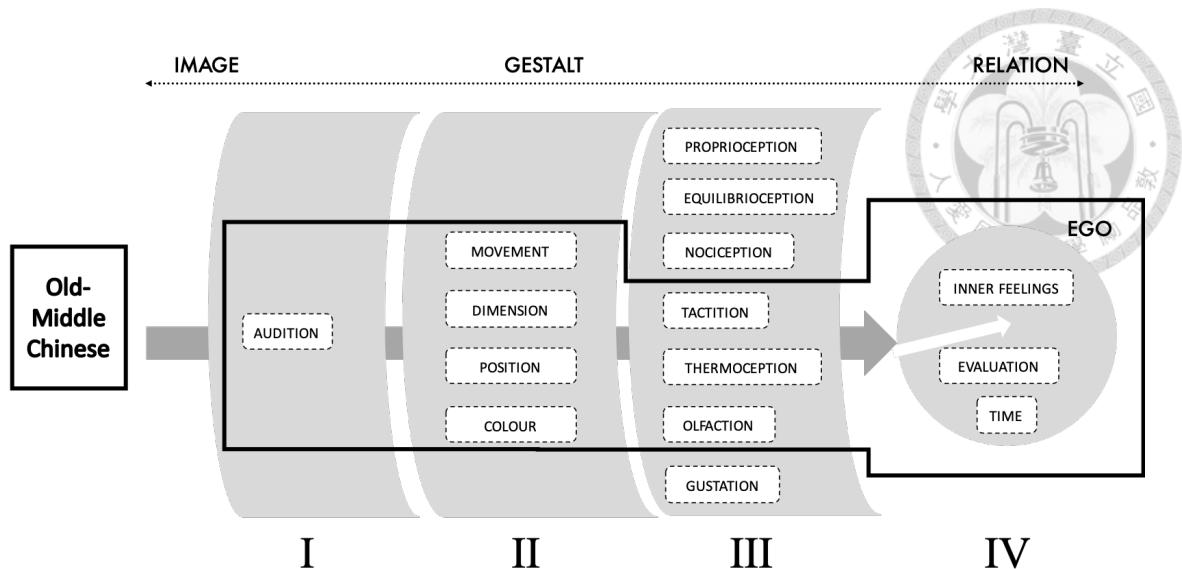


Figure 4.6: Semantic map for Old and Middle Chinese ideophones

However, related to this are new thorough explorations of language-internal classifications related to the clustering of certain sensory domains (Nuckolls 2019) or the ‘cut-off’ point for the detectability of iconicity (McLean 2019). These studies act as a reminder that typological generalizations can only claim so much; language-particular synesthesia for instance cannot be readily addressed currently.

Let us return to the criterion of sensory imagery for the cross-linguistic concept of ideophones. It is clear that ideophones do not exclusively own this criterion – in the same way they did not have the monopoly on the other criteria. In fact, research into sensory imagery and the related sensory vocabulary has been on the rise in recent years, with a current culmination in Winter (2019). He uses sets of vocabulary that have been rated by a number of people for their sensoriality, i.e. the degree to which a given word in the data set evokes a certain sense. The data sets, or ‘modality exclusivity norms’ were first proposed by Lynott & Conell (2009) and since then a

number of follow-up studies have been performed (Lynott & Connell 2013; Connell, Lynott & Banks 2018; Strik Lievers & Winter 2018), including a first adaption to Mandarin Chinese (Chen et al. 2019). An example is provided in Figure 4.2, which shows how *yellow* has a modal exclusivity of 95.1%, notably to VISUAL, while *harsh* has a very low modal exclusivity of 11.6%. This is not surprising because expressions like *harsh words*, *harsh sound*, *harsh chemicals*, *harsh winter*, *harsh colour* etc. readily come to mind, and these evoke all kinds of different senses.

Table 4.2: Modality norms for *yellow* and *harsh* (Adapted from Winter 2019:143)

word	VISUAL	TACTILE	AUDITORY	GUSTATORY	OLFACTORY	Exclusivity
<i>yellow</i>	4.9	0.0	0.2	0.1	0.1	95.1%
<i>harsh</i>	3.2	2.5	3.3	2.3	1.8	11.6%

In his “manifesto for norms”, Winter (2019) 132-135 lists some of the key-advantages of using these rated data sets as a way of exploring sensory vocabulary:

- (a) They are collected from hundreds of individuals, thus avoiding the pitfalls of complete subjective interpretation by the analyst.
- (b) They are collected from native language users and thus are reduced in their bias concerning a given linguistic framework.
- (c) They provide an alternative to conceptualization research, next to the traditional study of linguistic patterns by themselves.
- (d) They are collected before the analysis, and thus cannot be influenced

during the actual analysis part.

- (e) The collection process requires that the researchers operationalize their concepts and translate these into clear instructions.
- (f) The data can possibly be transformed into continuous data (rather than categorical data), enabling shades of meaning rather than a pure black-white dichotomy.
- (g) If the data is continuous, other statistical analyses are possible.
- (h) These data sets of modal exclusivity norms can be shared with the research community, improving further analyses.

However, Winter does admit that there are some flaws with the methodology, e.g., the wrong interpretation of homonyms such as *firm*, where the meaning ‘company’ can be confused with ‘sturdy feeling’ (Winter 2019:146). Thus, if used well, the benefits of ratings certainly outweigh the costs. These research benefits include among others a reanalysis of the classical studies on ‘linguistic synesthesia’ (cf. Ullmann 1957; Williams 1976) in terms of exact calculations of the attraction between the five classical senses in noun-verb constructions in English. This is the meso-level, but Winter also analyses the norms on a macro-level, where it is shown that there are actually only three clusters: VISUAL-TACTILE, OLFACTORY-GUSTATORY, and AUDITORY. However, a micro-analysis is also possible. Winter identifies 12 clusters based on the different ‘sensory profiles’, shown in 47.



(47) Winter (2019:166–173)

a.	PURE SIGHT	<i>gray, red, brunette</i>
b.	SHAPE AND EXTENT	<i>triangular, conical, curved</i>
c.	GROSS SURFACE PROPERTIES	<i>crinkled, prickly, sharp</i>
d.	MOTION, TOUCH AND GRAVITY	<i>ticklish, low, branching</i>
e.	SKIN AND TEMPERATURE	<i>lukewarm, cool, chilly</i>
f.	CHEMICAL SENSES	<i>bitter, sour, salty</i>
g.	TASTE	<i>cheesy, sweet, meaty</i>
h.	SMELL	<i>smelly, stinky, scented</i>
i.	SOUND 1	<i>noisy, deafening, bleeping</i>
j.	SOUND 2	<i>squeaking, booming, buzzing</i>
k.	IMPRESSION-RELATED	<i>radiant, misty, mellow</i>
l.	MULTISENSORY	<i>beautiful, clean, strong</i>

The consequences of these three-level analyses (Winter 2019) is that we are reminded that previously assumed ‘normal’ words are actually more multisensorial and often do have preferences for a given sense (cf. especially Strik Lievers & Winter 2018), but also that the findings do largely still depend on the sense categories that the researchers posit. Nevertheless, there seems to be somewhat of a consensus between different language-particular studies as to the different inventories of the senses that ideophones can contain. For example, Diatka’s (2014) study of Hindi ideophones, Van Hoey’s (2015) semantic hierarchy of Middle Chinese, Nuckolls’s (2016b; 2019) semantic clustering of senses in Pastaza Quichua, and McLean’s (2019) semantic map for Japanese resemble the senses in the hierarchy posited by Dingemanse (2012).



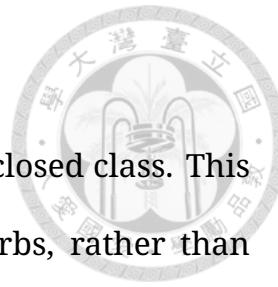
4.2.5 Ideophones belong to an open lexical class

Ideophones belong to an open lexical class, rather than a closed class. This puts them on a par with nouns, verbs, adjectives, adverbs, rather than prepositions, articles or determiners, conjunctions etc. This criterion is the most important addition in Dingemanse's revised definition contributes (2019:15–16). For him, the evidence comes from the sizes that are occasionally reported for ideophone inventories. Previously, he reported the magnitude of some well-documented languages (2018:15), see Table 4.3.

Table 4.3: Reported magnitude of some well-documented ideophone inventories (Dingemanse 2018:15)

Language	Reported magnitude of ideophone inventory
Basque	“more than 4,500” (Ibarretxe-Antuñano 2006:150)
Gbeya	“over 3,000” (Samarin 1971:161)
Japanese	“4,500” (Ono 2007)
Korean	“several thousands” (Sohn 2001:96)
Semai	“same order of magnitude” as nouns and verbs (Diffloth 1976:249)
Turkish	“one to two thousand” (Jendraschek 2002:39)
Zulu	“3,000” (von Staden 1977:200)

In many African languages ideophones are treated as a major word class (Kulemeka 1995), similar to the status they have acquired in Japanese and Korean linguistics. In Chinese grammars, ideophones have thus far not ac-



quired a status of part-of-speech³⁷, and whether that is a goal we should be aiming for is a question that remains without answer. However, the fact does remain that there is a large set of words that are sufficiently similar in terms of the previous four criteria, so it can definitely be argued that this also concerns an open class of lexical items. This in fact is one of the main motivations of a database like CHIDEOD (Section 3.2).

Apart from inventory size, Dingemanse sees processes of so-called IDEOPHONISATION (ideophone creation) (cf. Westermann 1937; Kunene 1965; Dingemanse 2014) and DEIDEOPHONISATION (loss of ideophones, cf. Flaksman 2017) as an indication that the class is open, rather than closed. For instance, the examples in (44d-44f) show how a PROSAIC WORD can become marked. Would one call it an ideophone? That is not easy to answer, but the way it is used suggests that maybe *pàng~pàng(de)* 胖胖的 is a quasi-ideophone or quasi-mimetic (see Section 4.1), although in Chapter 7 I will treat such occurrences as examples of ideophones. Another example is (48), where the reduplication of *màn* 慢 ‘slow’ indicates a special meaning. However, the expression in (48) as a whole has achieved unit status, i.e., it is processed as one cognitive chunk. As a result, it is hard to maintain its depiction or markedness.

- (48) 慢慢 來
màn~màn lái
slowly.RED.IDEO? come
“Take it easy, take your time.”

³⁷as far as I am aware of, see also Wu (2014).

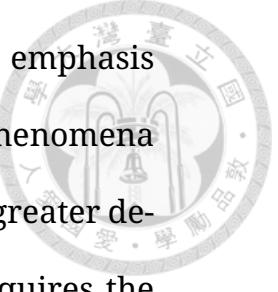
To summarize, these five criteria of the new definition for the cross-linguistic concept of ideophones (29) appear to befit Chinese ideophones. But this does not mean that ideophones are a homogeneous category in Chinese, as we have also seen they were not for Japanese. In fact, one of the fundamental points of this chapter is the claim that ideophones are also prototypically structured in Chinese. In the next sections we will first follow Dingemanse's idea that the five criteria can follow the tenets of a Canonical Typology, and then address and logically categorize a few groups of words that come close to the canonical ideal of ideophones, yet are not. After this, we will analyze the data available to us using statistical methods.

4.2.6 Dingemanse's (2019) criteria in a Canonical Typological framework

For Dingemanse, one of the main goals of treating IDEOPHONES as a comparative concept is arriving at better typological generalizations. This requires, however, that everybody is using the same terms in the same way. He mentions that the criteria he posited are well-suited for this – in a way that resembles Canonical Typology (Dingemanse 2019:20).

But what exactly is Canonical Typology? This framework, originally developed by Corbett (2007; 2011; 2015), takes as its main thesis that it is hard to arrive at typologically valid conclusions if the categories two researchers are working with share the same label, but a different meaning. In the words of Brown, Chumakina & Corbett (2013):

Canonical Typology seeks to avoid the tendency to use linguistic



terms with vague and shifting definitions by placing emphasis on the criteria used to associate particular linguistic phenomena with cross-linguistic categories. It therefore demands greater detail and rigour in terms of description, because it requires the typologist to be clear about the basis on which a phenomenon might be considered an instance of a particular concept.

Brown, Chumakina & Corbett (2013:3)

The goal, then, of Canonical Typology is to describe a set of criteria that are all available in the logically hypothesized canonical ideal – which may not exist. But this is not the point of this framework; what is important is that typological researchers are made aware of the features and criteria they implicitly assumed for a given linguistic phenomenon and explicitly state how two languages differ in the treatment of that category. This is perhaps best understood through an example, e.g. CASE (Corbett 2008; cf. Forker 2016). Corbett (2008) lists ten criteria and the resulting overlap between these makes it possible to attribute a canonicity measure to different cases in a given language or across languages. Forker (2016) summarizes:

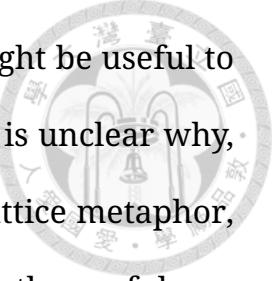
Corbett discusses the Russian case system as an example. According to criteria 1 and 2, the Russian instrumental case is far more canonical than the Russian accusative because it has less syncretism. Criteria 3 and 4 assert that canonical features and their values are distinguished consistently across relevant word classes and across lexemes within relevant word classes. In this respect, Russian cases are canonical because they are expressed

on all relevant word classes such as nouns, pronouns, adjectives, etc. Criterion 5 says that the use of canonical morphosyntactic features and their values is obligatory and Russian cases are also canonical in this respect. With respect to the Russian case system, Corbett concludes that the six traditionally assumed cases are relatively canonical with, e.g., the instrumental being more canonical than the accusative. Other “cases” such as vocative and second genitive are less canonical and thus often not assumed to be independent cases.

Forker (2016:78)

The idea of Canonical Typology is certainly intriguing, and applications have been made to e.g. a phenomenon that is often discussed in relation to ideophones: phonesthemes (see also Section 5.1.1). Kwon & Round (2015) have shown what a canonical phonestheme would look like. Later, Kwon (2017) applied Canonical Typology to the language-particular category of total reduplication in Japanese, again arriving at satisfactory conclusions in the framework of Canonical Typology.

Yet, Canonical Typology is not without criticism. The main problems that were identified by Forker (2016) include that (a) it is hard to identify what exactly is a canonical criterion or feature and what is not; (b) it is hard to discover new phenomena that cannot be conceived because they haven't been encountered yet; (c) the bond between real language data and linguistic concepts is weakened considerably; (d) frequency of instantiation is virtually neglected. For us too, this last issue is the most serious. Defining cri-



teria as a philosophical exercise is a noble venture, and might be useful to cover one's bases when doing typological theorizing, but it is unclear why, when the criteria have generated their groupings (in the lattice metaphor, see Figure 4.7), they are not quantified, so we can evaluate the usefulness of Canonical Typology against the better understood and received notion of prototypicality.

That being said, Dingemanse's suggestion that this theory may be useful for defining the scope of ideophones is a positive evolution in our understanding of the nature of this language phenomenon. Especially because he reflects on these five criteria of his definition as follows:

Together they generate a multidimensional space in which we can locate ideophone and ideophone-like phenomena within and across languages. So a given linguistic resource can be more or less class-like, structurally marked, word-like, depictive, or sensory in meaning, and the further it deviates on these dimensions from the canonical prototype, the less reason there is to identify it with the comparative concept of ideophones. There is broad agreement that Japanese, Basque, Quechua, Semai and Siwu are good examples of languages with open lexical classes of marked words that depict sensory imagery, i.e., ideophones. But what about items that do not clearly form coherent lexical classes, or languages realised in different modalities?

Dingemanse (2019:20)

In particular, the items Dingemanse (2019:20–27) envisions as not

clearly forming a coherent lexical class include (a) the status of phones-themes (in English), which he argues not to belong to an open class; (b) structurally marked words in Jahai (Burenhult & Majid 2011) and Maniq (Wnuk & Majid 2014) that unlike ‘real’ canonical ideophones in Semai (Diffloth 1976) “‘can be negated, relativized, and nominalized’ (Burenhult & Majid 2011:25–26): all properties not normally connected to ideophones” (Dingemanse 2019:24); (c) the (im)possibility of ideophones in sign languages – since they already use depiction as part of their ‘normal’ usage, it is hard to differentiate how this criterion works out in the prosaic vs. the mimetic lexicon. Furthermore, Dingemanse finds the inventory too small in American Sign Language to really count as fulfilling the open lexical class criterion. A summary of the issues discussed in this section can be visualized in Figure 4.7.

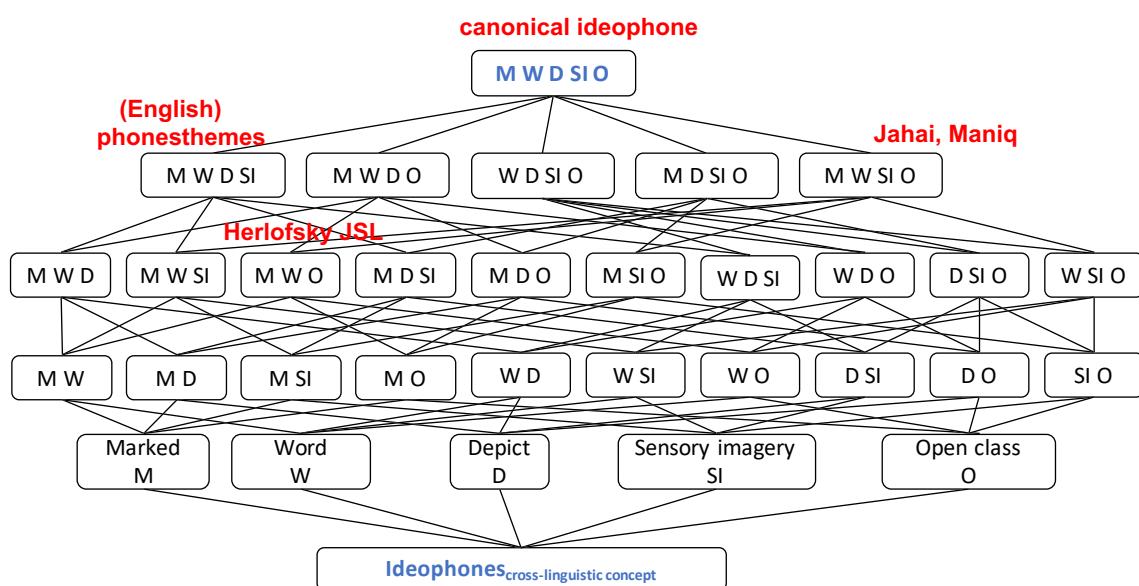


Figure 4.7: Lattice representation of the Canonical Typological criteria in Dingemanse (2019)'s definition. M = marked, W = words, D = depiction, SI = sensory imagery, O = open class; JSL = Japanese Sign Language

Thus, it is certainly possible to see the advantage of treating a cross-

linguistic definition as an assembly of canonical criteria. It certainly allows for a more mindful categorization of related phenomena and avoids over-claiming. For example, it has been argued that Japanese Sign Language also possesses ideophones because marked words (signs) of an open lexical class were identified (Herlofsky 2019). However, these three criteria do not appear to capture ‘real ideophonicity’, i.e., what we treat as the canonical ideophone. As a consequence, this phenomenon of Japanese Sign Language may be allotted to the second row of Figure 4.7: overlapping with the canonical ideophone, but deviating enough from it to not be accepted as an instance of the same phenomenon either. This shows that ideophones, as a cross-linguistic concept, do need to fulfill some formal and semantic criteria, and if these are presented in a Canonical Typological framework, then that is a step forward. What, on the other hand, the status of Chinese ideophones is with regard to these cross-linguistic criteria is a question that (currently) falls outside of the scope of this dissertation.

4.2.7 Non-canonical Chinese ideophones

In the preceding sections, we have given a case study of how the ideophone category may be defined and analyzed in Japanese, and provided reasons for logically in- and excluding certain sets of words in the category of ideophones. In what follows, we present a short overview of the kinds of words that were left out or at a later point taken out of the Chinese Ideophone Database (CHIDEOD).

The first group concerns REFERENTIAL REDUPLICATIONS, like the distribu-

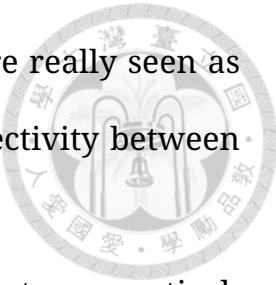
tive reduplication (49a) or lexicalised reduplications (49b). These are not qualificative, a criterion that Dingemanse seems to relate to depiction. There is, however, a group of mostly plants and animals (in Classical Chinese) that traditionally are categorized as binome (49c), but which may actually be derived through (regressive) reduplication and metonymy (Sun 1999:54–55). Since these are referential, rather than qualificative, they are usually not treated as good instances of ideophones, nor discussed in depth. An English equivalent would be to in-/exclude *cuckoo*, the bird. That being said, the example Zhang (2016:66) gives (49d) in which a qualificative ideophone metonymically stands for a referential usage of ‘woman’ (AT-TRIBUTE FOR PERSON) would usually be included in an ideophone inventory and subsequent discussion.

(49) Referential reduplications

- a. *rìrì* 日日 ‘every day’, *rénrén* 人人 ‘everybody’
- b. *gǒugǒu* 狗狗 ‘dog, doggy (pet)’, *bàbà* 爸爸 ‘dad, father’
- c. *fúyóu* 蝶蝣 ‘larval mayfly’ < MC *buw~juw* < OC **bèw~lèw*
- d. *yáotiáo* 窺窕 ‘(of a woman) gentle and graceful > beautiful woman’

The second group concerns words that do have the right semantics – sensory imagery – (50a), but they are not marked in any special way, such as adjectives (Dixon & Aikhenvald 2004).

The third group that one would not include in an ideophone inventory are repetitions (cf. Gómez 2009; Dingemanse 2015), like those presented in



(50b) – which may be very common in daily usage, but are really seen as pragmatic repetitions that somewhat soften the intersubjectivity between two interlocutors.

Fourth are quasi-mimetics like (50c) which have somewhat grammaticalized (cf. discussion above), or those that are totally derivable “from existing compounds or some kind of free combination of noun (N), verb (V), and adjective (A)” (Mok 2001:7–8). A Cantonese example from Mok is presented in (50d, repeated from 23). What would be included in an ideophone inventory, however, is this *pan2 pan2* part, especially if there is or used to be some other collocations to which this *pan2 pan2* would contribute a vivid meaning.

Lastly, there is a certain appeal to words that rhyme or alliterate – the two traditional categories of *shuāngshēng* 雙聲 ‘alliteration’ and *diéyùn* 疊韻 ‘rhyme’ which are often cited in binome-related research (see Chapter 2) – to be marked. However, this rhyme or alliteration can of course be purely accidental (50e). So without the ‘right’ semantics or other criteria it is more prudent to argue for their exclusion from the ideophone inventory.

(50) Other words that were excluded from CHIDEOD

- a. *hóng* 紅 ‘red’, *dà* 大 ‘big’
- b. *duì duì duì* 對 對 對 ‘yes, yes, yes’, *lái lái lái* 來 來 來 ‘come, come, come!’
- c. *màn màn lái* 慢慢來 ‘come slowly > take it easy’
- d. *sei2 pan2 pan2* 死板板 ‘dead board board > stubborn’

- e. *qiántiān* 前天 ‘the day before yesterday’, *hēi hé(zì)* 黑盒子 ‘black box’



This section has attempted to summarize a number of reasons why certain sets of words would not fit in with the definition of an ideophone, using the criteria established by Dingemanse (2019), and following the logic of Canonical Typology. What then *is* included in an inventory of Chinese ideophones will be further explored in the next section.

4.3 Finding the prototype with Multiple Correspondence Analysis

Thus far, the hard boundary of the Chinese ideophonic lexicon has been explored. It has become clear that on the basis of the features of Dingemanse’s definition, a large number of words can be excluded. At the same time, it has been argued that these features follow a prototypical structuring for the canonical ideophone. As we saw for Japanese in Section 4.1, on the language-particular level, a prototypical conception of the structure of its ideophone inventory fits the data better than a traditional differentiation based on necessary and sufficient conditions. It is not surprising that we will claim that this observation is also valid for the Chinese ideophone lexicon. Based on the discussion of the different variables contained in CHIDEOD (Section 3.2), it is clear that the numbers differ significantly depending on which variable is being looked at. That is to say, the groupings of values of the variable #morphological_template differ considerably from those based on #sensory_imagery. This has the consequence that it becomes hard to



state unambiguously what the prototype of the Chinese ideophone lexicon is. Based on intuition, one would probably imagine that the prototype for Chinese ideophones is a fully reduplicated ('BB' template), or patterns like 'AB', 'XAA' (see Mok 2001 and the table of patterns identified, presented in Section 2.3.3), but this only takes a structural perspective. Is it SOUND then, as we have seen for Japanese above (Figure 4.3)? From this perspective of sensory imagery this certainly seems a good candidate, as onomatopoeia (SOUND ideophones) seem to hold an important place in ideophone inventories across languages. Logically, one would then expect that the prototype conforms to one of the aforementioned patterns and the depiction of SOUND. But what about other correlations? Perhaps NON-SOUND ideophones also highly correlate to the same patterns.

It is this issue that we aim to address in this section: how can we chart the structure of ideophones in such a manner we find the most salient correlations between variables? One possible approach is performing a statistical analysis that looks at all values of the different variables, calculates the strength of correlations, and reduces the problem of many variables with many values into smaller number of dimensions that have the highest predictive value.

An analogy is in order. Let us say that we want to learn about tea consumption (Lê, Josse & Husson 2008; Sanchez 2012). If a sample of 300 people fill out a questionnaire containing the following questions: (1) the kind of tea that is consumed, e.g., black, green, or flavored; (2) how the tea is had, e.g., alone, with milk, with lemon, or other; (3) the form tea is bought in, e.g.,

in tea bags, loose, both; (4) the addition of sugar, e.g., yes or no; (5) where it is bought from, e.g., supermarket, local shop, or both. In this case we have five dimensions for which it is hard to predict which values of the variables will co-occur. For example, is tea bought in the supermarket always had with lemon and sugar? So, we want to reduce the five variables to a lower number, ideally two dimensions, which explain as much of the variance of the data as possible. Such a reduction can be performed through a family of statistical methods containing Principled Component Analysis, Correspondence Analysis, Multiple Correspondence Analysis, Multiple Factor Analysis. Principled Component Analysis is most suited for numerical data, (Multiple) Correspondence Analysis for categorical data, and Multiple Factor Analysis for a combination of numerical and categorical data. In light of these differences, Correspondence Analysis (in the case of two variables and their interactions) and Multiple Correspondence Analysis (three or more variables) are the best approaches. After the technique is performed, we will end up with two main dimensions (with correlating eigenvalues, see below) that can be projected two-dimensionally. Subsequently, the distance between the different answers of the questionnaire informs us which values co-occur and gives insight in the habits concerning tea.

However, we are not dealing with tea; we want to know how the ideo-phonic lexicon is structured. Since our data is also qualitative, we will rely on Multiple Correspondence Analysis (MCA) and to a lesser degree on (simple) Correspondence Analysis (CA). For the statistical background, see Benzecri (1973; 1984) and Greenacre (2006; 2007). A helpful illustration of the

technique in the field of medicine can be found in Soares Costa et al. (2013).

From a Cognitive Linguistics point of view, we find extensive discussions on CA and MCA by Glynn (2014), who states that the methodology of Correspondence Analysis “is a multivariate exploratory space reduction technique for categorical data analysis” (Glynn 2014:443). In other words, it “is an exploratory tool that helps one find which usage-features co-occur with other usage-features, giving a map of their overall patterning. Assuming that one is adopting a cognitive or functional approach to language, these usage-patterns can be interpreted as grammatical description, operationalized in terms of relative frequency” (Glynn 2014:444). This chapter also provides a hands-on tutorial on how to perform a (M)CA analysis with R.

Other applications within the Cognitive Linguistics tapestry of approaches ensued. Levshina (2015), for instance, uses this technique also for visualizing and exploring the prototypical exemplars of Gipper’s (1959) well-known study of the difference between the two German lexemes *Sessel* ‘comfortable chair’ and *Stuhl* ‘chair’. Another study along that line would be an MCA interpretation of Labov’s (1973) study on the variation of the semantic category of CUP. Glynn (2015) uses MCA to add more empirical foundations to the Idealized Cognitive Model theory first advanced by Lakoff (1987). Later, he uses MCA to further delve out how these methods can be applied study prototypes of polysemous words (Glynn 2016). And lastly, Deshors (2017) uses CA to study the different usage of the English Progressive Verbs constructions in different world Englishes, such as the British, American, Indian, and Singaporean variants.

As for the methodology, Glynn (2014) warns us about four different pitfalls when using Correspondence Analysis: ‘fishing’, over-simplicity, over-complexity and data sparseness. **FISHING** is the (almost) arbitrary selection of factors to analyze through CA, hoping for a significant correlation between factors. **OVER-SIMPLICITY** refers to analyzing the correlation between two variables that would not need a complicated method like CA but could simply be explored through bar charts, histograms etc. **OVER-COMPLEXITY** is the opposite; namely, when too many factors are included in the analysis. As Glynn (2014:451) states: “For example, there is obviously no point analyzing, simultaneously, 22 factors, each with 16 features, even if one has thousands of examples. Without even considering the impossibility of accounting for the variation (inertia), in such a data set, the results would not be interpretable for the simple reason that the visualization of so many factors becomes impossible to decipher.” **DATA SPARSENESS**, finally, is about having at least a certain number of examples per possible feature value. The rule of thumb states ten examples or more. If this is not possible, one could leave out these examples from the analysis.

The last methodological point before the two case studies of the CHIDEOD database and the ASBC corpus concerns the representativeness of the resulting analysis. This is described in a unit of measure called **INERTIA**. As Glynn (2014) explains, it is calculated on observed and expected frequencies of co-occurrence. A high inertia means that more variation of the data is represented by the model. With two variables, in CA, the inertia is high when column and row profiles have large deviations from their

averages. However, in MCA, these scores are not normally interpretable, which is a major drawback for this form of the technique. Glynn (2014) thus uses Greenacre's (2006) correction to these scores.

To operationalize the MCA analyses of CHIDEOD and ASBC, we will make use of R (R Core Team 2019), the tidyverse set of packages (Wickham 2017), FactoMineR (Lê, Josse & Husson 2008), factoextra (Kassambra & Mundt 2017), and ca (Nenadic & Greenacre 2007).

4.4 Case study 1: Ideophones in CHIDEOD

Let us now first investigate the decontextualized data as it is represented in CHIDEOD. This will provide us with a preliminary idea of what the structure of ideophones looks like.

4.4.1 Data and feature selection

The first issue we need to address is the scope of the data. The assumption made in this case study is that IDEOPHONE as a category has remained stable over time, and that the whole database of CHIDEOD can thus be considered data.

The second issue that needs to be dealt with is the selection of features to include in the analysis. In this first case study, we argue that the interaction between form and meaning is the direction to focus on. For this we can turn to the folk model of Chinese (see Section 1), presented below in (51):

(51)

$$\frac{\text{phonological form}}{\text{orthographic form}} \mid \text{meaning}$$



This folk model (51) can be operationalized in CHIDEOD by looking at the variables of (a) #morphological_template for the structure of the form; (b) #radical_support for the written form; and (c) #sensory_imagery for the meaning.

In preparing this dataset for the analysis, we need to address Glynn's (2014) concerns regarding fishing, over-simplicity, over-complexity and data sparseness (cf. Section 4.3). FISHING does not seem to be a problem, since the selection of the three criteria is motivated and not arbitrary.

OVER-SIMPLICITY is also not an issue, because we are interested in the interaction between these multiple variables, and MCA is an adequate way of approaching this problem.

OVER-COMPLEXITY may pose a problem for the variable of #radical_support, given that there are just so many possible radicals which are recorded here. A first possible solution to this problem is to simplify it to a binary variable, which contains if #radical_support is “present” or “absent”³⁸. A second, arguably better, one involves the regrouping of radicals into ontological categories. Here we will provide a rough regrouping, since we are dealing with the issue of over-complexity. However, more sophisticated models in the future are suggested to make use of the advances in the Hantology research program (Chou 2005). This research program aims to investigate the structure and ontology of Chinese characters from a lexical semantics

³⁸This approach was used first, in the manuscript that was defended.

perspective, based on Pustejovsky's Generative Lexicon (Pustejovsky 1995), with the goal of implementation in computational approaches and in language teaching. Advances include the description of different sensory faculties (Hong & Huang 2013), four-hoofed mammals (Huang, Yang & Chen 2008), and most recently a study of native speaker perception of radicals (Yang et al. 2018). While future research then should study the influence of radicals in ideophones in this Hantology framework, as mentioned before, here we perform a rough recategorization. This recategorization involves, for instance, converging the following radicals into one value "body": PERSON 亼, FOOT 足, HAND 扌, GOING 去, WALKING 行³⁹; converging different variants of WATER 水 and 氵 into "water"; 灬 and 火 into "fire", etc. Table 4.4 shows the 30 highest type frequency items for the new 'radical' variable. A full overview of all recategorizations can be found in Appendix 3.

The last of the four issues raised by Glynn (2014), DATA SPARSENESS, is handled by combining the morphological patterns of RU ($n = 11$), RAN ($n = 314$), YAN ($n = 5$) and ER ($n = 5$) together into one value called COMP (for compositional; $n = 335$). Examples of these patterns were presented in Table 3.8, see Section 3.2.3.1.

There is also DATA SPARSENESS to be found within sensory imagery. The values "SMELL" ($n = 10$) and "TASTE" ($n = 2$) will be conflated into one variable SMELLTASTE ($n = 12$) to reduce these problems.

To summarize, in terms of type combinations, we now are looking

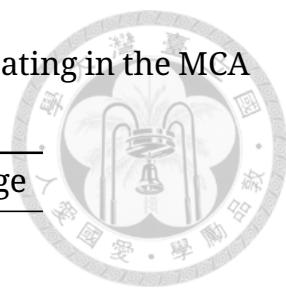
³⁹This is based on a suggestion from the committee. Note that this may not be the most ideal regrouping, as mentioned in relation to the Hantology program. For instance, EYE 目 or MOUTH 口 appear different enough from the other "body" items to be excluded from that category, although of course they are meronymic as opposed to the body in real life.



Table 4.4: Distribution of top radicals participating in the MCA of CHIDEOD

radical	radical_n	percentage	radical_support	radical_support_n
norad	2379	43.4%	NA	2379
mouth	926	16.9%	口	926
water	358	6.5%	氵	346
grass	123	2.2%	艸	123
heart	130	2.4%	忄	98
mountain	82	1.5%	山	82
body	278	5.1%	亼	80
body	278	5.1%	足	69
body	278	5.1%	扌	64
wood	61	1.1%	木	61
woman	51	0.9%	女	51
silk	46	0.8%	糸	46
jade	40	0.7%	王	40
body	278	5.1%	辵	40
metal	40	0.7%	金	40
fire	63	1.1%	火	38
stone	37	0.7%	石	37
speak	36	0.7%	言	36
sun	35	0.6%	日	35
heart	130	2.4%	心	32
moon	32	0.6%	月	32
eye	31	0.6%	目	31
feather	27	0.5%	羽	27
body	278	5.1%	彳	25
fire	63	1.1%	灬	25
otherrad	707	12.9%	虫	25
otherrad	707	12.9%	門	21
otherrad	707	12.9%	阝	21
otherrad	707	12.9%	車	20
otherrad	707	12.9%	雨	19

Table 4.5: Distribution of morphological templates participating in the MCA of CHIDEOD



morphological_template	count	percentage
BB	1726	31.5%
RR	1676	30.6%
RRRR	585	10.7%
BR	409	7.5%
COMP	335	6.1%
A	293	5.3%
ARR	212	3.9%
RB	117	2.1%
BBB	67	1.2%
RRR	27	0.5%
RRA	21	0.4%
BBBB	14	0.3%

at 5482 combinations of values concerning #morphological_template, #sensory_imagery, and #radical_support. The latter was already presented in Table 4.4; the values for #morphological_template are shown in Table 4.5, and those for #sensory_imagery are provided in Table 4.6. The values that will participate in the analysis can be found those three tables. However, it may be important to note that the ideophone *labels* – the actual items – are kept out of the analysis itself, because they are so-called **SUPPLEMENTARY VALUES**. They are later be projected on the distributions found through the calculated correlations between the variables that do participate in the analysis, the so-called **ACTIVE VARIABLES**.

4.4.2 The MCA of CHIDEOD

Let us now investigate the inertia of the dimensions. That is, we want to transform many variables into a smaller number of dimensions such that the highest amount of the variance within the data is accounted for. Thus

Table 4.6: Distribution of sensory imagery participating in the MCA of CHIDEOD



sensory_imagery	count	percentage
SOUND	2177	39.7%
VISUAL	1653	30.2%
EVALUATION	596	10.9%
MOVEMENT	589	10.7%
INNER_FEELINGS	343	6.3%
TEMPERATURE	44	0.8%
TIME	37	0.7%
TEXTURE	31	0.6%
SMELLTASTE	12	0.2%

the values of the active variables eventually can be grouped in a number of dimensions, that are ordered from high to low, according to their EIGENVALUES. Because the normal calculation of eigenvalues is deemed too conservative (Glynn 2014), we use Greenacre's adjusted scores (see Section 4.3). As Figure 4.8 shows, Dimension 1 has an adjusted inertia of 65.9% and Dimension 2 the value 17.7%. In other words, if we visualize the data in Dimesion 1 and Dimension 2 on the x-axis and y-axis, 83.6% of the variance within the data will be shown.

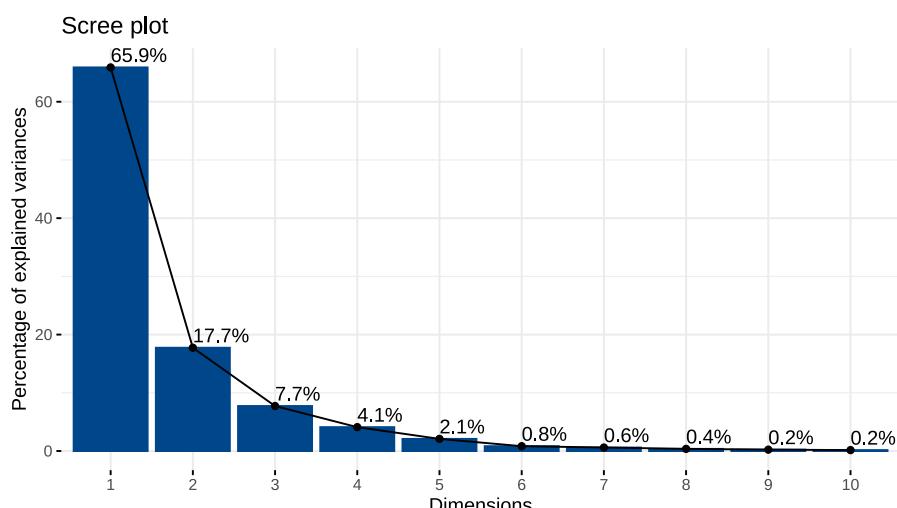


Figure 4.8: Eigenvalues for the MCA analysis

The first plot (Figure 4.9) shows the distribution of all data points. These clouds of points can best be thought of as each representing an ideophone in the database (with some overlapping). The barycenters (black) show where the mathematical center of each cloud of data points for each value is situated. For example, for the value ‘SMELLTASTE’ in Figure 4.9 represents the mathematical center for all values that had ‘SMELLTASTE’ as a value in the variable ‘sensory imagery’. A less cluttered version of the barycenters is provided in Figure 4.10.

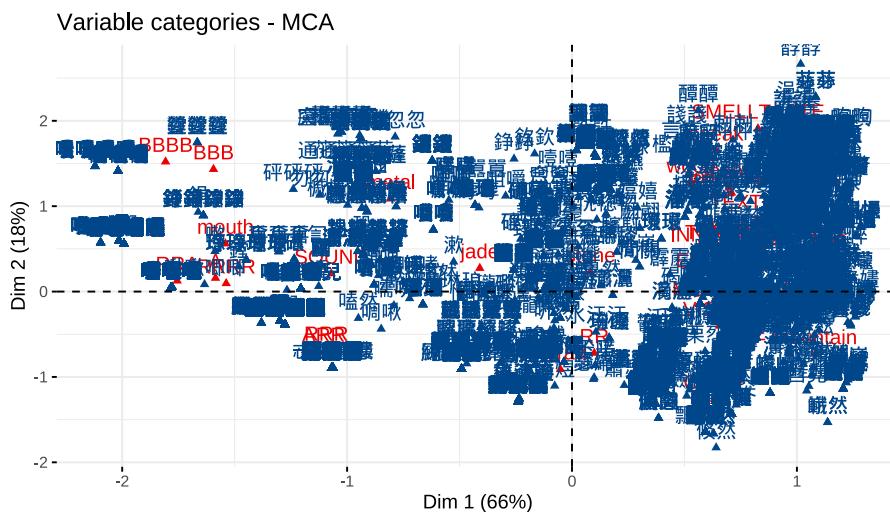


Figure 4.9: MCA plot of the barycenters of the CHIDEOD data with the supplementary values (the ideophones themselves) showing. A landscape version of this figure is provided in Appendix 4 (Figure 8.1).

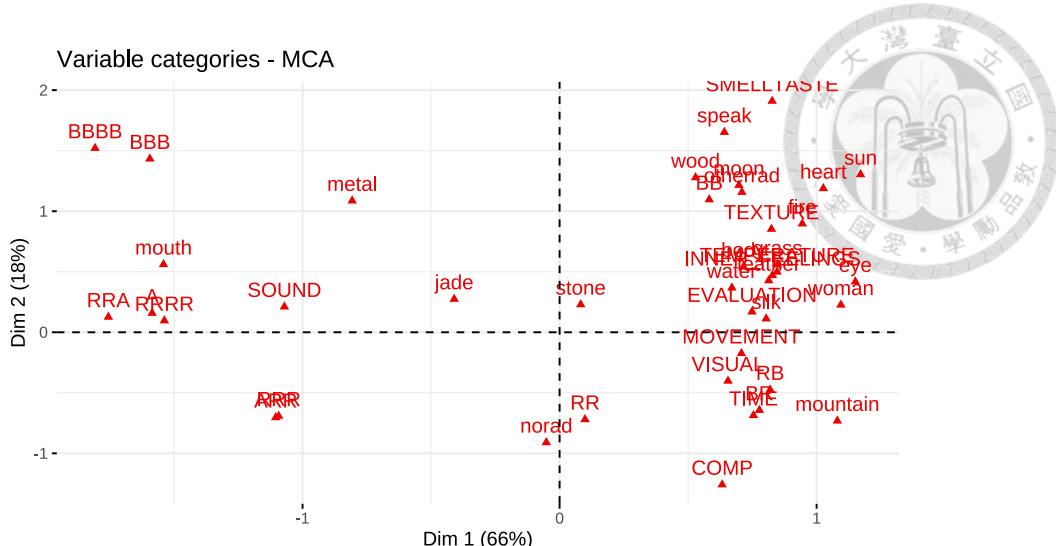


Figure 4.10: MCA plot of the barycenters of the CHIDEOD data

From Figure 4.10 it is clear that there are a number of constructions that cluster together with the sensory modality of SOUND, while the rest clusters together with the other modalities. Dimension 2 shows that there is a tendency for RR and COMP morphology values to not have radical support ('no'). But how different are these seemingly two clusters in Figure 4.10? It is possible to use ellipses to investigate the prototypicality of sensory modality. According to Levshina (2015:380), the 95% confidence ellipse around the barycenter shows the prototype of each category. We will go over the distribution of all barycenters for each variable. While the distribution of the barycenters remains the same for each plot, the clustering highlighted by the ellipses paints a different picture for each variable. In the case that two plots should share the same clustering, that would indicate that there is no difference between the two variables. If the two ellipses of a binary variable are projected on top of each other, it indicates that there is no difference between the two values of this binary variable.

Let us first inspect the variable SENSORY IMAGERY. By plotting the 95%

confidence ellipses around the centroids of this variable, as shown in Figure 4.11, it is clear that there are discrete prototypical centers, see Levshina (2015:380) for another demonstration of this method. Since this will be true for all subsequent analyses, we will not show them for the other variables.

Instead, we will focus on the fuzziness of the boundaries, which can be found by plotting the 95% confidence ellipses around all exemplars of the variable SENSORY IMAGERY. As shown in Figure 4.12, we can see that there is a light overlap between SOUND and the rest of the categories, which overlap quite heavily. This is to be interpreted that, in terms of the variable sensory imagery, there are two main clusters with fuzzy boundaries. While the prototypical centroids of these two groups may be discrete, the boundaries between them are not at all.

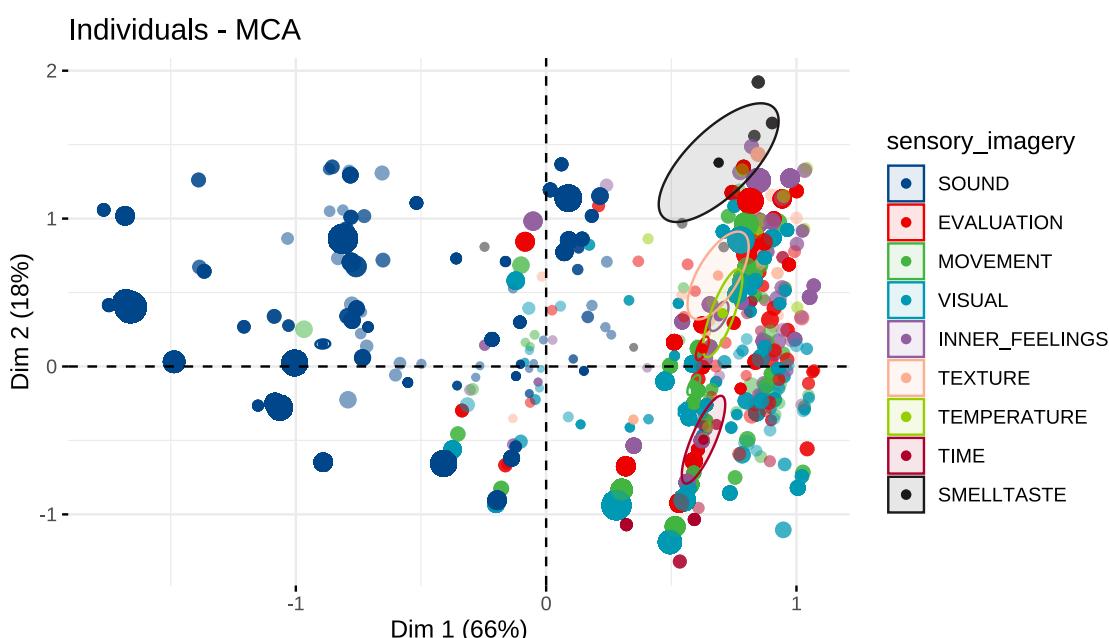


Figure 4.11: Confidence ellipses around the centroid of sensory modality

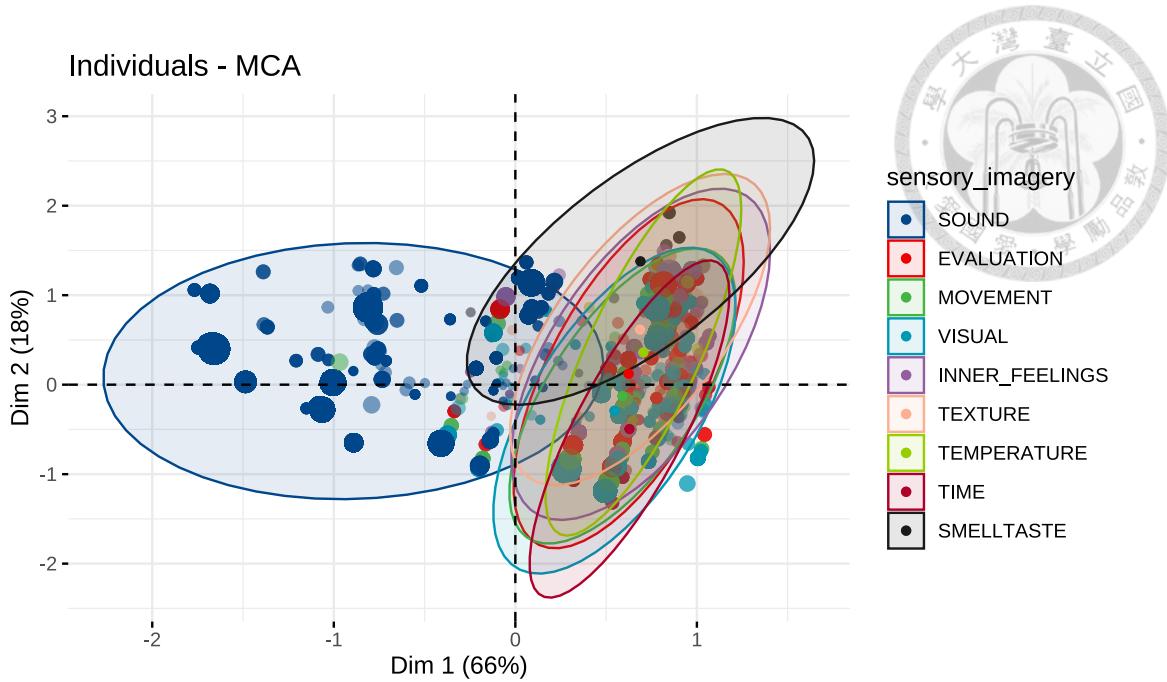


Figure 4.12: Confidence ellipse around the exemplars of sensory modality

However, when inspecting the MORPHOLOGICAL TEMPLATES, this dichotomy does not occur. Figure 4.13 clearly shows that it is not just a simple SOUND versus NON-SOUND that provides all of the information one needs to know to identify the prototypical ideophone.

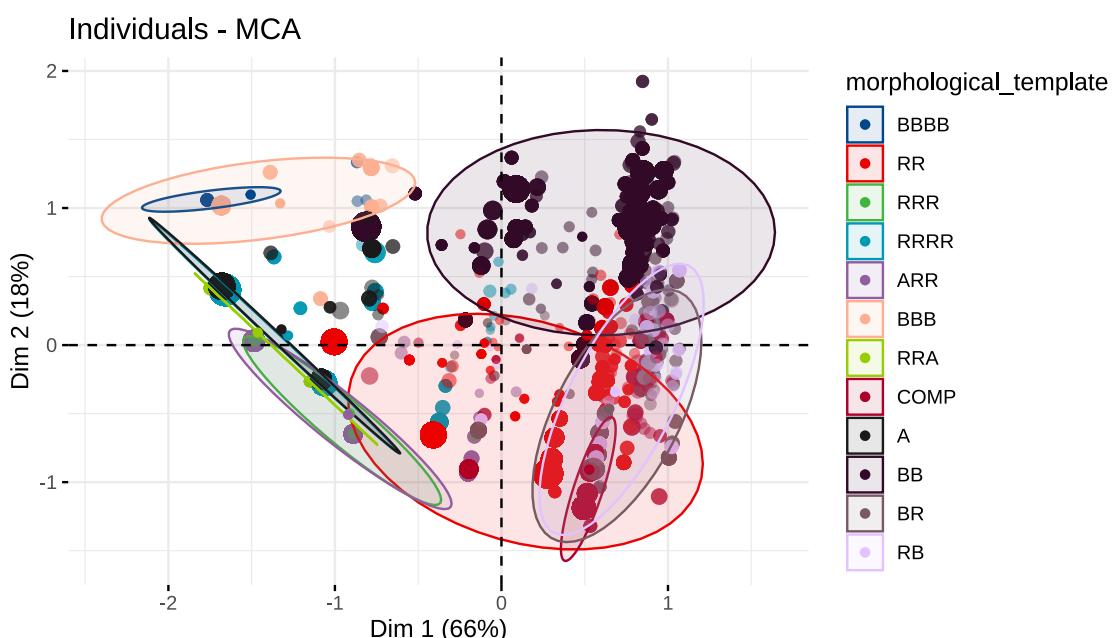


Figure 4.13: Confidence ellipse around the exemplars of morphological template

Adding Figure 4.14 on top of that, we can see that the RADICAL is yet another factor that confounds what might be the prototypical ideophone in Chinese.

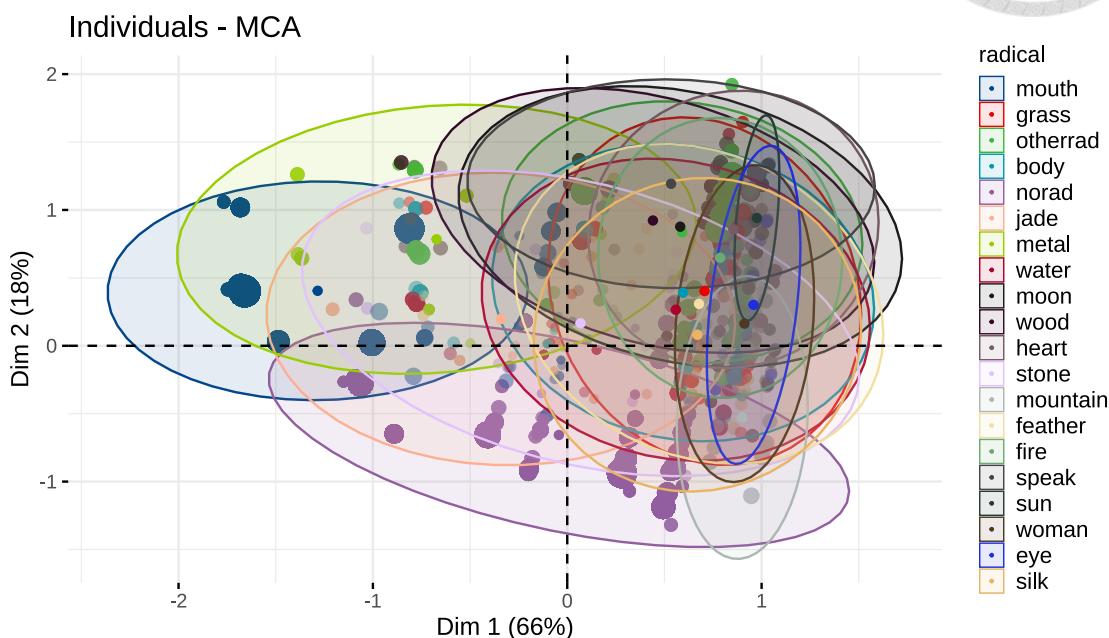
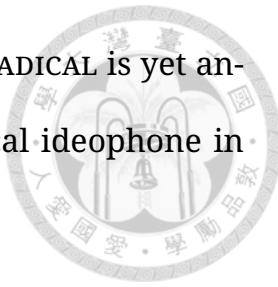


Figure 4.14: Confidence ellipse around the exemplars of semantic radicals

After the separate highlight of all of these variables, one can wonder how significant they are. In other words, how highly should we value the distributions found so far, and should we dismiss any of the three variables taken into consideration? There are two options to address this issue (Levshina 2015). The first is to reduce the number of dimensions, which will be shown below with a simple Correspondence Analysis (CA) of morphology and sensory modality. The other option is to perform a logistic regression analysis. Here we follow Levshina's (2015:383–385) demonstration for calculating the concordance index C between a variable and the two first dimensions (Dim 1 and Dim 2). As is shown in 52, the model, which contains the C value, it can be calculated by taking the logistic regression model ('lrm') from the

formula that takes the variable as the outcome and the sum of Dimensions 1 and 2 as the predictors.

(52) model = lrm(variable ~ dim1 + dim2)

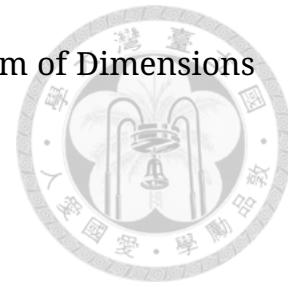
The C values, shown in Table 4.7, can be understood as follows. For “radical support binary ~ dim1 + dim2” (paraphrasing Levshina 2015:259): if you take all possible combinations of ‘yes’ and ‘no’ for this variable, the statistic C will be the proportion of the times when the model predicts a higher probability of ‘yes’ for the combinations with ‘yes’, and a higher probability of ‘no’ for the combinations with ‘no’. The C values found in this regression analysis all meet the threshold of acceptability⁴⁰ (Levshina 2015). It can be inferred that the first two dimensions produced by the MCA are relatively good predictors for the variables, and can be trusted.

Table 4.7: Regression analysis of the variables and the MCA model

Variable	C	Acceptability
morphology	0.72	acceptable
sensory modality	0.84	excellent
radical support binary	0.72	acceptable

It is worth considering if perhaps only the variables morphology and sensory modality are enough (the first option mentioned above). After all, the binary radical support factor is related to the Chinese writing system

⁴⁰Hosmer & Lemeshow (2000) state that a concordance index of $C = 0.5$ indicates ‘no discriminations’; $0.7 \leq C < 0.8$ ‘acceptable discrimination’; $0.8 \leq C < 0.9$ ‘excellent discrimination’; and $C \geq 0.9$ ‘outstanding discrimination’.



and perhaps one can argue that it should be left out in favor of morphology and sensory imagery⁴¹. Let us then perform a simple Correspondence Analysis (CA).

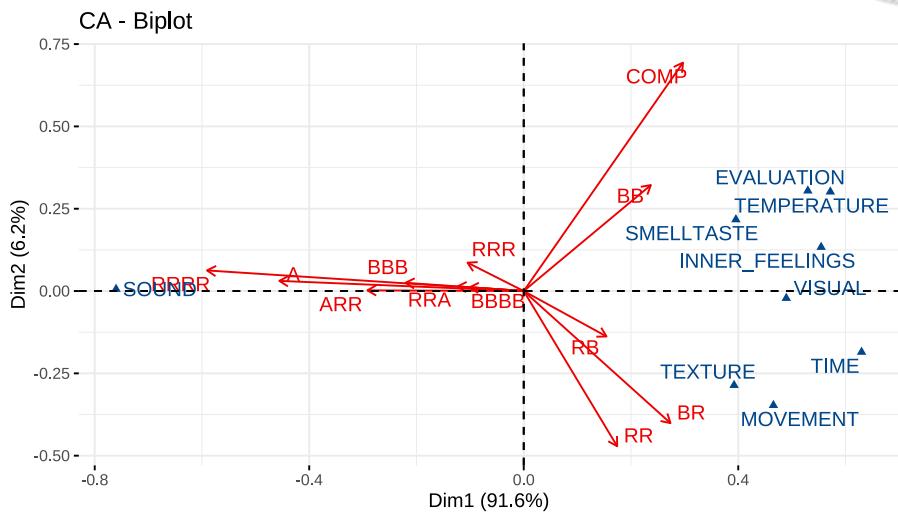


Figure 4.15: Biplot of the simple CA

In Figure 4.15 it can be seen that similar correlations still hold between SOUND and the previously mentioned constructions, as well as on the other side, the other modalities and those constructions. Furthermore, there is high inertia (90.6), meaning that the model is quite sturdy.

4.4.3 Interim summary

From the previous analyses, it is clear that a rather non-arbitrary relationship between the morphological constructions, sensory meanings and written forms of Chinese ideophones exists. As we have mentioned above, leaving out the radical support (motivation in the written form) results in even

⁴¹Van Hoey & Thompson (2019) also discussed this using a previous version of CHIDEOD. We showed that there was a tendency for morpho-phonological construction to be more important in its contribution to the ideophonicity of an item than whether or not the written form motivates an ideophone.

more valuable correspondences between values for morphology and sensory modality.

Some notable findings include that there are two prototypical cores – one around SOUND and one around the other modalities. This dyadic prototype structure of Chinese ideophones is, in hindsight, not entirely unexpected, yet was never really empirically shown before with statistics. This unexpectancy has to do with the type of iconicity that these sensory modalities provide: SOUND typically is a form of imagic iconicity, while the other modalities are more shifted towards diagrammatic iconicity (Dingemanse 2012).

This also provides a convincing reason why other modalities generally have been left out of the discussion in Chinese treatments – they were perhaps perceived as different enough from SOUND ideophones. Yet, as other treatments have shown, they are also similar enough to be treated through the same framework. This needs a resolution (in the future).

What is eventually left to be done is to diagram the results from the statistical explorations, in order to show the most important tendencies. This is shown in Figure 4.16. It is important to note that this just shows the tendencies, and not full one-on-one correlations – exactly the point we have been trying to make through the statistical methodology.

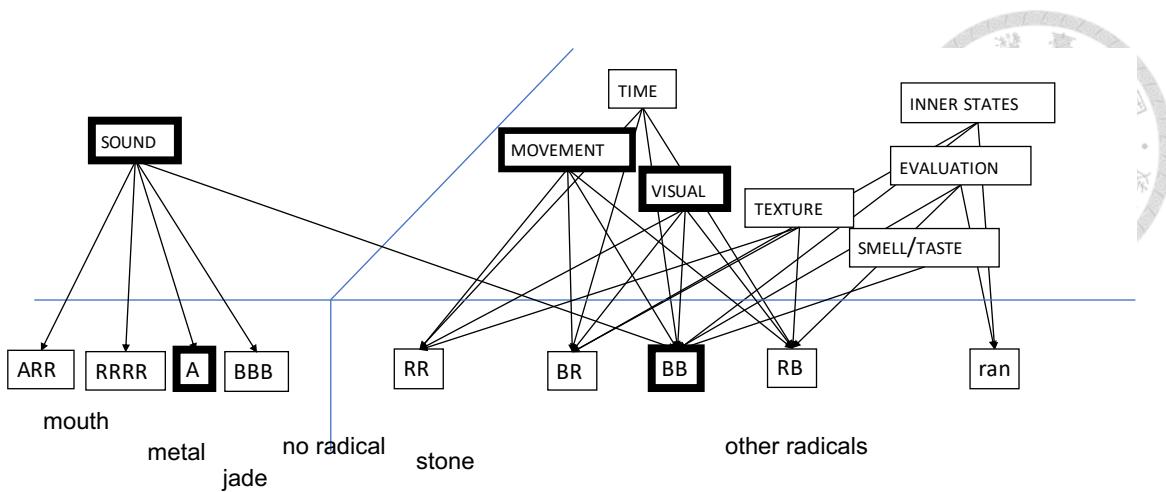


Figure 4.16: Diagram based on previous correspondence analyses

4.5 Case study 2: Analyzing the structure of ideophones in the ASBC

The case can be made that we should not ignore the variable of #language_stage, as done in the previous case study. After all, it is unlikely that ideophones used today are the same as ideophones in the past, even if they use the same orthography. They may differ in terms of grammatical constructions in which they appear, or the orthographic representations may have become deideophonized (Dingemanse 2017), or their semantics may have changed. Therefore, it is worthwhile to inspect more current usage, i.e. inspect tokens and types, in a balanced corpus of Mandarin Chinese.

4.5.1 Data and feature selection

As laid out in Section 3.3.3, the balanced corpus of choice is the Academia Sinica Balanced Corpus of Modern Chinese 4.0, referred to here as ASBC

or ASBC 4.0. The items that will participate in the MCA analysis consist of those items in CHIDEOD that occur in the ASBC. In other words, the corpus is reduced so as to contain only the sentences which possibly contain an ideophone. Based on the different tags (cf. Huang, Hsieh & Chen 2017) more values are taken out of the dataset, such as nouns (ASBC tags starting with ‘N’). Other checks implemented are making sure the ideophone is the same as the word found in the ASBC and not just a part of it, e.g., verbs (ASBC tags starting with ‘V’). Lastly, quantifiers and function words are taken out, as well as tags that occurred fewer than 50 times, to satisfy Glynn’s (2014) data sparseness criterion. This still leaves a sufficiently large enough dataset (38,981 observations) which may contain some noise which should be filtered out through MCA.

Now that we know which items to include in the analysis, it is time to focus on which variables. Given the balanced nature of the ASBC corpus, a host of other possibilities open up. So, apart from (a) the morphological template, (b) sensory imagery and (c) (binary) radical support variables, we can include (d) the ASBC tag allocated to the item; (e) the modality of an item (written vs. spoken); (f) what topical genre, or as it is called in ASBC, the class, it occurs in; and we can include information on the (g) frequency of an item.

Let us first have a look at the frequency. An obvious possibility would be to include the absolute token frequency as it occurs in the corpus. That this may not be the best possibility is argued by Gries and colleagues in a number of studies relating to collostructional analyses. For instance, when

Stefanowitsch & Gries (2003) investigated the English Imperative Construction, they found a number of verbs that were obvious candidates for high occurrence in this construction, e.g. *let*, *see*, *look*, *listen* etc. but they also found *process* and *fold* – two verbs one would not intuitively expect in this construction. They showed that the high occurrence of these verbs is due to an extremely high token frequency in just one file of their corpus. The frequency of occurrence was thus extremely skewed, or in their words UN-DERDISPERSED.

As a remedy for this, Gries (2008) first surveyed a number of statistical measures of DISPERSION that could be used to identify problematic cases as *process* and *fold* in a corpus. Finding himself not completely agreeing with any of the previous measures, he proposed his own, called Deviation of Proportions (DP). The formula for this measure is shown in (53). It takes the relative frequency of an item per corpus part (p_{1-i}) over all corpus parts, with s_{1-i} the percentages of corpus parts (Gries 2019a).

(53)

$$\frac{1}{2} \sum_{i=1}^n \left| \frac{p_i}{s_i} \right|$$

Or, following Levshina's (2015:82–85) demonstration, one first needs the total frequencies of each corpus part to calculate the proportions of each corpus part. These are the expected proportions. Then, to obtain the DP measure for each item, the absolute frequencies per corpus part and their proportions, called the observed proportions, are calculated. Next, the expected proportion is subtracted from the observed proportion. The absolute

value of the results are added to each other and multiplied by 0.5, in order to obtain the DP.

The resulting DP value for any item is a value between 0 and 1. A value close to 0 indicates that the item is evenly distributed given the size of the corpus parts. A DP close to 1 means that there is a strong preference for some corpus parts. In the example above, *process* and *fold* have a $DP > 0.995$ (Gries 2019a:394), indicating that their absolute frequency (or even relative frequency) needs to be corrected somehow.

Let us now return to the problem at hand, what kind of frequency should we include in the MCA analysis of ideophones found in the ASBC? By calculating the DP of each item and then correct that with the so-called NORMALIZED DEVIATION OF PROPORTIONS (DP_{norm}). This can be calculated by dividing DP by 1 minus the minimum value for an expected proportion s (Lijffijt & Gries 2012; Levshina 2015), shown in (54). In this case a value close to 1 is evenly dispersed and a value close to 0 is skewed towards one corpus part.

(54)

$$DP_{norm} = \frac{DP}{1 - min(s)}$$

Let us illustrate this with an example. In Table 4.8 the observed frequencies are shown per topical class (philosophy, literature, life, society, science, and art) for the item *wāng~wāng* 汪汪 ‘woof-woof’. The observed proportion and expected proportion are then provided. Based on these, the DP and DP_{norm} are then calculated. Notice how these measures are the same for the item across all topical classes. Lastly, the absolute frequency is cor-

Table 4.8: DP, DP_{norm} and corrected frequency for *wāng~wāng* 汪汪 ‘woof-woof’

ideophone	class	observed	observed_prop	expec_prop	DP	DPnorm	correcfreq
汪汪	哲學	2	0.095	0.100	0.515	0.557	11.691
汪汪	文學	15	0.714	0.200	0.515	0.557	11.691
汪汪	生活	3	0.143	0.200	0.515	0.557	11.691
汪汪	社會	1	0.048	0.323	0.515	0.557	11.691
汪汪	科學	0	0.000	0.101	0.515	0.557	11.691
汪汪	藝術	0	0.000	0.076	0.515	0.557	11.691

rected through a multiplication of the observed values and the normalized DP, and then added together. There is thus a penalty for heavily skewed ideophones and a reward for evenly distributed ideophones.

Now that we have the corrected frequency per item, we can turn these into a ‘frequency class’ factor. I have opted to count corrected frequencies higher than 200 as very high; higher than 100 as high, more than 50 as mid-high and everything below that as low. These quantitative values are converted to qualitative ones because it allows for the continued usage of the MCA. If numerical data were also to be included in the analysis, then a Multiple Factor Analysis should be performed.

To create the data set that contains all elements for the analysis, the following variables are extracted: (a) the morphological template, (b) sensory imagery and (c) (binary) radical support from CHIDEOD, and join these together with the items identified in ASBC. More specifically, we are keeping the item, and the features containing (d) the tag, (e) the modality of an item (written vs. spoken), (f) their topical class. Lastly the (g) corrected frequency is added to this group.

Now the data set is ready for a first inspection of the distribution of each variable, as was demonstrated in Section 4.4. Looking at (a) the MORPHO-

LOGICAL TEMPLATES (Table 4.9), BB, A and RR are the most common ones, accounting for 77.9% of the data. Since the frequency of the RRA pattern is 1, it does not meet the data sparseness criterion, and thus cannot participate in the MCA. The single item in question is *dīng~dīng~dōng* 叮叮咚. Consequently, it is dropped from the dataset.

For SENSORY IMAGERY, shown in Table 4.10 seems mostly in order, with SOUND and VISUAL ideophones leading the way, which is not that surprising. However, upon inspection it seemed that for SMELL the only type was *xīn~xīn* 欣欣, which may be a mis-categorization in CHIDEOD, at least in this case, see (55). For this reason, we drop it rather than merge it with TASTE, as we did above.

(55) ASBC (nº 105568)

還	有	茶樹	新芽	飽啜	朝露	的
hái	yǒu	chá-shù	xīnyá	bǎo-chuò	zhāo-lù	de
also	EXIST	tea-tree	sprout	full-suck	morning-dew	LNK
欣欣 春意。						
xīnxīn chūnyì						
delightful.IDEO springtime						

“There is also the delightful springtime of the tea tree sprouts full of the morning dew.”

The (c) BINARY RADICAL SUPPORT is better represented and nothing should be left out to avoid data sparseness or to reduce over-complexity. This is shown in 4.11.

The (d) TAGS IN THE ASBC may also provide a clue as to the prototypi-



Table 4.9: Distribution of morphological templates participating in the MCA of ASBC

template	count	percentage	examples	pinyin_tone
BB	15278	36.1%	濛濛	cóng~cóng
A	10138	24.0%	咚	dōng
RR	7632	18.0%	演漾	yǎn~yàng
RAN	4628	10.9%	茫然	máng~rán
BR	3199	7.6%	嘈啐	cáo~cuì
RB	835	2.0%	汗漫	hàn~màn
RRRR	251	0.6%	滴滴答答	dī~dī~dá~dá
RU	165	0.4%	恬如	tián~rú
ARR	131	0.3%	忒楞楞	tè~léng~léng
ARR	131	0.3%	忒楞楞	tēi~léng~léng
BBB	43	0.1%	嚓嚓嚓	cā~cā~cā
RRA	1	0.0%	當當丁	dāng~dāng~dīng

Table 4.10: Distribution of sensory imagery participating in the MCA of ASBC

sensory_imagery	count	percentage
SOUND	14638	34.6%
VISUAL	14082	33.3%
MOVEMENT	3996	9.4%
EVALUATION	3657	8.6%
INNER_FEELINGS	3542	8.4%
TIME	2008	4.7%
TEXTURE	239	0.6%
TASTE	76	0.2%
TEMPERATURE	61	0.1%
SMELL	2	0.0%



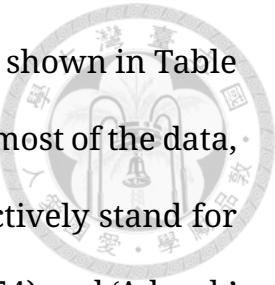
Table 4.11: Distribution of radical support (binary variable) participating in the MCA of ASBC

radical	count	percentage
norad	14320	33.9%
mouth	6711	15.9%
otherrad	4975	11.8%
body	4911	11.6%
water	3873	9.2%
heart	2244	5.3%
silk	1497	3.5%
woman	1113	2.6%
grass	745	1.8%
sun	438	1.0%
feather	380	0.9%
fire	292	0.7%
jade	167	0.4%
mountain	128	0.3%
moon	127	0.3%
wood	126	0.3%
eye	76	0.2%
stone	64	0.2%
speak	61	0.1%
metal	53	0.1%



Table 4.12: Distribution of ASBC tags participating (Huang, Hsieh & Chen 2017) in the MCA of ASBC

tag	count	percentage
VH	17671	41.8%
D	14041	33.2%
I	2731	6.5%
VC	2250	5.3%
VA	1517	3.6%
Da	758	1.8%
VE	610	1.4%
T	589	1.4%
Nv	347	0.8%
A	335	0.8%
VG	263	0.6%
VCL	256	0.6%
VK	244	0.6%
VB	230	0.5%
VJ	119	0.3%
VAC	116	0.3%
VI	110	0.3%
VHC	58	0.1%
FW	56	0.1%



cal nature of ideophones. A first distribution of the dataset, shown in Table 4.12, shows that the tags VH (42.0%) and D (33.1%) make up most of the data, and are illustrated in (56-57) and (58-59). These two respectively stand for ‘State Intransitive Verbs’ (Huang, Hsieh & Chen 2017:151–154) and ‘Adverb’ (Huang, Hsieh & Chen 2017:201–215), without any further subdivisions into what kind of adverb. It is not surprising that these two would make up the bulk of the data in the corpus, since ideophones are well-known to either function as a predicate or modify that predicate in a larger verbal complex (Nuckolls 2014) in other languages. Perhaps surprising is the relatively low frequency of I, ‘Interjections’ (Huang, Hsieh & Chen 2017:232) and T ‘particles’ (Huang, Hsieh & Chen 2017:230–231). Given that ideophones are often said to occur on the edge of the clause (Dingemanse & Akita 2016), this would be the tags that best fit them. On the other hand, as stated in the Section 3.2.3.3, there is a conceptual distinction between ideophones and interjections and it can be argued that these elements are better left out. Nevertheless, we will keep them in.

(56) ASBC (nº 100508)

路-上 有 黃-澄澄 的 路-燈，
lù-shàng yǒu huáng-chéng~chéng de lù-dēng
street-on EXIST yellow-bright.IDEO LNK street-light
“On the street there are bright yellow street lights.”



(57) ASBC (nº 100664)

一 個 人 就 文-質-彬彬
yī gè rén jiù wén-zhì-bīn~bīn

one CLF person already.EMPH literary-quality-refined.IDEO LE
“One person already was culturally refined.”

(58) ASBC (nº 100747)

倒 豆子 的 聲音 也 是 嘶哩嘩啦；
dào dòuzi de shēngyīn yě shì xīlī~huālā
pour beans LNK sound also COP clattering.IDEO
“The sound of pouring beans also was clattinger.”

(59) ASBC (nº 100801)

不 知道 究竟 為了 什麼 事 而 默默-不-語。
bù zhīdào jiùjìng wèile shénme shì ér mò~mò-bù-yǔ.
not know actually for what.Q ting CONJ silent.IDEO-NEG-speak.
“[I] don’t know why [he’s] so quiet.”

There are five (e) MODALITIES present in the ASBC dataset, see Table 4.13. These are either towards the written ('written', 'written-to-be-read', 'spoken-to-be-written') or towards the spoken ('spoken', 'written-to-be-read'). By far the most common one is 'written' (92.1%), which is unsurprising as spoken corpora are still lagging behind in general.

The (f) TOPICAL CLASSES of ASBC in descending order are 'literature', 'society', 'life', 'philosophy', 'art', and 'science'. As can be seen in Table 4.14,

Table 4.13: Distribution of ASBC modalities participating in the MCA of ASBC

mode	count	percentage
written	38930	92.0%
spoken	1926	4.6%
written-to-be-spoken	786	1.9%
written-to-be-read	416	1.0%
spoken-to-be-written	243	0.6%



Table 4.14: Distribution of ASBC tags participating (Huang, Hsieh & Chen 2017) in the MCA of ASBC

class	count	percentage	goalpercentage
文學	18250	43.1%	13%
社會	8531	20.2%	38%
生活	7318	17.3%	28%
哲學	3425	8.1%	8%
藝術	2657	6.3%	5%
科學	2120	5.0%	8%

there is a major over-representation in the remaining data set (43.2%) versus the global goal percentage for this corpus part in the ASBC as a whole (13%). This thus already suggests that a major number of ideophones we have found is more towards the literary, especially since the majority of them is of a written modality (see feature e above). This is not wholly unsurprising, since many members of CHIDEOD stem from sources that cover Pre-Modern Chinese, and there is also the issue of “language contact across time” (Eifring 2019), i.e., there are many Literary Chinese words still used in Modern Standard Chinese.

4.5.2 The MCA of ASBC and CHIDEOD

Now that we have discussed the first round of distributions, we now turn to the MCA analyses. Like before, we should first inspect the eigenvalues (see

Section 4.4.2), which show how much of the variance in the data is covered by the analysis. As Figure 4.17 shows, Dimension 1 and Dimension 2 are responsible for 68.3% ($49.9\% + 18.4\%$) which is still acceptable. Of these the Dimension 1 is clearly the most important one.

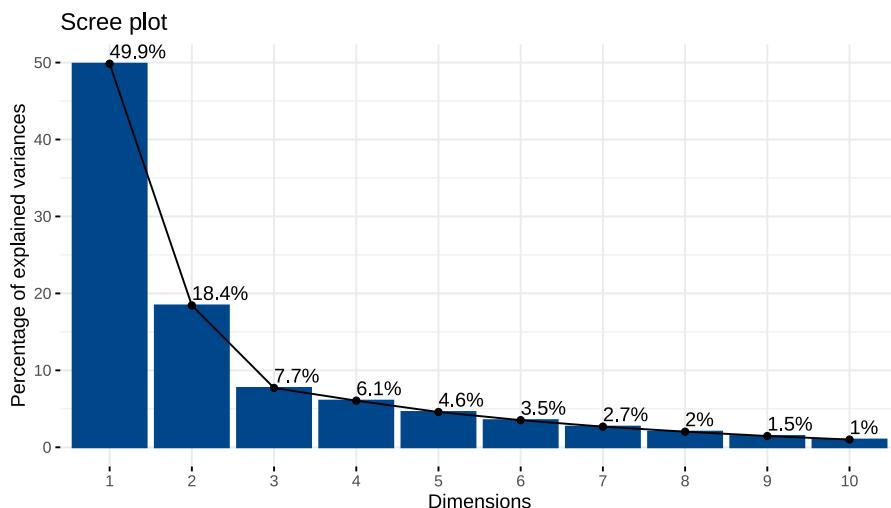


Figure 4.17: Eigenvalues for the MCA analysis after correction

On to the Multiple Correspondence Analysis of the data then. Figure 4.18 shows the main distribution of the features in the data set and how the supplementary variables (the ideophones and unaltered radical support) project on those. Not plotting the supplementary variables gives us Figure 4.19, which also suggests two clusters, as we have seen above. Note that this is based on the barycenters and thus does not contain the full spread of all points, which the plots below will make clearer. Inspecting the correlation plot between the principle dimensions and the variables, visualized in Figure 4.20, it can be seen that sensory imagery and morphological template are more explained by Dimension 1 than by Dimension 2, and frequency class is more explained by Dimension 2.

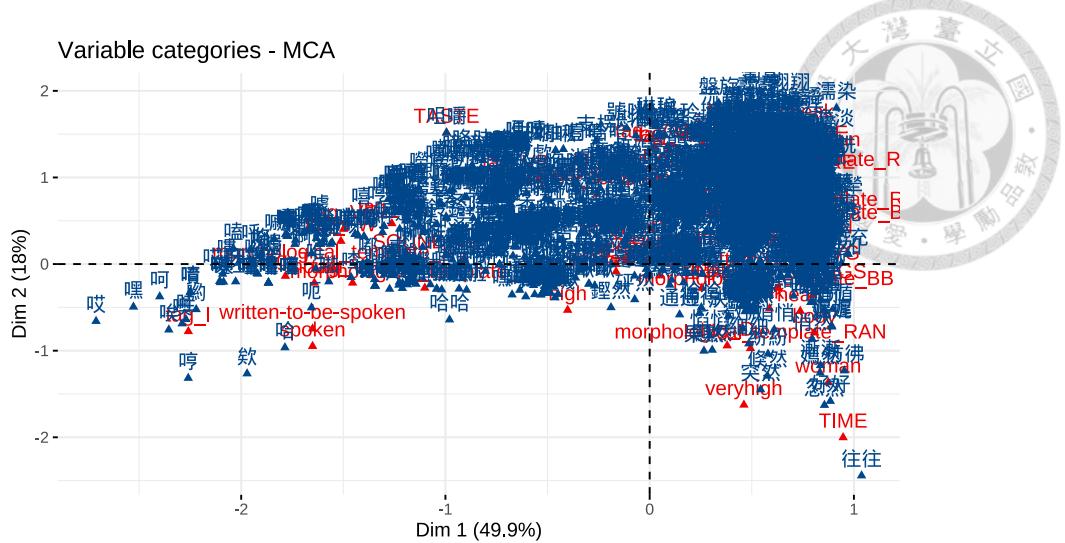


Figure 4.18: MCA plot of the barycenters of the CHIDEOD-ASBC data with the supplementary values (the ideophones themselves) showing. A landscape version of this figure is provided in Appendix 4 (Figure 8.2).

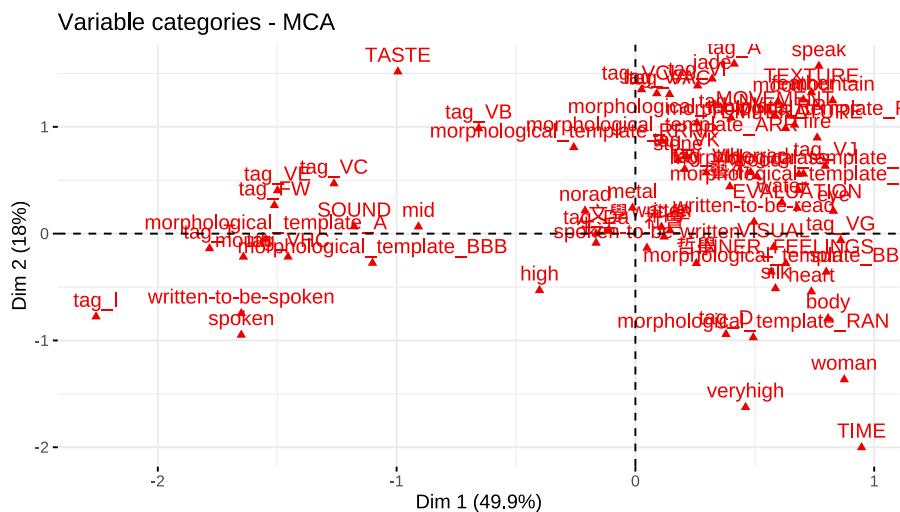


Figure 4.19: MCA plot of the barycenters of the CHIDEOD-ASBC data

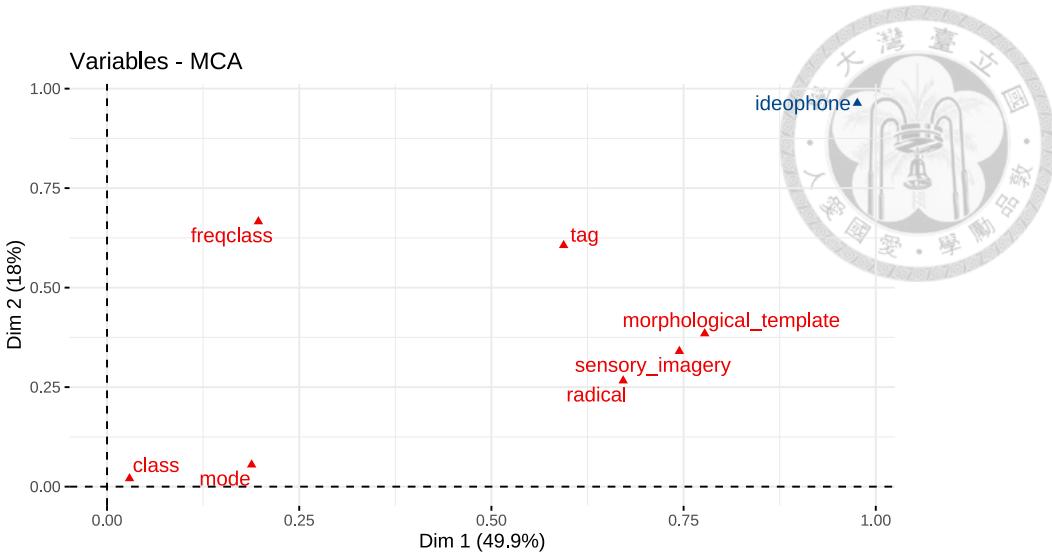


Figure 4.20: Correlation between variables and principal dimensions

Let us first look at the distribution of the variable MORPHOLOGICAL TEMPLATE in the MCA (Figure 4.21). ‘A’, ‘BB’ and ‘RR’ are the most common patterns, as we know from the distribution above. But it can be seen that ‘A’ is situated in its own separate cluster, yet still overlaps with other patterns (‘RB’ and ‘RR’). The two clusters are thus aligned along a gradient rather than discrete.

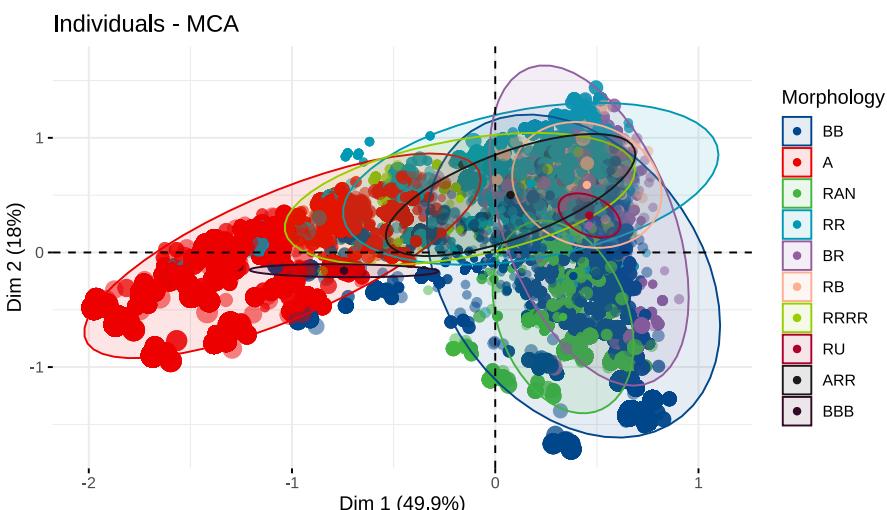


Figure 4.21: Confidence ellipse around the exemplars of morphological template

Looking next at the spread of SENSORY IMAGERY (Figure 4.22), we can see

once again that SOUND covers most of the horizontal dimension, i.e. Dimension 1. However, the 95% confidence ellipse still shows considerable overlap with the other values, suggesting that yes, there are two clusters for sensory imagery, but they are prototypically structured with fuzzy boundaries (cf. Levshina 2015).

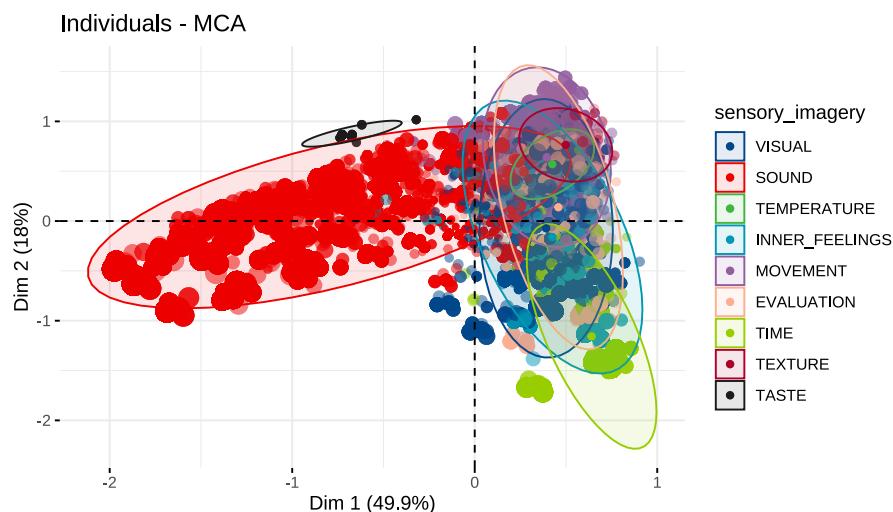


Figure 4.22: Confidence ellipse around the exemplars of sensory imagery

The binary feature of PRESENCE OR ABSENCE OF RADICAL SUPPORT (Figure 4.23) shows a very interesting pattern as compared to the MCA analysis of CHIDEOD above. In this case, the 95% confidence ellipses almost completely overlap, they are just aligned slightly differently along the Dimension 1. This suggests that radical support in practice does not play a huge differentiating factor for the marking of the prototypical ideophone – maybe it is just a curiosity of the Chinese writing system? Or perhaps it serves as a motivating factor but not deciding factor for ideophonicity?

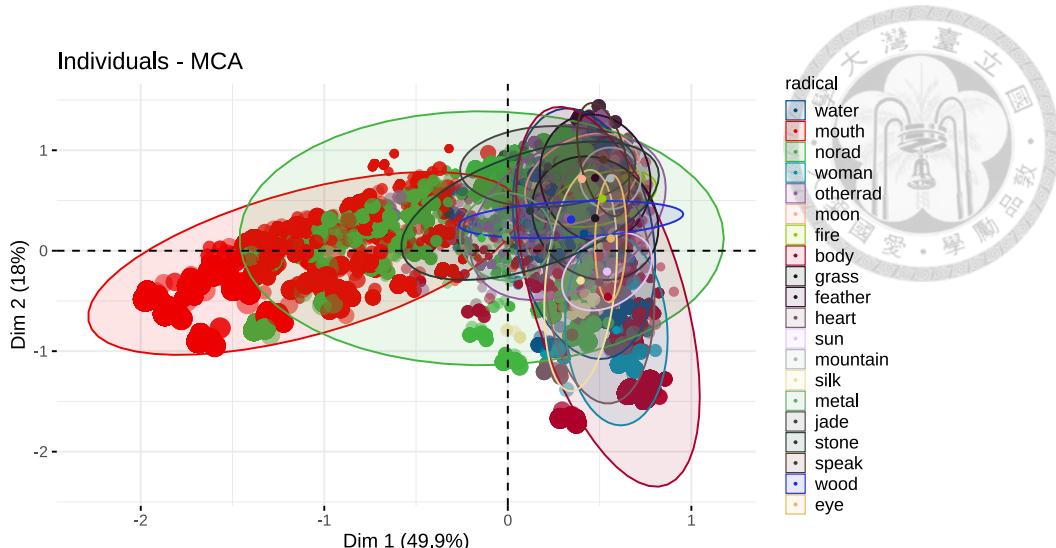


Figure 4.23: Confidence ellipse around the exemplars of semantic radicals

Continuing along the selected features that participate in this MCA, we could show the tags. However, upon inspection, it turns out that the ellipses did not yield any meaningful correlations. We thus decide not to plot this one, but let us instead look at the modality, visualized in Figure 4.24. The written vs. spoken orientation mentioned above is very present in the groups as the left ellipses align with the former and the right ones with the latter.

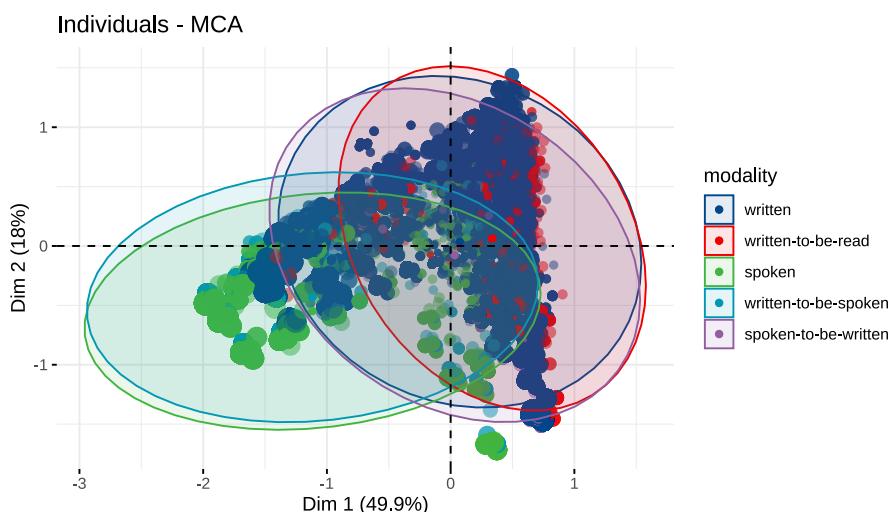


Figure 4.24: Confidence ellipse around the exemplars of modality

The last two variables to inspect are TOPICAL CLASS (Figure 4.25) and FRE-

QUENCY CLASS (Figure 4.26). Topical class does not reveal any particular distinction (the plot for the tag feature looked similar), although it can be seen that there is somewhat of a cline. The cline is also present for frequency class, but more ordered, with the highest frequency in the lower half of the plot and the lowest on top in a nice gradience of ellipses.

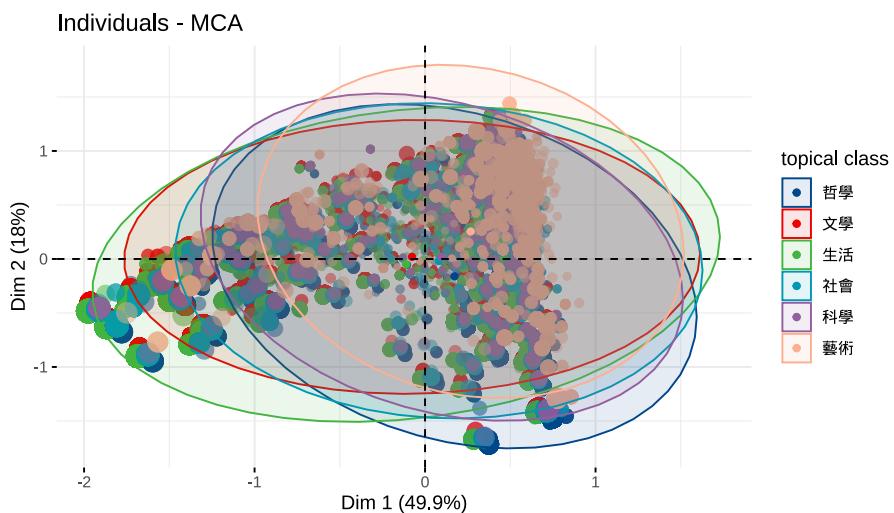


Figure 4.25: Confidence ellipse around the exemplars of topical class

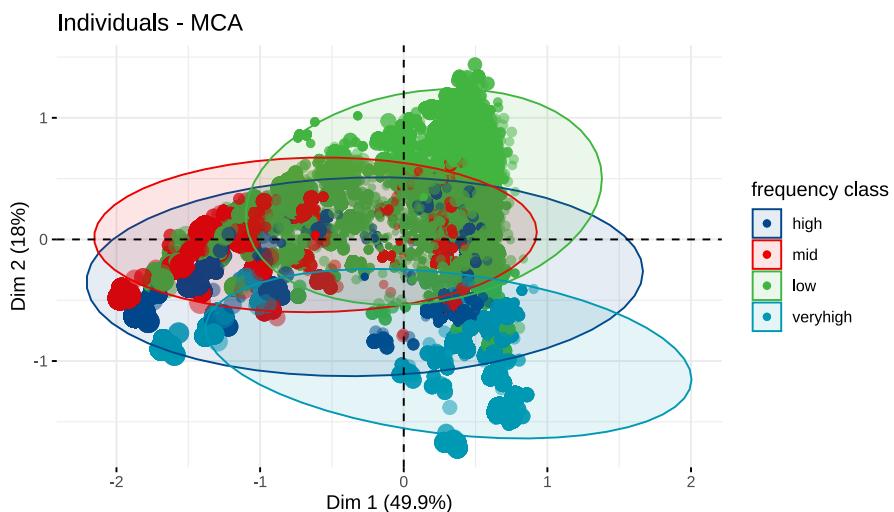


Figure 4.26: Confidence ellipse around the exemplars of frequency class

After all of these plots, we still need to check the significance of the correlations between dimensions and variables with a regression analysis, as

Levshina (2015:383–385) suggested. The concordance index C is summarized in Table 4.15 for the different variables, where it can be seen that most of the correlations are weak in predicting the variance of ideophones in the data set, i.e. the goodness-of-fit of this model is not the best, but at least acceptable.

Table 4.15: Regression analysis of the variables and the MCA model for ASBC

Variable	C	Acceptability
morphology	0.67	weak
sensory modality	0.53	very weak
radical	0.73	acceptable
tag	0.63	weak
modality	0.82	excellent
topical class	0.55	very weak
frequency class	0.69	weak

4.5.3 Interim summary

After applying Multiple Correspondence Analysis to the ideophones in the ASBC, it seems that the two clusters SOUND and NON-SOUND are still present in the data. The dyadic prototype clusters still overlap in multiple ways, suggesting that it is still valid to treat them as different realizations of the same phenomenon. Based on the contributions to the two dimensions and a series of simple Correspondence Analyses, it is possible to provide a similar diagram as Figure 4.16 above. This is shown in Figure 4.27. The sig-

nificant categories have been marked with a thicker line (based on their contribution to Dimension 1 to 3, cf. the discussion on eigenvalues above).

The arrows represent the interpreted result form the CA analyses between different variables. It is not a discrete representation, because there is no such discrete cut between SOUND and NON-SOUND ideophones, even though previous literature does suggest exactly this, by treating these phenomena as different lines of research.

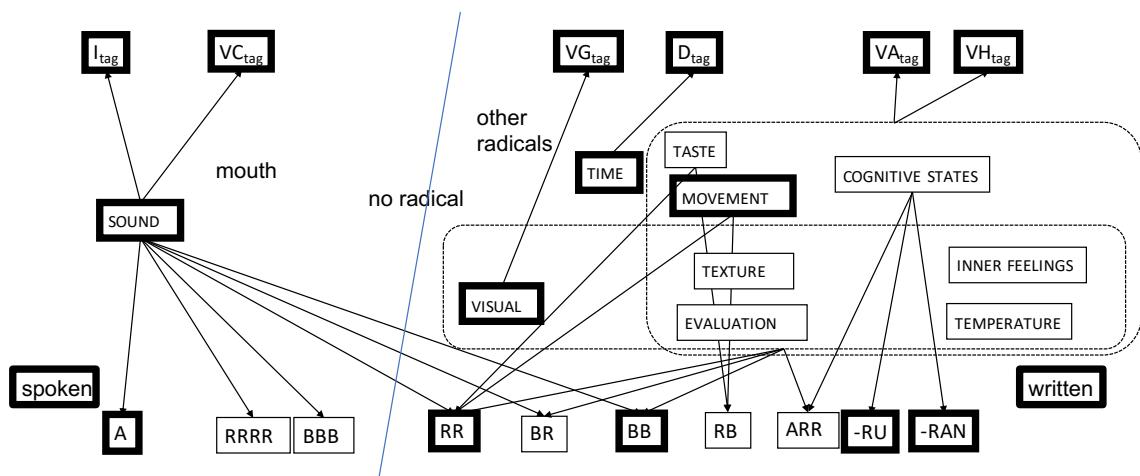
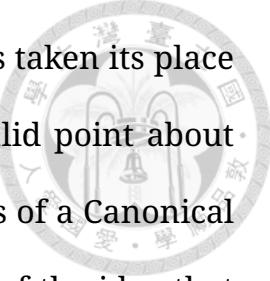


Figure 4.27: Diagram based on previous correspondence analyses

4.6 Conclusion

As we have seen in this chapter, we can approach the definition and scope of IDEOPHONES from multiple perspectives. On the cross-linguistic level, there was Doke's definition, which was formulated in impressionistic terms based on his experience with Bantu languages (Doke 1935). This definition is often cited, and one may still observe its influence in recent works, e.g., Haiman (2018:77) and Sasamoto (2019:5). Other definitions have been proposed (cf. Dingemanse (2011a)'s extensive overview), but in the last decade it seems that Dingemanse's *marked words that depict sensory*



imagery (2012), *belonging to an open lexical class* (2019) has taken its place in the typological spotlight. I believe he makes a very valid point about treating the keywords of this definition as possible features of a Canonical Typology framework. Dingemanse is also a big supporter of the idea that ideophones are best phrased in terms of prototypes, just like for example Childs (1994), Gabas & van der Auwera (2004), Lu (2006), and Akita (2009), although not all linguists agree with this, e.g., Heath (2019); Haspelmath.

We have now seen how the prototype works for Japanese in Lu (2006) and Akita (2009) – by no means the only authors who adopt this perspective. Both of these started from a list of types, and found that the ideophones of the type *koro~koro* 口口口口 ‘small thing rolling’ are the prototype. These are either called the ABAB type (based on morae), or the CV([^])CV-CVCV type (Akita 2009), based on consonants, vowels and the pitch accent. There are good reasons, related to the high type frequency, to consider this the prototypical structure of Japanese ideophones, with extensions to other types.

For Chinese, however, things are not that simple – which is the main reason for adopting a more statistical approach. There are some theoretical and featural arguments to exclude a number of items from the category CHINESE IDEOPHONE, as shown above. But even then the remainder does not clearly show preference for one structure or the other. In Mok’s (2001:45–46) inventory, onomatopoeia most frequently belong to the AB type (based on syllables)⁴², while Mandarin ideophones supposedly all belong to the XAA type, here treated as BB – which of course is not true but perhaps due to a sparseness of data in her study.

⁴²Such notation is confusing, see Section 3.2.

Sam-Sin (2008) and Meng (2012) both study Chinese ideophones in the Beijing dialect and both propose a prototypical approach, based on morphology, phonology, semantics, constructions etc. Taking this line of thinking to an extreme, we decided to adopt Multiple Correspondence Analysis to analyze the morphological template, sensory imagery and orthographical motivation (radical support) in the CHIDEOD database. This resulted in the dual clustering structure between SOUND ideophones and NON-SOUND ideophones, which may lead one to suggest that ideophones are completely different from onomatopoeia. However, as argued multiple times above, there was enough overlap between the different clusters to strongly suggest that they are on a cline rather than demarcated with a hard border. Next, we decided to investigate the ideophones in a balanced corpus ASBC of Standard Chinese, which confirmed the two clusters with fuzzy boundaries and prototypical cores.

Unfortunately, *the prototype* within the Chinese lexicon has not been found. As the multivariate analyses show, in order to do justice to the phenomenon in Chinese, a number of variables show that there are interrelated clusters, but not one item that is privy to them all. Being prototypically structured entails that there are better candidates and worse candidates of ideophones. Based on the interpreted diagram in Figure 4.27, we could, however, suggest exemplars like (60) in the ASBC as being prototypical for the SOUND cluster, and (61-62) for e.g. MOVEMENT and VISUAL ideophones. There is a clear tendency for the former to belong to the spoken modality, and the latter to the written – further showing that there is a difference be-

tween COLLOQUIAL and LITERARY ideophones (Van Hoey 2018a).



(60) ASBC (nº 103846)

發出 嘿 嘿 嘿 的 聲音，
fāchū bì bì bì =de shēngyīn,
emit beep.IDEO beep.IDEO beep.IDEO =LNK sound
“Emitted a beeping sound; beep, beep, beep.”

(61) ASBC (nº 100622)

悄悄的 退了 出去。
qiāo~qiāo=de tuì-le chū-qù
silently.IDEO-LNK retreat-LE exit-go
“[He] silently went back out.”

(62) ASBC (nº 101754)

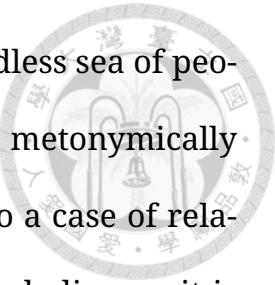
燦爛的 陽光。
càn~làn=de yáng-guāng.
bright.IDEO=LNK sun-light
“The bright sunlight”

Let us revisit the matter of the two clusters. Why do these two clusters exist? Akita (p.c.) has suggested that perhaps aspect may be involved in the difference between the two clusters, as this is the case in Japanese. Akita (2017a) surveys four proposals to distinguish Japanese reduplicated ideophones and suffixed mimetics: iterativity (Kita 1997; Hamano 1998), bound-

edness (Toratani 2005), telicity (Tsujimura & Deguchi 2007), and durativity-punctuality (Akita 2009). As it currently stands, we are not aware of satisfactory explanations for Chinese ideophones based on differences between TAM-marking.

Perhaps this is just due to the nature of the language, which flexible categories – a phenomenon called precategoriality (for Late Archaic Chinese) by Bisang (2008). This means that so-called zero-derivation is very common in Chinese, and Chinese should be considered as a prime candidate to support Construction Grammar approaches, such as Cognitive Grammar (Langacker 1987a; 1991); Construction Grammar (Goldberg 1995; 2006) or Radical Construction Grammar (Croft 2001). Outside the scope of this dissertation, yet a very interesting future research project, would be indeed to study the *Aktionsart* or lexical aspect of ideophones. Possible frameworks are a revised Vendlerian (Vendler 1957) paradigm such as Croft (2012) or Frame Semantics (Fillmore, Kay & O'Connor 1988; 2003). The former has been explored for Japanese, yet barely touched upon mimetics (Taoka 2000); the latter has been well explored by Akita and colleagues (Akita 2009; 2012b; Kiyama & Akita 2015).

Is there another aspect about aspect in relation to the two clusters that has been missed then? The iconicity aspect appears to be of great importance. Dingemanse (2011b) divides Peirce's ICONICITY into IMAGIC ICONICITY, GESTALT ICONICITY and RELATIVE ICONICITY. Imagic iconicity occurs when the phonological form of the ideophone mimics a sound in the real world, like the barking of a dog, expressed by *wāng~wāng* 汪汪. Gestalt iconicity depicts



even structure, such as *máng~máng rén-hǎi* 茫茫人海 ‘boundless sea of people’. In this case, the markedness of the reduplication (BB) metonymically suggests that the event lasts. However, this example is also a case of relative iconicity, which pertains to phenomena like sound symbolism as it is most well-known. In this example, the phonemes /an/ could suggest boundlessness by virtue of being an [+OPEN] and [+BACK] vowel and a [+NASAL] coda, making /an/ a candidate for a phonestheme in Chinese (cf. Chan 1996) – note that the phonesthemic explanation does not really work for the first example of the dog barking, see Section 5.1.1.

In any case, Dingemanse’s tripartite distinction, and subsequent research such as Dingemanse et al. (2015) or Sidhu & Pexman (2017), can aid in our understanding of aspectual relations. It seems that when the sensory imagery is SOUND and the modality towards the spoken continuum, imagic iconicity plays a more important role: one *dōng* 咚 means only one bounce; three *dōngs* means three bounces. This also explains morphological patterns such as B, BB, BBB, and BBBB and the less frequent occurrence of three and four reduplicated syllables – it is just not that often that we need to really mimic three or four bounces. By far the most common pattern in CHIDEOD is ideophones composed of two syllables (BB, BR, RB, RR and the compositional ones like -RAN). If they are situated near the SOUND on Dingemanse’s (2012) proposed hierarchy, such as MOVEMENT, there will be a greater chance of them being imagic iconic as well. An example that comes to mind is *pāi~pāi* 拍拍. This can either mean ‘tapping’ or ‘clapping’. However, this fully reduplicated item is polysemous from an aspectual

point of view as well. Are we dealing with an ATTENUATION IN INTENSITY ‘gently (keep on) tapping for a short while’ (Li 2015; Lǐ & Ponsford 2018), or is it loudly clapping twice (IMAGIC ICONICITY) or a metonymy for an ITERATIVE ‘keeping on tapping or clapping’. For the latter, see an analysis of reduplication in Afrikaans (van Huyssteen 2004).

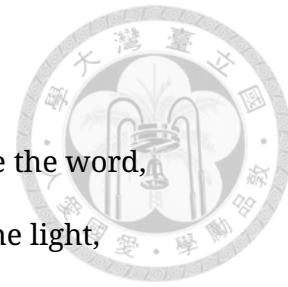
This means that future research should look further into the interplay between aspect and ideophones in Chinese and the context in which they are used. It seems that metonymy will provide an important role, as it is economic for the speaker to have to mark a depiction with two syllables (full or partial reduplication), rather than repeat it over and over, conforming to the even structure. That is not to say that repetition never occurs, clearly it does, but it unsurprisingly seems to be mostly in spoken interactions rather than in written material. The hearer or addressee also plays an important role in the communication process. As the conversation incrementally progresses, and ideophones are somehow formally marked by phonation, boundaries (cf. above) and supported by gesture (often depending on the semantics as well), they perceive the markedness and apprehend it as a depiction, through its modes of iconicity. It should be remembered, that ideophones, like other linguistic elements, often occur with the same collocates, as prefabs or entrenched constructions.

This brings us to another point: what exactly is the meaning of an ideophone? In Chapters 5 and 6 this question will be explored through diachronic prototype theory and semasiology and onomasiology, as we trace the semantics of ideophones depicting LIGHT and their collocates

through time. We will investigate how the variational phenomena in the synchronic data evolved over time by focusing on the prototypical structure and other instances of salience. But for now, it can be concluded that Chinese ideophones most certainly are prototypically structured as a language-particular category. To select a number of necessary and sufficient conditions, while only investigating one parameter, e.g., morphology, without including others like e.g., depiction, does not do justice to them.



5 Diachronic prototype semantics



Only in silence the word,
only in dark the light,
only in dying life:
bright the hawk's flight
on the empty sky.

Ursula K. Le Guin

5.1 Introduction

The argument that has been developed thus far has posited that it is beneficial to investigate the ideophonic lexicon of Chinese (Chapter 2). By collecting data from multiple sources (Chapter 3), it has been shown that this group of words is not homogeneous; there are many instances of variation observable within the data. Using statistics we have tried to delineate the prototypical structure of the ideophonic lexicon in a synchronic stratum (Chapter 4). Let us now then adopt a diachronic perspective, and in doing so explore research question 6b: how did the prototypical variation of ideophones change over time?

The sources used in this chapter consist of corpus material, more specifically the Scripta Sinica (Section 3.3.1). Fortunately, Chinese historical linguistics currently is in the position of having an abundance of corpora, although its historicity means that many multimodal elements cannot be adequately studied – there is no record of that. On the other hand, Chinese lexicography has a long tradition, and the script contains semantic informa-

tion beneficial for our understanding of ideophones and their development through time.

This chapter⁴³ rests on the semiotic folk model (see Section 1.3), in the sense we will investigate variation in the <orthographic> pole, <phonological> pole, and <semantic> pole. In other words, we take the *xíng yīn yì* 形音義 ‘form-sound-meaning’ of Chinese words seriously. Bearing this model in mind, the chapter is organized as follows: First phonesthemes will be treated (Section 5.1.1). The well-known case of *gl*-phonesthemes will guide us to the selection of ideophones situated in the semantic domain of LIGHT. The notion of phonesthemes will be investigated for this group of words. However, their results will indicate that instead of a strong phonestheme claim, a weaker clustering of groups within LIGHT ideophones is maintainable. Second, the group of LIGHT ideophones will then be used to illustrate the different aspects of Diachronic Prototype Semantics, and how analysis of ideophones is also dependent on the granularity or specificity of the units one is investigating, namely as Mental Spaces, as Frame, as Domain / Idealized Cognitive Model, or as Image Schema. Let us now first turn to phonesthemes then.

5.1.1 Phonesthemes: one meaning for multiple forms

There is a quite well-known tradition of investigations into phonesthemes during the 20th century and the last two decades. Especially the English *gl*-phonestheme has caught the attention of numerous scholars, such as

⁴³This paper was first presented at ICPEAL 19 - CLDC 9 (Van Hoey & Lu 2018) and further explored at ICLC 15 (Van Hoey 2019b).

Jespersen (1922); Reid (1967); Jakobson & Waugh (1979); Magnus (2001); Sadowski (2001); Boussidan, Sagi & Ploux (2009); Kwon & Round (2015); Thompson (2017). The analysis of this phonestheme usually involve categorizing words containing the *gl-* phonestheme into different groups. For instance, Magnus (2001) categorizes them according to semantic differences, as shown in 5.1.

Table 5.1: The *gl-* phonestheme according to Magnus (2001)

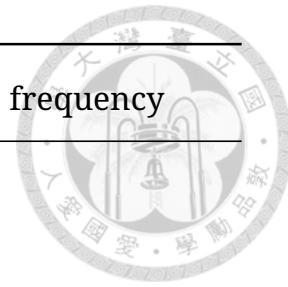
meaning	types	frequency	type
reflected or indirect light	<i>glare, gleam, glim,</i> <i>glimmer, glint, glisten,</i> <i>glistter, glitter, gloaming,</i> <i>glow</i>	10	
indirect use of the eyes	<i>glance, glaze, glimpse,</i> <i>glint</i>	4	
reflecting surfaces	<i>glacé, glacier, glair, glare,</i> <i>glass, glaze, gloss</i>	7	

This kind of inventorizing work is compelling, although it is hard to claim that we are dealing with ‘real iconic’ mappings between *<sound>* and *<meaning>* for every case, as Magnus often suggests. This point is also made in a more extensive study of this phonestheme, presented by Sadowski (2001), who categorizes all 47 independent words starting with *gl-* as can be found in the *Oxford English Dictionary* (Simpson & Wiener

1989), presented in Table 5.2. As can be seen, Sadowski is able to discern 9 groups of meanings. He shows that these meaning clusters can be traced back to Middle English and some of them also to Old English. While he is able to posit a form-meaning mapping for this phonestheme across different Germanic languages, a cross-linguistic mapping is ruled to be out of the question. In other words, this phonestheme belongs to the category of conventional sound symbolism (Hinton, Nichols & Ohala 1994). This has later been supplemented by the observation that the *gl-* phonestheme derives from the Proto-Indo-European stem *GHEL ‘yellow’ (Boussidan, Sagi & Ploux 2009; Thompson 2017), although this does not necessarily rule out an iconic relation (*pace* Thompson & Do (2019)), it is just unlikely to be based on imagic iconicity.

Table 5.2: The *gl-* phonestheme according to Sadowski (2001)

meaning	types	type frequency
light, brightness	<i>glad, glade, glaik, glance,</i> <i>glare, glass, gleam, glee,</i> <i>gleed, gleg, glent, glimmer,</i> <i>glisten, glitter, glow</i>	15
looking, seeing	<i>glance, glare, glent, glint,</i> <i>gloat, gloom, glower, glut</i>	8
moving lightly	<i>glace, glaive, glance, glent,</i> <i>glide, glint, glissade</i>	7
deceiving	<i>glaik, glaver, gleek, glib,</i> <i>gloze</i>	5



meaning	types	type frequency
dark light	<i>gloaming, gloom, glower,</i> <i>glum</i>	4
smoothness	<i>glaborous, gleg, glib,</i> <i>glossy</i>	4
slimy substance	<i>glair, gleet, glue</i>	3
joy	<i>glad, glee</i>	2
splendour	<i>glamour, glory</i>	2
other	<i>slack, glen, gladiator,</i> <i>gland, glean, glebe, gleg,</i> <i>gloss, glove, gluttony</i>	10

These examples of categorizing phonesthemes (Magnus 2001; Sadowski 2001), shown in Tables 5.1-5.2, lead us to two crucial questions: how many form-meaning mappings would one need to postulate a certain phonestheme? And, from an epistemological and methodological point-of-view, how could we reliably find this out? A demonstration of how *not* to approach this issue will be instructive. Is it enough to, for example, flip through a dictionary of English to get data, and then redefine words in such a way that a commonality is found? In the caricatural example (63) this⁴⁴ is illustrated. Note that all these words end in /eɪs/, and can be defined in terms of a common element, namely LOCATION. In this case we have 7 positive cases. Is it enough to say there is a phonestheme at work?

⁴⁴This is based on a simple search for words ending in the rhyme /eɪs/ on en.wiktionary.org, which yielded these results.

(63) Example of ‘possible’ phonesthemes

<i>face</i>	/feɪs/	‘location on forehead’
<i>space</i>	/speɪs/	‘wide location, esp. outside of Earth’
<i>place</i>	/pleɪs/	‘location’
<i>base</i>	/beɪs/	‘location on which things are built’
<i>bass</i>	/beɪs/	‘low pitch sound location when singing’
<i>case</i>	/keɪs/	‘location in which to put stuff, box’
<i>trace</i>	/træɪs/	‘location where something has been’



It certainly seems plausible at first sight, and they also seem to adhere to a number of Canonical Typological (see Section 4.2.6 for a short introduction to this framework) characteristics for phonesthemes (Kwon & Round 2015): (Criterion 1) they occur in a greater number of lexical stems, (Criterion 2) across many parts of speech, namely verbs and nouns, (Criterion 3) but they are weakly sound-symbolic (cf. below). (Criterion 4) It is hard to interpret if the meaning of the proposed phonestheme /eɪs/ is multifold or singular, as well as the other way around, (Criterion 5) if LOCATION is expressed through other forms as well. According to Criterion 6 of Kwon & Round (2015), it is non-canonical, as /eɪs/ occurs with many other elements as well. However, for their Criterion 7, it is canonical for phonesthemes, since it agglutinatively combines with its residue (in this case the initial). So in this approach, /eɪs/ is somewhat phonesthemic, yet not completely canonical.

But is this a believable assessment? After all, we do have seven items in example (63), and one could argue it *feels* like a decent number to base phonesthemicity on. In Chinese, for instance, Chan (1996) lists eight exam-

ples, based on Pulleyblank (1973), of words starting with /m-/ to posit the phonestheme DARK, COVER, BLIND, shown in (64).

(64) Chan (1996:9)

- a. *máng 盲* ‘blind; deluded’
- b. *mào 冒* ‘to go forward with eyes covered; to cover’
- c. *mèi 昧* ‘obscure, dark’
- d. *mēn 憶* ‘to cover; mournful, sorrowful’
- e. *méng 蒙* ‘to cover; to conceal’
- f. *mǐ 眇* ‘blind, as with dust’
- g. *míng 夥* ‘dark, obscure’ (*mìng 夜* ‘night’)
- h. *mò 默* ‘dark, secret; silent’

She calls the /m/ for conveying DARK, COVER, BLIND language-specific, but states that the [+ LABIAL] *feature* of /m/ is “motivated, and [...] by no means arbitrary” (Chan 1996:9). Because [+LABIAL] is a [grave] segment, and evidence from Russian poetry (Priestly 1994) connects gravity to sadness⁴⁵, there is some darkness and blindness in Chinese /m/. Furthermore, /m/ “has a relatively long duration of lip closure, during which time the oral cavity is thrown into total darkness” (Chan 1996:9–10). The counter example for BRIGHTNESS *míng 明* is “countered” by Chan by noting that there are two other candidates for this meaning, namely *liàng 亮* and *guāng 光*, both with a low vowel in various dialects. This issue settled, she argues for a new

⁴⁵This cross-modal mapping is presumably a conceptual metaphor.

type of sound symbolism to be added to Hinton, Nichols & Ohala (1994:2–6)'s four-fold typology, shown in (65), where local synesthetic sound symbolism (65d) is the new type.



- (65) a. corporeal sound symbolism, e.g., *hēng* 哼 'humph (doubt)'
- b. imitative sound symbolism, e.g., *wēng(wēng)* 嗡 嗡 'bzzz bbzzz (buzzing of bees)'
- c. (universal) synesthetic sound symbolism, e.g., /a/ is big, /i/ is small
- d. (local) synesthetic sound symbolism, e.g., labialized consonants like /t^w/ in *tuán* 團 'round, circular'
- e. conventional sound symbolism, e.g., /gl-/ maps onto LIGHT

For the toy example (63), where it could be argued that /eɪs/ might be categorized as a local synesthetic sound symbolic type or a conventional sound symbolic type. To find out, one would probably need to look at cross-linguistic data, which has been explored in a number of recent studies. To our knowledge, the largest-scale study is Blasi et al. (2016), which investigated word lists of basic vocabulary (of the Swadesh type) in 6,452 languages. They identified a number of positive and negative correlations between structural elements and the items on the lists, e.g., BREASTS positively correlates with /u/ and /m/ but repulses /a/, /h/ or /r/. They explain this by referring to (a slightly misquoted) Jakobson (1960) and Traunmüller (1994), stating that the occurrence of the bilabial nasal consonant /m/ and high back vowel /u/ resembles the mouth configuration of suckling babies or the sounds feeding babies produce (Blasi et al. 2016:3).

On a smaller scale, but arguing for a featural analysis, there is work by Joo (2018; 2019). With a sample of 66 genealogically distinct languages, yet with little to no consideration for their areality, he investigates how phonological features correlate with meanings or abstractions of meanings of items occurring on the Leipzig-Jakarta list (Tadmor 2009). He finds a similar positive correlation as Blasi et al. (2016) for BREAST, i.e. [+LABIAL] and [+NASAL], but no negative correlation (and also does not mention Blasi et al. (2016)'s /a/, /h/, /r/ in the overview). If these two case studies had included LOCATION, they might have revealed more about the cross-linguistic occurrence of features with /eis/, or other elements for that matter, by relying on a more robust methodology.

However, the matter fundamentally hinges on the understanding of the term iconicity. Is it a discrete and categorical property (*do items have iconicity, yes or no?*), or does it involve semiotic relations in a Peircian jacket, with imagic and diagrammatic signs (cf. the approach taken by Dingemanse 2012 for ideophones), or is it scalar (cf. Winter 2019's norms for sensory items)? Dingemanse, Perlman & Perniss (2020) identify these three perspectives and do not consider them mutually exclusive. Even stronger, they consider the bridging of these fundamental positions to be fertile ground for new experimental evidence. A recent discussion on the methodology for such experiments and inferences can be found in Motamed et al. (2019). Especially relevant to our toy example (63) is their discussion on going from data-driven correspondences to iconicity. It is clear that just listing possible candidate phonesthemes, i.e., iconic form-meaning mappings, is insuffi-

cient. The seven /eɪs/ words in (63) would probably not make the cut. There is more hope for Chan's (1996) /m/ phonestheme (64), given that methods like Blasi et al. (2016) and Joo (2019), or even Winter (2019), provide more robust statistics to consider it iconic.

Yet, they forego a crucial issue, namely the historical aspect. Even a cursory reflection on the etymology of the /eɪs/ words in (63) would reveal discongruencies for this currently congruent phonological form. In other words, just because a number of words have the same form in a given language stage, this should not be taken at face value and added to one's database. Ideally, Blasi et al. (2016) and Joo (2019) should find a way to incorporate these ideas in their methodology, in order to make their statistics even more reliable. An ideal study, then, must investigate for every item what previous forms were, as to mitigate for accidental sound changes that led to the same form.

Now, Chan (1996) is in better waters, as Table 5.3 shows. For the items we could find in Baxter & Sagart (2015), the Middle Chinese and Old Chinese reconstructions show that /m/ is postulated as respectively /m/ and /m[˥]/, which is still a labial nasal. But is there then nothing more to say? Is /m/ effectively a submorphemic element that belongs to the (local) synesthetic sound symbolism category, i.e., displays iconicity (the cross-linguistic phenomenon)?

Perhaps another approach could statistically reveal more about the phonesthemic status of /m/. For instance, Smith (2015) successfully studies fully reduplicated forms in the *Shījīng* 詩經, focusing on phonesthemes and meanings. While he adopts a sound statistical approach, very much

Table 5.3: /m/ phonestheme for items present in Baxter & Sagart (2015)

item	Mandarin	MiddleChinese	OldChinese	gloss
盲	máng	mæŋ ¹	*m ^č <r>aŋ	blind
冒	mào	maw ³	*m ^č uk-s	to look down on; covetous
冒	mào	maw ³	*m ^č uk-s	covering
昧	mèi	mwoj ³	*m ^č [u][t]-s	dusk; dark
蒙	méng	muwnŋ ¹	*m ^č onŋ	cover (v.)
冥	míng	menŋ ¹	*m ^č eŋ	dark

in line with the collostructional analysis methodology (see Chapter 7; Stefanowitsch & Gries 2003), his data sample appears quite small, as he limits the total number of items to 320. On the other hand, he nuances his findings considerably. The main conclusion from his study is that there are certainly statistically significant mappings between form and meaning, or rather semantic feature, classifying them as phonesthemes. However, more crucially, he does not necessarily treat these phonesthemes as iconic (in the cross-linguistic sense), but argues that many of them can also just be conventional. As for Chan (1996), she also shows some restraint for her phonestheme /m/, devising the new type of *local* synesthetic sound symbolism – yet also arguing for its iconic motivations.

But does that mean that Chinese ideophones do not make use of sound symbolism? With certainty, it can be said that some Chinese (with Chinese in the broad sense) ideophones rely on sound symbolism, as has been argued before (Mok 2001; Arthur Lewis Thompson 2019a). However, the examples used to drive home this point consist of SOUND ideophones, of which we know that they mostly belong to the colloquial stratum and are used in spoken language – where other devices, like gesture and phonation, can support their markedness. And another form of support they make use of

is sound symbolism, mostly belonging to the type of imitative sound symbolism, but not without excluding other types from Hinton, Nichols & Ohala (1994)'s typology like synesthetic sound symbolism, and conventional sound symbolism. In the next section, we want to focus on a group of ideophones which depict sensory imagery different from SOUND, namely VISION. More specifically, ideophones in the semantic domain of LIGHT will be probed for phonesthemes.

5.1.2 LIGHT syllables through time

Now that we have discussed some possible issues with phonesthemes (Section 5.1.1), it is time to begin the case study of ideophones in the semantic domain of LIGHT. These also belong more to the literary stratum, as opposed to the colloquial stratum. Therefore, they occur mostly in written language and rely more on radical support than colloquial ideophones do. In this section, we do make the assumption that phonesthemes do not play an important role in these LIGHT ideophones. The reasons are that it is better to be 'iconicity-critical' and rely on empirical evidence, because it is actually quite complex to prove that statistical correlations are also iconic and not just arbitrary or systematic, see the overview in Dingemanse et al. (2015).

Table 5.4: Sample of 35 ideophones expressing LIGHT

<i>Hànyǔ pīnyīn</i>	character	<i>Hànyǔ pīnyīn</i>	character
yìyì	熠熠	yíngyíng	瑩瑩
yuèyuè	爚爚	yíngyíng	熒熒
yàoyào	耀耀	xuānxuān	軒軒



<i>Hànyǔ pīnyīn</i>	character	<i>Hànyǔ pīnyīn</i>	character
yào yào	耀耀	shàn shàn	剡剡
yì yào	熠熠	xióng xióng	熊熊
yì yào	熠耀	shuò shuò	燦燦
yù yù	煜煜	huàng huàng	晃晃
yù yì	煜熠	càn làn	爍爍
yì yù	熠煜	càn càn	熒熒
yè yè	燁燁	làn làn	爛爛
yè yè	曮曮	làn mǎn	爛熳
jīng jīng	晶晶	wēi wēi	煒煒
wéi yè	煒煒	guāng guāng	光光
zhuó zhuó	灼灼	zhuó shuò	灼燦
hào hào	皓皓	shuò shuò	鑠鑠
jiǎo jiǎo	皎皎	shǎn shǎn	閃閃
luò luò	瑩瑩	zhēng zhēng	錚錚
hù hù	扈扈		

It will be shown that the phonology of LIGHT ideophones groups them differently according to major language stage (Mandarin Chinese – Middle Chinese – Old Chinese). A sample of 35 ideophone types was selected from CHIDEOD, presented in Table 5.4⁴⁶. Most of these are of the full reduplication type (morphological template BB) or partial reduplication (RR or BR). For their meanings, the reader is referred to the examples shown in Table

⁴⁶It is not claimed that these 35 types exhaust the semantic field of LIGHT ideophones.

Table 5.5: *Hànyǔ dà cídiǎn* definitions for some LIGHT ideophones

traditional	pinyin_tone	definitions
灼灼	zhuó~zhuó	1. 明亮貌。2. 鮮明貌。3. 明白貌。4. 彰著貌。5. 盛烈貌。6. 炙熱貌。 7. 思念殷切貌；熱切貌。8. 蜀美女名。
熒熒	shuò~shuò	1. 光芒閃動貌。2. 酷熱貌。
瑩瑩	luò~luò	1. 分明貌；顯著貌。2. 卓絕貌。3. 擊石聲。
鑠鑠	shuò~shuò	1. 光芒閃耀貌。2. 油光潤澤貌。

Table 5.6: Kroll (2015) definitions for some LIGHT ideophones

traditional	pinyin_tone	definitions
灼灼	zhuó~zhuó	evident, brilliant, aglow, vivid and vibrant, brightly blazing, plain and patent
熒熒	shuò~shuò	flashing, flaring, effulgent, alight, rutilant; splendid
瑩瑩	luò~luò	evident, apparent; conspicuous, outstanding

5.5 (*Hànyǔ dà cídiǎn*) and Table 5.6 (Kroll 2015). It can be seen that they are all in the domain of LIGHT, although some are considered considerably more polysemous, like *zhuó~zhuó*.

To study the phonology of these items, it is useful to list all the different syllable types of the sample. It is a methodological choice to abstract the ideophones to their syllabic forms ('single characters') and perform analyses based on these abstractions. However, the limitation in scope does not conflict with further abstractions to syllabic elements as the main units for analysis. The single syllables of these ideophones were then reconstructed to Middle Chinese and Old Chinese using Baxter and Sagart's systems (Baxter 1992; Sagart 1999; Baxter & Sagart 2014; 2015), see Table 5.7. We will use *Hànyǔ pīnyīn* and Baxter's Middle Chinese transcription in the remainder of this section.

Immediately, a number of groups in the data in Table 5.7 stand out. These can be further categorized: based on the Old Chinese reconstruction, some belong to the same word family, a term for words that share the same root



Table 5.7: Syllable types with their Mandarin, Middle Chinese and Old Chinese reconstructions

character	Mandarin		Middle Chinese		Old Chinese
	pinyin	IPA	Baxter	IPA	IPA
爚	yuè	j ^w eɿ	yak	yak	*lewk
燿	yào	jauɿ	yewH	yewH	*lewk-s
耀	yào	jauɿ	yewH	yewH	*lewk-s
灼	zhuó	tʂwɔɿ	tsyak	t̪eak	*tewk
犖	luò	l ^w oɿ	lak	lak	*r̪ewk
熒	shuò	ʂwɔɿ	syak	ʂak	*ʂrewk
鑠	shuò	ʂwɔɿ	syak	ʂak	*ʂrewk
熠	yì	jiɿ	yik	yik	*Gʷəp
煜	yù	jiɿ	yuwk	yuk ^w	*Gʷrəp
燁	yè	jeɿ	hip	hip	*Gʷrəp
熀	yè	jeɿ	hip	hip	*Gʷrəp
暉	yè	jeɿ	hip	hip	*Gʷrəp
煥	wěi	weiɿ	hjwɪjX	yjwɪjX	*Gʷəj?
暉	wěi	weiɿ	hjwɪjX	yjwɪjX	*Gʷəj?
熯	càn	tʂanɿ	tsanH	t̪sanH	*tsʰan-s
爛	làn	lanɿ	lanH	lanH	*r̪an-s
熳	màn	manɿ	manH	manH	*m̪an-s
漫	màn	manɿ	manH	manH	*m̪an-s
熔	màn	manɿ	manH	manH	*m̪an-s
晃	huǎng	xwaŋɿ	hwangX	ywaŋ	*Gʷiaŋ?
光	guāng	kwaŋɿ	kwang	kwaŋ	*kʷiaŋ
煌	huáng	xwaŋɿ	hwang	ywaŋ	*Gʷiaŋ
亮	liàng	ljaŋɿ	ljangH	ljaŋH	*raŋ-s
暉	huī	xweiɿ	xjwɪj	xjwɪj	*qʷhər
輝	huī	xweiɿ	xjwɪj	xjwɪj	*qʷhər
熒	huī	xweiɿ	xjwɪj	xjwɪj	*qʷhər
熑	yíng	jinɿ	hweng	ywen	*N-qʷəŋ
熒	yíng	jinɿ	hweng	ywen	*N-qʷəŋ
熚	zhēng	tʂəŋɿ	tsreang	t̪ʂəŋ	*ts̪əŋ
晶	jīng	t̪ɕiŋɿ	tsjeng	t̪seŋ	*tseŋ
軒	xuān	ɛwɛnɿ	xjon	xjon	*q̪har
剡	yǎn	jɛnɿ	yemX	jemX	*N-ram?
熊	xióng	ɛiuŋɿ	hjuwng	yjuŋ ^w	*C.Gʷəm
扈	hù	huɿ	huX	yuX	*m-q̪a?
閃	shǎn	ʂanɿ	syemX	ʂemX	*s.tem?
皓	hào	hauɿ	hawX	hawX	*g̪u?
皎	jiǎo	f̪iauɿ	kaewX	kæwX	*k̪raw?

but have different affixes⁴⁷ (cf. Sagart 1999; Baxter & Sagart 2014); another group of the Old Chinese data has a nasal coda: splitting up the two groups into [+NASAL] and [-NASAL] codas seems therefore a good start. Next, the nucleus should also be taken into consideration. Following these organisational principles, we can arrive at Table 5.8.

Table 5.8 gives the impression that we are not far from postulating a number of phonesthemes. For instance, we might state that, for Old Chinese, groups like /ewk/, /Gʷrəp/ or the schwa /ə/, /an/, /aŋ/ and /eŋ/ each *directly* represent ‘light’⁴⁸. But how could we say that it is directly? After all, there is no SOUND correspondence, i.e., no imagic iconicity between the source domain (LIGHT) and the target (the phonesthemes).⁴⁹ What is worse, is that the groups seem to change across different language stages. For instance, the [-NASAL] codas splinter in /Vk/, /ak/ but also /ew/; the [+NASAL] codas largely stay the same: /an/, /aŋ/ and /eŋ/. In Mandarin, then, groupings like /jV/ and /Cwo/, vs. /an/, /aŋ/ and /eŋ/ are possible.

Admittedly, the preceding phonology has some rough edges, but we demonstrated that it is not hard to create the illusion that one is dealing with phonesthetic mappings, based on iconicity, without this actually being the case. Certainly, types of iconicity may be identified. For example, if it [+OPEN] and [+LOW] significantly correlate with ‘more light, super

⁴⁷A number of ideophones were hard to classify in these larger groups. They are: 軒 xuān < MC xjon < OC *qʰar, 眺 yán < MC yemX < OC *N-ram?, 熊 xióng < MC hjuwng < OC *C.Gʷəm, 扱 hù < MC huX < OC *m-qʰa?, 閃 shǎn < MC syemX < OC *s.tem?, 瞩 hào < MC hawX < OC *gʷu?, and 眇 jiǎo < MC kaewX < OC *kʷraw?.

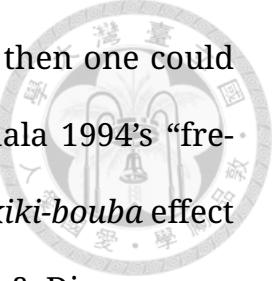
⁴⁸For what it is worth, one could also say that /-art/ in English *light*, *bright*, *sight* etc. are phonesthetic, but they are not. They come from different etymological roots, respectively Proto-Germanic *leuhtą, * berhtaz and *sihtiz.

⁴⁹But do note that a number of phonestheme proponents would gladly call it a day after grouping things like this and running some cursory statistics.



Table 5.8: Reconstructions to Old and Middle Chinese

	coda = obstruent	coda = nasal
nucleus = e	<p>爚 <i>yuè</i> < MC yak < OC *lewk</p> <p>熿 <i>yào</i> < MC yewH < OC *lewk-s</p> <p>耀 <i>yào</i> < MC yewH < OC *lewk-s</p> <p>灼 <i>zhuó</i> < MC tsyak < OC *tewk</p> <p>犖 <i>luò</i> < MC lak < OC *r^fewk</p> <p>燦 <i>shuò</i> < MC syak < OC *r^fewk</p> <p>鑠 <i>shuò</i> < MC syak < OC *r^fewk</p>	<p>瑩 <i>yíng</i> < MC hweng < OC *N-q^{w^fenj}</p> <p>熒 <i>yíng</i> < MC hweng < OC *N-q^{w^fenj}</p> <p>晶 <i>jīng</i> < MC tsjeng < OC *tsenj</p> <p>錚 <i>zhēng</i> < MC tsreang < OC *ts^fenj</p>
nucleus = ə	<p>熠 <i>yì</i> < MC yik < OC *G^wəp</p> <p>煜 <i>yù</i> < MC yuwk < OC *G^wrəp</p> <p>燁 <i>yè</i> < MC hip < OC *G^wrəp</p> <p>燐 <i>yè</i> < MC hip < OC *G^wrəp</p> <p>瞞 <i>yè</i> < MC hip < OC *G^wrəp</p>	
	<p>煇 <i>wěi</i> < MC hjiw^fX < OC *G^wəj?</p> <p>韙 <i>wěi</i> < MC hjiw^fX < OC *G^wəj?</p>	
nucleus = a	<p>暉 <i>huī</i> < MC xjwij < OC *q^{wh}ər</p> <p>輝 <i>huī</i> < MC xjwij < OC *q^{wh}ər</p> <p>暉 <i>huī</i> < MC xjwij < OC *q^{wh}ər</p>	<p>燦 <i>càn</i> < MC tsanH < OC *ts^{h^fan-s}</p> <p>爛 <i>làn</i> < MC lanH < OC *r^fan-s</p> <p>熳 <i>màn</i> < MC manH < OC *m^fan-s</p> <p>漫 <i>màn</i> < MC manH < OC *m^fan-s</p> <p>縵 <i>màn</i> < MC manH < OC *m^fan-s</p>
		<p>晃 <i>huǎng</i> < MC hwangX < OC *G^{w^fanj?}</p> <p>光 <i>guāng</i> < MC kwang < OC *k^{w^fanj}</p> <p>煌 <i>huáng</i> < MC hwang < OC *G^{w^fanj}</p> <p>亮 <i>liàng</i> < MC ljangH < OC *ranj-s</p>



bright', and [+CLOSED] or [+HIGH] with 'a speck of light', then one could convincingly argue that synesthetic sound symbolism (Ohala 1994's "frequency code"), also known as size sound symbolism or the *kiki-bouba* effect (Köhler 1929; Ramachandran & Hubbard 2001; Lockwood & Dingemanse 2015a) underlies these two opposite tendencies.

Similarly, for [+NASAL] codas one could say that even in this small sample, there is a lot of /n/ and /ŋ/ going on. Unfortunately, these two codas are also very common in Sinitic languages, meaning that even if we identify a small cluster of words or features that relate to a given meaning, there will be a host of other meanings which also could be argued to be phonosemantic. The take away, then, is that it is hard to prove iconic relations in modern languages, let alone in different stages. Now this does not mean we should despair. It is not because we were unable to find such motivated, statistically significant correlations that such networks are without linguistic or cognitive value. In fact, this kind of systematicity or maybe even conventional sound symbolism can be of great interest, the main reason being cognitive mechanisms like entrenchment and frequency effects. As a recent overview states:

A fundamental insight, which is paralleled by evidence from the study of language change [...], is that the repetition of identical tokens in the input (known as *token frequency*) results in increased entrenchment in terms of the strength of the corresponding specific representation, whereas repetition of varied items sharing commonalities of form or meaning (*type frequency*) facilitates

categorization, abstraction, generalization, and the emergence of variable schemas (Abbot-Smith & Tomasello 2006; Goldberg 2006; 2009; Lieven & Tomasello 2008:174; Matthews et al. 2005; Tomasello 2003:173–175).

Schmid (2017:14)

In other words, while we may not have found the strong phonesthemic links between form and meaning that studies in that field often claim, there definitely are the smaller scale clusters, which facilitate the formation of small-scale schemas. Perhaps this suffices. After all, if the basic tenet of construction grammar approaches is correct, most of our linguistic knowledge and usage is mediated through constructional assemblies of form and meaning, often with open or half-open slots, which allow for degrees of creativity (Goldberg 2013).

In sum, in this section we have briefly explored syllables belonging to ideophones situated in the semantic domain of LIGHT. In other words, I have touched upon phonological formal variation versus a fixed (abstracted) meaning, in search of possible phonesthemes. However, as the theoretical sketch of this field in Section 5.1.1 already showed, it is hard to prove that phonesthemic links exist, especially when the clustering groups change over time. It may be argued that these two clusters somewhat resemble prototypes, but this is perhaps not maintainable, as the division in [+NASAL] and [-NASAL] was a methodological choice. Perhaps it is better, then, to just stick to the smaller scale systematic networks of types as they were roughly identified above, given their cognitive-psychological

benefits in the form of phenomena like prototypicality, entrenchment, schematization, and frequency effects. In the following sections of this chapter, we will examine these phenomena, while taking the written form as the constant (as opposed to meaning), and the referential meanings as the variation.

5.1.3 Data of the current study

As mentioned at the end of last section, we want to explore the notions of prototypicality and frequency effects in the diachronic semantic development of LIGHT ideophones. Instead of swallowing this whole field in one gulp, we will recycle a number of fully reduplicated items from the last case study. More specifically, we will only treat those ideophones that have an obstruent coda in their Old Chinese reconstruction. The focus lies on their full reduplication forms, since full reduplication (BB) is the most prototypical way of forming ideophones in Middle Chinese (and also Old Chinese), as we know from before (Chapters 3 and 4, but see also Mok 2001; Van Hoey 2015)

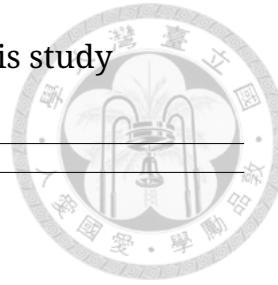
Let us then first turn to the types that will be investigated in this study, shown in Table 5.9. For brevity in illustration, we have opted for the definition provided by the Taiwanese Ministry of Education (MOE) Dictionary (in Hǎndiǎn 漢典 2004–2018), rather than the more extensive *Hànyǔ dà cídiǎn*

⁵⁰ or other dictionaries.⁵¹ It is clear that these dictionaries treat these words

⁵⁰ Although these *Hànyǔ dà cídiǎn* definitions are available in CHIDEOD.

⁵¹ The *Shuōwén jiězì* 說文解字 glosses most of the characters ('words') as either meaning 'light' (*guāng yě* 「光也」) or 'shining' (*zhào yě* 「照也」), with only occasionally mentioning the source of the light, e.g. *diànguāng yě* 「電光也」 lightning beam'. The *Kāngxī zìdiǎn* 康熙字典 adopts these explanations.

Table 5.9: Ideophone types used as material in this study



characters	Hanyu pinyin	MOE dictionary	Kroll (2015)
熠熠	yìyì	閃亮光耀的樣子。	vividly bright
煜煜	yùyù	光明照耀的樣子。	burning brightly, flamboyant
燿燿	yàoyào	光明的樣子。	flashing, sparkling, glittering
耀耀	yàoyào	光明的樣子。	flashing, sparkling, glittering
爚爚	yuèyuè	光明的樣子。	flashing and flickering, blazingly bright
灼灼	zhuózhuó	花茂盛鮮明。明亮。	evident, brilliant, aglow, vivid and vibrant, brightly blazing, plain and patent
熳熳	shuòshuò	光閃動的樣子。	flashing, flaring, effulgent, alight, rutilant; splendidous
鑠鑠	shuòshuò	光明閃耀的樣子。	polished, gleaming; shining; glittering, flashing
聳聳	luòluò	事理分明的樣子。	manifestly evident; conspicuous, outstanding
燁燁	yèyè	光明磊落的樣子。	光明磊落的樣子。
輝輝	yèyè	光鮮明亮的樣子。	brightly shining; flashing, flaring, gleaming
顯赫		顯赫的樣子。	
熒熒	yèyè	NA	NA
曄曄	yèyè	盛大的樣子。	brightly shining; flashing, flaring, gleaming
煒煒	wéiwéi	光彩極盛的樣子。	swirling or globed flames; bright shimmer
韁韁	wéiwéi	光明盛大的樣子。	vividly dazzling, gorgeously glistening
暉暉	huīhuī	NA	fire-red, blazing brightly; splendid; brilliant
輝輝	huīhuī	NA	radiant illumination, glow; splendor, brilliance
暉暉	huīhuī	晴朗的樣子。	radiant, gleaming; vividly white, cendent; gloze, spread light; dazzle

as (near-)synonyms. However, below it will become clear that their meanings do differ in their semantic preference or collocational habits, prime indicators of their meaning.

5.2 Methodology

The methodology of this case study, relies on the convergence of a number of lexical semantic frameworks. We first survey three language-particular treatments of ideophones. That is, we will look at one case of Pastaza Quichua ideophones and two of Japanese mimetics. Next, the theoretical foundation is expanded by incorporating Diachronic Prototype Semantics. The resulting blend will provide fertile ground for four case studies, displaying frequency effects, variation and prototype effects across time.

5.2.1 Unifying three Cognitive Linguistics definitional frameworks

One of the first Cognitive Linguistics-oriented frameworks for discussing the semantics of ideophones is represented by Nuckolls's ongoing studies of Pastaza Quichua ideophones (Nuckolls 1996; 1999; 2001; 2010; 2014; 2019; Nuckolls et al. 2017). She has long been a proponent of using IMAGE SCHEMAS as the main way of representing an ideophone's most basic meaning. In Nuckolls et al. (2017), she revisits some of her case studies to stress their dynamic and multimodal nature. For instance, *polang* is shown to be afforded by either a 'glide across water' or 'glide up to the surface of the water' schemas, as is shown in Figure 5.1. It is important to note that image schemas here seem to be interpreted more according to Johnson's (2005) rich semantic scenarios than the slightly more abstract usage found in most discussions on the term (Hampe & Grady 2005; Oakley 2007). Nevertheless, the difference between the two is only one of granularity and abstraction.

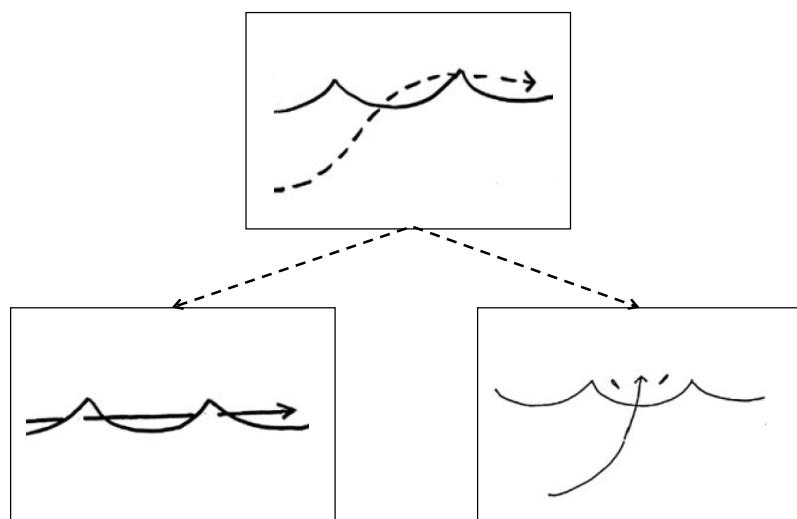


Figure 5.1: The image schema representation of *polang* (adapted from Nuckolls et al. 2017:163–168)

A second framework is Lu's (2006) treatment of Japanese (and Mandarin

Chinese) mimetics or ideophones. Lu stresses the scenario or script nature of ideophones and thus uses an IDEALIZED COGNITIVE MODEL (ICM) approach (Lakoff 1987). Her main examples are Japanese *korokoro* 可口可口 ‘(something small) rolling’ and *gorogoro* 戈口戈口 ‘(something large) rolling’. She shows that the dimensions of the object, as well as the manner of movement etc. can be abstracted into an ICM that could be used to define the meaning of these lexemes, see Figure 5.2.

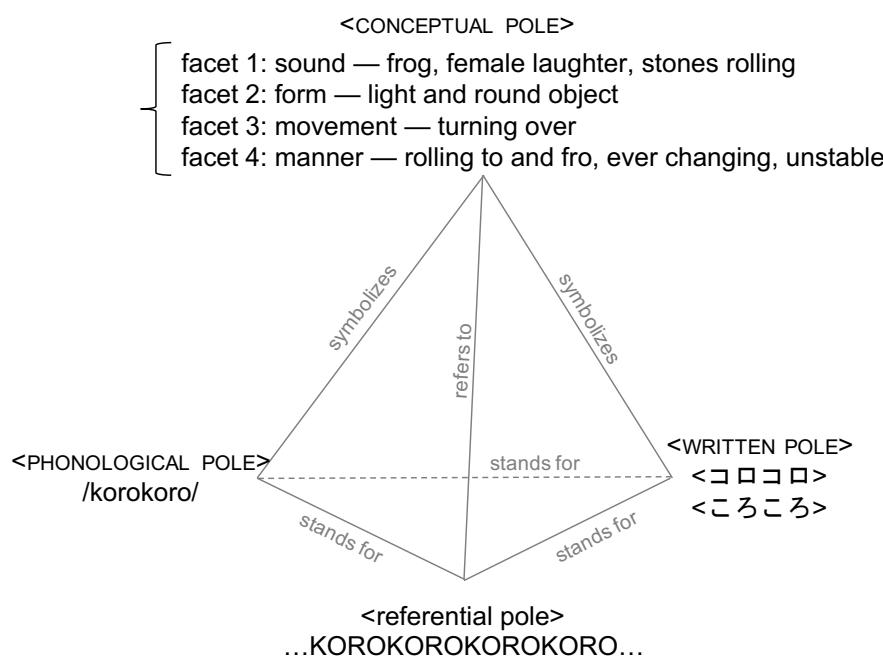


Figure 5.2: The ICM of *korokoro* (adapted from Lu 2006:133)

The third framework is represented by Akita: FRAME SEMANTICS. Akita argues that Japanese ideophones tend to be highly specific in their semantics and invoke highly specialized frames, with little semantic extension (Akita 2012b). More specifically, Akita demonstrates that Japanese mimetics have a higher frequency of instantiating related frame elements (67-68) than the non-mimetic words that were considered (Akita 2012b:83). This is contrasted with the way SOUND ideophones are depicted in Chinese, slightly

vague and with many possible referents (Akita 2013b), as illustrated in (66).

(66) Vague Chinese onomatopoeia (Akita 2013b:26)

cī~lā~cī~lā 刺啦刺啦 means ‘ripping a thin wooden board’ ‘sound of a saw’

jī~jī~chā~chā 嘍嘍喳喳 means ‘chatter’ ‘chirp’

pā~pā~pā~pā 啪啪啪啪 means ‘keys dangling’ ‘noisy footsteps’

However, as soon as we leave the imagic iconicity of SOUND ideophones to more diagrammatic end of the iconicity spectrum (Dingemanse 2012), it will become untenable to state that Chinese ideophones are only vague, and not polysemous, as we will illustrate with LIGHT ideophones. In fact, even for SOUND ideophones the line between polysemy and vagueness is not clear at all, as has been shown before with prosaic lexical items like *to pain* (Tuggy 1993; see also Geeraerts 2006c). Related to this is that it seems highly unlikely that Japanese would not display vagueness for its LIGHT ideophones, especially for the two well-known ideophones *kira~kira* キラキラ ‘sparkling, twinkling’ and *pika~pika* ピカピカ ‘sparkling, twinkling’.

(67) *sutasuta* スタスタ ‘(walking) briskly’ collocates with the following through an Inheritance relation:

- a. the verb *aruku* 歩く ‘walk’, an instantiation of the SELF_MOTION frame;
- b. the verb *iku* 行く ‘go’, an instantiation of the general MOTION frame.

- (68) *kotsukotsu* コツコツ ‘tap’ collocates with the following through Core-Element-Instantiating relation:

- a. the noun *tataku* 烫く ‘hit’, an instantiation of the IMPACT frame;
- b. the verb *tobira* 扇 ‘door’, a core element of the IMPACT frame, i.e. IMPACTEE.

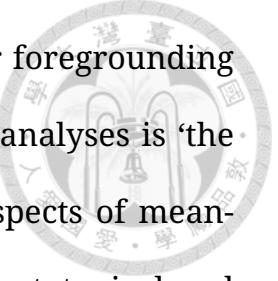
In a later version, Kiyama & Akita (2015) represent the meaning of mimetics through a bundle of frame elements (attribute value matrix notation), based on Osswald & Van Valin (2014). For *suta~suta* ‘walking briskly’ they give Figure 5.3. Since the goal of Kiyama & Akita (2015) is to explore the gradability of mimetics⁵², they identify these frame elements using statistical tests. They also note that it is confidence which causes the speed to be quick, and the sound inconspicuous (quiet) which disallows noisy shoes to be used with this specific mimetic.

BRISK_STEPS	
SELF_MOVER	[1] [+sentient]
AREA	[2]
MTR_PTN	walking
INNER_STATE	confident
SPEED	quick
PATH_STABILITY	stable
SOUND	inconspicuous
SHOES	normal

Figure 5.3: Frame semantic representation of *suta~suta* walking briskly, adapted from Kiyama & Akita (2015)

The three frameworks briefly discussed above all agree on the encyclo-

⁵²For Chinese ideophones, this unfortunately falls outside the scope of the current dissertation, although there is much future potential in the identification of frame specific elements.



pedic nature of ideophones, their multimodality, and their foregrounding markedness. However, it is hard to say that any of these analyses is ‘the right framework’, because they all emphasize different aspects of meaning description. Image schemas try to capture the most prototypical and essential core meaning (similar to Tyler & Evans’ (2003) notion of the so-called PROTOSCENE for prepositions); frame semantics, conversely, stressed the high specificity of ideophones; and idealized cognitive models (ICMs) occupy the middle ground and are perhaps most flexible, but only because they are flexible by nature.

However, maybe we do not have to choose. In the field of conceptual metaphor, it has been proposed that metaphors can be studied on different levels of granularity (Kövecses 2017): on the lowest level there are MENTAL SPACES, which connect the metaphors online (working memory) as the discourse happens dynamically. One level higher there are frames and domains. Domains are “not analogue, imagistic patterns of experience but propositional in nature in a highly schematic fashion” (Kövecses 2017:325), while frames elaborate particular aspects of the domain matrix. They are all elaborations of the highly abstract image schemas, which are directly meaningful pre-conceptual structures, that are highly schematic Gestalts, consist of continuous analogue patterns and have an internal structure consisting of only a few parts (Kövecses 2017:324). Figure 5.4 below attempts to capture these different levels of metaphor in a diagram.

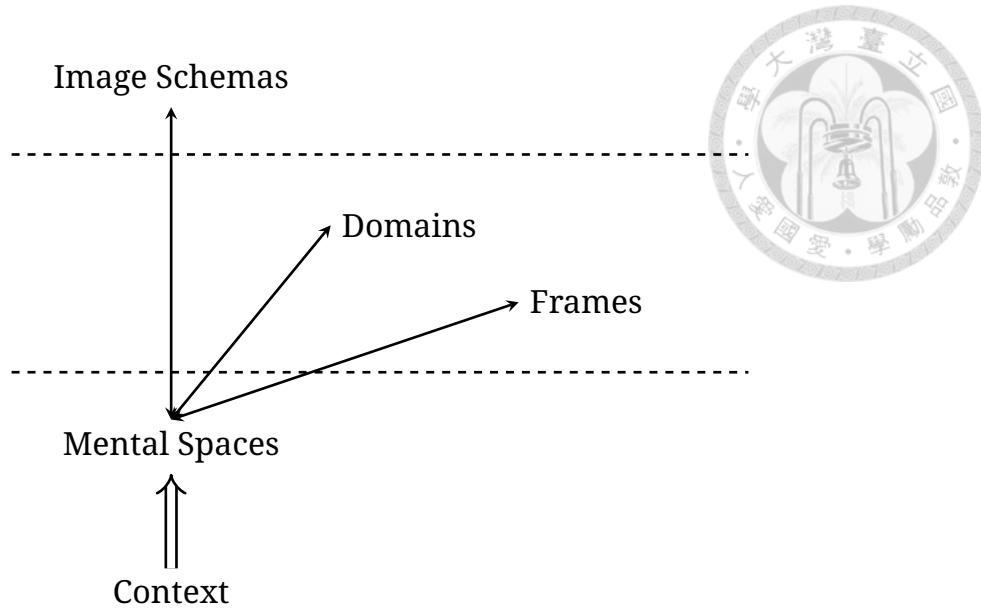


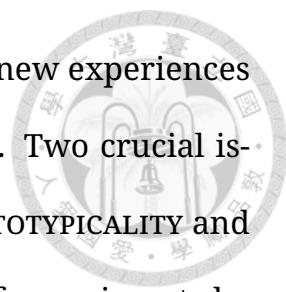
Figure 5.4: Levels of Metaphor (adapted from Kövecses 2013)

There is enough⁵³ indication that this model (Kövecses 2017) may be of use to the current study. It should be noted, though, that his domains are somewhat similar to Lu's ICMs, which is why I will use both terms interchangeably. One piece of the methodology is still missing: how can we identify and represent data, especially in the lowest level (mental spaces). For this issue we must turn to diachronic prototype semantics.

5.2.2 Diachronic prototype semantics

One of the big themes that propelled the Cognitive Linguistics movement is the attention it devotes to semantics. As the semanticist Dirk Geeraerts explains: “language is seen as a repository of world knowledge, a structured

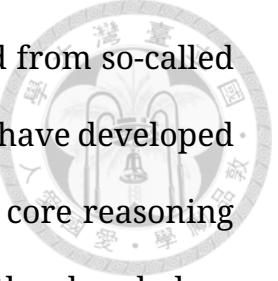
⁵³Increasingly, it seems that ideophone studies can learn many things from advances made in metaphor theory, as they both deal with imagery (in the Langackerian sense, cf. Langacker 1987a) and the interplay between perception and conception (cf. Talmy 2000a on “ception”). This can also be seen by the sheer volume of studies that explore ideophones (as well as sound symbolism) in genres like poetry, like Webster’s (2014; 2017) analyses of the Navajo poet Rex Lee Jim, or Hiraga (2005), who interweaves blending theory with poetry and mimesis. My near future work also includes the exploration of multimodal metaphor theory and ideophones, which will be presented together with Iju Hsu at the 13th RaAM virtual conference (June 2020). Also in other fields, e.g. humor, connections with ideophones are being made (Dingemanse & Thompson 2020).



collection of meaningful categories that help us deal with new experiences and store information about old ones" (Geeraerts 1997:8). Two crucial issues in this theory of categorization are the notion of **PROTOTYPICALITY** and that of **POLYSEMY**. Prototypicality emerged from a series of experiments by psycholinguist Eleanor Rosch (1975; 1975) as a theory of categorization that opposed the traditional categorical definitions which required necessary and sufficient conditions. Rather, it was shown that some members of a lexical field are better representatives of it than others, e.g. the robin is the most prototypical bird in Anglo-Saxon culture, because it has wings, feathers etc. On the other hand, penguins, ostriches and the like are less typical representatives of 'bird'. Now the question that relates this very brief summary to our study is, can we find the same prototypical structure in the semantics of **LIGHT ideophones**?

This question presupposes that an ideophonic item is **POLYSEMOUS**, i.e., a series of interrelated meanings that are activated according to the situation they are used in. Most Cognitive Linguists appear to reject the idea that there is a single meaning for a given lexical item (the monosemy hypothesis), instead opting for the polysemy hypothesis. But if there is polysemy, how can we discern it from vagueness? Useful discussions of the different approaches to the phenomenon can be found in Tuggy (1993) and Geeraerts (2006d; 2010b), to name just two.

Let us illustrate this with a classic example: fruit. In a monosemous approach, such as the idealist Natural Semantic Metalanguage developed by Wierzbicka (1992; Goddard & Wierzbicka 2014b), which we have briefly



met in Section 3.1, they would give a definition constructed from so-called semantic primes, concepts that exist in every language they have developed the theory for. The idea is to get at the core experience or core reasoning people use when they talk about fruit. Geeraerts, on the other hand, does not agree with this and stresses the prototypical polysemous structure of a word such as fruit. It is polysemous because it has at least the meanings ‘something that people can eat and that grows on a tree or a bush’ and ‘the result or effect of something’. However, it is also vague, with regard to the differences between oranges and watermelons, because those differences do not lie at the basis of a distinction between senses (Geeraerts 1997:18–19).

We will not go in depth into the distinctions between the two approaches, as this is a well-known fundamental difference in perspective (see for instance Geeraerts 2006c for a detailed discussion), but follow Geeraerts’s position of “polysemy with prototypicality features”, as it is more practically applicable to our dataset. This similarity is especially evident when compared to the approaches used in his volume on diachronic prototype semantics, in which he discusses the prototypical nature of the concept of ‘prototypicality’ itself (Geeraerts 1997:22), and some of the differences in semantic changes that follow from it by stressing different parts of the concept. We present it here in the revised version (Geeraerts 2010b:189), in Table 5.10 (first shown in Table 1.2).

These four different types of prototypicality effects are put in a diachronic perspective by Geeraerts (1997) and in a synchronic perspective

Table 5.10: Four types of prototypicality effects (as shown in Table 1.2)

	Extensional characterization (on the level of exemplars)	Intensional characterization (on the level of definition)
Non-equality <i>(salience effects, core/periphery)</i>	(a) differences of typicality and membership salience	(b) clustering into family resemblances
Non-discreteness <i>(demarcation problems, flexibility)</i>	(c) fuzziness at the edges, membership uncertainty	(d) absence of necessary and sufficient definitions

in Geeraerts (2010b). For diachronic studies (1997:23), the four effects are as follows (69a-69d):

- (69) a. By stressing the extensional non-equality of lexical-semantic structure, prototype theory highlights the fact that changes in the referential range of one specific word meaning may take the form of modulations on the core cases within that referential range.
- b. By stressing the intensional non-equality of lexical-semantic structure, prototype theory highlights the clustered set structure of changes of word meaning.
- c. By stressing the extensional non-discreteness of lexical-semantic structure, prototype theory highlights the phenomenon of incidental, transient changes of word meaning.
- d. By stressing the intensional non-discreteness of lexical-semantic structure, prototype theory highlights the encyclopedic nature of changes in word meaning.

The effect that is of most interest to our current study is number (69b). When this aspect of prototypicality is stressed, the overall configuration of

the various readings of a word comes to the fore. This highlights two phenomena: first, the overlapping and interlocking of the different readings, with attention for the different starting points in existing meanings a novel meaning may have. Second, the differences in structural weight among the different meanings of an item. Some meanings do not survive very long, other, more important (and prototypical) meanings persist through time. Geeraerts (1997:47–62) shows this through the analysis of the Dutch verb *ver-grijpen* ‘mis-take’. After showcasing different meanings in different historical contexts, he summarizes the data in a very nice visualization, first shown in Geeraerts (1983) and later in Geeraerts (1997), presented in Figure 5.5.

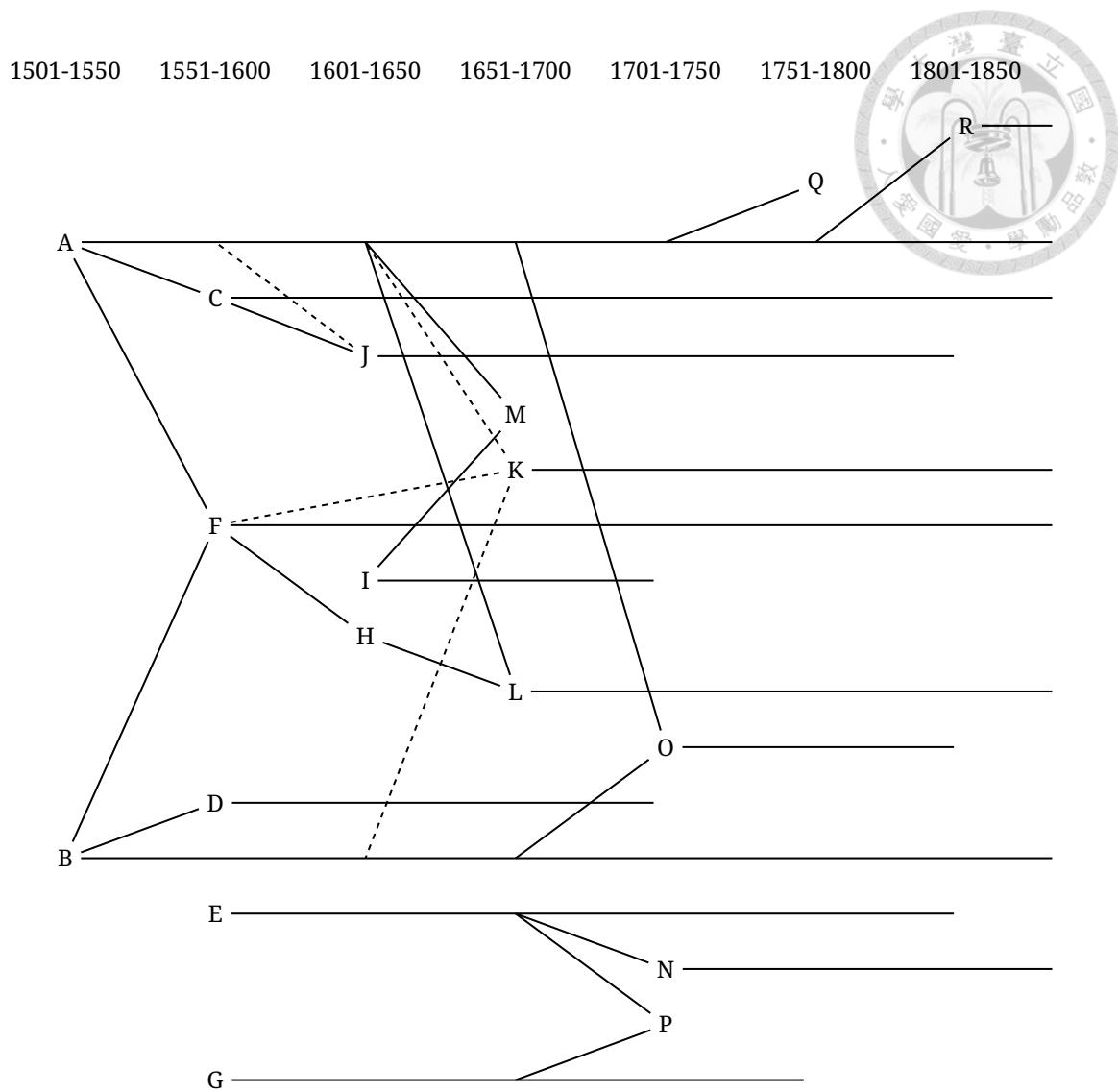
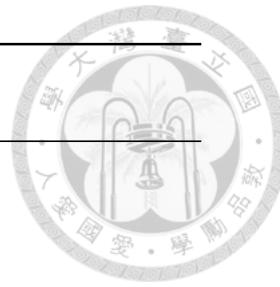


Figure 5.5: *Vergrijpen* (Adapted from Geeraerts 1997)

Table 5.11: the different meanings of *vergrijpen* in Figure 5.5

label	definitional gloss
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



label	definitional gloss
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)

From Figure 5.5 above, three main observations can be made. First, new meanings arise from the joint influence of several meanings, e.g., meaning F ‘to do something forbidden’ has its conceptual starting points in A ‘to use physical violence against (someone)’, as well as B ‘to oppose someone to whom one owes respect and obedience’. Second, some meanings crop up occasionally, but do not persist in time, such as Q ‘to commit suicide’ can be considered as an extension of A, but it is used in only one time period, 50 years in this case. Third, not all concepts are equally important in the process of semantic change. For instance, A, B, C and F are more important than meanings E and G. This diagram reveals much about the development

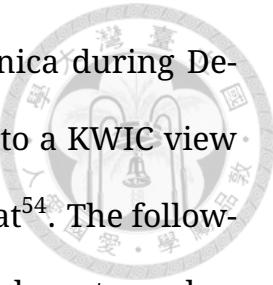
of semantic structure, and it will be used as a model to further investigate the data on LIGHT ideophones.



5.3 Mental spaces and Frames: corpus-based case studies

To study the items presented in Table 5.9 the Scripta Sinica (Academia Sinica 中央研究院 2015) was consulted. In the visualizations below, a timeline is added to every visualization to orient the reader who is not familiar with Chinese dynastic history, since this corpus largely follows the traditional periodization, see Section 3.3.1. This has two consequences: in comparison with Geeraerts, it was not possible to do the same fine-grained analysis as he did (periods of 50 years). On the other hand, it did make it possible to see the bigger ‘macro evolutions’ of the semantic networks. In the context of the bigger Chinese historical corpus, the second evolution is the most relevant for current purposes.

To determine different senses for different items, we keep in mind Tyler & Evans’ (2003:38–45) methodology, because their method allows for polysemy, while trying to avoid the so-called ‘polysemy fallacy’, i.e., positing more meanings than there actually are. In this regard, the basic meaning of LIGHT may be maintained on the most abstract level of image schemas but in the lowest level of mental spaces it is untenable, as semantic preference (Geeraerts 2010b:170–173) clearly demonstrates, e.g., *yè~yè* 瞳瞳 co-occurs mostly with ‘plant’-like meanings, as well as ‘light’ sources. This would count as at least two frames (a PLANT frame and a LIGHTSOURCE frame), divided in many mental spaces. This will be illustrated below.

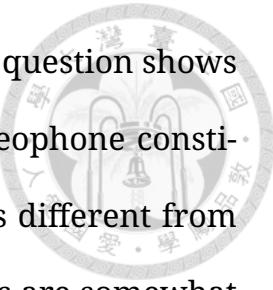


The data were manually extracted from the Scripta Sinica during December 2016 - January 2017. After collecting these data, into a KWIC view of them was obtained, which I saved in a spreadsheet format⁵⁴. The following case studies are based on some 3500 tokens, which is a decent number of examples to analyze manually. However, the distribution of token frequency was not equal, e.g., *zhuó~zhuó* 灼灼 was the most frequently used ideophone in our data set. See Figure 5.11 for a comparison against three other ideophones. In order to consider a meaning as prototypical we had three criteria that interacted with each other: whether the meaning persisted through time; whether it was productive for other meanings; and whether it occurred in the data with high frequency. Given the unevenness in the distribution, we have generally considered a tally of 5 occurrences of a given meaning in a given period as ‘high(er) frequency’. Of course, compared to really frequent items, e.g. closed-class items, this pales in comparison. However, for ideophones it counts as high enough; it should be borne in mind we rely on historical sources.

5.3.1 Divergent networks

Below in example (70), some examples of *yuè~yuè* 燮爚 are presented. The *Hànyǔ dà cídiǎn* 漢語大詞典 gives as the sole meaning of this ideophone ‘radianc’ 光采耀目貌。. Some of the meanings found in the dictionary and in the data are shown in (70). The dictionary definition seems adequate to capture the core meaning, but the point is that this in some ways is under-specified. The meaning of *yuè~yuè* is elaborated by the collocate. In each

⁵⁴This has the consequence that we took collocates of 15L and 15R into consideration.



case, the slightly different translation for our ideophone in question shows how the interplay between different collocates and the ideophone constitute subtle differences in meaning: the radiance of stars is different from that of the moon, or the metaphor in eyes. However, things are somewhat complicated because the collocate will not always occur right next to the ideophone.

- (70) a. “The six masters went into pursuit, the beasts were shocked. Shooting and *flashing*, they ran like thunder and sped like lightning.” 六師發逐，百獸駭殫。震震爚爚，雷奔電激。in 後漢書 · 班彪列傳 (Jin dynasty)
- b. “The eyes radiant, *brilliantly shooting out*” 眼有光芒，爚爚外射。in 《搜神記》 (Jin dynasty)
- c. “Clouds floating over, almost indescernable, *gleaming* moonlight, full mellow wine.” 浮雲吹盡數秋毫，爚爚金波滿滿醪。in 中秋夕寄平甫諸弟 (Song dynasty)
- d. “In the court of the Southern Pole, the Old Man’s star [= Canopus], it’s glittering, *twinkling*, shining and gleaming.” 南極之庭老人之星 煙煙爚爚煌煌熒熒 in 老人星賦

Based on these data it can be seen (Figure 5.6) that there is a clear prototypical core from its earliest occurrence. The main item *yuè~yuè* describes, is ‘lightning’ or ‘thunder’, in example (70a) a metaphor for the speed of pursuit. At the same time, there is a meaning of ‘eyes (shooting)’, which can be interpreted as a metaphor. Around the Tang dynasty, more ‘light’ meanings

develop, related to the prototypical ‘lightning’: ‘light’ (in general) and ‘stars’.

From these onward there are usages that involve ‘pearls’ and ‘moonlight’.

The latter comes through a metaphor of *jīn-bō* 金波 ‘golden waves > moon-light’. There is a large hiatus between the Yuan dynasty and the resurfacing usage of LIGHT in the beginning of the 20th century. We presume LIGHT has taken over the main prototypical core of *yue~yue* and was still in use, but unrecorded, until it resurfaced.

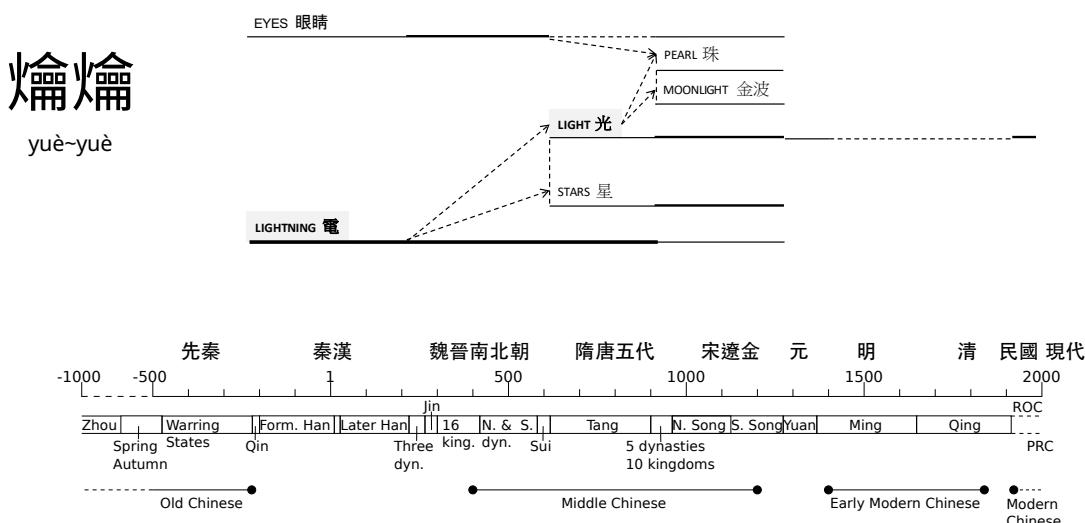


Figure 5.6: (ref:yueyue)

The visualization format in Figure 5.5 follows that of Geeraerts (1997). However, it departs from this method in a number of crucial ways. First, the labels indicate collocations of the ideophones, which elaborate the conceptualization of the studied ideophone. In other words, meaning here is interpreted as an ideophonic item’s relation to a referent. Furthermore, this visualization attempts to improve upon Geeraert’s *vergrijpen* model by including information on token frequency, depicted by line thickness: higher token frequency is thicker and lower frequency is thinner. It should be re-

ported that this has been done somewhat subjectively, depending on how much material was available for each item or group of items. The dotted lines indicate a presumed continuation of a certain meaning. The dotted arrows are used to represent meaning extensions, following Langacker's (2008b:37) convention, see Heath (2014) for a handy overview of his symbolism.

This case study of *yue~yue* shows how the visualizations should be read and what can minimally be inferred from all of them. The network that results from the tokens is quite small. It will become bigger in subsequent case studies, with Figures 5.7 to 5.13 all following the same vizualation format.

5.3.2 Conflation of forms and semantic transfer

In this section, we will investigate three case studies to see how semantic transfer between variants can occur throughout time. The three case studies include a set of two ideophones that share the same phonology, but have different orthography and are marked as synonyms in the dictionary; a set of ideophones that are near-synonyms with the same phonology and somewhat differing orthographic variants; and lastly a set of ideophones that are orthographically dissimilar and only phonologically near-homophones in the earliest stages but have diverged more throughout time.

The first case study is a pair of synonyms, *yào~yào_{FIRE}* 燿 燿 and *yào~yào_{LIGHT}* 耀 耀. As can be seen, there is some variation in the <orthography> of the form *yào~yào*. The main difference is the semantic radical (Section 3.2.2.1) that distinguishes the two: either FIRE 火 or

LIGHT 光. For clarity, I will mark the radicals with subscripts in the English text. Is the variation of the <orthographic forms>, with the <phonological form> as a constant, related to a difference in meaning? It seems that is the case, although one would not be able to infer that from the *Hànyǔ dà cídiǎn*, which lists them as near-synonyms, see (71).

(71) *Hànyǔ dà cídiǎn* definitions for the *yào~yàos*

- a. *yào~yào_{FIRE}* 耀耀: ‘*yào~yào_{FIRE}*, *yào~yào_{LIGHT}*: brilliant, shining’ 耀耀, 耀耀: 明亮閃光貌。
- b. *yào~yào_{LIGHT}* 耀耀: ‘see *yào~yào_{FIRE}*: brilliant’ 參見耀耀: 光明貌。

Using the same identification process as before, we can trace how the different collocating meanings emerge as a network over time. This is presented in Figure 5.7.

Originally, the *yào~yào_{FIRE}* variant is elaborated more by ‘lightning’, while the other variant has ‘light’. They both have ‘stars’ at this point. Interestingly, at some point (ca. Táng dynasty), ‘lightning’ also crops up in the semantic matrix of *yàoyào_{LIGHT}*. So there is a conflation between the two near-synonyms. In data from the same period, different kinds of ‘jewelry’ (such as ‘pearl’, ‘crown’, ‘treasure’) also are described with both *yào~yàos*, although it is more plausible to see a transfer from the LIGHT variant to the FIRE variant, since it has a higher type frequency (cf. next subsection).

There is also the development of metaphorical extensions, describing the illustriousness of people, e.g., the face (72a) for *yàoyào_{LIGHT}* and later a prince (72b) for *yàoyào_{FIRE}*. Such metaphorical extensions have been subsumed un-

der the term ‘posture’ in this and other case studies, because even though they are quite varied in real referents, they are comparable. In any case, this curious exchange of meanings shows the mutual influence related semantics and a related phonological form may have on each other.

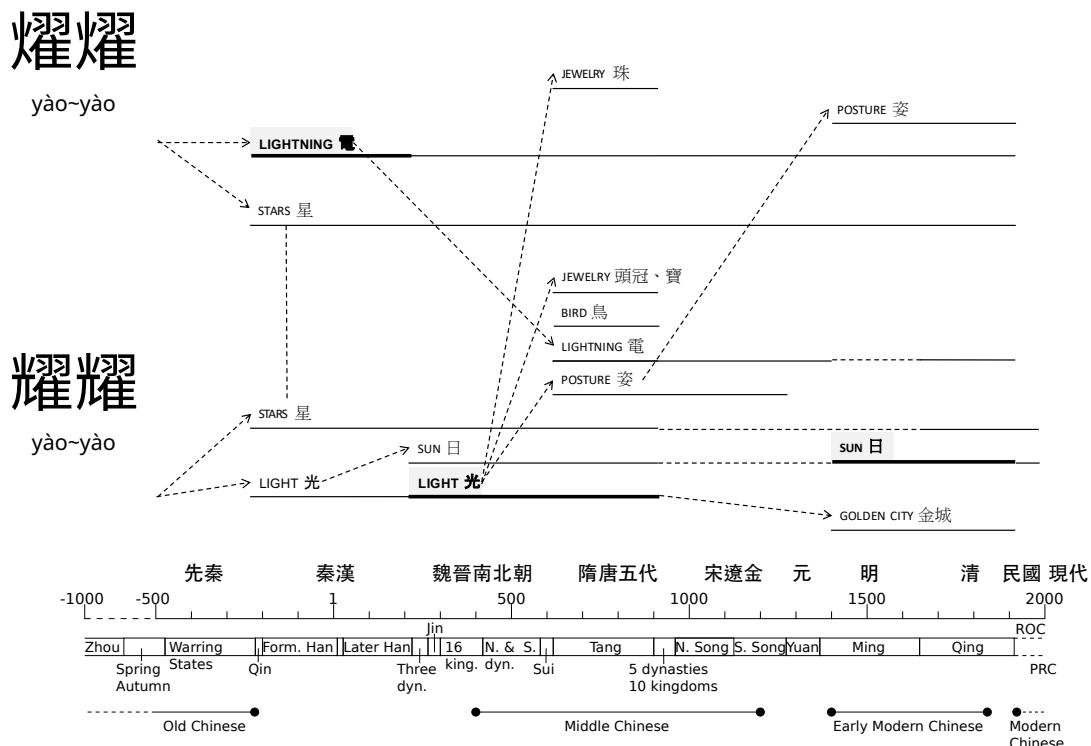


Figure 5.7: **yàoyào** 燿耀 and **yàoyào** 耀耀

- (72) a. “[...about female] shining face, [...]” **耀耀面子** in **遊仙窟**
 b. “[A woman with the name Wu:] in the second year she gave birth to a shining prince.” **吳氏 [...] 二年生耀耀封世子** in **國朝獻徵錄**

Let us explore if other ideophones with the same <phonology> and different <orthography> display similar mutual transfer effects. Instead of two LIGHT-like semantic radicals, we will look at FIRE and METAL, in the case study of *shuò~shuò_{FIRE}* 燦爛 and *shuò~shuò_{METAL}* 鐸鑸. Their dictionary defi-

nitions are shown in (73).

(73) *Hànyǔ dà cídiǎn* definitions for the *shuò~shuòs*



- a. *shuò~shuò_{FIRE}* 燥熾: 1. ‘flashing, brilliant’ 光芒閃動貌; 2. ‘scorching hot’ 酷熱貌。
- b. *shuò~shuò_{METAL}* 鐸鑑: 1. 1. ‘shining, brilliant’ 光芒閃耀貌。; 2. ‘glossy shining’ 油光潤澤貌。

Based on these, we can see that they are near-synonyms, although the FIRE radical variant has a secondary meaning involving heat, and the METAL variant one of a glossy kind of brilliance. Investigating the visualization, shown in 5.8, it becomes clear that in the early part of history, their meanings were quite separate. However, at some point the FIRE variant increases in types, presumably as an extension from the ‘lantern’ meanings. These extended meanings later then also occur in the variant with the METAL semantic radical, where they have limited productivity.

熒熒

shuò~shuò

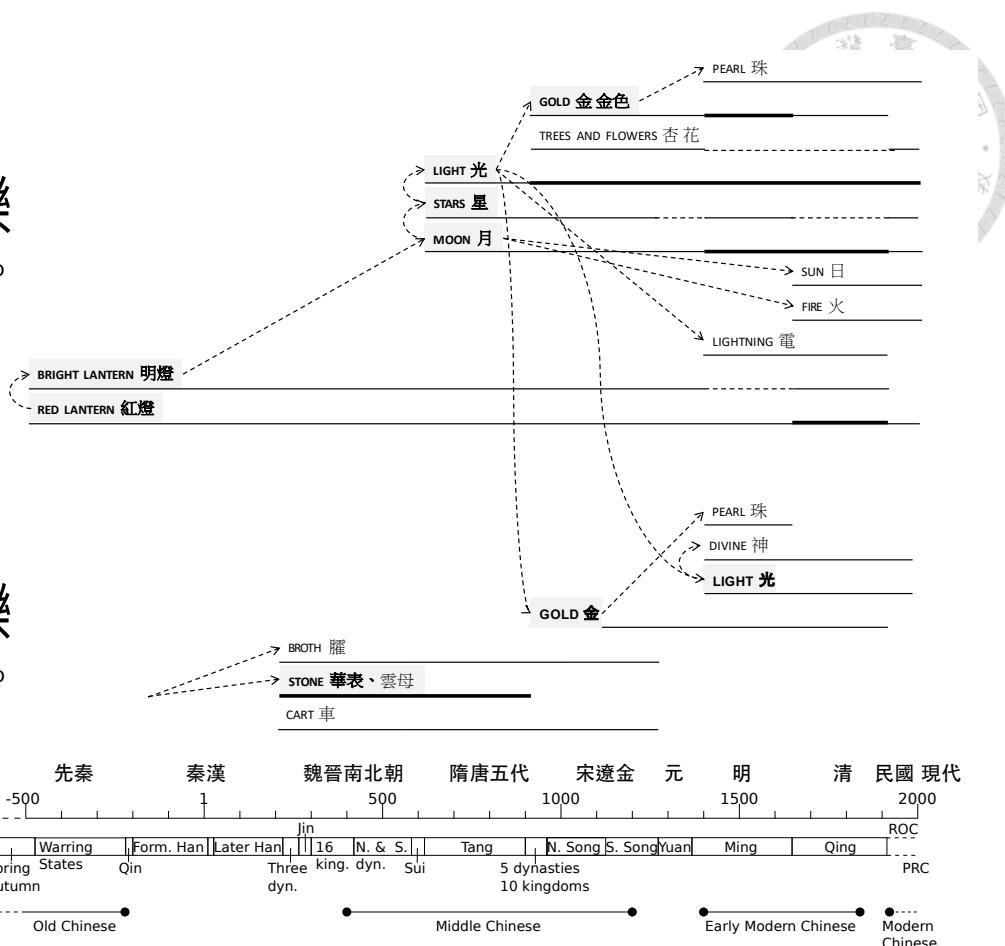


Figure 5.8: *shuò~shuò* 熒熒 and *shuò~shuò* 鐸鐸

In this case, the semantic transfer is not mutual, but directional. This makes sense since METAL is not a radical that one would intuitively associate with LIGHT. If there is conflation between orthographic variants with identical phonology, we may also wonder if similar effects can be noticed for near-homophones as well. In other words, we will be looking at a case study where the orthography is somewhat similar and the phonology too. Geeraerts (1997:163–175) discusses such an example of two near-synonyms in Dutch, *vernielen* ‘to destroy’ and *vernietigen* ‘to destroy’, whose similarity in phonological form also influenced their usage and semantic transfer. But we will be investigating *yì~yì* 熒熒 and *yù~yù* 煙熑.



(74) *Hànyǔ dà cídiǎn* definition for *yì~yì* and *yù~yù*

- a. *yì~yì* 煙熠: ‘bright, twinkling’ 鮮明貌；閃爍貌。
- b. *yù~yù* 煙燭: 1. ‘bright, blazing’ 明亮貌；熾盛貌。; 2. ‘open-hearted and forthright’ 形容胸懷坦蕩。⁵⁵

(75) Reconstructions for *yì~yì* and *yù~yù*

- a. *yì~yì* < MC yik~yik < OC *gwəp~gwəp
- b. *yù~yù* < MC yuwk~yuwk < OC *gwrəp~gwrəp

The dictionary definitions for these two ideophones can be seen in (74), which slightly overlap. With respect to the historical phonology, the two items have very similar forms in Old Chinese, with only an /-r-/ differentiating them, as illustrated in (75). As Figure 5.9 shows, the earliest attestations share a collocate in ‘light’, but also in vegetation-related items. For *yì~yì* this is pulled more towards flower-like collocates, and for *yù~yù* more towards trees, although later ‘flower’ collocates are observed. ‘Stars’ is a later usage for both items⁵⁶, occurring in the same broad period. However, ‘fire’ jumped from *yì~yì* to *yù~yù*, if indeed these two ideophones were still mutually intelligible enough in that period. The two definitely seem to have had their own clusters of meanings, but when a collocate like ‘fire’ appears in the semantic matrix of one of them and then later in that of the other, where it is found for a given duration and then disappears without a trace, this *can* be interpreted as a dynamic conflation between these two items.

⁵⁵I did not find this meaning attested in my data.

⁵⁶A curious observation, since STARS is a normal collocate in contemporary Mandarin, as I found out when I attended a concert called *xīng-guāng yì~yì* 星光熠熠 Twinkling starlight in 2016.

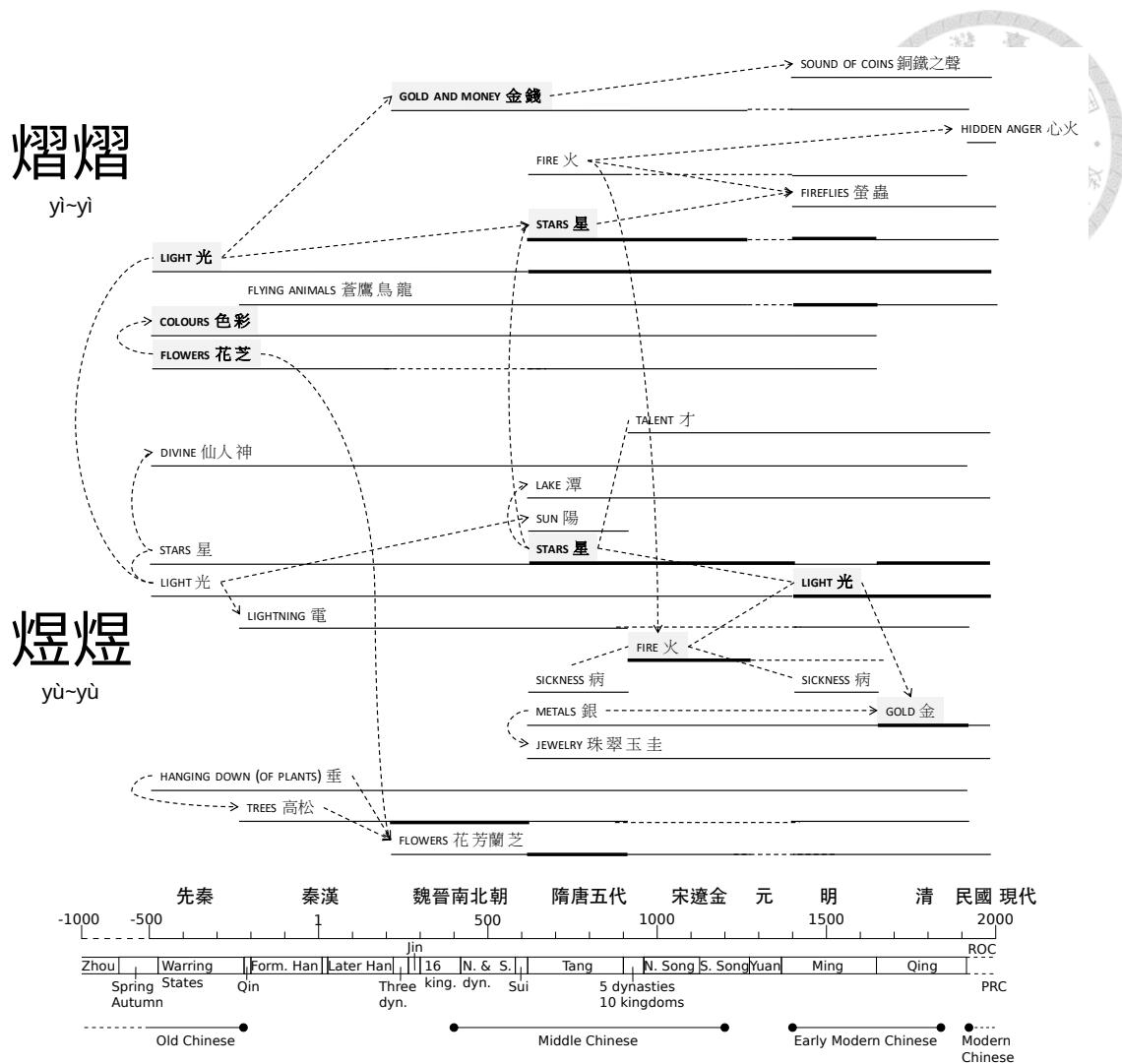
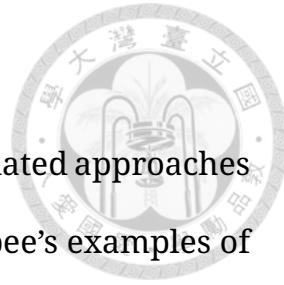


Figure 5.9: *yì~yì* 煙 and *yù~yù* 煙

These two case studies showcase how similar forms can interact with each other in terms of referential meanings. For $yào-yào_{\text{FIRE}}$ and $yào-yào_{\text{LIGHT}}$, or $shuò-shuò_{\text{FIRE}}$ and $shuò-shuò_{\text{METAL}}$, the lesson is that this can be found in <orthographically> similar items with identical <phonology>. For *yì~yì* and *yù~yù* the takeaway is that different <orthography> but similar <phonology> can also display some of these interactions and conflations in the semantic transfer.



5.3.3 Frequency effects

One of the perduring claims of Cognitive Linguistics and related approaches to language is that frequency plays an important role. Bybee's examples of the English past [Bybee & Hopper 2001a; but see also Bybee (1985); Bybee (1988); Bybee & Hopper 2001b among others] provide a clear illustration of the effects frequency can have. Some examples are shown in (76). Most verbs in English form their Past Tense by adding *-ed* to the stem. Consequently, the type frequency of this construction is very high. According to Bybee, the frequency effect associated with type frequency is that it reinforces the productivity of that paradigm.

(76) Regular and irregular verbs in English

- a. regular verbs: *watch-watched, look-looked, play-played*
- b. irregular verbs: *run-ran, be-was, think-thought*

However, compared to (76), most token frequencies of irregular verbs will be higher than those of many regular verbs. According to Bybee & Hopper (2001a), this high token frequency can have two effects. First, phenomena like phonetic change progresses more quickly in items with high token frequency. In the case of these verbs that might mean a reduction in the phonetic form, if the intended meaning is clear enough from context. Second, somewhat paradoxically, token frequency effect is that items with high token frequency may also be more resistant to change. Because the irregular verbs in our example are so frequent, they resist analogical leveling, i.e., forming their Past Tense through *-ed*.

Below we will observe similar things for two sets of ideophones. A set of phonologically identical and orthographically similar variants, reminiscent of the set of two *yào~yàos* discussed in the preceding section, includes three variants of *huī~huī*: *huī~huī_{FIRE}* 燥燥, *huī~huī_{LIGHT}* 輝輝 and *huī~huī_{SUN}* 暉暉.

(77) *Hànyǔ dà cídiǎn* definition for the three *huī~huīs*

- a. *huī~huī_{FIRE}* 燥燥: ‘simplified as *huī~huī_{LIGHT}*: bright’ 燥燥, 辉辉 : 明亮貌。
- b. *huī~huī_{LIGHT}* 輝輝: 1. ‘outstanding’ 顯赫貌。; 2. ‘bright’ 光耀貌。; 3. ‘light’ 亮光。; 4. ‘shiny, glossy’ 光澤，潤澤。
- c. *huī~huī_{SUN}* 暉暉: 1. ‘heat of blazing sun’ 形容日光灼熱。; 2. ‘bright-colored’ 艷麗貌。; 3. ‘clear’ 清輝貌。; 4. ‘onomatopoeia [thunder]’ 象聲詞。

From the dictionary definitions in (77) it can already be seen that there is some overlap between these three variants, but that the FIRE variant has a more general meaning, i.e. it is not very productive, while the LIGHT and SUN variants do seem to have a higher productivity. In terms of collocations⁵⁷, we can visualize the data in Figure 5.10.

⁵⁷I find it hard to translate all of the near-synonymous definitions provided to English, but in this case study, I agree with the *Hànyǔ dà cídiǎn* entries and have not found many different collocates, although I might have grouped them differently in the diagram.

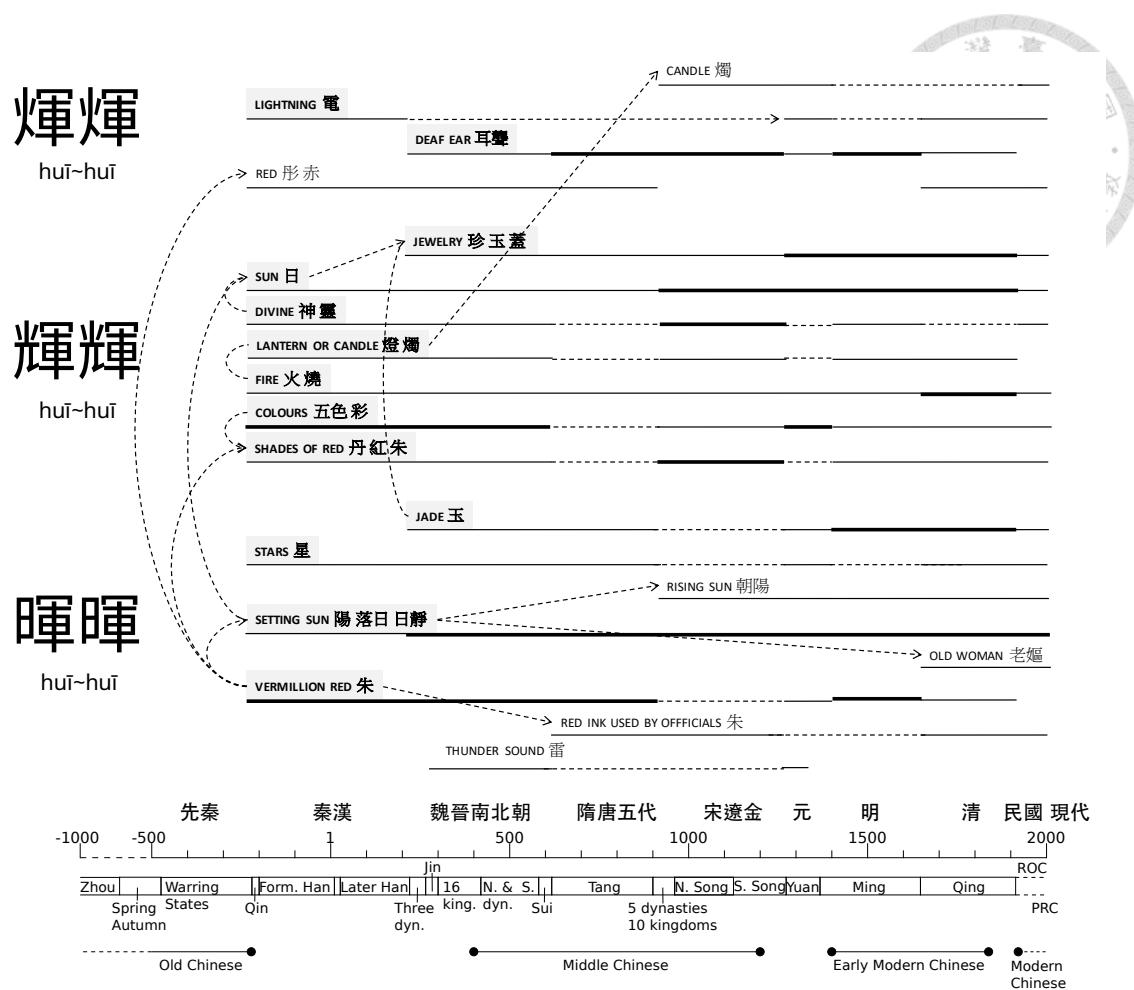


Figure 5.10: *hui~huī* 煙暉, *hui~hui* 輝暉 and *hui~hui* 暉暉

In Figure 5.10 it can be seen that the meanings of *hui~hui*_{FIRE} are definitely fewer in number than those of *hui~hui*_{LIGHT} and *hui~hui*_{SUN}. All three, however, have their own special elaborating collocates, with e.g., *hui~hui*_{FIRE} used mostly for ‘red’⁵⁸. *Hui~hui*_{SUN} collocates more with ‘vermillion red’ and the ‘setting sun’. The meanings of *hui~hui*_{LIGHT} appear more diversified. What binds these three together, however, is their ability to depict different kinds of light, as well as shades of ‘red’.

This interplay between the three items is asymmetrical, precisely because *hui~hui*_{LIGHT} has the most productive set of interrelated meanings. It

⁵⁸Also the ‘deafness’ meaning from Chinese traditional medicine stands out.

has a higher type frequency, the effect of which is the reinforcement of its productivity, similar to the regular English Past Tense formation, briefly touched upon above. This polysemy was already found in the earliest period the ideophones were encountered, so it can only be assumed that they were extended from even earlier meanings. That being said, does this high type frequency also translate into high token frequency? Figure 5.11 below shows that it does, especially during the latter half of the Chinese imperial history. The effect of this high token frequency for *huī huī*_{LIGHT} entrenches certain meanings the more it is used. But on the other hand, *huī~huī*_{SUN} has relatively high token frequency effect for a bundle of related meanings, namely the ‘setting sun’ and ‘vermillion red’, two colours that may be conceptually related. They are also productive over time – a type frequency effect.

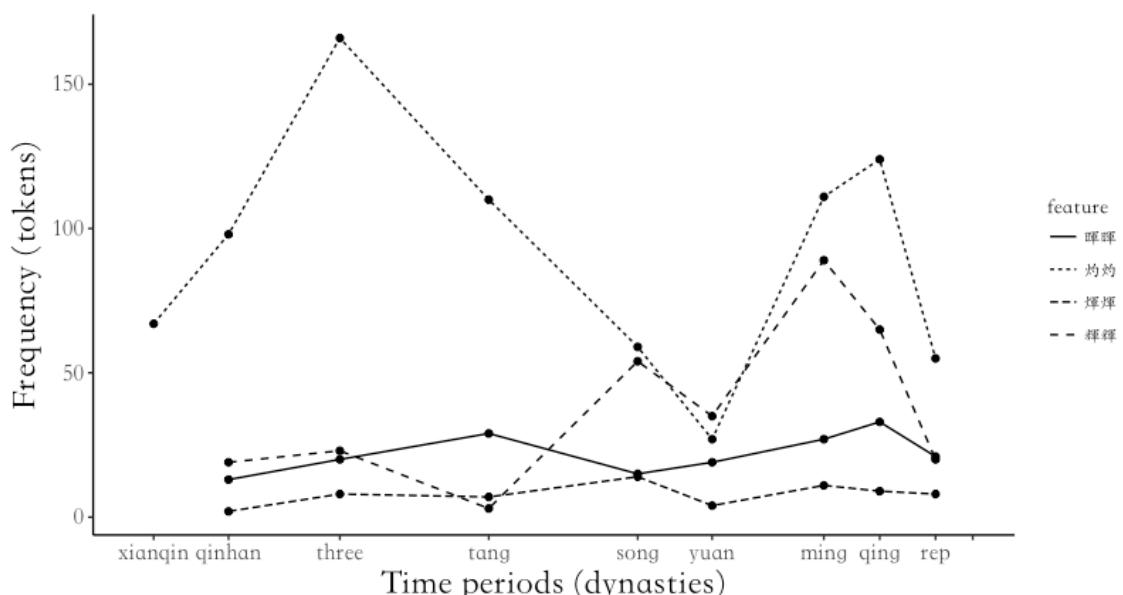


Figure 5.11: Token frequencies of *huī~huī* 煙暉, *huī~huī* 輝輝 and *huī~huī* 暉暉 vs. *zhuó~zhuó* 灼灼

Let us take another ideophone, *zhuó~zhuó* 灼灼, which had the highest

token frequency in the sample throughout history, see Figure 5.11. This horizontal axis follows the same division in time periods as used in the *Scripta Sinica* corpus, as discussed in Section 3.3.1, e.g., ‘three’ stands for the period of the Three Kingdoms, and ‘rep’ for the Republic period. We will take this as a case study of how token frequency, entrenches certain meanings, and leaves dynamic usage for other ones, especially when there do not seem to be immediate formal variants to compete with. Once again we depart from dictionary definitions to get a feel for the different senses, shown in (78).

(78) *Hànyǔ dà cídiǎn* definitions for *zhuó~zhuó* 灼灼

- a. ‘bright’ 明亮貌。晉傅玄《明月篇》：“皎皎明月光，灼灼朝日暉。” 唐韓濬《清明日賜百僚新火》詩：“灼灼千門曉，輝輝萬井春。”《封神演義》第七五回：“余元的寶劍，光華灼灼。” 魯迅《野草·臘葉》：“今夜他卻黃蠟似的躺在我的眼前，那眸子也不復似去年一般灼灼。”
- b. ‘bright color’ 鮮明貌。《詩·周南·桃夭》：“桃之夭夭，灼灼其華。” 晉陸機《擬青青河畔草》詩：“粲粲妖容姿，灼灼美顏色。” 唐楊衡《寄贈田倉曹灣》詩：“芳蘭媚庭除，灼灼紅英舒。”《花月痕》第十三回：“隔水望芙蓉，芙蓉紅灼灼。”
- c. ‘clear’ 明白貌。漢董仲舒《春秋繁露·郊祭》：“天下福若無可怪者，然所以久弗行者，非灼灼見其當而故弗行也。”《漢書·外戚傳下·孝成許皇后》：“咎敗灼灼若此，豈可以忽哉！”顏師古注：“灼灼，明白貌也。”
- d. ‘showing’ 彰著貌。晉潘岳《夏侯常侍誄》：“英英夫子，灼灼其俊。”唐李賀《公莫舞歌》序：“會中壯士，灼灼於人。”叶蔥奇注：“昭昭在人耳”

目。”明胡應麟《少室山房筆叢·經籍會通一》：“名藏書家，代有其人，漢則劉向桓譚……皆灼灼者。”王闔運《陸建瀛傳》：“陸之治江，灼灼有能。”



- e. ‘vigorous’ 盛烈貌。《文選·陸雲〈漢高祖功臣頌〉》：“灼灼淮陰，靈武冠世。”李周翰注：“灼灼，盛烈貌。”
- f. ‘scorching’ 炙熱貌。《醫宗金鑑·四診心法要訣上》：“熱無灼灼，寒無滄滄。”
- g. ‘eager appearance’ 思念殷切貌；熱切貌。晉王羲之《問慰諸帖上》：“足下晚各何以？恒灼灼。”唐喬知之《定情篇》：“更憶娼家樓，夫婿事封侯，去時恩灼灼，去罷心悠悠。”
- h. ‘Name of a beautiful woman in the Shǔ-Hàn dynasty’ 蜀美女名。前蜀韋莊《傷灼灼》詩：“嘗聞灼灼麗於花，雲髻盤時未破瓜。”舊注：“灼灼，蜀之麗人也。”

On the one hand, the data manually extracted from Scripta Sinica does not seem to fully conform to these dictionary definitions. On the other, the dictionary does not really reflect the usage or token frequency either – something that is really apparent with this particular item: as Figure 5.12 shows, the semantic matrix of *zhuó~zhuó* revolves mostly around the shining brilliance of things associated with spring and the blossoming of ‘flower leaves’, ‘pinkish red’ and ‘peach trees’, meaning (78b) in the definitions. This BLOSSOMING frame, as we call this bundle of related meanings, is illustrated in example (79).

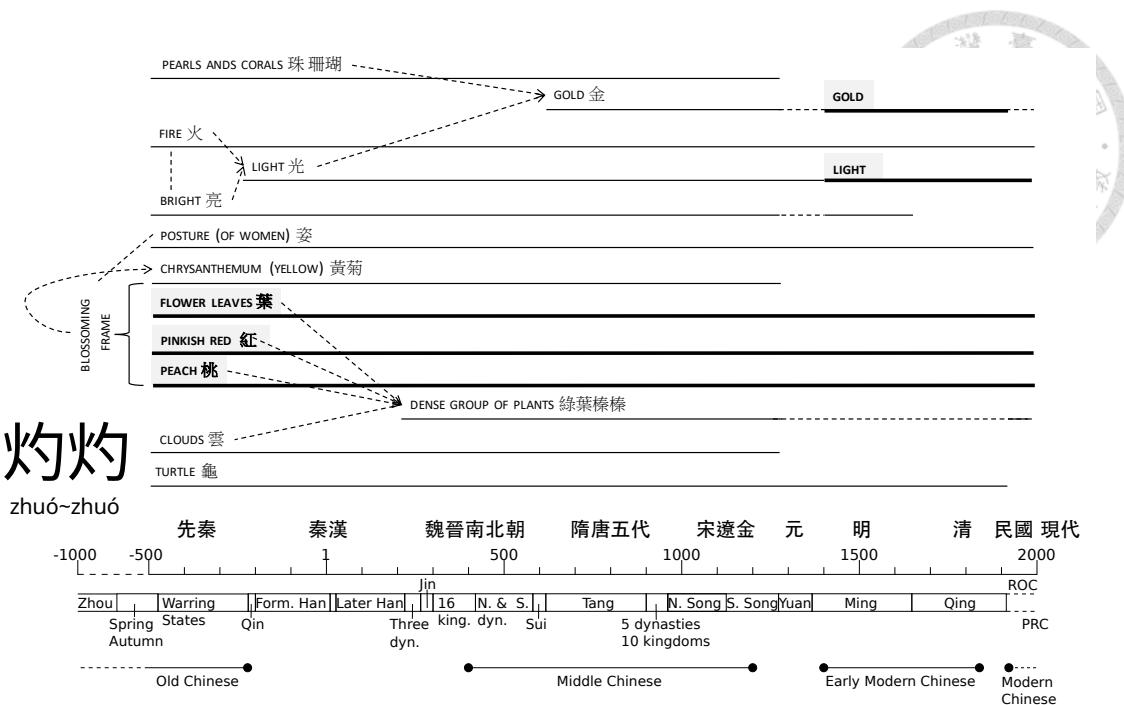


Figure 5.12: *zhuó~zhuó* 灼灼

(79) Flowery usage of *zhuó~zhuó* 灼灼

- “The peach tree is young and elegant, *brilliant* are its flowers” 桃之夭夭灼灼其華, originally in *Shījīng* 詩經, but repeated throughout history
- “*Brilliant* the hundred leaves red” 灼灼百朵紅, in *Quán Táng shī* 全唐詩
- “Bright, The moon amidst the clouds; *shining*, the flowers between the leaves.” 明明雲間月, 灼灼葉中花。in *Wénxuǎn* 文選

This frame extends to depict the dazzling colors of ‘chrysanthemums’, but also to ‘normal LIGHT’ items. Another extension from the prototypical bundle is that to the concept of ‘densemess of plants’⁵⁹. This is indeed a token

⁵⁹There is also a seemingly anomalous usage of *zhuó~zhuó* to depict turtle(shells?), which we were not able to relate to any of the other meanings.

frequency effect, on the semantic side: the very high usage of some authoritative phrases from the *Shījīng* 詩經 crystallizes this group of meanings and is extensionally productive to related meanings. As a summary, this section on the three *huī~huīs* and *zhuó~zhuó* has showed that frequency effects can be identified in the diachronic usage of ideophones.

5.3.4 Transient prototypicality

The last case study shows the development of $yè\sim yè_{\text{FIRE+SUN}}$ 燥燁, $yè\sim yè_{\text{FIRE}}$ 燥燁 and $yè\sim yè_{\text{SUN}}$ 暉暉. Like *yào~yào* and *huī~huī* the current three are phonological homophones and orthographic near-homographs. As before, let us start with the dictionary definitions (80).

(80) *Hànyǔ dà cídiǎn* definition for the three $yè\sim yès$

- a. $yè\sim yè_{\text{FIRE+SUN}}$ 燥燁: NA
- b. $yè\sim yè_{\text{FIRE}}$ 燥燁: 1. ‘bright’ 明亮；燦爛；鮮明。; 2. ‘scorching’ 灼熱貌；顯赫貌。
- c. $yè\sim yè_{\text{SUN}}$ 暉暉: 1. ‘beautiful’ 美盛貌。; 2. ‘radiant’ 光芒四射貌。; 3. ‘brilliant talent’ 指才華外露。

For $yè\sim yè_{\text{FIRE}}$ and $yè\sim yè_{\text{SUN}}$ there seems to be enough overlap in the definitions; for $yè\sim yè_{\text{FIRE+SUN}}$ we were not able to find a definition. However, it does occur in the Scripta Sinica. The diagram of the data, shown in Figure 5.13, suggests that the complex form of $yè\sim yè_{\text{FIRE+SUN}}$ might be a specialized variant of $yè\sim yè_{\text{FIRE}}$.

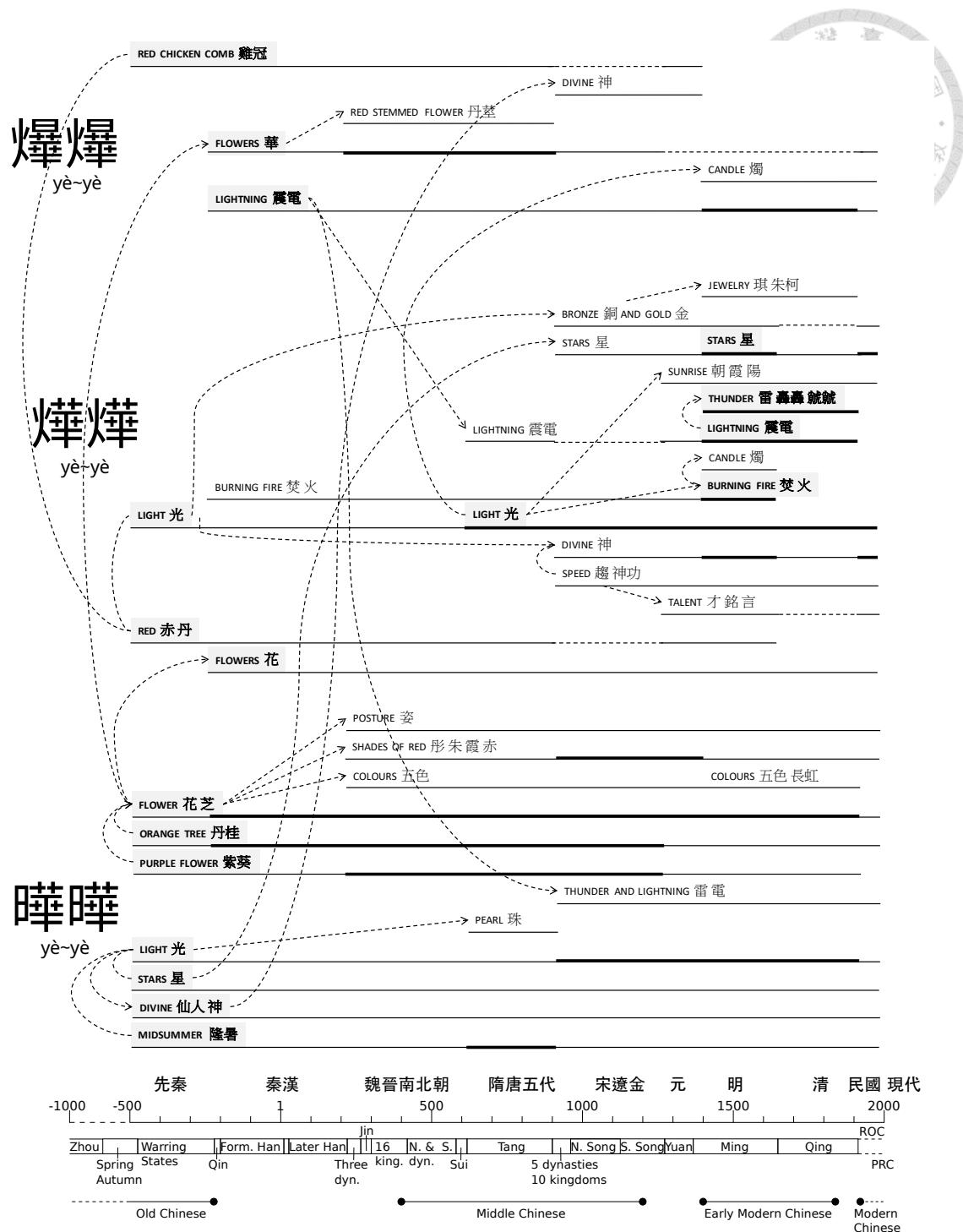


Figure 5.13: $yè\sim yè$ 煙燶, $yè\sim yè$ 燁燶 and $yè\sim yè$ 瞳瞩

There is also a large overlap here between collocates that elaborate the meaning of $yè\sim yè_{FIRE}$ and $yè\sim yè_{SUN}$. Therefore, if $yè\sim yè_{FIRE+SUN}$ is a specialized variant of $yè\sim yè_{FIRE}$, it is not surprising that there is also some semantic transfer from $yè\sim yè_{SUN}$ to $yè\sim yè_{FIRE+SUN}$, namely in collocates that mean

‘flowers’ and ‘the divine’. But the transfer also happens in the other direction: ‘lightning’ starts in $yè\sim yè_{FIRE+SUN}$ and makes its way to $yè\sim yè_{FIRE}$ and also to $yè\sim yè_{SUN}$. Another collocate, ‘stars’, starts in $yè\sim yè_{SUN}$ and ends up in $yè\sim yè_{FIRE}$, where it is used more often.

Due to the strong core meanings and fuzzy overlap a choice needs to be made about which $yè\sim yè$ to use. It is here that we see some effects that frequency may have had. For type frequency, $yè\sim yè_{SUN}$ clearly comes out on top of these three. Once again, it is most productive in terms of extensions, as the left half of the figure shows. For token frequency, $yè\sim yè_{SUN}$ is the winner as well. But here something curious is going on. $Yè\sim yè_{SUN}$ with the highest token frequency in the first half of dynastic history has a bundle of meanings – let us call it a COLORFUL FLOWER frame – that is solidified over time, and stays that way. This is a typical token frequency effect. The general collocate of ‘light’ becomes more used over time, but does so too for $yè\sim yè_{FIRE}$. The FIRE variant takes over the token frequency in the right half of the diagram, where this higher token frequency also seems to correlate with a renewed extensional productivity, resulting in a higher type frequency on that end. If token frequency is somehow related to prototypicality (as some researchers argue), then we are witnessing a shift in prototypicality from the SUN variant to the FIRE variant. In other words, prototypicality in ideo-phones can also be transitory.

As a summary, this case study brings it all together. It underscores the dynamicity of language usage and choices that have to be made, which is related to the fuzzy boundaries that exist between the competing forms.

These each do have somewhat of a prototypical usage or bundle of meanings, taking the form of a polysemous network; but this prototypicality is transient, so also dynamic. Furthermore, we can see some frequency effects at play. While this concludes the case studies, in the next section we will dwell briefly on the types of extensions that can readily be seen in the diagrams.

5.3.5 Types of extensions

In the preceding case studies, we have been studying things from a mostly semasiological point of view. That means that we have been looking at the meanings that belong to a given label, and how they are connected to one another and through diachrony, how they change over time. However, a brief comment is in order about the types of extensions that are available in the case studies. The case studies have highlighted how semantic transfer happens from one orthographic variant to another, but the more mundane, or more typical, types of extensions of course also occur within the case studies. More specifically, denotational semasiological extensions (Geeraerts 2010b:26–28) are readily observable in the data.

There are four basic types of extensions that take up the bulk of extensions in language: SPECIALIZATION, GENERALIZATION, METONYMY and METAPHOR. An example of specialization is when the English word *queen* originally means ‘woman, wife’ but later on specializes to ‘female sovereign, or king’s wife’. Generalization can be illustrated by *moon*, which at first referred to the earth’s satellite, but now can be used to describe the

Table 5.12: Types of semantic extensions present in the preceding figures

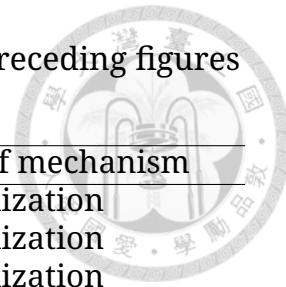


Figure	Ideophone	Meaning 1	Meaning 2	Type of mechanism
5.6	<i>yuè~yuè</i>	light	moonlight	specialization
5.6	<i>yuè~yuè</i>	light	pearl	specialization
5.7	<i>yào~yào</i>	light	sun	specialization
5.7	<i>yào~yào</i>	light	posture	metaphor
5.8	<i>shuò~shuò</i>	lantern	moon	metaphor / metonymy
5.8	<i>shuò~shuò</i>	moon	sun	metonymy
5.8	<i>shuò~shuò</i>	lantern	stars	metaphor / metonymy
5.8	<i>shuò~shuò</i>	lantern	light	generalization
5.9	<i>yì~yì</i>	fire	anger	metaphor
5.9	<i>yì~yì</i>	coins	sound of coins	metonymy
5.10	<i>huī~huī</i>	vermillion	red ink	metonymy
5.10	<i>huī~huī</i>	setting sun	rising sun	metonymy / metaphor
5.10	<i>huī~huī</i>	setting sun	old woman	metaphor
5.12	<i>zhuó~zhuó</i>	blossoming	dense plants	metonymy
5.12	<i>zhuó~zhuó</i>	clouds	dense plants	metaphor
5.13	<i>yè~yè</i>	light	candle	specialization
5.13	<i>yè~yè</i>	light	fire	specialization
5.13	<i>yè~yè</i>	lightning	lightning	generalization

satellites of other planets as well. Metonymy is largely based on contiguity, e.g., when you say you *drink a glass*, you actually mean the liquid inside of that glass. Metaphor, then, is based on (figurative) similarity. A very typical phrase would be *time is money*⁶⁰. Of course, the border between metaphor and metonymy is not always clear-cut, cf. for instance Goossens's (1990) discussion on METAPHTONYMY.

These types of extensions are illustrated in Table 5.12, as they can be found in the diagrams above, with the *caveat* that some of them are definitely up for debate. To be clear, what we are looking for are extensional mechanisms in the types of collocates, which serve as proxies for types of extensions in meaning.

⁶⁰This of course also acts as a conceptual metaphor (Lakoff & Johnson 1980) to which many expressions are connected.

It can be seen in Table 5.12 that this non-exhaustive list of semasiological changes covers most of the case studies above. It is not wholly unsurprising that ideophones are also subject to these semantic changes. While in synchronic usage it is easy to get the feeling that one ideophone depicts one thing or at least a fixed set of things, the interplay of written record, the particular ontology of the Chinese writing system and the NON-SOUND modality we are looking at here, i.e. VISUAL ideophones all contribute to the dynamicity of the system over time. After all, ideophones are conventionalized words and words display these semantic mechanisms of change. Presumably there is less flexibility if the iconicity relation that affords ideophones is warped more towards imagic iconicity – there would just be less leeway to go astray, similar to how English had the verb *pipen* /pi:pən/ for the peeping sound of young birds, which changed through /paipən/ by virtue of the Great Vowel Shift, but then fell out of usage because it was no longer iconic, and a new written form *peep* /pi:p/ was introduced. But since our case studies are all in the domain of VISUAL sensory imagery, we are dealing with diagrammatic iconicity, seen in the markedness of reduplication and motivated by the writing system. Coerced iconicity (Dingemanse 2011b) can only constrain a bit, but these words mostly act as normal words, and seem to traditionally have been interpreted as such as well. With this side note dealt with, it is time to leave the lower level of language usage and see if we can abstract a bit, beyond mental spaces and their sanctioning frames.

5.4 Domains/ICMs and Image Schemas



In the preceding sections, we have been fruitfully adopting Kövecses's (2017) levels of metaphor theory to ideophones for mental spaces, i.e., raw language usage (Fauconnier 1994; Fauconnier & Sweetser 1996). We have also observed highly specialized frames when these meanings became entrenched, as Akita (2012b) suggested. Let us now abstract it one level higher, to domains or ICMs (Lu 2006). This is done by taking the most prototypical meanings as they occur in the case studies in a network and grouping them by their conceptual similarity.

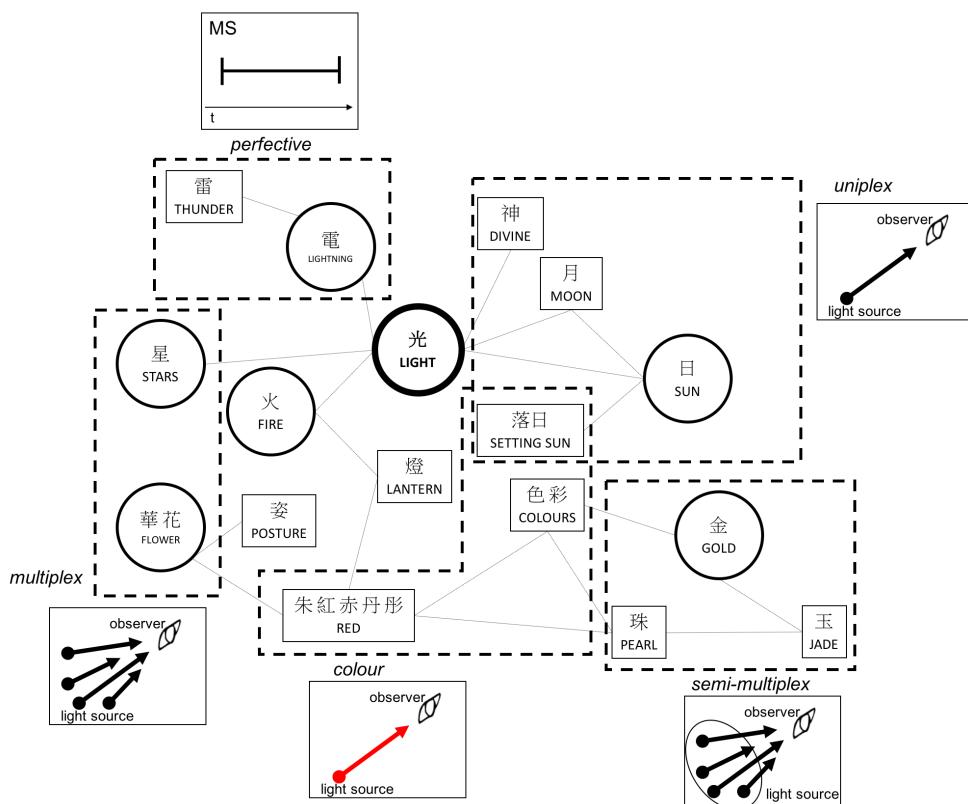


Figure 5.14: Frames and Domains / ICMs

Figure 5.14 shows the result of such an exercise. When the different frames, e.g., the prototypical LIGHTNING frame (in a circle), and their extended frames, e.g., THUNDER (in a rectangle) are grouped like this, the do-

mains or ICMs that bind these frames together, highlight different aspects of an instantiating image schema. For the SUN related frames, the uniplex domain shows how there is a single light source, which reaches the observer. Light sources can also be multiplex (Lakoff 1987; Talmy 2000b), as in the case of STARS or FLOWERS, where it is the distributed collectivity of light sources that depict the ‘LIGHT’ in this case. It can also be somewhere in between, as in the case of metals like GOLD and related frames, where (presumably) it is different smaller sparkling places within a bounded entity that depict the LIGHT. For this reason, we have used the term semi-multiplex. Another aspect that can be highlighted is the colours, with a strong preference for RED. Since this is related to the reddish glow of the SUN, FIRE as well as FLOWERS, it is motivated yet an unexpected productive domain of extension. Finally, the boundedness or perfectivity (Langacker 2008b:151, 156–157) of a LIGHT-related event may be highlighted as well, as in the case of ‘lightning’ and ‘thunder’. In the next chapter, a slightly revised version of these domains will be presented, based on computational aid, see Section 6.6.

Now, the reader may wonder why the most prototypical sense of LIGHT is not included in any of these ICMs in Figure 5.14. Since in a way it is the result of the overlapping meanings, it may be more useful to consider it a proto-scene or Image Schema of ‘LIGHT FALLING INTO THE OBSERVER’S EYES’. Notice how this differs from our usual way of understanding how light works. From a scientific point-of-view (pun intended), a light source emits beams of light that are caught by the observer’s retina and translated into elec-

trical signals that are sent to the brain for decoding, or that is how basic optics in physics works. However, our folk model of light thinks of seeing and looking at as something volitional or intentional. This is made clear by certain conceptual metaphors such as SEEING IS TOUCHING (Lakoff & Johnson 1980), KNOWING IS SEEING (Sweetser 1990) or special instances of the CONDUIT metaphor (Reddy 1979), e.g., the Mandarin *wǒ míngbái le* 我明白了 ‘I clear PFV > I understand’. In contrast to this way of thinking, there is at least one important category where this folk model is reversed: ideophones. This is not a surprise, because ideophones highlight the multimodal, mimetic and performative nature of the experience, which is often unintentional. The different models are presented in Figure 5.15, and capture the essence of the highly abstract Image Schemas very well.

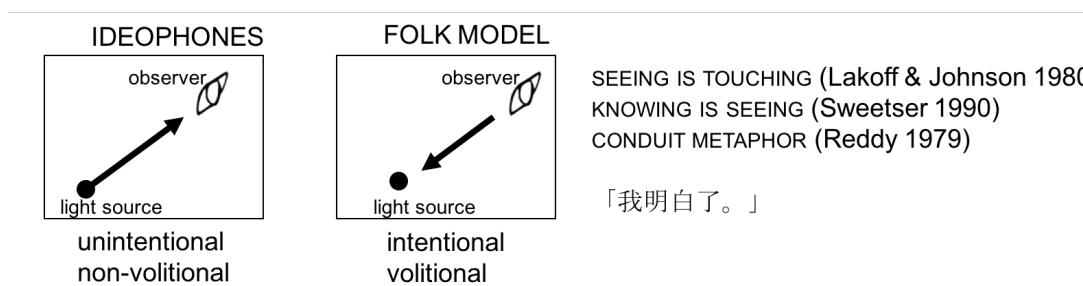


Figure 5.15: Image Schema of LIGHT ideophones vs. the standard folk model

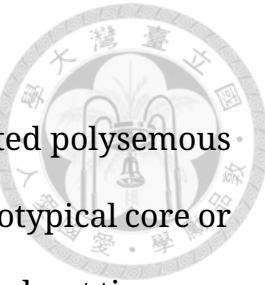
5.5 Conclusion

Using literary Chinese ideophones in the semantic domain of LIGHT as case studies has revealed that the application of diachronic prototype semantics in combination with the stratified ‘levels of metaphor’ approach to the study of ideophones is a fruitful venture. In metaphorical terms, we have been

looking at the data both horizontally and vertically.

Horizontally, it was found that meanings form interrelated polysemous clusters that are dynamic throughout time, with a clear prototypical core or prototypical bundles as a core, semantically extending throughout time, see section 5.3.1 on *yue~yue*. The items investigated also seemed to influence one another, as the mutual influence of the *yao~yao* case study displayed in section 5.3.2. In the subsequent section (5.3.3) *huihuī* and *zhuó~zhuó* showed that type and token frequency effects (Bybee & Hopper 2001a) also can play important roles in the semantic development of these items. However, prototypicality is not only dynamic between different meanings of an ideophonic item, it can also be among different variants. The case study of *yè~yè* in section 5.3.4 showed that some variants may be the most productive for a while, but others might take over this role at one point. In other words, prototypicality is transient.

Vertically, these notions were all studied on the level of mental spaces and, upwards, frames. The interplay between dynamicity and stability is a direct consequence from the flexibility afforded through mental space theory, which provides a theory of online meaning construction, and the entrenching, schematizing, force of frames. However, such an abstraction can be pushed even further. Domains or ICMs were identified that each highlight something different of the embodied Image Schema underlying most of these tokens. LIGHT FALLING INTO THE OBSERVER'S EYES appears a good Image Schematic summary of what happens when people use LIGHT ideophones to depict their experience. It does, however, stand in stark contrast with the



other folk model we have of SEEING (cf. section 5.4).

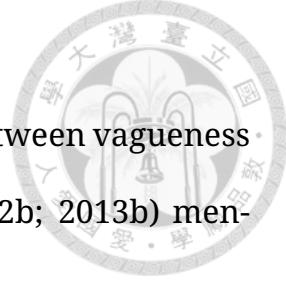
It is now also possible to generalize about the debate between vagueness and polysemy. As mentioned in section 5.2.1, Akita (2012b; 2013b) mentions that Mandarin SOUND ideophones are vague, while Japanese ones are highly-specific. As our investigations on the VISUAL LIGHT ideophones show, it is not a black and white case. Take for example $yè$ ~ $yè_{\text{SUN}}$ in section 5.3.4, which has a prototypical bundle of ‘light’, but also of ‘flower’-related meanings. With regards to polysemy at least these two big meanings will have to be distinguished; with regards to vagueness, however, it can be argued that the different readings of ‘light’, ‘stars’, the ‘divine’ and ‘midsummer’ are somewhat vague in the prototypical bundle of LIGHT. This resembles the example of fruit discussed in section 5.2.2.

And now we turn back to the semiotic folk model of Chinese section 1.3, repeated in ((81)).

(81)

$$\frac{\textit{phonology}}{\textit{orthography}} \mid \textit{semantics}$$

We found considerable variation and dynamicity on the three poles of the symbolic assembly. On the <phonological> pole, networks of word families were briefly presented, as well as reanalysis in these networks, embedded in a discussion on phonesthemes (Section 5.1.1). As for <meaning>, this case study has shown the flexibility of meanings from a diachronic perspective. On the <orthographic> pole, the functional components seem to play an important conceptual role as well: FIRE 火, LIGHT 光, SUN 日, and



METAL 金 all are prototypical frames identified in this study, and that is exactly a distinguishing factor between literary Chinese ideophones and colloquial ones: the ‘radical support’ and motivation in the writing system, as argued before (Van Hoey 2017; Van Hoey 2018a), but also nuanced in Chapter 4. This chapter is not devoid of limitations. For example, only a small sample of an arguably much larger semantic field has been investigated. Studying more will chart more of the variation and prototypicality effects at play, although it seems that the most important effects have already been displayed in this chapter. So more data would enrich, and presumably nuance, some of the effects. In the next chapter we approach this issue by scaling the data and using a computational method to explore the interplay between orthographic variants and different meanings.



6 Variational salience and LIGHT ideophones



If there's something strange,
in your [semantic vector]
neighborhood,
who you gonna call?
Ghostbusters!

Ghostbusters theme song

6.1 Introduction

In Chapter 5, a manual study of LIGHT ideophones revealed that their semantics takes the form of prototypical clusters that change over time. It has been argued before (Geeraerts 2000) that prototypical structure is an expression of salience. But what exactly is SALIENCE? The term warrants further explanation, before research question 6c (“How did other instances of variational salience [within ideophones] evolve over time?”) can be addressed.

Schmid & Günther (2016) approach the term through four linguistic lenses, in order to unify them in a socio-cognitive framework that stresses two central elements: EXPECTANCY and CONTEXT. The first of the four positions they survey includes that things may be salient because they are familiar and entrenched. This corresponds with the idea that salience is coded in the mental lexicon (Giora 2003). The second position is that words are salient because they have high relative frequency vis-à-vis the experience they denote (Geeraerts 2017). In this interpretation, salience is related to entrenchment. The third scenario is that salience is understood

as surprisal: segments stand out in a cognitively salient manner if they are unexpected given the context (Rácz 2013). Lastly, salience can also mean that things are completely unfamiliar (Barto, Mirolli & Baldassarre 2013).

Table 6.1 summarizes these four perspectives.

Table 6.1: Overview of the four positions on saliency

	expected	unexpected
with context	Geeraerts: entrenchment	Rácz: surprisal
context-free	Giora: mental lexicon	Barto et al.: unfamiliarity

Saliently speaking, it is somewhat unsurprising and bound to the context of this dissertation that we will follow Geeraerts's treatment of salience phenomena, i.e., his typology of lexicological salience phenomena (see Geeraerts 2000; recontextualized in 2006b; 2010b; 2017).

In his taxonomy, he discerns two major branches of salience: perspectival salience and variational salience. The well-known example of the Commercial Transaction frame (Fillmore 1977; Lawler 1988; Croft, Taoka & Wood 2001), with *to buy* and *to sell* as the two main contrasting verbs, illustrates perspectival salience nicely. That is, when *to buy* is used, the BUYER is more salient than the SELLER or the THING BEING SOLD. For *to sell*, it is the SELLER who gets most of the spotlight. In Langacker's (1987b) terms, the seller gets accorded trajector status. However, the two verbs of the Commercial Transaction frame can also be used to illustrate variational salience. If one wants to express a commercial transaction, one is forced to choose between these two alternatives *to buy* and *to sell*. This constitutes

an onomasiological choice, in which one option will be more likely than the other, and by extension will have a higher salience value.

This chapter⁶¹ focuses on variational salience, which is further split between semasiological salience, onomasiological salience and structural salience. While Geeraerts further subdivides his taxonomy, we will restrict ourselves to these three here. The three types of variational salience can be understood in relation to a quote by Baldinger:

“Semasiology [...] considers the isolated word and the way its meanings are manifested, while Onomasiology looks at *the designations of a particular concept*, that is, at a *multiplicity of expressions which form a whole*.”

Baldinger (1980:278), in Geeraerts (2000:83); emphasis mine

While at face-value semasiology and onomasiology appear to be two sides of the same coin, Geeraerts maintains that we should distinguish the ambiguous definitional phrasing of onomasiology. If the term is understood as ‘designation’, i.e., *naming* a relationship between meanings and lexical items named by them is studied. If it is a ‘multiplicity of expressions’, onomasiology is concerned with the relations between related lexical items (Geeraerts 2000). He therefore interprets the latter in a structural manner. Applied to the term of salience, this gives three types: semasiological salience, onomasiological salience and structural salience.

Examples of semasiological salience include prototypicality effects, such as the uneven structure of items belonging to the category *bird* (Rosch 1975;

⁶¹ Parts of this chapter were presented at the 15th International Cognitive Linguistics Conference (ICLC 15), see Van Hoey (2019b).

Rosch & Mervis 1975), or the meaning of the preposition *over* (Lakoff 1987; Dewell 1994; Tyler & Evans 2003). For onomasiological salience, a clear illustration is sociolinguistic prevalence. In a Taiwanese context, when presented with a picture of a bowl of rice with some sort of diced braised pork on top, people from the northern part of Taiwan are more likely to call this dish *lüròufàn* 滷肉飯. People from the southern part will probably categorize this dish as *ròuzàofàn* 肉燥飯⁶². Structural salience, then, involves difference in salience between the “weight of different dimensions that distinguish various categories from each other” (Geeraerts 2006b:88). The example Geeraerts uses concerns ingredients of beer types and the name they are given when marketed (see Section 6.6).

Applied to ideophones, we can now say that the previous chapter has been mostly concerned with semasiological salience, and minimally with onomasiology. The latter occurred when multiple variants were contrasted and then probed for their meanings. In this chapter we will develop the three types of salience further, by also addressing a question asked at the end of the previous chapter: can we reach similar conclusions through computational approaches, and can we scale the data?

The implementation of choice for this chapter is distributional semantics, in which a semantic vector space is calculated. In such a space, linguistic elements are placed along a n number of dimensions, with their so-called cosine value indicating the conceptual distance. In Section 6.2, the methodology will be explained and put in the context of linguistic approaches. This is followed by the adaption to the current chapter in Section 6.3. It aims

⁶²I want to thank Hsu Iju for pointing this out to me.

to answer the question on how to create a semantic vector space for a historical corpus that contains many Chinese ideophones. This will then be followed by a case study on the topic of LIGHT ideophones in relation to semasiological salience (Section 6.4), onomasiological salience (Section 6.5), and structural salience (Section 6.6).

6.2 Distributional relational semantics

According to Geeraerts's (2010b:165–178) classification of lexical theories, computational distributional analyses take a syntagmatic perspective in the relational approaches in what he calls 'Neostructuralist semantics'.⁶³ They are inspired by developments in the 1950s, such as the famous Firthian dictum "You shall know a word by the company it keeps" (1957:11), which implies that words that co-occur tend to have similar meanings, or Harris's (1954:156) argument that differences in meaning correlate with differences in distribution. Most of these approaches⁶⁴ are radically corpus-based and thus take very seriously the usage-based criterion for linguistic evidence. With the astounding progress made in corpus technology – both in the size of corpora as well as processing speed – there has been a steady rise in computational linguistic approaches that accompanied it.

Lenci (2018) surveys the field: most studies seem to focus on the 'right' statistical measures that need to be taken to turn a corpus into a collection of data that deliver meaningful results. The basic idea is that words

⁶³Paradigmatic perspectives are exemplified by the WordNet project and its cross-linguistic offshoots (Fellbaum 1998), and lexical functions approaches such as Mel'čuk's Meaning-Text Theory (1988).

⁶⁴Geeraerts (2010b:166) refers to Levin's (1993) classification of English verbs, which is based on the different alternations semantically similar groups of verbs can undergo.

(or the units one is interested in) can be given a score, based on their co-occurrence rate, which allows them to be visualized in the word space.

Sahlgren (2006:18) introduces the idea as follows:

The word-space model is, as the name suggests, a spatial representation of word meaning. Its core idea is that semantic similarity can be represented as proximity in n -dimensional space, where n can be any integer ranging from 1 to some very large number – we will consider word spaces of up to several millions of dimensions later on in this dissertation. Of course, such high-dimensional spaces are impossible to visualize, but we can get an idea of what a spatial representation of semantic similarity might look like if we consider a 1-dimensional and a 2-dimensional word space, such as those represented in Figure 6.1.

Sahlgren (2006:18)

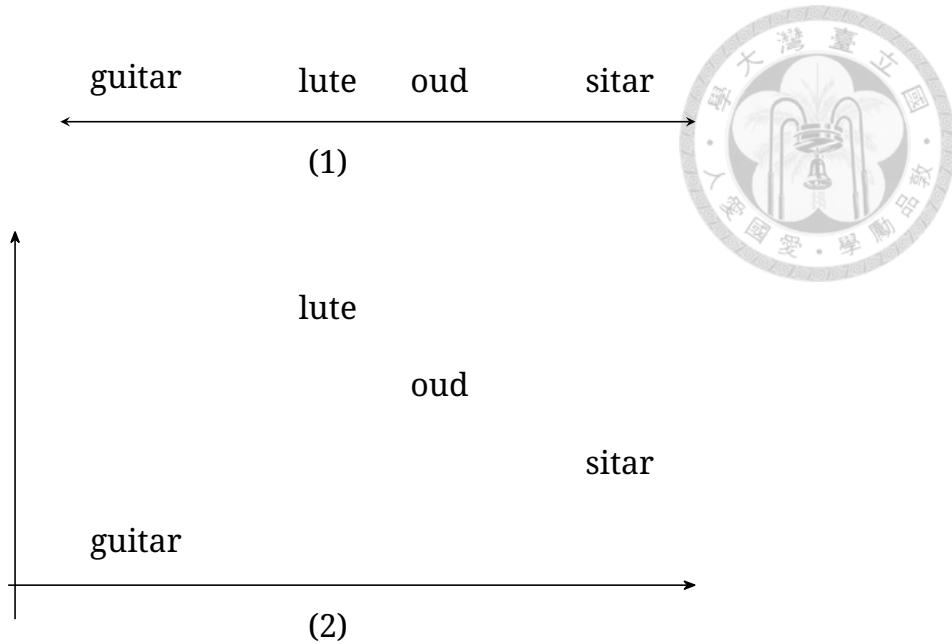


Figure 6.1: (1) A 1-dimensional word space, (2) A 2-dimensional word space
(adapted from Sahlgren 2006:18)

As Figure 6.1 shows, four musical instruments are investigated: the guitar, the lute, the oud⁶⁵, and the sitar. The spatial proximity between words indicates how similar their meanings are: the *oud* is closer to the *lute* than to the *guitar*. This is analogous to saying that the meaning of *oud* (OUD)⁶⁶ is more similar to the meaning LUTE than to the meaning GUITAR. Sahlgren convincingly shows that there are conceptual metaphors underlying this visualization: SIMILARITY IS PROXIMITY (see also Lakoff & Johnson 1999), and ENTITIES ARE LOCATIONS. The reasoning is as follows:

When we think about meanings as being *close* to or *distant from* each other, we inevitably conceptualize the meanings as locations in a semantic space, between which proximity can be measured. However, the ENTITIES ARE LOCATIONS metaphor is com-

⁶⁵This is an Arabic short-necked, lute-type, pear-shaped stringed instrument.

⁶⁶A notation we will borrow from Langacker's Cognitive Grammar (Langacker 1987a; 1991; 2008a) is that forms (whether phonological or written) are italicized, and meanings are given in (small) capitals.

pletely vacuous in itself. Conceptualizing a sole word as a lone location in an n -dimensional space does nothing to further our understanding of the word. It is only when the space is populated with other words that this conceptualization makes any sense, and this is only due to the activation of the **SIMILARITY IS PROXIMITY** metaphor.

Sahlgren (2006:19), emphasis in original

He then states that the conceptual metaphor underlying these distributional approaches is a combination of those two metaphors – the Geometric Metaphor of Meaning: **MEANINGS ARE LOCATIONS IN A SEMANTIC SPACE**, and **SEMANTIC SIMILARITY IS PROXIMITY BETWEEN THE LOCATIONS** (Sahlgren 2006:19). He further notices that even though word space models typically compute more than one dimension to represent similarities, which can be counter-intuitive to “beings such as us who live in a spatially low-dimensional environment”, the prime metric that is visualized, is **PROXIMITY**.

Now that the basic idea of these distributional approaches has been discussed, it is worth to reflect on the impact of these ideas. In the field of computational linguistics the approach has been very popular (as Lenci 2018 shows), but in theoretical linguistics, the impact has been limited (Boleda 2020). This means that in the amalgam of Cognitive Linguistic approaches these methods also have made less of a splash, with one main exception: research developed at or in collaboration with the Quantitative Lexicology and Variational Linguistics research unit at the University of Leuven (Heylen, Speelman & Geeraerts 2012; Ruette 2012; Wielfaert,

Heylen & Speelman 2013; Heylen et al. 2015; Peirsman, Geeraerts & Speelman 2015; De Pascale 2019).

The main argument of these studies is to show that distributional approaches are applicable to lexicology and lexicography. Generally, they advocate the study of semantics at both the TOKEN LEVEL and the TYPE LEVEL (Heylen, Speelman & Geeraerts 2012). The type level aggregates over all occurrences of a word, giving a representation of a word's general semantics, mostly used to retrieve semantic relations between words, e.g., synonyms in the task of thesaurus extraction. In this sample, the token level represents the semantics of each individual occurrence of a word, and is typically used to distinguish between the different meanings within the uses of one word, notably in the task of Word Sense Disambiguation or Word Sense Induction. They state that lexicological studies “typically combine both perspectives: their scope is often defined on the type level as the different words of a lexical field or the set of near-synonyms referring to the same concept, but they then go on to do a fine-grained analysis on the token level of the uses of these words to find out how the semantic space is precisely structured” (Heylen, Speelman & Geeraerts 2012:17).

It should be noted that this methodology is basically a black box⁶⁷ (Heylen, Speelman & Geeraerts 2012:21; Wielfaert, Heylen & Speelman 2013), so it is still necessary to manually check the results and conclusions that are drawn from it. However, that the empirical evidence provided by these usage-based and quantified studies is fruitful, is beyond doubt.

⁶⁷ While it is a black box when used in AI, the implementation we are using is based on linear algebra which renders the box more transparent.



6.3 Methodology

Thus, if distributional semantic relations are fruitful within Cognitive Linguistics, then the most urgent issue is: how does such distributional semantic analysis work? This question will be answered by relying Peirsman et al.'s (2015:58–61) summary of the distributional semantic methods, then pushing along the route of machine learning (mostly pursued by computational linguists), and then returning to more conservative ‘count-based methods’. Peirsman et al.’s first step is defining the context model to use. However, this seems to take Standard Average European language processing for granted. That is, not all languages split their words using white spaces. For this reason, our first step is to actually segment the text. The second step is then the definition of a context model and selecting the units for analysis. The third step is the computation of frequencies. The fourth and final step is the statistical formula for the assessment of similarity.

6.3.1 Step 1: Segmentation

The first problem pertaining to our Chinese data is segmentation, as was touched upon in Section 3.3.2. As is shown in (82), *Táiwān-dàxué* is taken here as one unit ‘National Taiwan University’.

(82) 我 來到 台灣-大學

wǒ lái-dào Táiwān-dàxué

1.sg come-to Taiwan-university

“I come to National Taiwan University”

However, let us reflect on what we want from the computer? Should it actually split the two in *Táiwān* and *dàxué*? The argument for doing so would be that we might be interested in seeing what other universities are mentioned in the corpus. On the other hand, maybe the researcher is interested in e.g., adjectives specifically modifying National Taiwan University. This would need a different grouping: [*Táiwān-dàxué*] as one group.

Luckily there are workarounds around this problem. Popular segmenting tools such as the `jieba` library for Python or its twin package in R, `jiebaR`, provide the option to get all possibilities (both *Táiwān-dàxué*, and *Táiwān*, and *dàxué*) or make an ‘educated guess’. Recently, a new segmentation tool for Python was introduced, `ckiptagger` (CKIP group 2019). This library uses a pretrained model which allows for higher accuracy, and can provide named-entity recognition (Li, Fu & Ma 2019) – although we will not be using that feature. Applied to the example, `ckiptagger` can adequately guess that *Táiwān-dàxué* can be one unit. Currently, these tools do a good job for Standard Chinese, but for Pre-Modern Chinese the case is a bit different. Still, `ckiptagger` has proved itself to be useful for constructing a subcorpus from the Scripta Sinica corpus, namely a Diachronic Chinese Ideophone Corpus (DIACHIC), as detailed before in Section 3.3.2. This segmented subcorpus allows us to follow the rest of the steps in the methodology.

6.3.2 Step 2: Context models and units

After clearing away the main difficulty of Chinese, segmentation, it is useful to discuss the different context models and units we need for further

analysis. In general three context models can be discerned: document-based models, syntax-based models, and word-based models (Geeraerts 2010b:174–176; Peirsman, Geeraerts & Speelman 2015:58–59).

First, document-based models, like latent semantic analysis (Landauer & Dumais 1997), take textual entities as features. Their context vectors take into account what documents, sections, articles, sentences, or similar stretches of text a target word appears in. Instead of the most intuitive approach, where one looks at words that are characteristic of a given text, this model treats the texts in which a word appears as characteristic features of that item. Word meanings are assumed to be similar if they appear in similar documents. Because these take a very big *context window* (the document), the results are to be interpreted as taxonomic similarity, or general association (Sahlgren 2006). Peirsman, Geeraerts & Speelman (2015) give the examples of *wave* and *sea*, or *doctor* and *hospital*.

Second, syntax-based context models are on the other end of the spectrum: they take words as semantically similar if they often appear in the same syntactic relation, e.g., *bike* and *bicycle* as the direct object of *ride*. These models have a very small context window, and allow for fine-grained models of semantic similarity, ordered in a strict taxonomy that distinguishes hyponyms, hypernyms, co-hyponyms and synonyms.

Third, word-based models lie somewhere in between syntax-based and document-based models. The units of analysis are context windows that are slightly bigger than in syntax-based models, so they just take into consideration the words that appear in the context of the target word, without

considering the syntactic relations between them. Word meanings are assumed to be semantically similar if they often have the same context words in that window. Examples that this approach would yield are, *bike* and *bicycle* that both often co-occur with words like *ride*, *wheel*, and *car*.

Peirsman, Geeraerts & Speelman (2015) stress that “this choice of context definition is extremely important” (2015:59), because the adopted model greatly influences the results. Since we are mostly interested in the different meanings of LIGHT ideophones in a given period, rather than the exact document or syntactic configuration, it follows that WORD-BASED MODELS are the best choice for this study. However, this leads to a following question: how do we get these units, and what is the context window around target words?

One straightforward answer would be to take as the context window n number of (segmented) words to the left and / or to the right of the target word. In a given sentence like (83), we could be interested in the kind of words that co-occur with the verb *hit*.

(83) I hit the ball hard

If the window is then set to 1 to the left, and maybe 2 to the right, we could capture *I hit* and *hit the ball* very easily. However, computationally, the larger the corpus gets, the harder it becomes to make these educated guesses of what the context window should be. Another way that tries to address this problem uses N-GRAM grouping (see Cavnar & Trenkle 1994) – a model that has been popularized by huge corpus projects like Google n -gram (Michel et al. 2011). It works as follows: by setting n to e.g., 1 we would

look at the different words in a sentence, and taking these as the basic units for further quantitative analysis. Applied to our example sentence in (83), the *tokens* become *n*-grams based on the value given to *n*: unigrams (84a), bigrams (84b), or trigrams (6.3.2).

- (84) a. Unigram ($n = 1$): *I, hit, the, ball, hard*
- b. Bigram ($n = 2$): *I hit, hit the, the ball, ball hard*
- c. Trigram ($n = 3$): *I hit the, hit the ball, the ball hard*

Of course, the point is not to just divide 1 sentence or string of words into tokens and then be done with it; the point is to quantify this over many strings and then eventually be able to predict the probability of combinations of target words or group of words.⁶⁸ Without doubt, such statistical measures have had many important applications, especially in the digital humanities. For example, *n*-grams have provided quantitative arguments for solving genre classification (Stamatatos, Fakotakis & Kokkinakis 2000), authorship attribution (Kešelj et al. 2003; Hung, Bingenheimer & Wiles 2010), and stylometric problems (Van Hoey 2014). However, let us return to our example in (83). While *n*-grams may be able to capture many relations of the ones we postulated, an interesting grouping such as *hit hard* (verb and adverb – although these syntactic relations are not the main point) may not be found, simply because there are too many intervening words. This is a problem that can be solved by using so-called SKIP-GRAMS.

⁶⁸The mathematical formula for finding *n*-grams where x_i is a target, is $x_{i-(n-1)}, \dots, x_{i-1}$. The chance of finding a certain combination, then is expressed as $P(x_i | x_{i-n+1}, \dots, x_{i-1})$.

Guthrie et al. (2006) define skip-grams in a sentence composed of the words $w_1 \dots w_m$ as follows (85), where skip-grams reported for a certain skip distance k allow a total of k or less skips to construct the n -gram.

(85)

$$\{w_{i_1}, w_{i_2}, \dots w_{i_n} \mid \sum_{j=1}^n i_j - i_{j-1 < k}\}$$

For our example sentence in (83), this would give the following permutations when the skip k is set to 2 (86a) or to 3 (86b) (leaving out the period for ease of comparison).

- (86) a. *I hit, I the, I ball, hit the, hit ball, hit hard, the ball, the hard, ball hard*
- b. *I hit the, I hit ball, I hit hard, I the ball, I the hard, I ball hard, hit the ball, hit the hard, hit ball hard, the ball hard*

As can be seen in (86), many more tokens (the units for further analysis) are generated in this way. In fact, as Guthrie et al. (2006:1222–1223) show, for 2-skip-trigrams (86b) this number is at least three times higher than in adjacent trigrams . Below we give their computations for bi- and trigrams with different skips⁶⁹ .

⁶⁹This is generated by the formula

$$n \sum_{i=1}^{k+1} i - \sum_{i=1}^{k+1} i(i+1), \text{ for } n > k + 2 = \frac{(k+1)(k+2)}{6}(3n - 2k - 6) \quad (1)$$



Table 6.2: Bigrams

Sentence Length	Bi-grams	1-skip	2-skip	3-skip	4-skip
5	4	7	9	10	10
10	9	17	24	30	35
15	14	29	30	50	60
20	19	37	54	70	85

Table 6.3: Trigrams

Sentence Length	Bi-grams	1-skip	2-skip	3-skip	4-skip
5	3	7	10	10	10
10	8	22	40	60	80
15	13	37	70	110	155
20	18	53	100	160	230

This shows that k -skip- n -grams overcome one of the problems we noticed above, namely finding *hit hard* as a combination in our example sentence. On the other hand, they do over generate (like most computational approaches): there are many combinations that we may not find interesting from an analytic point of view. This problem is tackled by providing a large amount of data – a sentence should not be the sole object of computation. Rather, a body of sentences or texts is what should be run through the process. Based on the frequency of certain combinations it then becomes clear what the useful groupings are, and what units can be used in

the word-based context model. More specifically, we need to compute the frequencies.



6.3.3 Step 3: Frequencies and co-occurrence strength

Peirsman et al. (2015:59–60) illustrate this step by looking up three target words in their syntactic contexts. However, this is where the path of this study starts to diverge: all the co-occurrence frequencies for all skip-grams need to be computed. This will lead to a contingency table like the one given Table 6.4, where the corresponding values of every word are given based on the skip-grams found in (86b).

Table 6.4: Contingency table that takes all skip-grams in (86b) into account

	I	hit	the	ball	hard
I	6	3	3	3	3
hit		2	2	2	
the			1	1	
ball					
hard					

However, because words usually do not occur with the same frequency in a corpus, these values need to be weighted. A method that has gained decent popularity in the last three decades is the Pointwise Mutual Information index (PMI, Church & Hanks 1990), which “expresses whether a context feature c co-occurs with target word w more or less often than expected by chance. If the target word and a contextual feature occur inde-

pendently from one another, their probability of co-occurrence $P(w, c)$ equals $P(w) \cdot P(c)$ " (Peirsman, Geeraerts & Speelman 2015:60–61). This is computable, because we can divide their actual probability (or relative frequency) by the product of the probabilities of w and c – assuming independence.

(87)

$$\begin{aligned} PMI(w, c) &= \log \frac{P(w, c)}{P(w) \cdot P(c)} \\ &= \log \frac{\text{relfreq}(w, c)}{\text{relfreq}(w) \cdot \text{relfreq}(c)} \end{aligned}$$

Next, this value is transformed on a logarithmic scale, with values closer to 0 meaning more coincidental, and values farther away from 0 indicating possible collocational strength. This is expressed as follows in (87), and would give $PMI(I, hit) = \log((3/26)/(1/5 * 1/5)) = 1.06$ for I *hit*, and $PMI(I, I) = \log((6/26)/(1/5 * 1/5)) = 1.75$ for I and I . That a higher value is computed for the target word and the target word is a logical consequence of using this computational method.

6.3.4 Step 4: Similarity

The computation of similarity between word meanings depends on a number of methodological choices. In computational linguistics, the basic method of skip-grams has been gaining traction in recent years, most notably within the domain of machine learning. A very widely accepted

method has been the so-called continuous skip-gram with negative sampling approach, developed at Google (Mikolov, Yih & Zweig 2013; Mikolov et al. 2013), which was popularized through the Word2Vec application (Goldberg & Levy 2014). The tokens that are attained by continuous skip-grams (similar to the approach explained above) are quantified over a large body of texts and then thinned out by a process they call negative sampling in the hidden layer of the neural network⁷⁰. This processing-heavy method is very good at predicting similar words. One famous example is the *word math* that such a system can do, Mikolov et al. (2013) were able to compute numbers for the words *king*, *queen*, *man*, and *woman* in the word space. They then asked the model to subtract *man* from *king*, and add *woman* to it, resulting in a value that is very close to the value for *queen*. So that would be $KING - MAN + WOMAN \approx QUEEN$, or visualized in Figure 6.2.

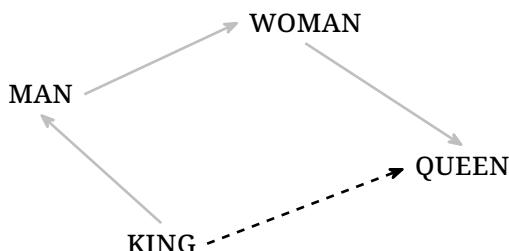


Figure 6.2: Word math

The results they attain have proven inspiring, as can be seen by the large number of computational studies that have tried to improve their methodology⁷¹ (Lenci 2018). Specifically of interest to our study are problems like

⁷⁰This method updates the weights with a small number of “negative” words, meaning that the output of the network will be 0 for these contexts.

⁷¹The paper by Mikolov et al. (2013) has been cited more than 10,000 times by the time of writing.

Polysemy and semantic change over time.

First, there is the problem of polysemy. Both computational linguists (Reisinger & Mooney 2010; Huang et al. 2012; Tian et al. 2014) and Cognitive linguists (Peirsman, Geeraerts & Speelman 2015:70–72) have argued that the modelling of polysemous words is a pervasive problem in distributional semantics. If words are polysemous (or homosemous), like *bank*, *left*, *doctor*, they all will influence the contextual distribution, resulting in a vector value in the word space that is located somewhere between the different meanings, weighted by frequency. If one meaning occurs more frequently than other ones, this frequent meaning will be reflected by the contextual distribution and nearest neighbors. Reisinger & Mooney (2010) and Huang et al. (2012) have tried to tackle the problem by taking two-steps approaches. The first step pre-trains the vectors to get single prototype word representations through a multi-layer neural network, and the second step identifies multi-word embeddings for each polysemous word by clustering all the context window features, which are usually computed as the average of single prototype embeddings of its neighboring words in the context window (Tian et al. 2014). However, a multi-prototype word embedding model can also be created with machine learning if more probabilistic statistics are introduced in the neural net. For example, Tian et al. (2014) are able to differentiate *apple* in *apple_1* and *apple_2*; the former having *strawberry*, *cherry*, *blueberry* as its most similar words, and the latter *iphone*, *macintosh*, *microsoft* as its highest similarity. Of course, it should be remembered that they use a huge corpus, implying that in small focused corpora (like



DIACHIC) this approach may not be as necessary.

Second, the issue of semantic change has been addressed by Hamilton, Leskovec & Jurafsky (2016), who compare three similar computational methods: positive point-wise mutual information (PPMI)⁷², singular value decomposition (SVD)⁷³, and skip-gram with negative sampling (SGNS ‘Word2Vec’). They recommend using singular value decomposition in synchronic accuracy tasks and ‘detection tasks’, but skip-grams with negative sampling for ‘discovery’ tasks; they recommend not using positive point-wise mutual information because it performs worse compared to the other methods. In other words, if the task is to compare if words became more similar over time (their meanings became more similar and thus their mutual distance decreased), and the findings were put next to well-known case studies such as *gay* moving from HAPPY, SHOWY to HOMOSEXUAL, LESBIAN, then singular value decomposition was the better task. The ‘discovery’ task focused on the top 10 words that changed between 1900 and 1990 in English, where skip-grams with negative sampling came out on top.

So, it is clear that the field of machine learning is advancing rapidly. However, as Levy, Goldberg & Dagan (2015) show, the high performance of these methods that depend on machine learning may be due to hyper parameters, rather than the machine learning component. They refer to skip-

⁷²This method replaces negative values obtained by PMI with 0.

⁷³This is a mathematical method to reduce the high n -dimensionality of the word space matrix to a lower number of dimensions, a method that is generalized from latent semantic analysis (Deerwester et al. 1990; Landauer & Dumais 1997) and that is found to make the correlations more robust (Levy, Goldberg & Dagan 2015; Hamilton, Leskovec & Jurafsky 2016)

grams with negative sampling and GloVe⁷⁴ as “neural” or “prediction-based” embeddings, as opposed to positive point-wise mutual information (PPMI) and singular value decomposition (SVD), which are “count-based” methods.

Moreover, the word math function (Figure 6.2) that gave the Word2Vec its popularity, can also be achieved through linear algebraic methods (Levy & Goldberg 2014). This is an important point, because it means we can continue along the trend of computational cognitive linguistics as discussed above, without relying too much on the black box. However, what the previous discussion shows is that matrix reduction through singular value decomposition is a useful step that needs to be taken.⁷⁵

How, then, is similarity of this reduced matrix calculated? Peirsman et al. (2015:61) suggest finding the cosine between the target word and its context vector and those of all other words in the corpus, shown in (88). The result of this calculation is a list that can be ordered by decreasing cosine, which results in the *nearest neighbors* appearing at the top.

(88)

$$\cos(v_1, v_2) = \frac{\sum_i (v_{1i}, v_{2i})}{\sqrt{\sum_i v_{1i}^2} \cdot \sqrt{\sum_i v_{2i}^2}}$$

⁷⁴GloVe is short for “Global Vectors”, a method that introduces context vectors in the machine learning algorithm (Pennington, Socher & Manning 2014).

⁷⁵Following Levy et al., “singular value decomposition factorizes M into the product of three matrices $U \cdot \Sigma \cdot V^\top$ where U and V are orthonormal and Σ is a diagonal matrix of eigenvalues in decreasing order. By keeping only the top d elements of Σ , we obtain $M_d = U_d \cdot \Sigma_d \cdot V_d^\top$. The dot-products between the rows of $W = U_d \cdot \Sigma_d$ are equal to the dot-products between rows of M_d . In the setting of word-context matrices, the dense, d -dimensional rows of W can substitute the very high-dimensional rows of M . Indeed, a common approach in NLP literature is factorizing the PPMI matrix M^{PPMI} with SVD, and then taking the rows of $W^{SVD} = U_d \cdot \Sigma_d$ and $C^{SVD} = V_d$ as word and context representations, respectively” (Levy, Goldberg & Dagan 2015:213). In our study we will rely on the `widyr` package for R to achieve this reduction.

6.3.5 From DIACHIC to semantic vectors for ideophones

For the operationalization of the methodology outlined above, I have chosen to make use of the programming languages python and R. It follows the steps mentioned above.

Step 1: gathering segmented data. The data for this chapter can be found in the Diachronic Chinese Ideophone Corpus (DIACHIC), which is a subcorpus of the Scripta Sinica, see details in Section 3.3.2. Two issues still remain. On the one hand, we might be able to cope better with the polysemy problem raised above. The idea is that since this is a somewhat specialized and smaller corpus, problems relating to the polysemy of *apple* as ‘fruit’ and ‘technology company’ will be limited. On the other, it can be argued that the subcorpus is just not big enough, or does not reach the critical mass to calculate semantic vector spaces. This stems from our arguably being spoiled in this day and age when it comes to corpus size, as most computational approaches definitely have come from data belonging to the big data era, and historical corpora are not comparable in terms of size. However, that does not mean we should not try. After all, it is impossible to invent new material that wasn’t recorded. Nevertheless, this issue does raise a potential problem with the results and their validity. So we propose to treat this issue as a disclaimer, and see how far we can get. As a helpful guide, we have reproduced Figure 3.5 without genre divisions in Figure 6.3.

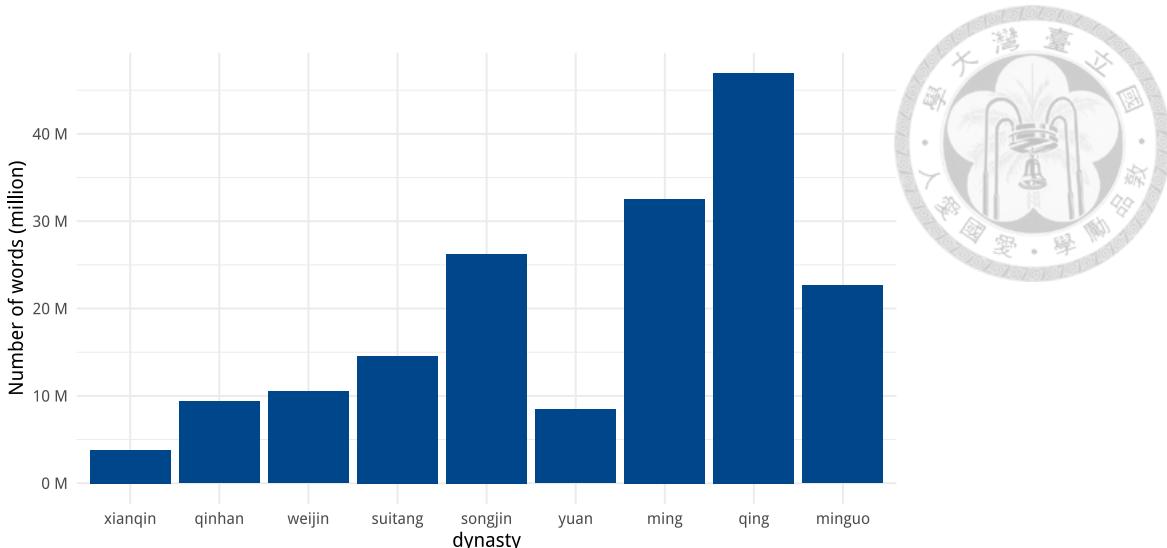


Figure 6.3: The number of words per dynasty in DIACHIC

Step 2: deciding on the model and units. It is important to mention that from this step onwards we rely heavily on R scripts written by the creator of the `tidytext` package in R, Julia Silge, and her adaptation of linear algebraic methods to create word space models in R. Our scripts are based on two of her blog posts: (1) *Word vectors with tidy principles*⁷⁶ and (2) *Tidy word vectors, take 2*⁷⁷, published on 30 October and 17 November 2017 respectively.

As mentioned above, the context model that fits our goals best, is the word-based context model. The units in our model, then, are skip-grams that are identified by a function `slide_windows()`, which works as a context window that calculates the skip-grams starting on each word by metaphorically sliding over the segmented text. I set the size of the sliding window to 8 words, which I believe gives reasonable contingency in Chinese.

Step 3: frequencies and co-occurrence strength. Next, the strength of the co-occurrence of the skip-grams that resulted from the previous func-

⁷⁶<https://juliasilge.com/blog/tidy-word-vectors/> Silge (2017a)

⁷⁷<https://juliasilge.com/blog/word-vectors-take-two/> Silge (2017b)

tion needs to be weighted. For this we will also calculate the PMI value for each pair of words. This is then stored in a TIDY data frame.⁷⁸ This function is called `tidy_pmi()` and results in an extremely long and thin table. A constraint I added is that every skip-gram has to occur at least 20 times, to increase the computational efficiency.

Step 4: similarity. To reduce the dimensionality the `widely_svd()` function from the `widyr` package is used. In her example, Silge uses 256 dimensions. We have used 100, because the results are deemed adequate enough. However, this was also done out of practical consideration, as we could not get values for 256 dimensions for the smallest period groups in DIACHIC. This operation eventually results in a data frame object `tidy_word_vectors`, which can be queried for the nearest neighbors of a target word.

6.4 Semasiological salience

Our first task is to computationally replicate findings from Chapter 5. For this we are looking at *zhuó~zhuó* 灼灼, the item in the case study that had the most tokens. After the calculation of the semantic vectors per period, it is possible to identify the nearest neighbors of this item, what in essence is a semasiological conception of meaning. That is to say, we look at an ideo-phone, and try to find the most salient meanings, understood here as the collocates with the highest cosine value (see Peirsman, Geeraerts & Speelman 2015 for an introduction).

As an extra check, however, we are relying on the *Hànyǔ dà cídiǎn* 漢

⁷⁸This means that every variable has its own column, every row is an observation (per variable), and that each type of observational unit is a table (Wickham 2014).

Table 6.5: Top collocates for *zhuó~zhuó* 灼灼 per period in the DIACHIC

xianqin	qinhan	weijin	suitang	songjin	yuan	ming	qing	minguo
灼	火	玉	露	曾祖	露	光	排印本	真言
煌煌	反注	光	華	張氏	去	玉	謂疏	香水
灼灼	反下	綠	光	墓誌	英	龍	即	燒
翠葉	燒	灼	灼	陳氏	將	火	易堂	一百八
光	又	露	貌	作	灼	天		燒香
嬌	音	華	雲	郎知	傷寒			塗香
照	內	煌煌	音	子男	象金	作		香花
翠	灼	照	賦	曰	已	和		芥子
露	爛	貌	照	夫人	露	生		方
生	雲	雲	煌煌	太君	亭亭	燒		香
火	于偽	英	綠	長適	翠	故		沈水香
英	同	雜	爛	孺人	貞	日月		香華
熱	或	翠	玉	李氏	兩	內		氣
熾	反本	離	也	大父	若	陽		一千八
炎	汗出	賦	楚辭	太夫人	仲景	星		三時
華	復扶	若	草木	氏	在	五色		蘇
霞	若	灼灼	英	曾祖諱	曰			和
靈	風	爛	陸離	承事	三			香湯
參差	化為	翠葉	灼灼	公諱	當			日
雲	易	參差	彩	孫女	玉			檀香

語大詞典 for an extra filter. A pre-study showed that a number of grammatical elements had high cosine values. In order to focus just on lexical collocates, we decided to get all the definitions in the *Hànyǔ dà cídiǎn* for the 35 ideophones in the original sample of Chapter 5. From these, all characters occurring 5 times or more in those definitions were kept as a proxy for lexical relatedness to the collocates. On top of this, Chinese punctuation signs were left out, e.g., < <>,< > > and “empty” characters like < 兮 >.

Consequently, Table 6.5 shows the top 20 nearest neighbors for *zhuó~zhuó* 灼灼 per period in the DIACHIC. This a moderate success: many of these items clearly are in the semantic domain of LIGHT or related domains we identified before in the previous chapter. As a comparison, let us look at the diachronic extensional figure made before, reproduced in

Figure 6.4.

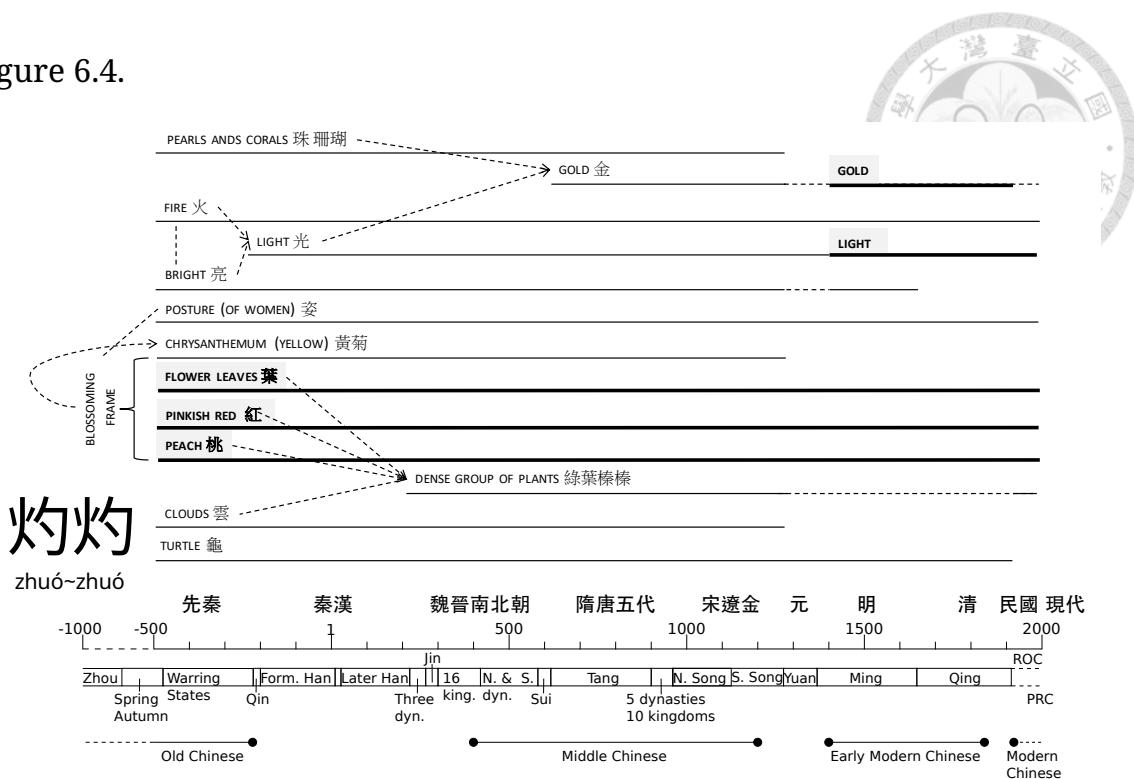


Figure 6.4: The semasiological analysis of *zhuó-zhuó* 灼灼 (see Chapter 5)

Table 6.5 and Figure 6.4 share many frame elements, e.g., BLOSSOMING FLOWERS frame. These include ‘flower’ 花, 花, ‘(pinkish) red’ 紅, ‘spring’ 春, ‘tree’ 樹 etc. Curiously, also the colors ‘green’ 綠 and ‘white’ 白, which were not identified before. Another group of words relates to ‘beautiful woman’ 美人, 佳人, 夫人. There are the more general LIGHT words, e.g., ‘shine’ 照, ‘light’ 光, as well as a number of ideophones that seem to be conceptually close to *zhuó-zhuó* 灼灼, like *huáng~huáng* 煌煌, *tíng~tíng* 亭亭, *qīng~qīng* 青青 etc.

But this computational approach not only allows us to get the top n nearest neighbors. We can also visualize the distance between these items. After all, every neighbor is related to a number that expresses cosine distance. In Figure 6.5 the top neighbors are shown. To improve clarity, overlapping items are not shown. The cosine distance between most neighbors is mini-

mal, although a general trend is that it appears to decrease over time.

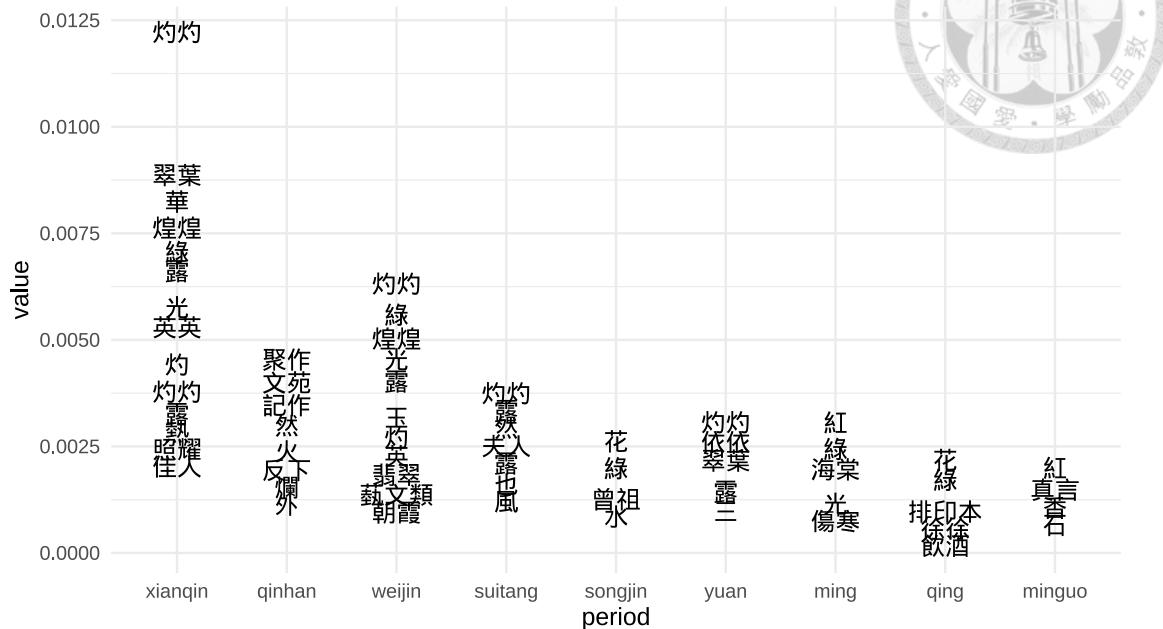


Figure 6.5: Semasiological salient collocates for *zhuó~zhuó* 灼灼

From this first case study it can be summarized that the manual method from Chapter 5 is corroborated by this computational approach, or vice versa, that this computational approach does not stray too far from the manual method. However, it needs to be borne in mind that corpus size may be an issue. On the upside, having this computationally quantified perspective next to a traditional analysis provides nuance, and can aid in lexicological in-depth studies of ideophones.

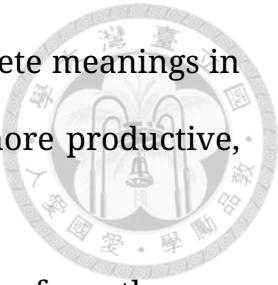
In terms of semasiological salience, we can see that the highest neighboring values (apart from the ideophone itself) are very variable, which reflects in a way the dynamicity of prototypicality. Looking at Figure 6.5 the long-term evolution of the items does suggest that there is at least a prototypical core or bundle of prototypical cores that remain throughout history in this ideophone's usage.

But is it possible to study new items with these methods? Three other LIGHT ideophone variants that have not received any attention so far are *càn~càn* 燦爛, *càn~làn* 燦爛 and *làn~làn* 爛爛. Just like in the previous chapter, let us first inspect the definitions provided by the *Hànyǔ dà cídiǎn*, shown in (89).

- (89) a. *càn~càn* 燦爛: 1. ‘glittering’ 閃閃發亮貌。; 2. ‘colorful appearance’ 色彩鮮艷貌。
- b. *càn~làn* 燦爛: 1. ‘bright appearance’ 明亮貌；鮮明貌。; 2. ‘gorgeous’ 華麗；絢麗。; 3. ‘beautiful poem’ 形容文辭華美。; 4. ‘beautiful things’ 形容事情或事業輝煌；美好。
- c. *làn~làn* 爛爛: 1. ‘bright’ 光亮貌；光芒閃耀貌。; 2. ‘colorful appearance’ 色彩鮮艷貌。

From the definitions we can see that these three formal variants have overlapping meanings, most of all in the depiction of ‘light’. *Càn~càn* and *làn~làn* are full reduplications, but the curious thing is that *càn~làn* also exists. I believe this is motivated by the onset of *làn*, /l-/ , which goes back to Middle Chinese /l-/ and Old Chinese */r^h-/. According to Sun (1999:67), these liquid consonants often occur in progressive reduplication in Old Chinese, so it makes sense from that phonological perspective that we get the evolution from the Old Chinese form */ts^han-s~r^han-s/ to the Middle Chinese /ts^han³~lan³/ to Standard Chinese /ts^han˥~lan˥/ (*càn~làn*). But it is also motivated by the existence of the full reduplications, which may well be derivative of the progressive reduplication through reanalysis⁷⁹, or the other way

⁷⁹This is reminiscent of the relation between *máng~máng* 茫茫, *máng~máng* 芒芒 and



around. The variant *càn~làn* with its relatively more discrete meanings in the dictionary definitions does suggest that it has been more productive, because of its higher type frequency.

As is visualized in Figure 6.6, the highest ranking tokens from the vector spaces partially confirm the higher entrenchment of the mixed form *càn~làn*. First of all, there is no data *càn~càn* in the broad periods of “weijin” and “suitang”, or “minguo”. The most obvious explanation for this is that it is an offshoot of one of the other two. However, it could also have been knocked out of the data if there just were not enough tokens. In any case, it strongly suggests that it is not a relevant competitor in the first two periods and maybe towards the end, but quite competitive in the middle part of the plot (“songjin” to “ming”). The other fully reduplicated variant, *làn~làn*, on the other hand, is present throughout time. Yet, its nearest neighbors are not as near as the mixed variant *càn~làn*, which also occurs throughout time and has higher cosine values.

cāng~máng 蒼茫, as argued previously (Van Hoey & Lu 2019a).

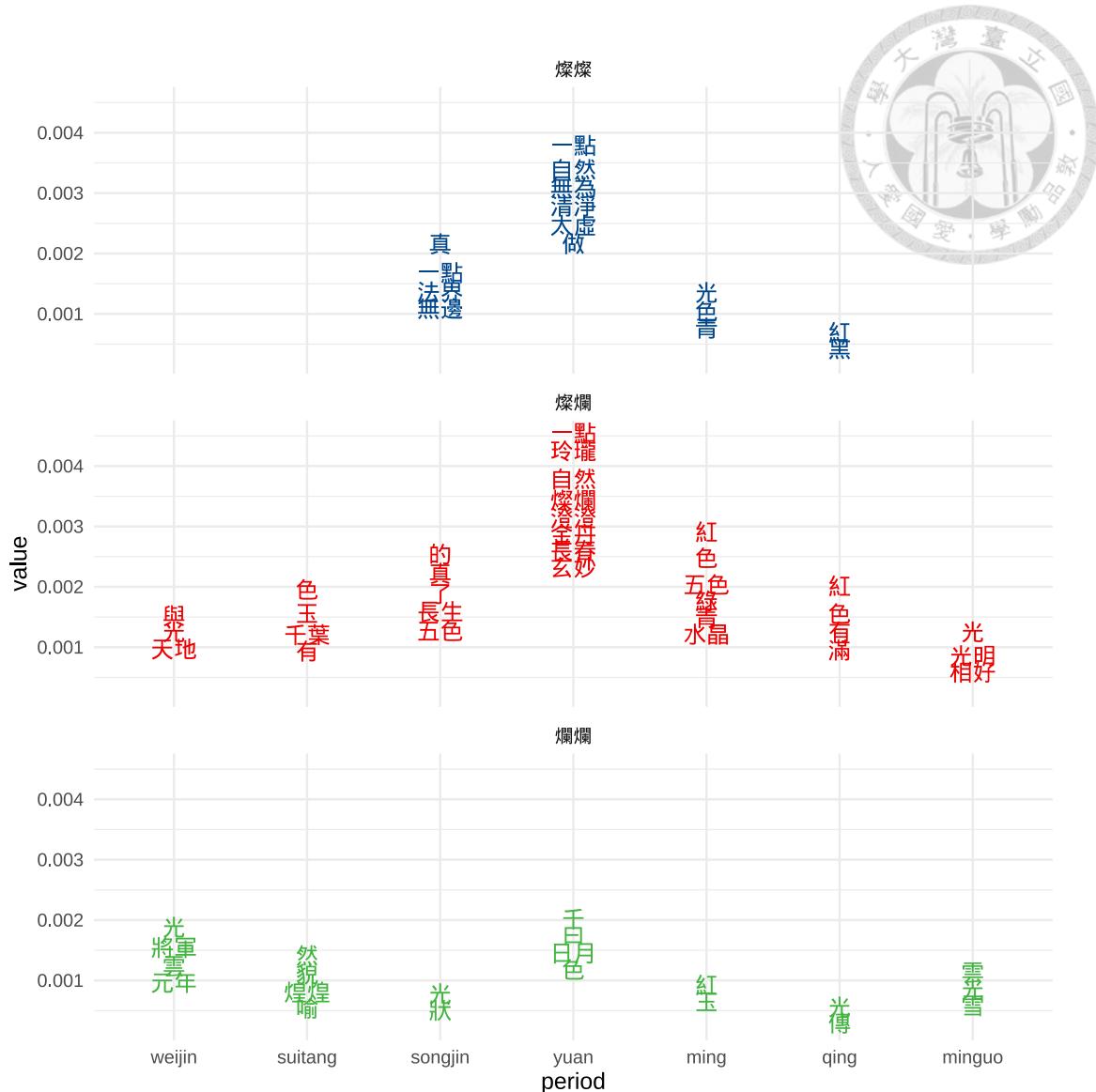


Figure 6.6: Comparing nearest neighbors for *càn~càn* 燦燢, *càn~làn* 燦爛 and *làn~làn* 爛爛

Additionally, note that these three variants are not fully compositional. Through a Modern Chinese lens, one could interpret *làn* as ‘rotten, tender’ (Chen et al. 2019). In their experimental materials it is stated that *càn~làn* is compositional, and more precisely a compound meaning ‘bright-tender’. However, it is quite clear from what we have seen so far that the *càn* and *làn* syllables in question belong to the ideophonic lexicon rather than the prosaic lexicon.

In terms of semasiological salience each of these three variants also displays the same dynamic meanings over time. Prototypicality is operationalized as higher cosine value meanings for the neighbors, and could potentially be the start for more detailed lexicographic studies. However, one element that was present in the previous chapter that we have in these visualizations is the extension between different collocates. That does not need to be a serious problem, though, since a full lexical semantic study now has two sources of evidence to approach the semasiological problem: given a form, what is its meaning structure, at one point in time or over time. Armed with quantified rankings, it becomes possible to outline a statistically driven approach to prototypicality phenomena, and to frequency effects as well. For type frequency, the researcher could make a cut-off point and count types per period. For token frequency, then, one could count the raw tokens that might still be available. In the case of DIACHIC, they are available, although the matter will not be further pursued here.

This example of *càn~càn* 燦燦, *càn~làn* 燦爛 and *làn~làn* 爛爛 also leads us to onomasiological salience, since it pertains to three competing forms. From the semantic vector space, it appears that the chance of *càn~làn* is higher than that of *làn~làn*, and most certainly *càn~càn*. *Càn~làn* is also the variant that still occurs regularly even today, so it has outshone its competitors. But how do all these different ideophones hold up against one another? That is a question that will be explored in the next section.

6.5 Onomasiological salience



Within (Cognitive) lexical semantics, onomasiology and semasiology are often considered to be two sides of the same coin. Semasiology investigates the structure of *meanings*. This includes familiar types of relations such as inclusion (i.e., generalization, specialization), literal similarity (i.e., family resemblances), figurative similarity (i.e., metaphor) and contiguity (i.e., metonymy). Onomasiology, on the other hand, is not as concerned with *meaning* but rather with *naming*. Related types of relations then include the same four: inclusion (i.e., taxonomies), literal similarity (i.e., lexical fields), figurative similarity (i.e., conceptual metaphor) and contiguity (i.e., frame semantics) (Geeraerts 2010b:284).

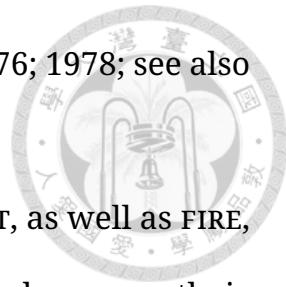
We have already seen examples of both perspectives in the previous and current chapters. For onomasiology, these examples mostly consisted of competing written forms. Typically, we have seen case studies related to the following question: given the same phonological form, what are the different variants that exist? This allowed us to study the meaning (semasiological perspective) of variants like two *yào~yàos* (Section 5.3.2), three *huī~huīs* (Section 5.3.3), three *yè~yès* (Section 5.3.4) and variants like *càn~càn*, *càn~làn* and *làn~làn* (Section 6.4).

It is now time to do more justice to the onomasiological perspective. In the vertical perspective adopted in the previous chapter, LIGHT was at the core of the different frames (Figure 5.14). This can arguably be thought of as a taxonomy as well, with LIGHT as unique beginner, and other frames as more specific meanings with basic-level status. This is analogous to natural

folk taxonomies (Berlin, Breedlove & Raven 1973; 1974; 1976; 1978; see also Geeraerts 2010b:199–203).

In the case studies here we will thus be looking at LIGHT, as well as FIRE, MOON, STARS and FLOWER, in order to investigate which ideophones are their nearest neighbors and, if the distributional hypothesis is correct, conceptually the most entrenched to these items. Fortunately, the semantic vector space per period has already been calculated. This greatly facilitates the operationalization of this issue: it suffices to set these frames as the target and filter the nearest neighbors for ideophones.

Figure 6.7 shows the result of this exercise for LIGHT, for which we used the item *guāng* 光. Displayed are all ideophones with the highest cosine value. Highlighted are the ones that are LIGHT ideophones. We can see some familiar items, such as *zhuó~zhuó* 灼灼, but the top items apparently are *huáng~huáng* 煌煌 and *líng~lóng* 玲瓏. The former can be glossed as ‘glittering and gleaming; sparkling and glistening, majestic and magnificent’; the latter as ‘tinkling of gems; clear and transparent’. These are two quite general ‘light’ meanings, namely the glittering aspect.



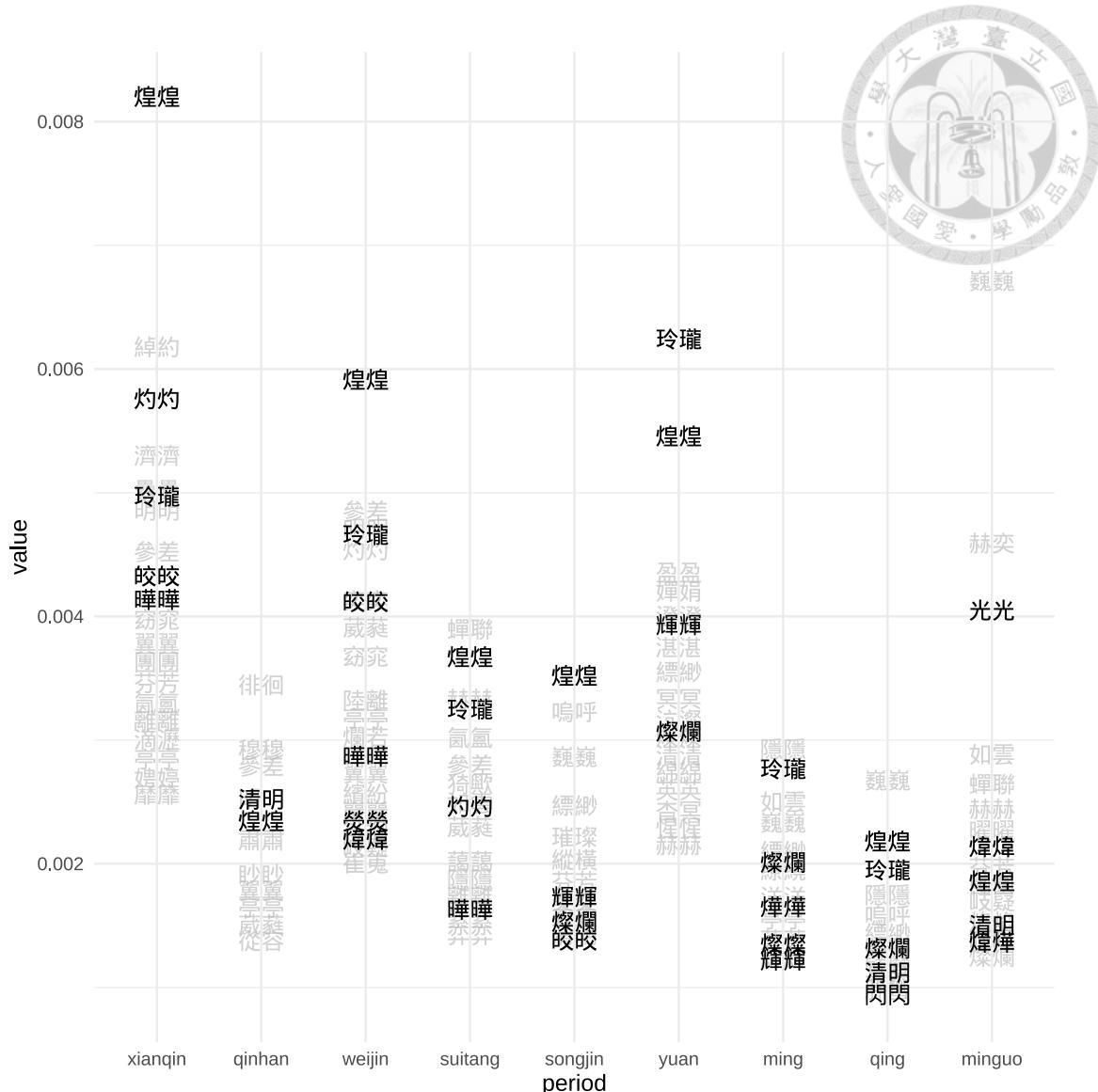


Figure 6.7: The nearest neighboring ideophones for *guāng* 光 LIGHT

More specific meanings should probably yield more specific aspects of LIGHT as well. For instance, the frame of FIRE, operationalized as the item *huǒ* 火 in the data, results in Figure 6.8. Curiously, while LIGHT ideophones do make the plot, they are not the nearest ideophonic neighbors by a stretch. This can be explained by the fact that fire is a well-known source of light, but also a source of heat, and has its own ontological status, that includes many other facets as well. The LIGHT ideophones that do occur with it, include among other the twinkling ones we just saw in the preceding paragraph,

but also another ‘twinkling’ ideophone, *yíng~yíng* 燔熔.

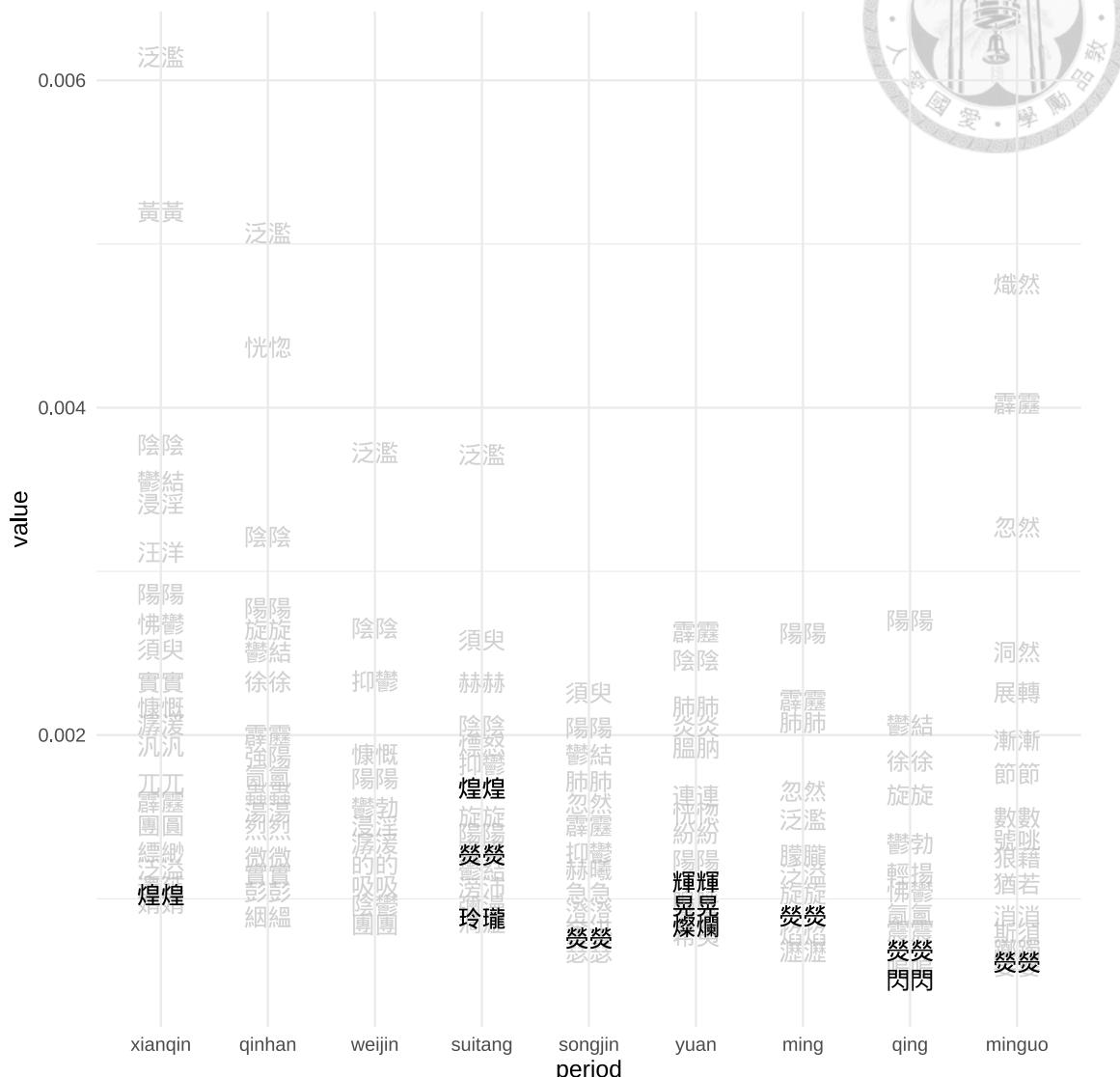


Figure 6.8: The nearest neighboring ideophones for *huǒ* 火 FIRE

And similar results can be found for MOON (*yuè* 月). Different aspects of the moon are highlighted through different ideophones, e.g., *juān~juān* 娟娟 ‘beautiful, graceful’. In terms of light ideophones, *jiǎo~jiǎo* 皎皎 ‘clear, glistening white’ is quite entrenched across time. Unsurprisingly, this highlights the whiteness and shine of the moon, especially when it is full. Other familiar ideophones are *yì~yì* 熠熠, *líng~lóng* 玲瓏 and *huī~huī* 輝輝. In this case, it might be better to gloss the twinkling ones as ‘brilliant’, although the

question of best gloss would turn the perspective back to semasiological.

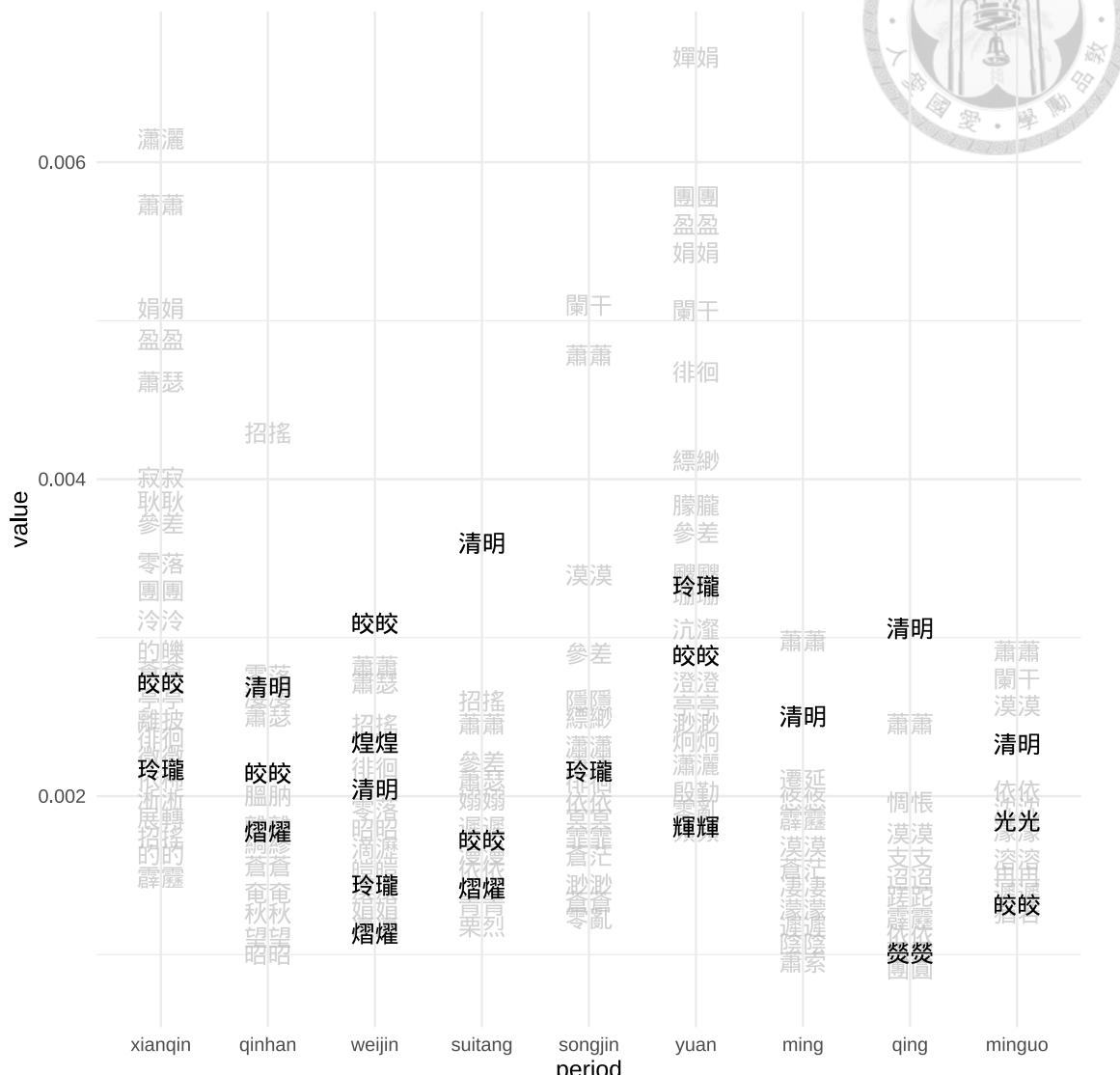


Figure 6.9: The nearest neighboring ideophones for *yue* 月 MOON

If we turn to another celestial frame, STARS, operationalized as *xīng* 星⁸⁰, we once again see a similar picture emerge. As demonstrated in Figure 6.10, the LIGHT ideophones for STARS are the top ones for the other frames we have inspected thus far, with a few additions, such as *yè~yè* 瞽瞽, familiar from a previous case study. Yet, *huáng~huáng* 煌煌 comes out on top again, throughout history.

⁸⁰*Xīng* not only means ‘stars’ but can also mean ‘planet’. In Modern Chinese a reduplicated form *xīng~xīng* 星星 is used more often to denote ‘stars’.

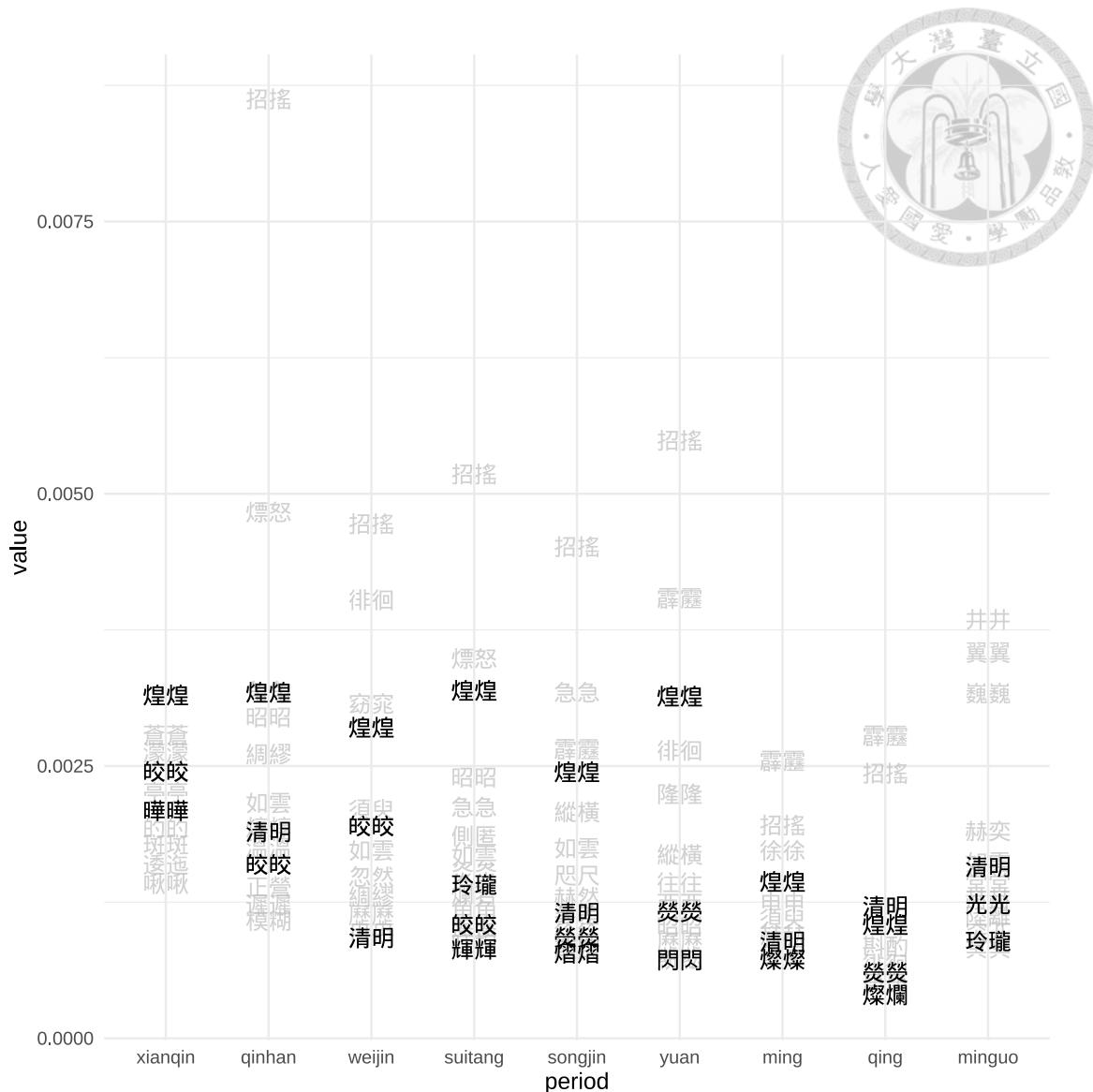


Figure 6.10: The nearest neighboring ideophones for *xīng* 星 STARS

The emerging picture strongly suggests that these LIGHT ideophones and the collocates are closely related in the semantic vector space, which was calculated based on DIACHIC and contains information not just for LIGHT ideophones, but all items in CHIDEOD. It thus corroborates the usefulness of semantic vectors to computationally approach issues for the elaboration of ideophones and their collocational meanings on the one hand, and target words and nearby ideophones on the other.

In the last case study of this application of the semantic vector space,

we investigate FLOWER, which can be operationalized as *huā* 花 or also (earlier) *huá* 華. In Figure 6.11 it can be seen that there is a difference in the LIGHT ideophones attracted by the FLOWER variants, despite them also having a significant overlap. Furthermore, there is a difference in relatedness to LIGHT in the two cases. For *huá* 華, LIGHT ideophones are ranked relatively higher than they are for *huā* 花. In terms of depiction of sensory imagery, this suggests that the way the latter is conceptualized, is more related to other aspects of FLOWERS. That is to say, aspects like fragrance (*fēn~fēn* 芬芬, *fēn~fēi* 芬菲) seem more important than the different shining and brilliant colors in the *huā* 花 variant than in the *huá* 華 variant. However, this is based on the vector space. While the tokens were not counted, the former variant is likely to co-occur more often with the meaning ‘flower’ than the latter; that is also how it survived in Mandarin Chinese.

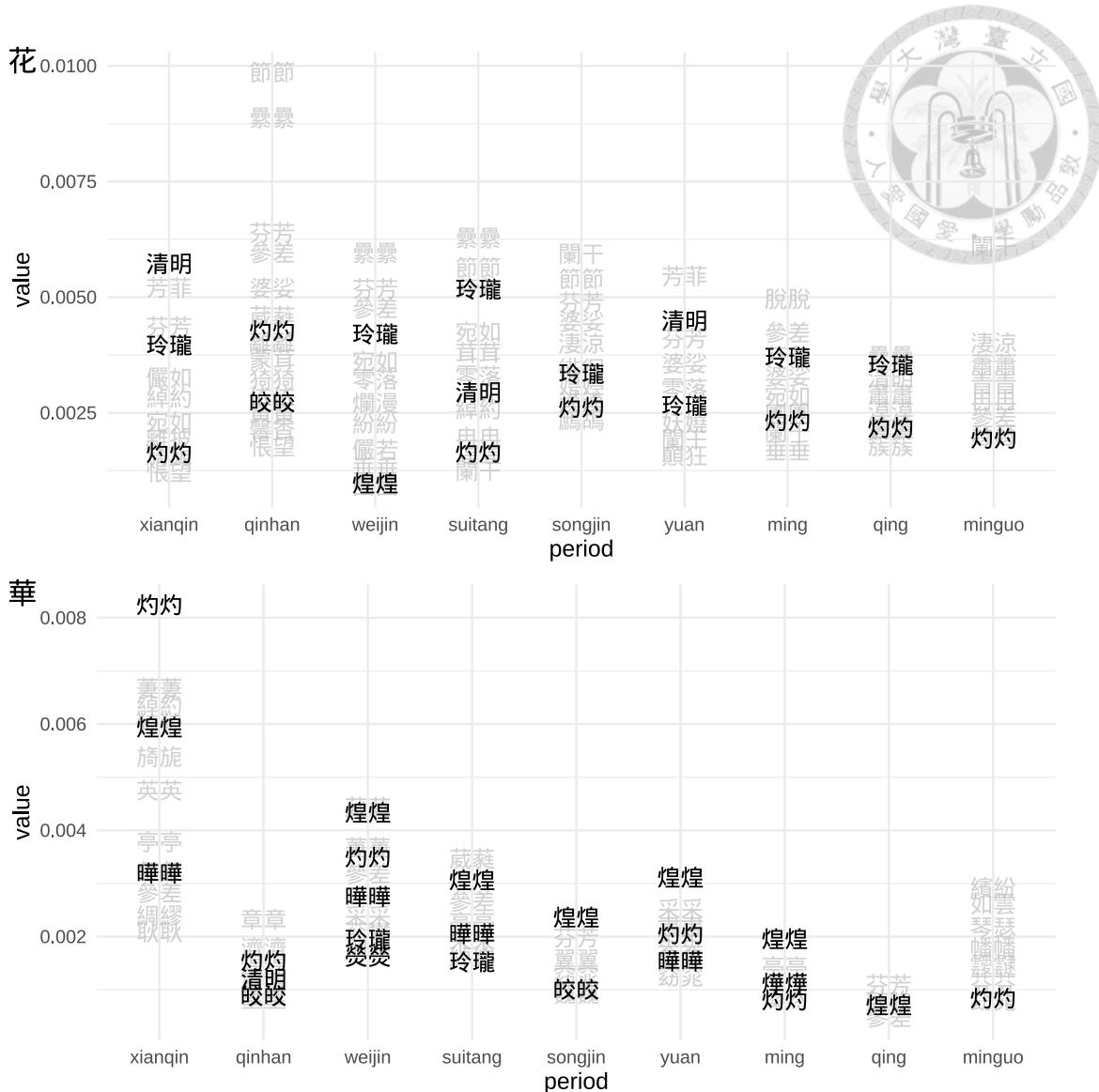


Figure 6.11: The nearest neighboring ideophones for FLOWERS: *huā* 花 and *huá* 華

These case studies also reveal the importance and potential of examining the variation of ideophones and their collocates. From the semasiological studies we learnt that the semantic elaboration of collocates is structured unevenly, with prototypical tendencies and fuzzy boundaries in a dynamic network. The lesson here is that LIGHT-related items seem to cluster in the same “corner” of the highly multidimensional semantic vector space, but that there are also differences in entrenchment (Geeraerts 2017).

But we can actually go beyond that: there are differences in directionality and reliance, or cue validity. For example, as the FLOWER case study shows, when you think of a flower, the brilliant colors may *not* be the first thing you want to depict (with ideophones). It might be the fragrance, or the denseness of a flower field etc. Consequently, LIGHT ideophones do not occupy the most entrenched spots in the ideophone ranking. Given the cue of FLOWER, it is not certain that one of our LIGHT ideophones, e.g., *zhuó~zhuó* 灼灼, will show up first.

On the other hand, when studying the semantic structure of an item like *zhuó~zhuó* 灼灼 in Section 6.4, we saw that items of a BLOSSOMING frame (which includes FLOWERS) displayed a high amount of salience. Such directionality is one of the consequences of taking the two sides of semasiology and onomasiology seriously. It will also play a very important role in Chapter 7, where a number of constructions are studied through association measures.

Lastly, the theoretical status of these collocates deserves a discussion as well. An important theoretical advance of onomasiology has been the differentiation between the conceptual layer and referential layer. Geeraerts and colleagues demonstrate the importance of deconflating these two levels in a number of studies that focus on clothing terminology in Belgian Dutch and Netherlandic Dutch (Geeraerts, Grondelaers & Bakema 1994; Grondelaers & Geeraerts 2003; Geeraerts 2017 among others). The idea is that, for example, a referential picture of a clothing item like JEANS can be conceptualized as a SPECIAL TYPE OF PANTS THAT IS BLUE AND LONG, or just as the hypernym

PANTS. From this referential level and the conceptual level, there is then the formal linguistic level, which in Belgian Dutch uses the word *jeans* for this type of pants, belonging to the former conceptualization and *broek* ‘pants’ in for the hypernym. In Netherlandic Dutch the two forms respectively are *spijker-broek* ‘nail-pants’ and *broek*. It is not a long stretch to see how such sociolinguistic entrenchment could be applied to the field of ideophones, if data for different groups of speakers are available, e.g., Taiwan Mandarin vs. Mainland Mandarin vs. Singaporean Mandarin etc. However, a more important note from this line of research is the issue of the three levels. In the case studies so far, we have dealt with two formal levels, the phonological and the orthographic – a consequence of the Chinese semiotic folk model. We have also used the conceptual level. But have we touched upon the referential?

I would argue that the border between the conceptual and referential levels is not clear-cut in many of our cases. We don’t have the real referent to compare ideophones to, like we could possibly do with current linguistic data, but many of the collocates have been of a relatively concrete nature: sun, moon, fire, flowers, lightning, etc. Because meaning and by extension language relies on encyclopedic semantics, we know that these concepts behave in similar manners per event. This is reinforced by the argument that ideophones depict sensory imagery, and rest upon notions of iconicity. That means that concepts are equaled in a strong reading, or that the boundaries are fuzzy, in a weak reading.

But how do the concepts or referents compare to one another, and to

the formal layers? To answer that question, we need to turn to structural salience.

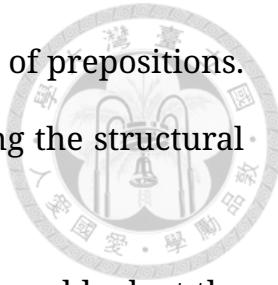


6.6 Structural salience

The third kind of variational salience, apart from semasiological and onomasiological salience, is what Geeraerts (2000) calls structural salience. In his example study of Belgian beer names and their lexical semantic structuring (originally Geeraerts 1999; reprinted in Geeraerts 2006b) he studies the relation between referential features, such as ingredients of beer or their provenance such as a monastery etc., and the names these beers are given. He finds that the feature of ‘seasonal’ has a very high structural salience, in the sense that 100% of the beers that were brewed seasonally were also marketed as being seasonal. Another feature is whether beers are brewed in a monastery, and again 100% of the beers that have this feature are marketed as such – called trappist beers. Contrast this to a feature like a low alcohol percentage: only 22.7% of those beers were marketed with this feature. So the structural salience of this feature is much lower compared to the other two.

Another, perhaps more accessible (for linguists) example is spatial terms, i.e., prepositions. Taylor (1988) compares three English and three Italian prepositions using Cognitive Grammar as a *tertium comparationis*. That is, he uses conceptual features like “the nature of the landmark”, “the nature of the trajector”, “contact”, “orientation of the trajector with respect to the landmark”, “static or dynamic relation”, and “the role of the

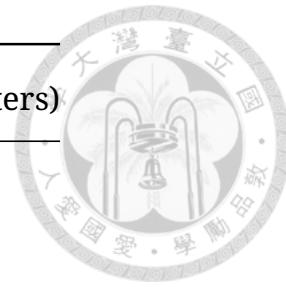
observer” (Taylor 1988:304–305) to compare these two sets of prepositions. Quantifying such conceptual features is a way of exploiting the structural salience.



In the domain of LIGHT ideophones, we might take a second look at the conceptual level (Section 5.4): the frame elements we have identified like LIGHT, SUN, MOON etc. are likely to be of differing distinctive weight; they have different degrees of structural salience. This idea can be operationalized by using the data set containing all LIGHT ideophones and their top collocates, as we have done above in Section 6.4. In this data set, the domains can be marked with proxies like those shown in Table 6.6. Note that this list is not claimed to be exhaustive, nor very precise, although it goes a long way in providing a coherent data set. Even more if we set a boundary to the minimum cosine value the neighbors need to have. I have (arbitrarily) chosen for a cosine value of at least 0.001. The remaining condensed data set still contains 4029 observations.

Table 6.6: Operationalization of structural salience

frames	operationalization (characters)
thunder	雷
metal	金
lightning	電
light	光
fire	火
stars	星
flower	華花



frames	operationalization (characters)
posture / beauty	姿佳美
lantern	燈
red	朱紅赤丹彤
color	色彩
pearl	珠
jade	玉
divine	神
sun	日陽
moon	月

Let us first pretend that our dataset is not diachronic in nature. This allows us to visualize the division of the different frames, i.e. the conceptual features. Figure 6.12 displays this distribution, both in relative frequency (percentages on the plot) and absolute frequency (marked with numbers on the plot). LIGHT, as operationalized through the character 光, is by far the dominant frame (of LIGHT as a semantic field), which is not that surprising as the manual study in Chapter 5 has already shown. This is followed by a shared second place for SUN and FLOWER. Next we have SHADES OF RED, then COLOR, METAL, JADE and so on. At the bottom we find LIGHTNING, LANTERN and THUNDER, which does appear somewhat odd, in the sense that lightning and thunder do seem quite marked in terms of light, yet here they are on the boundary.

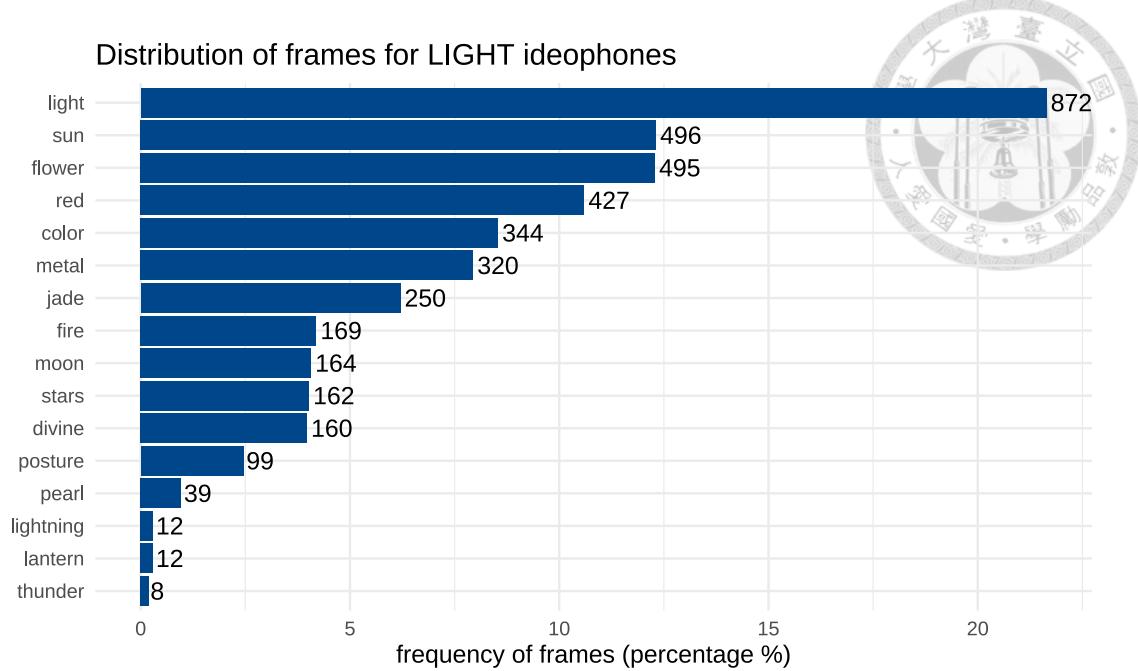


Figure 6.12: Distribution of frames for LIGHT ideophones

The structural salience in Figure 6.12 also has a few consequences for the network that was proposed in Chapter 5 (Section 5.4). For instance, we now have a quantified perspective on these frames, and can interpret their frequency as conceptual distance. We first reproduce the manually identified clusters in Figure 6.13. Compare this to the newer network in Figure 6.14.

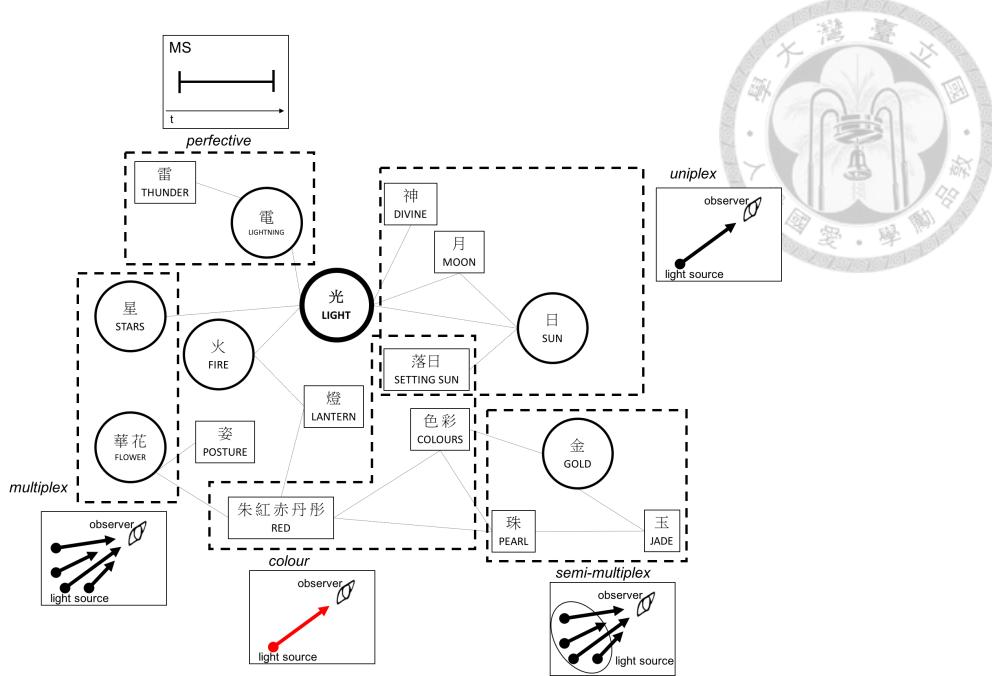


Figure 6.13: The original frames and domains (ICMs)

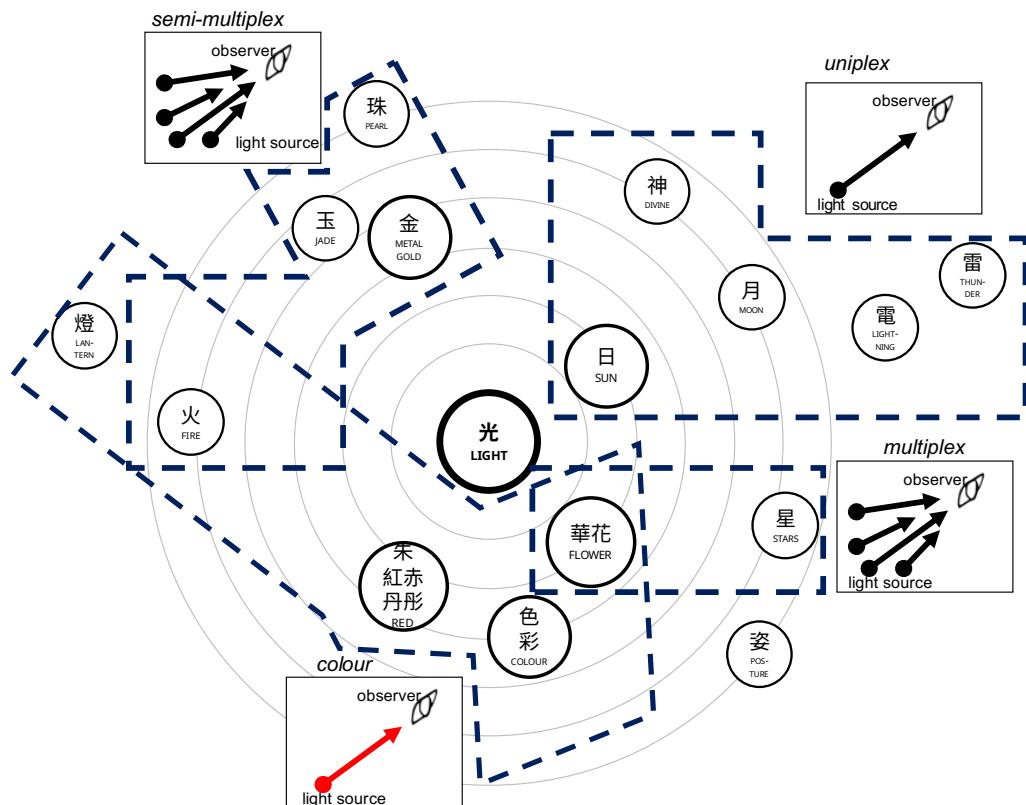


Figure 6.14: The revised frames and domains (ICMs)

We can see that the revised frames and domains (ICMs) in Figure 6.14 have the frames laid out across concentric circles with LIGHT in the middle

and the other frames further and further away from the center according to the frequencies identified above. This is a diagram drawn by me for the sake of clarity and does not represent any accurate mathematical distance; it reflects our conceptual distance.

It can also be seen that the grouping of the domains was reinterpreted, i.e., collections of frames, similar to Idealized Cognitive Models (see Kövecses 2017; Van Hoey 2018b). In Chapter 5 we found the following domains, apart from LIGHT itself: 1) UNIPLEX LIGHT SOURCE⁸¹, e.g., *the sun the moon*; 2) MULTIPLEX LIGHT SOURCE, e.g., *stars, flowers*; 3) SEMI-MULTIPLEX LIGHT SOURCE, e.g., *gold, pearls*, and *jade* – the idea being that the ‘sparkles’ occur on multiple places in the bounded entity; 4) COLOUR, e.g., *red* or different *colors*; 5) PERFECTIVITY, e.g., *lightning* and *thunder* which are semelfactive, which Talmy would classify as ‘uniplex’ (2000b:48–49). Here, on the one hand, the diagram discerns the following clusters: UNIPLEX, MULTIPLEX, SEMI-MULTIPLEX and COLOR. PERFECTIVITY has been reassigned to the cluster of UNIPLEX, based on theoretical grounds (Talmy 2000b) as well as frequency. On the other, it has become clear that the domains should probably overlap, as a way of visualizing componential analytic groupings (see Geeraerts, Grondelaers & Bakema 1994).

Apart from the four domains – UNIPLEX, MULTIPLEX, SEMI-MULTIPLEX, COLOUR – it may be possible to identify other features of the componential analysis as well. For example, the SEMI-MULTIPLEX domain can also be conceived of as being ARTIFICIAL, as opposed to the other frames in UNIPLEX and MULTIPLEX clusters, which are NATURAL. However, these do differ

⁸¹Plexity here refers to its usage by Talmy (2000b:48–50); see also Lakoff (1987:441–443)

from the frames that profile COLOR. So except for LIGHT, it seems that the frames elaborate either the shape or the color of their referent. Notably, every domain seems to have at least one frame that has a higher frequency.

FLOWER, RED, SUN, METAL and COLOR compose this set of more salient frames.

It is exactly in this way that we propose structural salience for the LIGHT ideophones can be identified.

In our current approach to structural salience we have been ignoring the factor of periodization. The picture that has emerged is perhaps slightly optimistic, as the frames have not had a uniform distribution across time. This can be seen in Figure 6.15. Fortunately, for the abstraction in the previous figure, the general distributions per period do not appear to deviate that much from those where the factor of time was ignored. That suggests that the conclusions still hold, but nuances that idea by adding some amount of dynamicity to the system.

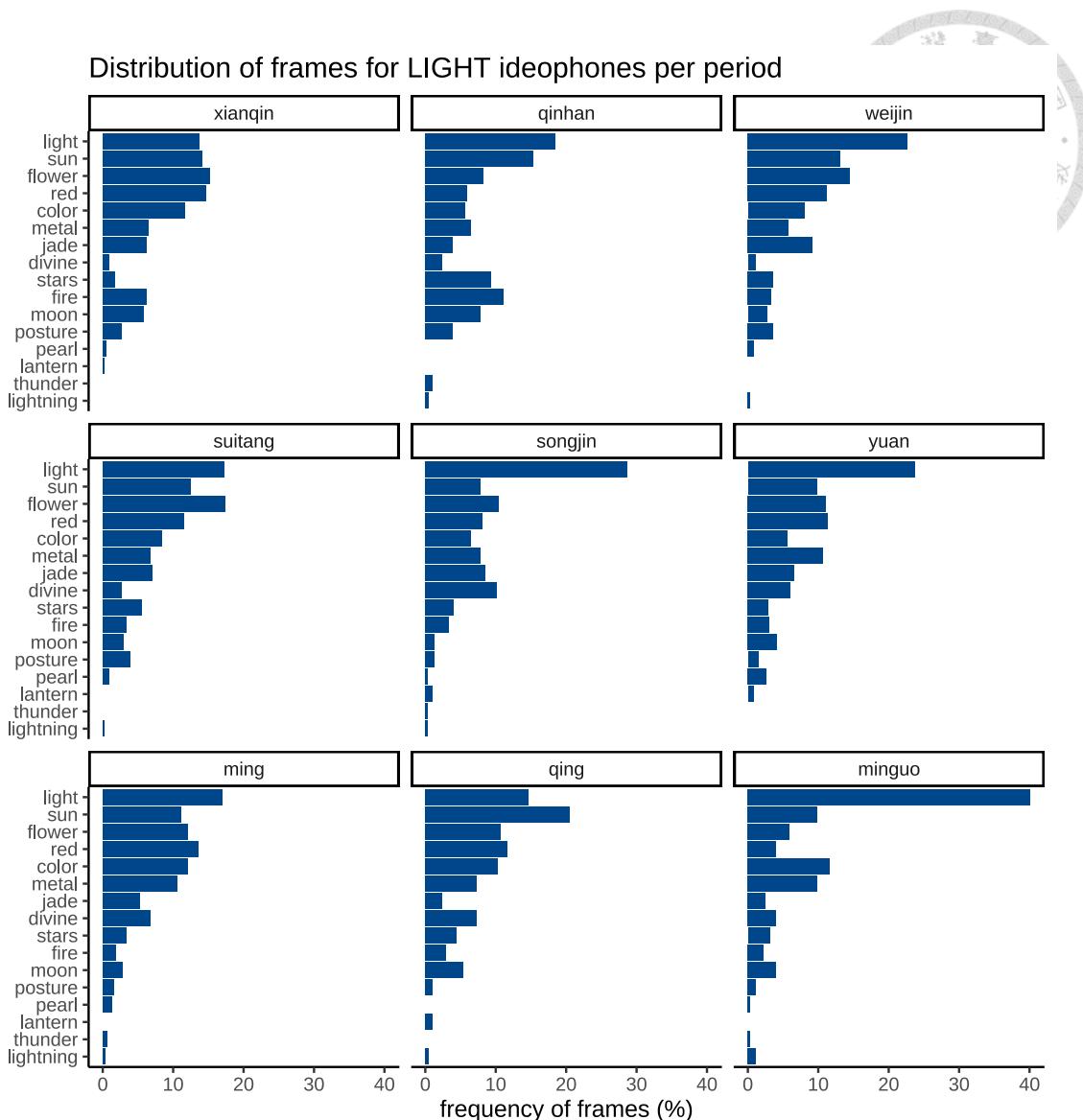


Figure 6.15: Distribution of frames for LIGHT ideophones per period

However, there is more structural salience to observe. We can exploit the nature of the Chinese writing system to not only look at phonological forms, but also at the orthography, as has been observed before. It certainly stands out that many LIGHT ideophones are written with characters that have recurring radical support (see Section 3.2.3.2). More precisely, a small set of semantic radicals occurs over and over in the items: LIGHT 光, FIRE 火, and SUN 日. There are some less frequently occurring semantic radicals as well: JADE 玉, WHITE 白, and METAL 金. Marking every item in our current data

set for its radical support, provides insight about the frequency of the combinations of frames and semantic radicals, both situated on the conceptual level. However, as argued before, frames have a somewhat fuzzy relation with the referential level, while semantic radicals have an opaque relation with the (orthographical) formal level.

We can calculate the directional attraction and repulsion between frame and semantic radical per period⁸². The methodology used for this depends on the current state of a family of methods called collostructional analysis (Gries 2019b), which will form the backbone of Chapter 7. In that chapter, we provide a step by step introduction to the methodology of collostructional analyses. At the risk of proceeding too hastily, methodologically wise, we propose to show the results of the association measures between frame and semantic radical per period and to interpret them. The reader is kindly referred to the next chapter for a detailed explanation of contingency and directionality measures (Section 7.2).

⁸²We will leave out “minguo”, however, because of purely practical reasons – it is easier to include eight figures than nine. Since “minguo” borders on Modern Chinese, it can be left out of the historical overview.

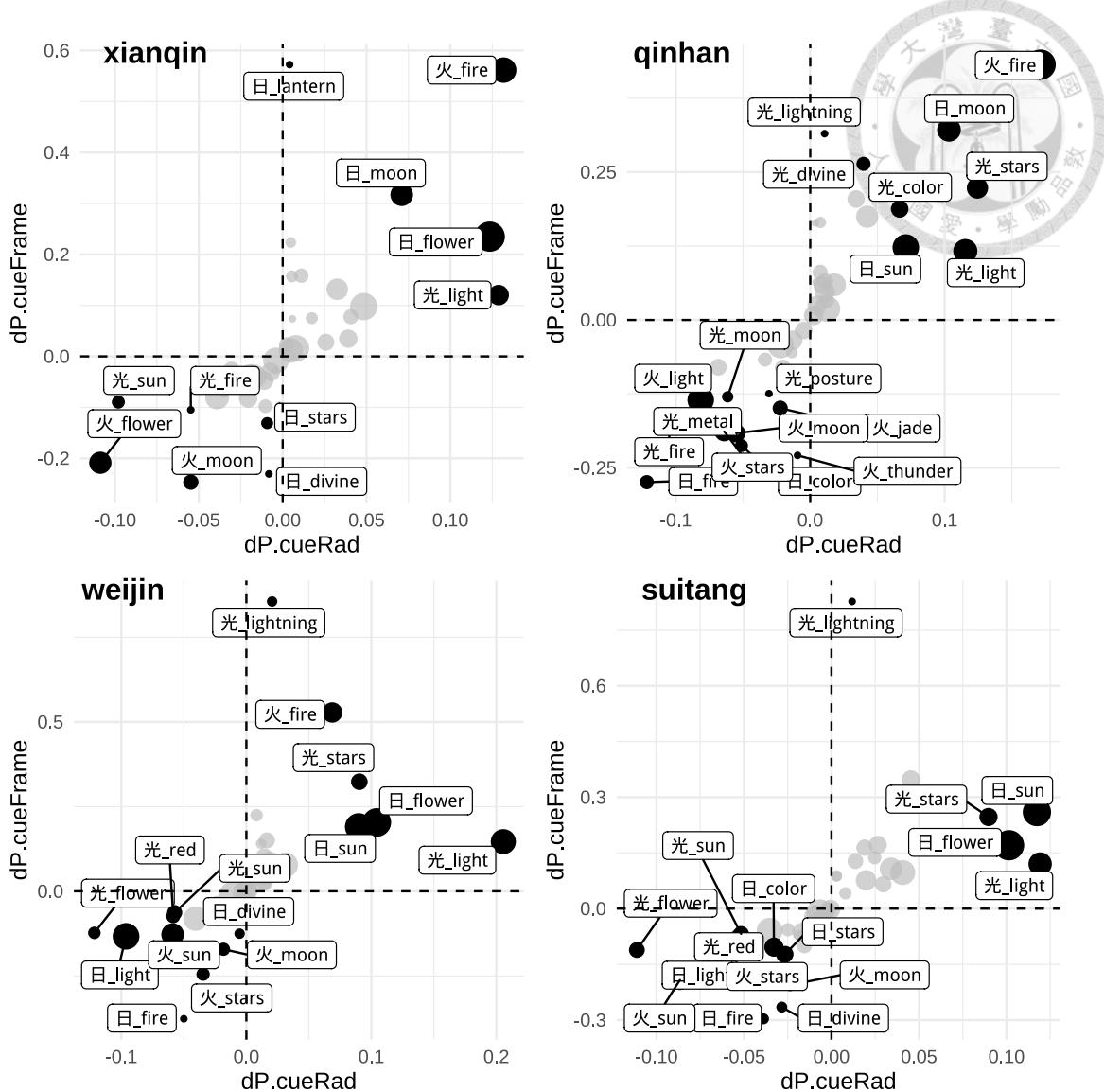


Figure 6.16: Structural salience: association measures between semantic radical and frame – xianqin to suitang

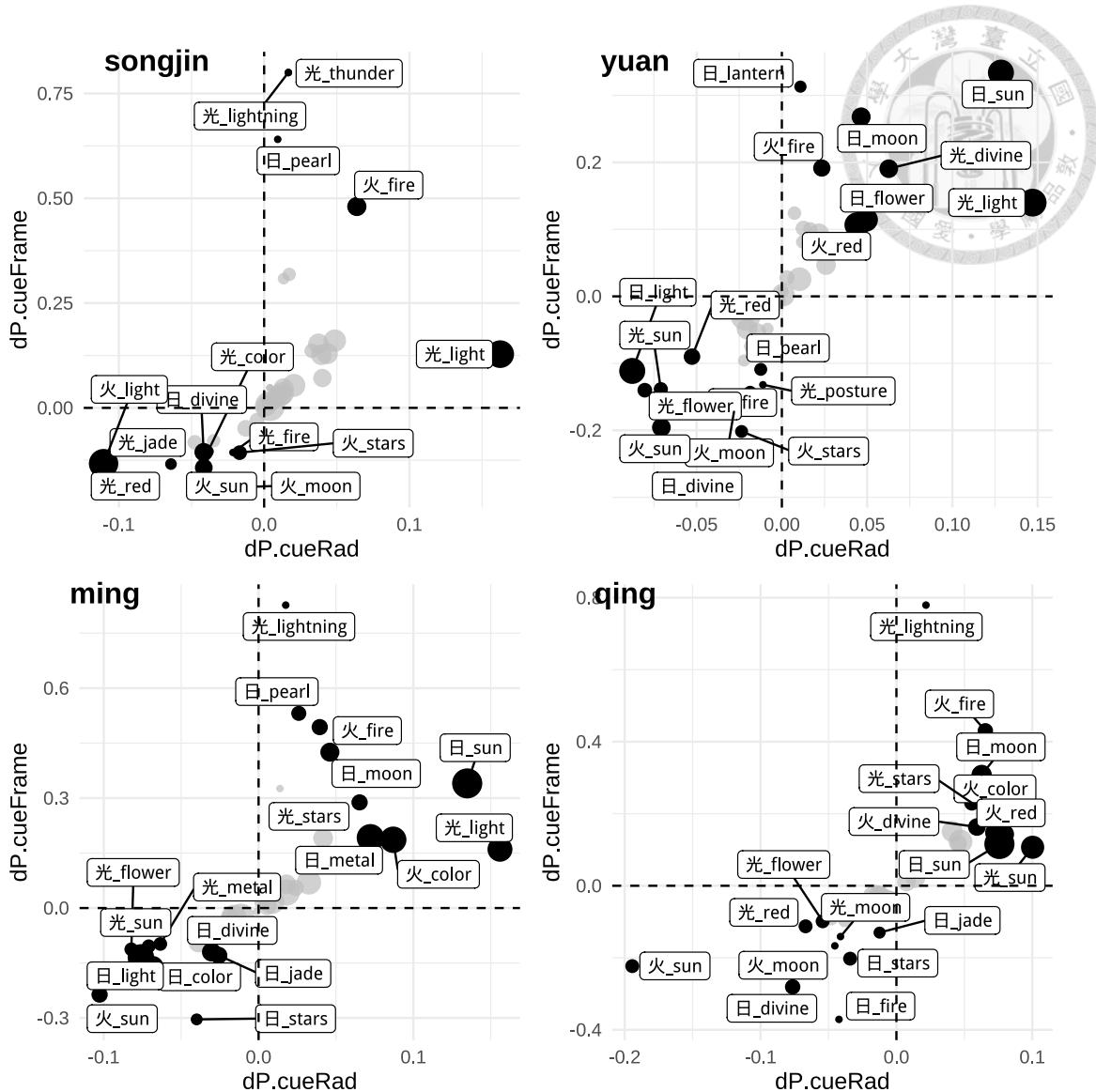
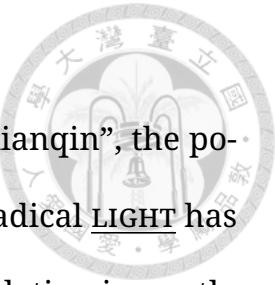


Figure 6.17: Structural salience: association measures between semantic radical and frame – songjin to qing

Let us inspect Figures 6.16 and 6.17. The most important quadrant per plot, i.e. per period, is situated in the top right corner. For both the periods “xianqin” and “qinhan” (Figure 6.16), we can see that “火 _fire” is located in the top right corner. This means that for the periods “xianqin” and “qinhan”, the radical FIRE and the frame FIRE attract each other relatively well. Given an ideophone with the FIRE radical, the meaning most likely will be something from the FIRE frame. Given the FIRE frame, most likely the corre-



sponding LIGHT ideophone will contain the FIRE radical.

However, this is not always the case. For instance, in “xianqin”, the position in the plot of “光 _light” indicates that the semantic radical LIGHT has a relatively strong attraction to the frame LIGHT, but this relation is mostly one-directional: the frame LIGHT also occurs with other radicals.

Focusing now on the position of “火 _sun” in the periods “xianqin”, “weijin”, “suitang”, “songjin”, “yuan”, “ming” and finally “qing”, it can be seen that it is highlighted consistently in the left bottom corner of these plots. This suggests that there is a mutual repulsion going on between the semantic radical FIRE and the frame SUN. This makes sense, because a mutual attraction between semantic radical SUN and frame SUN stands out in the plots of the periods “qinhan”, “weijin”, “suitang”, “yuan”, “ming”, and “qing”. So this means that throughout time, and in the arguably small lexical field of LIGHT ideophones, the repulsion between semantic radical FIRE and the frame SUN, and the attraction between semantic radical SUN and frame SUN have been constants over time.

This is not surprising: intuitively, one would guess that the SUN would indicate SUN rather than any other radical. However, I want to contend that we cannot intuitively predict the attraction and repulsion between semantic radicals and frames of different ontological statuses. For example, “火 _moon” is highlighted in all periods except for “ming”, and occurs in the left bottom each time. This strongly suggests that both the frame MOON and the radical FIRE repulse each other. Note that this does not mean that “MOON can never occur with LIGHT ideophones that have the FIRE radical”, but rather

that it is statistically uncommon⁸³ or unexpected.

As will be detailed in Chapter 7, these directional contingency-based association measures can provide a very nuanced view on the behavior of different linguistic phenomena. In this section, the value of determining structural salience has been made clear. While we have focused on a few examples, it is possible to trace the evolution of a given pattern, radical or frame throughout time and study its interaction. The methodology can of course be expanded to a plethora of other research objects, but for reasons of scope we will not pursue its application to structural salience any further here.



6.7 Conclusion

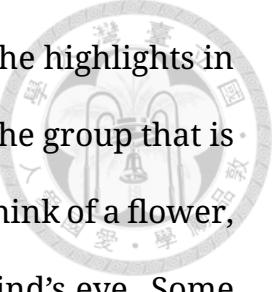
The case studies in this chapter have shown that variational salience phenomena can be identified within the lexical field of LIGHT ideophones, and are dynamic throughout time. In terms of semasiological salience, the results from Chapter 5 have been corroborated. Using distributional relational semantics has allowed for a more nuanced picture in the sense that the conceptual distance between lexical items has been quantified and nearness of neighbors has been apprehended as a sign of prototypicality. However, what has been gained in quantitative nuance, has been lost in the domain of qualitative categorization of the types of semantic extensions that were proposed before. To be more precise, in the current approach, we

⁸³This type of argument is often erroneously raised when talking about size sound symbolism, viz. *bouba* evoking bigger and rounder shapes vs. *kiki* sounding smaller and sharper. The fallacy lies in the argument that “English has *big* with an /i/ but meaning ‘big’ and *small* with /æ/ meaning ‘small’”, which runs contrary to the statistically well-proven size sound iconicity of *bouba* and *kiki*.

no longer see clearly how certain collocates hang together in networks. A possible remedy for this problem may involve tagging through projects like WordNet (Fellbaum 1998) and especially its Chinese implementation (Hsieh & Huang 2009–2019), although this falls outside the scope of the current dissertation.

The semasiological salience case studies should be interpreted as an addition to the lexical semantics toolbox: we can approach a study object like LIGHT ideophones by not only dictionary material, manual usage-based study, but also computational approaches. These each provide different perspectives and through convergent evidence result in a broader understanding of this kind of ideophones, which can relatively easily be transported to other groups of ideophones. I am not the first to come to such a conclusion. Peirsman, Geeraerts & Speelman (2015:75) state that “[The distributional method] gives researchers a more empirical way of establishing semantic equivalence that can moreover be based on the corpora they are using for their studies. This can complement the inherently limited capacity of dictionaries, taxonomies and researchers’ own intuition”.

Reversing the perspective of meaning to that of naming led us to study onomasiological salience within LIGHT ideophones. The same semantic vector spaces were queried for a collocate and return a ranked list of the nearest ideophones. In that section, a number of ideophones that we had discussed before kept popping up over and over for the frames that were investigated (LIGHT, FIRE, MOON, STARS, and FLOWER). However, other LIGHT ideophones that were not part of the detailed study in Chapter 5 also appeared,



most notably *huáng~huáng* 煌煌. And even more extreme, the highlights in every plot showed that LIGHT ideophones may not even be the group that is most entrenched for these frames. For example, when you think of a flower, it may not be its brilliant colors that first pops up in the mind's eye. Some sign languages, e.g., Flemish Sign Language, conceptualize flower first and foremost in terms of smell⁸⁴.

Further explorations of onomasiological salience in relation to ideophones is possible in two main ways. The first is through a serious database project where referential material is compared to ideophonic usage. This will allow researchers to propose quantified ways to measure the degree of entrenchment between the referential, conceptual and linguistic formal layers. The second is related to such an undertaking: if such a database also records data about sociolinguistic variables such as location, age, gender, an even more stratified understanding of onomasiological salience can be developed. Note that this line of research is heavily inspired by a seminal study performed by Geeraerts, Grondelaers & Bakema (1994). Currently, Akita's Multimedia Encyclopedia of Japanese Mimetics (2012a) and Nuckoll's (2019) Quechua Real Words dictionaries are two projects that could provide bases for such investigations, although they are limited in terms of sociolinguistic variables and the inclusion of variation.

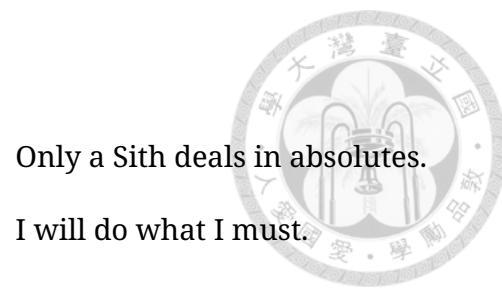
Lastly, structural salience has been interpreted in two differing ways. The first pertains to a distribution of frames, as we had identified them before, both in an achronic manner and a diachronic manner. This allowed us to revisit the diagram containing the relations between frames and domain-

⁸⁴<https://woordenboek.vlaamsegebarentaal.be/gloss/BLOEM?sid=1655>

ICMs (following the proposal for metaphor by Kövecses 2017). The second step was then to use association measures from the family of collostructional analysis methods (see next chapter) to visualize the attraction and repulsion of certain combinations of frames and characters. While the current scope of data is quite limited, some interesting evolutions can be deduced from the visualizations throughout time.

It can thus be concluded in this chapter that there are different salience phenomena at play within the smallish lexical field of LIGHT ideophones – a conclusion that most likely occurs in other lexical fields of ideophones as well. In the next chapter, we will leave behind the diachronic treatment of LIGHT ideophones and focus instead on the constructions in which ideophones in Modern Chinese occur in. However, we will take with us the lesson from these two chapters that not all items are equally representative. This will become even more obvious when the attraction and repulsion between ideophones and different constructions is studied.

7 Constructions



Only a Sith deals in absolutes.

I will do what I must.

Obi-Wan Kenobi

7.1 Introduction

Thus far, it has become clear that different instances of salience are present within the Chinese ideophonic lexicon (Chapter 6). But are salience phenomena also observable for the constructions in which ideophones occur? In other words, as research question 6d asked: how do ideophones interact with the constructions they appear in? This syntagmatic chapter will take as the point of departure Paul (2006) with her discussion of two classes of adjectives in Chinese (see also Zhū 1956). These will be compared to an analysis of ideophones in the Beijing dialect presented by Meng (2012). This will provide us with the different constructions that need to be investigated. The method chosen to investigate the interactions between construction and item is collostructional analysis. After the step-by-step introduction of this method, all constructions in Meng (2012) will be verified with data from ASBC 4.0. In the last section of this chapter, we revisit the infamous ABB-construction with this methodology.

7.1.1 Zhū and Paul's adjectives

As Paul (2006) relates, Zhū (1956) argues that there are two classes of adjectives in (Mandarin) Chinese: BASE FORMS (*jīběn xíngshì* 基本形式) and



COMPLEX FORMS (*fùzá xíngshì* 複雜形式), shown in (90-91)⁸⁵. They are differentiated on the basis of their semantics: base forms express qualities, while complex forms describe the state or mood of that quality. Paul demonstrates that the main criterion underlying this categorization is actually whether an adjective can be used in a modification structure in which the particle *de* 的 does not occur. Complex adjectives always require *de*, base forms optionally take this particle. The first step she takes is doing away with the confusion between bare adjective (base form or complex form) and adjective phrases (91c). This is important because adjective phrases always require the modifier *de* as well: in the case of base form adjectives occurring in a longer adjective phrase, *de* will occur. In other words, while *dà* is a base form adjective, *hěn dà* will take *de* as the linker to the item it modifies.

(90) base form adjectives

- a. monosyllabic: *dà* 大 ‘big’, *hóng* 紅 ‘red’
- b. disyllabic: *gānjìng* 乾淨 ‘clean’, *cōngmíng* 聰明 ‘intelligent, smart’

(91) complex form adjectives

- a. reduplicated adjectives: *xiǎo~xiǎo* 小小 ‘tiny’, *rè-hū~hū* 热呼呼 ‘warm’
- b. modifier-head adjectives: *bīng-liáng* 冰涼 ‘ice-cold’, *xuě-bái* 雪白 ‘snow-white’

⁸⁵We do not necessarily agree with all of Paul (2006)'s glosses. For example, *hěn dà* can mean ‘very big’ or just simply ‘big’, depending on the context. Also *rè-hū~hū* is of the type usually called ABB, which will be discussed below.

- c. adjectival phrases: *hěn dà* 很大 ‘very big’, *fēicháng piàoliáng* 非常漂亮 ‘extremely beautiful’



Furthermore, Paul notes that it is often overlooked in the (formal) literature on adjectives that Zhū (1956) includes monosyllabic *as well as* disyllabic adjectives in the base form category – both distinct from stative verbs⁸⁶. She argues that the behavior of this *de* 的 particle for adjectives plays a crucial role in delineating two different morpho-syntactic classes of adjectives in Chinese, in a way that is reminiscent of Croft’s (2001) Radical Construction Grammar⁸⁷.

This indicates that constructions play an important role in parts-of-speech research. It is also one main reason for this dissertation to refrain from assigning a specific lexical class to ideophones (other than ideophones). Some of the features Paul ascribes to complex adjectives seem familiar to us from our discussion on the scope of the category of ideophones (Chapter 4). For instance, she states that her paper is a proposal “in semantic terms. As noted by Zhū himself (p. 5-6), reduplicated adjectives introduce the speaker’s subjective evaluation of the property expressed by the adjective rather than solely referring to that property (as base form adjectives do)” (Paul 2006:306). Such a statement comes very close to, or is at least compatible with the phrasing that they “depict (sensory) imagery”, especially when that imagery is of the EVALUATIVE kind (as I have argued often occurs with Chinese ideophones). Later, she notes how complex

⁸⁶ Both in Paul (2006) and Paul (2015b) she expresses her surprise that many “studies make the claim that Mandarin Chinese does not have a class of adjectives distinct from intransitive stative verbs (cf. among others McCawley 1992; Larson 1991; Tang 1985 (Paul seems to have made a reference mistake), Lin 2004 (no reference given)).”

⁸⁷ However, Paul (2006) does not refer to general principles of cognition.

adjectives are incompatible with the negation particle *bù* 不 or comparative constructions. This is also reminiscent of Dingemanse (2019)'s exclusion of Mwaghavul and Semai from *the canonical ideophone* (cf. Chapter 4) based on his DEPICTION feature. Such incompatibility with negation has been noted before, of course. For instance, Johnson (1976) says about Bantu ideophones that only external negation is possible. Worded differently, Kita (1997) states about Japanese ideophones that there is an impossibility of logical negation. Tolskaya (2011) even devotes her whole PhD dissertation to driving this point of positive polarity home. In his discussion of Lithuanian ideophones Wälchli (2015) notices the same feature. As for Chinese, Hsieh (2017) remarks that the same can be said for Chinese, while Meng (2012) takes a slightly more nuanced view: “logical negation of a mimetic clause is seldom acceptable” (Meng 2012:3).

Using the same ASBC data we used before, a quick search (for disyllabic ideophones) showed that there are indeed some items recorded in CHIDEOD for which a pattern “NEGATIVE + ITEM” could be found. The most frequent ones⁸⁸ were *yóu~yù* 猶豫 ‘hesitate, waver’ (45 tokens), *hǎo~hǎo* 好好 ‘good, well’ (41 tokens), *yǒng~yuè* 蹦躍 ‘jump of joy’ (9 tokens), and *gū~dú* 孤獨 ‘lonely’ (6 tokens). On the one hand, the case can be made that these words are not very ideophonic. After all, they have high token frequency all over the corpus and may have lost some of their markedness. This de-ideophonization pattern has been observed before (Dingemanse 2017). They may also be part of idiomatic expressions. For *hǎo~hǎo* 好好, a frame defi-

⁸⁸Actually, instead of token frequency, we took their $\Delta P_{construction \rightarrow ideophone}$, see below. At this point it suffices to say that this is the significant cue validity from construction to ideophone.

nitely emerges (92), where *hǎo~hǎo* (92) is integrated in an idiomatic phrase related to correct behavior, in this case mostly studying. In other words, it is used when you want to express statements like “if you don’t behave well and study...”, or “if you don’t behave well and just talk trash...”. It should be noted that it is not claimed that *hǎo~hǎo* 好好 is still ‘a real ideophone’ in Mandarin Chinese, as it was in Premodern Chinese, exactly because it is so frequent and has deideophonized over the ages. However, it is recorded in CHIDEOD and thus is a potential ideophone, which is why it is included in the discussion here.

- (92) a. *bù hǎo~hǎo niàn-shū* 不好好唸書 ‘not well read-books > not study hard’
- b. *bù hǎo~hǎo xuéxí* 不好好學習 ‘not well study > not study hard’
- c. *bù hǎo~hǎo dú-shū* 不好好讀書 ‘not well read-books > not study hard’
- d. *bù hǎo~hǎo gàn-huà* 不好好幹活 ‘not well do-words > not talk trash’

For *yóu~yù* 猶豫 ‘hesitate, waver’, the data seems to suggest that the negation actually helps in establishing the positive polarity, in the sense that ‘hesitation and wavering’ is not an extreme sensory imagery, but ‘not hesitating, not wavering’ is, as is illustrated in (93).



(93) ASBC (nº 201853)

我 毫不 猶豫地 答應 離婚。

wǒ háobù yóu~yù=de dāyìng lí-hūn

1SG NEG hesitate.IDEO=ADV agree separate-marriage

“Without hesitation, I agreed to divorce.”

7.1.2 Meng's ideophonic constructions

Above, we saw that Paul (2006) provides some good arguments to inspect the different constructions adjectives occur in. There are obvious similarities between her complex adjectives and ideophones, such as the *de* 的 particle and the positive polarity (avoidance of negation). Therefore, the argument can be extended to ideophones. We turn to Meng (2012), whose study departs from ideophones, as opposed to adjectives. She makes a basic distinction between three classes of ideophones: onomatopoeia, which she calls O-IDEOS; A-IDEOS, ‘ideophonized’ prosaic words; and the basket category of M-IDEOS. The first is loosely correlated with COLLOQUIAL IDEOPHONES, the latter with LITERARY IDEOPHONES. In her clear discussion, she identifies constructions like the utterance construction, adverbial adjuncts, adverbial complements, predicative and attributive constructions. In the examples below (94-98), we supplement characters and *Hànyǔ pīnyīn* transcription to the examples of the different construction types.



- (94) Utterance / Holophrase construction (adapted from Meng 2012:35)

噠， 噠， 外面 有 敲門 聲。
dā dā wài-miàn yǒu qiāo-mén shēng.
knock.IDEO knock.IDEO outside-LOC EXIST knock-door sound
“Knock knock, there is a knocking sound on the door.”

- (95) Adverbial adjunct construction (ASBC n° 202869)

公車 吶——的 過 站 不 停 ，
gōngchē xiū=de guò zhàn bù tíng
bus whoosh=LNK pass stop NEG stop
“Whoosh, the bus rushed past the bus stop.”

- (96) Adverbial complement constructions (adapted from Meng 2012:37)

女孩子們 早已 哭得 嘤哩嘩啦。
nǚ-hái.zi=men zǎo-yǐ kū=dé xī~lī~huā~lā
female-girl=PL early-already.PFV cry=COMP bawl.IDEO
“The girls bawled and cried.”

- (97) Predicative construction (adapted from Meng 2012:37)

屋 門口 的 火苗 呼呼 = 的 。
wū mén-kǒu de huǒ-miáo hū~hū=de
room door-opening LNK fire-flame blazing.hot.IDEO=DE
“The fire in the doorway is blazing.”



- (98) Attributive construction (adapted from Meng 2012:52)

手 握著 那 條 綠油油的 毛巾。

shǒu wò=zhe nà tiáo lǜ-yóu~yóu=de máojīn

hand hold=DUR DEM CLF green-vibrant.IDEO=LNK towel

“[He] is holding that vibrant green towel in his hands.”

In her overview, Meng (2012) takes a similar stance as Paul (2006), by seeing the particle DE as a pivotal marker for complex adjectives and ideophones. She differentiates the modification marker DE, complementizer DE_C and ideophonic marker DE_I . As she relates, Zhū (1961) claimed that DE following simplex adjectives in attributive position is a nominalizing DE, but following complex adjectives in attributive position and predicative position is an adjectival marker. Meng (2012) suggests that DE as an “ideophonic suffix” needs to be differentiated from the other ones, as it foregrounds the expressiveness of ideophonic usage:

[I]deophonic words, when being used as predicates, tend to be followed by DE_I . Some scholars claim an inherent suffix -tə in lexical items such as ‘state adjectives’, and others consider it as constantly syntactic (the same as the modification marker DE). I propose tə as an ideophonic marker in this case, serving as a prosodic stopping point at the end of ideophonic items. It acts as a result of the abnormality of ideophonic prosodic pattern, which creates an unnatural stop at the end of a sentence.

Meng (2012:71)

It is our view that DE should not be seen as a suffix, but rather as a clitic that indeed can mark ideophonic markedness: Meng (2012) also regards ideophonic DE_I as a quotative particle similar to Japanese *to* と, which is well-known to occur in ideophonic constructions. To summarize, it can be said that there is no reason to just say that ideophones belong to a given wordclass, but rather interact with different constructions.

However, it can be noted that such occurrence in different constructions is not unique to Chinese ideophones, in two senses. In the first, many other classes of words can be used predicatively, attributively, adverbially and even as holophrases (utterances), since these are per definition defined in a functional manner and correspond to prototypical constructions (Croft 1991; 2001), especially in Chinese where wordclass assignment is sometimes called precategorial (Bisang 2008), although there are of course clear prototypical cases of words exclusively belonging to one wordclass, see for example Huang, Hsieh & Chen (2017).

In the second sense, it has also been noticed for other languages that ideophones are not wordclass-exclusive. Dingemanse (2017) lists the following constructions for Siwu, ordered according to relative occurrence in his data: adverbial (46%), complement (26%), holophrase (12%), adjectival (= attributive 6%), predicative (5%), and other (5%). For Japanese, too, the non-exclusivity of ideophones as a category in that language vis-à-vis their occurrence in different constructions has been noted (Sells 2017). Furthermore, for Quechua (Nuckolls 2014) and even in the earliest literature on ideophones in Bantu languages this flexibility was noticed. A line of re-

search (Akita 2009; Dingemanse & Akita 2016; 2017) generalizes this into the idea that ideophones occur in constructions on a cline between belonging to the core syntax and occurring as free elements, away from the core.

Akita formulates this in terms of a correlation between his Lexical Iconicity Hierarchy and Grammatical Functional Hierarchy (Akita 2009; but see Akita 2017b). The Lexical Iconicity Hierarchy, seen in (99-100), places ideophones on a cline, following the traditional Japanese tripartition into phonomimes, phenomimes and psychomimes and ranges from superexpressives to nonmimetics. The Grammatical Functional Hierarchy is based on layered structuring of the clause, as it is presented by Van Valin & LaPolla (1997). Akita explains that the “hierarchy shows how far from the predicate (or core) of a clause a grammatical function is located. Interjections are structurally independent of the predicate. Adjuncts have a modification relationship to the predicate. Arguments are selected by the predicate. Accordingly, relevance to the predicate increases in this order” (Akita 2009:88).

(99) Lexical Iconicity Hierarchy

Superexpressives > Phonomimes > Phenomimes > Psychomimes >
Nonmimetics

(100) Grammatical Functional Hierarchy

Periphery (non-arguments: adjuncts, interjections) > Core (predicate, arguments of the predicate)

Using token frequencies, Akita (2009) shows how the two hierarchies relate to each other, reproduced in Figure (7.1). This visualization inspired

me to do a similar analysis on Tang dynasty poetry (Van Hoey 2015). However, instead of the Lexical Iconicity Hierarchy, I used an adapted version of Dingemanse (2012)'s hierarchy for sensory domains in which ideophones occur. This is shown in Figure (7.2).

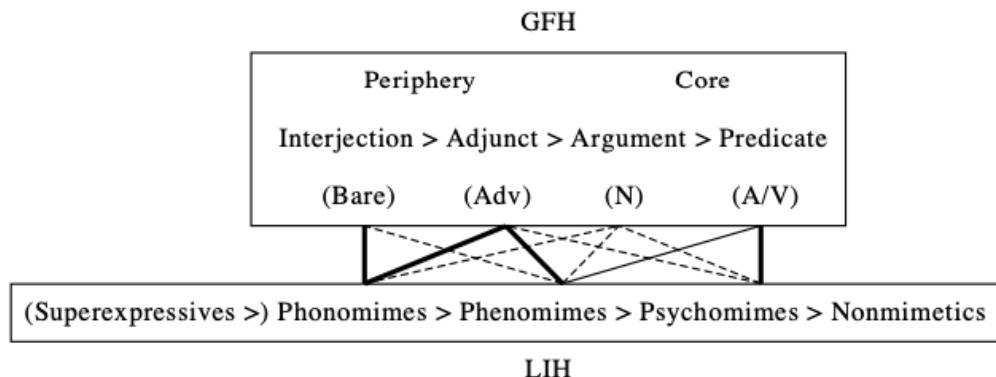


Figure 7.1: The iconic LIH-GFH mapping model for mimetic syntax in Japanese (Akita 2009:247)

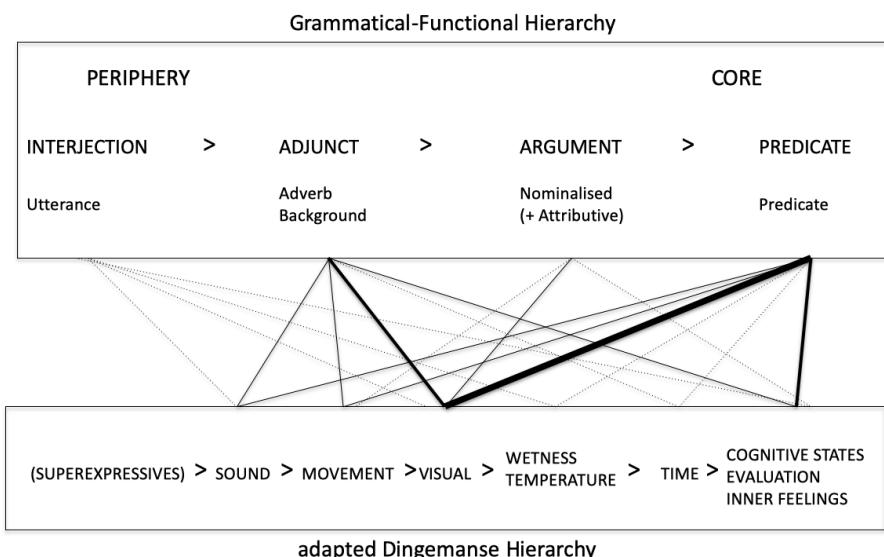


Figure 7.2: The iconic “LIH”-GFH mapping model for mimetic syntax in Middle Chinese (Van Hoey 2015:84)

As both figures show, there is virtually no connection to nominal constructions. Japanese tends to connect to bare (holophrases), adverbial and predicative structures. Middle Chinese (poetry), however, connects mostly

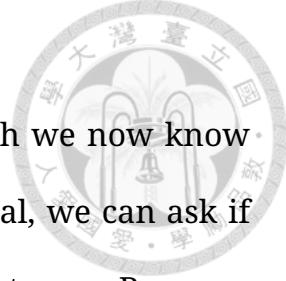
to predicative and adverbial constructions.

Since Meng (2012) identifies these constructions, which we now know are observed across languages as well as language internal, we can ask if there are ways of adding structure to these construction types. Because, surely, not all ideophones occur in all construction types in an equal manner, as indicated by Akita's (2009) hierarchies and the replication for Middle Chinese (Van Hoey 2015). However, perhaps we need to turn to a more advanced framework for calculating just how strong the attraction between construction and ideophone is, if we are to satisfactorily chart the variation and prototypicality of the data. For this, we can rely on collostructional analysis and the association measures derived through it. In the next section we first introduce this methodology (Section 7.2), and in the subsequent section (Section 7.3) these are applied to the constructions identified by Meng (2012).

7.2 Collostructional analysis

The family of collostructional analysis methods was first shown by Stefanowitsch & Gries (2003), and is based on the idea of the distributional hypothesis, which we have encountered before (Chapter 6). “You shall know a word by the company it keeps” (Firth 1957), or more explicitly by Harris (1970:785):

[i]f we consider words or morphemes A and B to be more different in meaning than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C. In other words, difference of meaning correlates with



difference of distribution.

Harris (1970:785), quoted in Gries (2019a:386)



Stefanowitsch & Gries (2003) originally developed the method as an extension of construction grammar theories, showing that it is a good first step to obtain token frequencies of a given construction or the frequency a given element occurs in a slot of a construction. As they show, for the N WAITING TO HAPPEN construction, an intuitive corpus exploration would list the concordances of the phrase *waiting to happen* and then sort according to items on the left. The naked eye would see how there are many different types, e.g., *accident waiting to happen*, *disaster waiting to happen*, *crisis waiting to happen* etc. A logical step would be to tally all nouns (the “N”) we are interested in, resulting in 14 tokens of *accident waiting to happen*. Ranking the types according to their token frequency already provides us with some information about the relation between the construction and the elements that occur in it. As Bybee (2010) has demonstrated at length, this is a good example of token frequency and token frequency effects.

Intuitively, one could stop here. But, as Stefanowitsch & Gries (2003) argue, there is the problem that it could just be the case that *accident* has a high token frequency in these constructions, because *it occurs frequently overall*. An analogy would be to investigate verbs in English and finding lots of forms of the verb *to be*. These do not really contribute that much to a construction or the way it stands out, because *to be* is everywhere. According to Gries and Stefanowitsch, to remedy this issue one needs information about how many times *accident* occurs in different constructions, as well as

Table 7.1: Crosstabulation of *accident* and the [N waiting to happen] construction,
adapted from Stefanowitsch & Gries (2003:219)



constructions	items		row totals
	<i>accident</i>	$\neg\text{accident}$	
[N waiting to happen]	14	21	35
\neg [N waiting to happen]	8,606	10,197,659	10,206,265
column totals	8,620	10,197,680	10,206,300

Table 7.2: Crosstabulation of items and constructions

constructions	items		row totals
	<i>item</i>	$\neg\text{item}$	
construction	<i>a</i>	<i>b</i>	$a+b$
\neg construction	<i>c</i>	<i>d</i>	$c+d$
column totals	$a+c$	$b+d$	$a+b+c+d$

information about other items occurring in that construction, such as *crisis* and *disaster*. These numerical data for *accident* can be cross-tabulated, as shown in Table 7.1, and are generalized in Table 7.2.

The next step after cross-tabulating is calculating the measure of association between the item and the construction. A host of methods exist, and Gries, as the most vocal representative of collostructional analysis has often argued in favor of the Fisher-Yates Exact test, because it is based on contingency, and takes into account token frequency as well as significance. The penultimate step in the collostructional analysis process, is to perform the test for all different items and ranking them based on the *p*-value. This indicates to what degree a nominal is attracted to the N WAITING TO HAPPEN construction. Not entirely unsurprisingly, *accident* comes out on top (Stefanowitsch & Gries 2003), with a reported *p*-value of 2.12E-34, followed by *disaster* (1.36E-33), and then a number of nominals which occur only once.

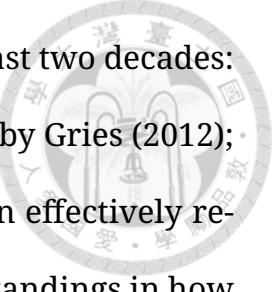
The final step is to explain why these nouns are so prolific in this construction. In this introductory example, they argue that this way, one can find a number of associations that one would perhaps not typically expect.

Let's say one was making a dictionary. In this case, this method suggests that you should mention the N WAITING TO HAPPEN construction at both the *accident* and *disaster* entries. It thus could provide a helpful tool in lexical semantic studies.

However, not only lexical semantics can benefit from this methodology. Following the recent overview (Gries 2019b), we can distinguish three different applications of collostructional analyses. COLLEXEME ANALYSIS is the term given for the case study we have just covered, which quantifies the degree of attraction and repulsion of words occurring in a constructional slot. Other examples from Stefanowitsch & Gries (2003) include investigating the English imperative or verbs that participate in the ditransitive.

DISTINCTIVE COLLEXEME ANALYSIS calculates the preference of words in slots of two functionally similar constructions, e.g., the ditransitive vs. the prepositional dative, or the *will* vs. *going to* future (Gries & Stefanowitsch 2004a).

CO-VARYING COLLEXEME ANALYSIS measures the attraction and repulsion between two items occurring in two slots of a given construction, illustrated by the V1 SOMEONE INTO V2-ING construction, e.g., *trick someone into buying* or *force someone into accepting*; or the V YOUR WAY PREP NP construction (Goldberg 1995), e.g., *make your way to the top* or *weave your way through the crowd* (Gries & Stefanowitsch 2004b).



The method has received a number of criticisms in the last two decades: first from Bybee (2010), but this was rebutted convincingly by Gries (2012); later also from Schmid & Küchenhoff (2013), but once again effectively refuted by Gries (2015). The criticism ranges from misunderstandings in how to construct the contingency table (especially “cell *d*”) to issues with the tests used. Here we first briefly discuss Schmid & Küchenhoff (2013) and the reply by Gries (2015).

Schmid & Küchenhoff (2013) argue that it is problematic to find the values of cell *d* (cf. Table 7.2), because these are based on double negatives (not construction and not item). Gries refutes this: it is an important methodological choice to select similar items to find the numbers. In other words, one needs to “choose a level of resolution on which to count constructions that is close to the phenomenon in question: If one does a C[ollostructional] A[nalysis] on an argument structure construction, then obviously using the number of letters of a corpus is useless, as is using the number of files – using the number of verbs or lexical verbs is probably more useful” (Gries 2015:511).

The second criticism relevant to us is the argument that the Fisher Yates Exact test is both over-complicated (because of cell *d*) and not maximally informative, because it is bidirectional (Schmid & Küchenhoff 2013): the obtained *p*-value is a measure for stating how much the construction and the item attract or repulse one another. They propose using two other measures, ATTRACTION and RELIANCE, based on Schmid (2000), to address the directionality issue. Attraction shows how much construction attracts the

lexeme, Reliance captures how much the lexeme relies on the construction⁸⁹. Discerning directionality is important because as Gries (2019b:392–393) later admittedly illustrates, in 2-grams (“groups of two words”) word 1 can attract word 2 (e.g., *according to, instead of*); word 2 can attract word 1 (e.g., *in vitro, de facto*); or word 1 and word 2 attract each other (e.g., *Sinn Fein, bona fide*).

The issue Gries (2015) takes is that Schmid & Küchenhoff (2013)’s Attraction and Reliance do not take the effect proportion into account, or in other words, they are not contingency-based (cf. Levshina 2015:234). An alternative that both parties seem to agree upon, however, is cue validity, expressed with ΔP (Ellis 2006; Ellis & Ferreira-Junior 2009). This measure can be calculated in two ways and is contingency based. $\Delta P_{construction \rightarrow item}$ is shown in Eq. (101); $\Delta P_{item \rightarrow construction}$ can be seen in Eq. (102).

(101)

$$\Delta P_{construction \rightarrow item} = \frac{a}{a+c} - \frac{b}{b+d}$$

(102)

$$\Delta P_{item \rightarrow construction} = \frac{a}{a+b} - \frac{c}{c+d}$$

All participants in the debate on which association measure to use of course have a number of valid arguments, but it is comforting to know that

⁸⁹Using the contingency table, they are calculated as follows:

$$Attraction = 100 \frac{a}{a+c} \quad Reliance = 100 \frac{a}{a+b}$$

strong correlations exist between Reliance and $\Delta P_{item \rightarrow construction}$ on the one hand, and Attraction, $\Delta P_{construction \rightarrow item}$, the (log-transformed) Fisher Exact test's *p*-value, as well as log-likelihood⁹⁰, as demonstrated by Levshina (2015:236). For our purpose, $\Delta P_{item \rightarrow construction}$ and $\Delta P_{construction \rightarrow item}$ are the most appropriate, because they are directional, as well as contingency based, i.e., take the proportion into consideration. Gries (2019b) first demonstrates how these measures can be plotted on the x-axis and y-axis to study the directionality. Later, he urges us to add a third dimension, for example token frequency or dispersion (cf. Section 4.5.1), effectively stating that we need *tuples*, bundles, of association measures and data to get a better handle on the phenomenon under investigation. In short, he argues against the conflation of frequency and effect size in the choice of association measure; for contingency-based measures; details on the directionality of the attraction and repulsion; and dispersion. Gries's final representations are 3D, but because these visualizations are hard to interpret in print, I will present only 2D plots below.

7.3 Collostructional analyses of ideophone constructions

Now that the methodology has been introduced, it is time to apply it to the constructions in which ideophones occur. We will take Meng's (2012) constructions as the point of departure: first predicative constructions are dis-

⁹⁰This is also a bidirectional, contingency-based test, cf. Dunning (1993).

cussed, then adverbial, and lastly attributive constructions. These are displayed in Table 7.3.

Table 7.3: Constructions identified by Meng (2012) and their abbreviations

Constructions	Abbreviation	Section
Predicate followed by <i>de</i>	IDEOPHONE _{PRED} DE	7.3.1.1
Predicate not followed by <i>de</i>	BARE IDEOPHONE _{PRED}	7.3.1.2
Adverbial followed by <i>de</i>	IDEOPHONE DE _{ADV}	7.3.2.1
Adverbial not followed by <i>de</i>	BARE IDEOPHONE _{ADV}	7.3.2.2
Adverbial complement	DE _{COMP} IDEOPHONE	7.3.2.3
Attributive (followed by <i>de</i>)	IDEOPHONE DE _{ATT}	7.3.3

The main source of data will be the ideophones identified in the ASBC, as they were used in Chapter 4 as well, rather than the whole ASBC. As a disclaimer, the data in this chapter is noisy in the sense that it contains items from CHIDEOD as they were found in ASBC, and may possibly include a number of items that can no longer be considered ‘ideophones’ in Mandarin Chinese, like 好好, see the discussion surrounding example (92). One solution to this problem can be a future study probing the iconicity and ideo-phonicity ratings of items found in the corpus, as mentioned in Section 3.2.7. Unfortunately, this is currently beyond the scope here.

Another solution is making the choice to include only certain subgroups of items in CHIDEOD, e.g., onomatopoeia proper, or differentiate between items from different periodical sources. This methodological choice is further complicated by the difficulties in the morphemes under investigation:

the three DE particles, namely *de* 的, *de* 地, and *de* 得. Since we are interested which items occur in these constructions from all different periods, and have been possibly inherited from previous periods, it makes sense to tolerate some noise within the largely written data. To further illustrate, however, that the skewedness is tolerable, we present the following Tables 7.4 and 7.5, which respectively show the type frequency and token frequency of items as they occur in older data sources vs. newer data sources, see Section 3.2.1 for a discussion. In this case, the data sources from Wáng (1987); Gōng (1991); and Lǐ (2007) count as Modern Chinese sources, while {*Shījīng* 詩經} (Van Hoey 2016a); {*Táng shī sān bǎi shǒu* 唐詩三百首} (Van Hoey 2015) and Kroll (2015) were consulted for Premodern Chinese. We made the choice to not use the *Hànyǔ dà cídiǎn* 漢語大詞典 as a discriminating factor because it spans both categories. Consequently, the items in the fourth cell (— Premodern data and — Modern data) actually reflect items that are present in the *Hànyǔ dà cídiǎn*, abbreviated as HYDCD in the tables.

Table 7.4: Types within the dataset as compared to Premodern and Modern data within CHIDEOD

	+ Premodern data	- Premodern data	total
+ Modern data	564	71	635
- Modern data	188	314 (= HYDCD)	502
total	752	385	1137

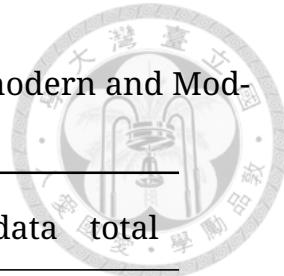


Table 7.5: Tokens within the dataset as compared to Premodern and Modern data within CHIDEOD

	+ Premodern data	- Premodern data	total
+ Modern data	17752	495	18247
- Modern data	12402	9614 (= HYDCD)	22016
total	30154	10109	40263

As can be seen from both these tables, about half of the data can be found in Modern data sources as can be found in CHIDEOD. For types, we observe a frequency of 635 (55.8%); for tokens the frequency is 18,247 (45.3%). If there is a bias in the data, it seems that it will be a balanced one. Nevertheless, we recommend that the iconicity and familiarity of the items in CHIDEOD is tested through experiments or surveys in the near future, so that the findings presented here can be put in perspective.

7.3.1 Predicative constructions

The first construction we investigate is the predicative construction for ideophones, as illustrated above in example (97). Meng (2012:37) states that for SOUND ideophones, ideophones here tend to be followed by *de* 的. More precisely, according to her, they are followed by the ideophone marking DE_I.

7.3.1.1 IDEOPHONE_{PRED} DE construction A rough first attempt at finding relevant examples of this construction is looking for IDEOPHONE DE followed by punctuation. With the regular expression "`({ideos})_\\w+\\b\\w_DE\\b`", we can find all different ideophone types ('ideos') extracted



from CHIDEOD, followed by their tag ('_\\w+'), followed by a space, followed by any of the three DE particles ('\b\\w_DE\b'). The resulting list has 276 possible tokens, of which upon inspection 149 seemed to capture the construction, as illustrated in (103). The scope of the data will be limited to disyllabic items.

- (103) a. (ASBC n° 106206)

結果 我 就 覺得 怎麼 褲子 覺得 乾乾 = 的 ?
jiéguǒ wǒ jiù juédé zěnme kùzi juédé gān~gān=de
so 1SG then think how pants think dry.IDEO=DE
“So, why do I feel these pants are still so dry?”

- b. (ASBC n° 100698)

現在 好好的，
xiànzài hǎo~hǎo=de
now good.IDEO=DE
“Now things are pretty good.”

- c. (ASBC n° 101234)

他 還是 懶洋洋的，
tā háishì lǎn-yáng~yáng=de
3SG still lazy-endlessly.IDEO=DE
“He’s still so lazy.”

These predicative constructions often come across as depicting the impressions and categorizations the speaker makes about a whole event (the current condition in 103b) or topic ('the pants' in 103a, 'he' in 103c). It is also

a construction that is learned very early. Van Hoey & Chiarung Lu (2019b) investigated such constructions in the Chinese CHILDES corpora (MacWhinney 2000), and found many occurrences, especially in relation to tasks such as the description of every day items like an apple. However, as stated before (Chapter 4), the ideophonyhood of these purely reduplicated tems is questionable: reduplication does not entail ideophonicity.

Interestingly, in the ASBC data, there is some conflation between attributive *de* 的 and adverbial *de* 地. From my own personal experience and interactions with native speakers of Taiwan Mandarin, the attributive *de* is often used to mark adverbial usage. The topic deserves further study, but unfortunately falls outside the scope of this dissertation. There are 83 tokens of attributive *de* 的 and 52 tokens of adverbial *de* 地⁹¹.

Now the collostructional analysis will be applied to these data. For all types in the data it is possible to construct a contingency table based on the token frequencies. As mentioned before, we are limiting these studies to disyllabic ideophones, leaving more than 22,000 observations (rows) available in the dataframe. One of the most significant ideophones in this IDEOPHONE_{PRED} DE construction is *jiàn~jiàn* 漸漸 ‘gradual’, as shown below in Table 7.6. It occurs 36 times in this construction (cell *a*). It furthermore occurs 508 times in different environments (cell *b*), bringing the total to 544 tokens of *jiàn~jiàn* in the data set. The construction under investigation occurs 139 times, of which 103 times (cell *c*) with other items. Knowing that the total of the dataframe is 22209 (*a* + *b* + *c* + *d*) gives us the logical value of

⁹¹One possible explanation may be that the adverbial usage in fact really is adverbial, but inspected samples indicate that this is not the case.



21562 for cell d .

Table 7.6: The contingency table for $jiàn\simjiàn$ 漸漸 in the IDEOPHONE_{PRED} DE construction

	construction	\neg construction	row totals
$jiàn\simjiàn$	$a = 34$	$b = 508$	542
$\neg jiàn\simjiàn$	$c = 103$	$d = 21562$	21665
column totals	137	22070	22207

In the next step, the association measures are calculated. As mentioned before, we take into account contingency and directionality, and as a result calculate association measures in terms of cues. If the construction is the cue, $\Delta P_{construction \rightarrow ideophone}$ can be calculated with the following formula: $\frac{a}{a+c} - \frac{b}{b+d}$. If, however, the ideophone is the cue, $\Delta P_{ideophone \rightarrow construction}$ follows the formula: $\frac{a}{a+b} - \frac{c}{c+d}$. For $jiàn\simjiàn$ this would mean that $\Delta P_{construction \rightarrow ideophone} = 0.24$ and $\Delta P_{ideophone \rightarrow construction} = 0.06$.

Remember that the most important function of the association measures is not the individual values, but rather that these measures can help rank the attraction or repulsion between ideophones and constructions, so as to understand which ones ‘jump out’.

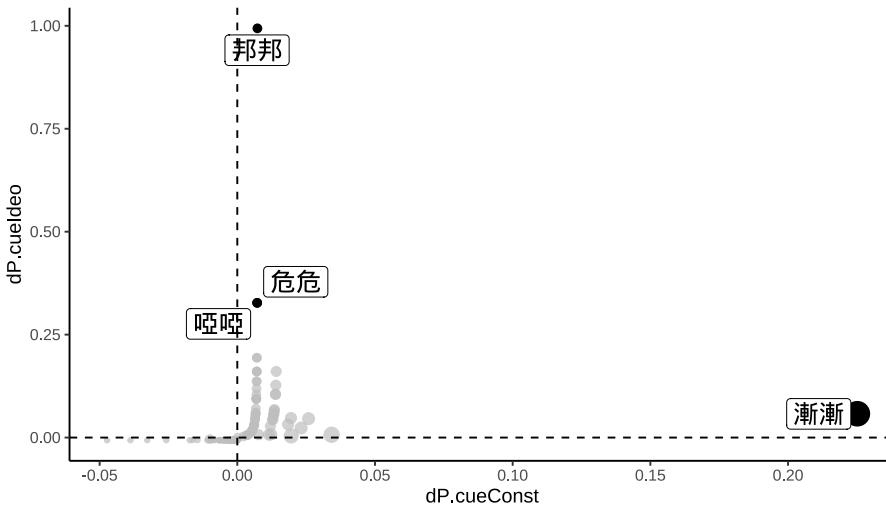
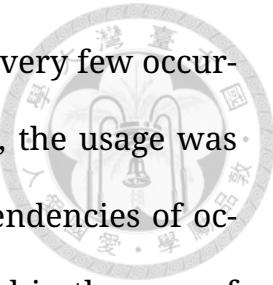


Figure 7.3: Association measures for ideophones in the $\text{IDEOPHONE}_{\text{PRED}}$ DE construction

As Figure 7.3 shows, *jiàn~jiàn* 漸漸 indeed has a *relatively high* $\Delta P_{\text{construction} \rightarrow \text{ideophone}}$, but quite a low $\Delta P_{\text{ideophone} \rightarrow \text{construction}}$, being situated in the lower right quadrant of the plot. This means that in the $\text{IDEOPHONE}_{\text{PRED}}$ DE construction, this item will be most likely to appear. It is also relatively frequent, with 36 tokens.

On the other end, we see that *bāng~bāng* 邦邦 is the highest value. There was only one token ($n = 1$) of this ideophone, used in the so-called ABB construction (see Section 7.4), embedded in the $\text{IDEOPHONE}_{\text{PRED}}$ DE construction. More specifically, the token was *yìng-bāng~bāng=de* 硬邦邦的 ‘hard, stiff’. Since *bāng~bāng* occurs only once with this construction, and nowhere else in the data subset of this construction, there is a very high $\Delta P_{\text{ideophone} \rightarrow \text{construction}}$. In other words, the ideophone *relies* on the construction to appear.

With this example it has been made clear how one can use association measures to explore the attraction and repulsion between ideophone and



constructions. It is somewhat atypical, because there were very few occurrences of this construction in the data set. But even then, the usage was not spread equally over the items. On the contrary, the tendencies of occurrences can be computed. Of course, this does not stand in the way of creative usages or even pretend to be *the one and only* representation of a given language user's usage of ideophones in this construction – but it does paint a picture. And it is not wholly unsurprising that *jiàn~jiàn* 漸漸 had the highest $\Delta P_{construction \rightarrow ideophone}$; I have heard it being used on multiple occasions when referring to the descriptions of something visual, be it a haircut where the hair does not abruptly go from short to longer, but rather gradually, or in a diagram, where an effect shows that the value gradually increases over the x-axis.

7.3.1.2 BARE IDEOPHONE_{PRED} construction Not all ideophones occur with *de* in predicative constructions, as (104) shows.

(104) a. (ASBC n° 103600)

他 有點 茫然。

tā yǒudiǎn máng.rán

3SG a.bit absent.minded.IDEO

“He's a bit absent minded.”

b. (ASBC nº 105538)

在	門口	徘徊，
zài	mén-kǒu	pái~huái,
at.LOC	door-opening	wavering
“Wavering in the doorway.”		



Following the same protocol as above, we can use a regular expression like "`({\ideos})_\\w+ ({punctuationregex})`" where '`punctuationregex`' contains all different punctuation categories in the data. This allows us to find all ideophones followed directly by punctuation, with the assumption that most of the resulting nearly 6000 rows are representative of the $\text{IDEOPHONE}_{\text{PRED}}$ construction. As an example, the contingency table for *máng-rán* 茫然 is shown in Table 7.7.

Table 7.7: The contingency table for *máng-rán* 茫然 in the $\text{IDEOPHONE}_{\text{PRED}}$ construction

	construction	\neg construction	row totals
<i>máng-rán</i>	$a = 43$	$b = 54$	97
\neg <i>máng-rán</i>	$c = 4238$	$d = 17872$	22110
column totals	4281	17926	22207

The calculation of $\Delta P_{\text{ideophone} \rightarrow \text{construction}}$ and $\Delta P_{\text{construction} \rightarrow \text{ideophone}}$ is shown in Figure 7.4. It can be seen that the ideophones are more divided than in the previous case study. The ideophone with the highest $\Delta P_{\text{construction} \rightarrow \text{ideophone}}$ is *móhú* 模糊 ‘vague’, illustrated in (105).

However, this ideophone is somewhat atypical, in the sense that it seems depictive of vagueness, but is not formally marked. In the terms laid out before, it would belong to morphological template of RR.



(105) a. (ASBC n° 202027)

都 是 那樣地	面目	模糊，
dōu shì nà-yàng=de	miàn mù	móhú,
all COP that-way=ADV	appearance	vague.IDEO
“It’s all so vague.”		

b. (ASBC n° 203641)

顯得 零碎 而 模糊。	
xiǎndé língsuì ér móhú.	
seem scrappy CONJ vague.IDEO	
“It looks scrappy and formless.”	

While the $\Delta P_{construction \rightarrow ideophone}$ for *móhú* 模糊 is relatively highest, it also belongs to a group of items for which the $\Delta P_{ideophone \rightarrow construction}$ is high as well, such as *fēn~yún* 紛 紜 ‘diverse and confused’ and *chōng~chōng* 怮 怮 ‘laden with anxiety’.

However, a number of items in this quadrant, e.g., *āiyá* 哎 呀 and *yún~yún* 云 云, are less clear in their depiction. The former is mostly an interjection and is hard to interpret as being more than a reaction to a situation. The latter originally was added to CHIDEOD for its ideophonic depiction, but *yún~yún* 云 云 has also grammaticalized to mean ‘and so on’. This is because *yún* 云 has a long-lasting functional usage as quotative

particle (Pulleyblank 1995:82), and its reduplication is very much in line of that usage.

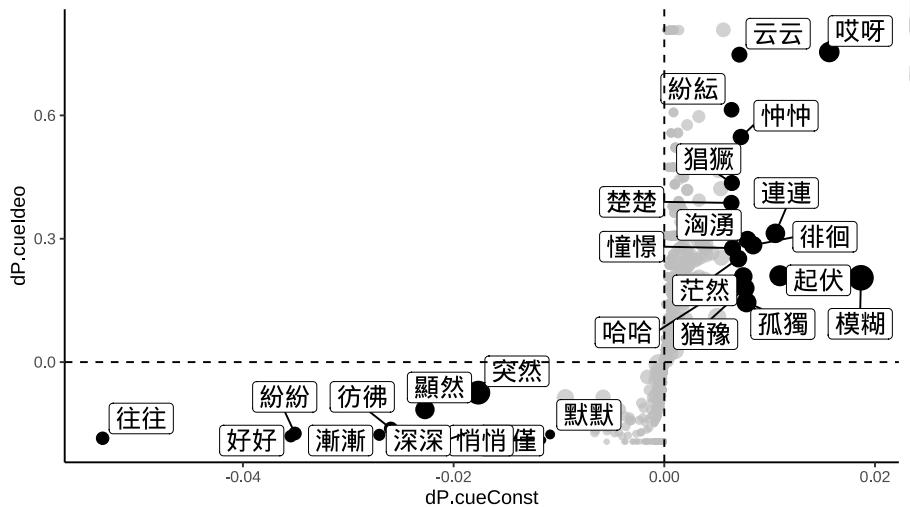


Figure 7.4: $\Delta P_{construction \rightarrow ideophone}$ in the BARE IDEOPHONE_{PRED} construction

Interestingly, the plot in Figure 7.4 also shows which ideophones are not attracted, i.e., repulsed by predicative usage as operationalized in this section, e.g., *wáng~wáng* 往往 ‘often’, *hǎo~hǎo* 好好 ‘good, well’ etc. The ones that are highlighted in the plot are all “ideophones”⁹² that occur more as adverbs, and should thus appear in the upper right quadrant when we look at adverbial usage (Section 7.3.2). Stefanowitsch (2006) has even proposed that items occurring in this lower left quadrant are actually corpus-based evidence of ungrammatical sentences, i.e., he proposes using association measures as a way to quantify grammaticality judgments. This can have important methodological consequences for future research on the constructions ideophones in which occur.

⁹²I want to point out that their ideophonehood is really not sure. They are frequent enough for their occurrence not to come across as marked anymore, even though they still depict some sensory imagery.

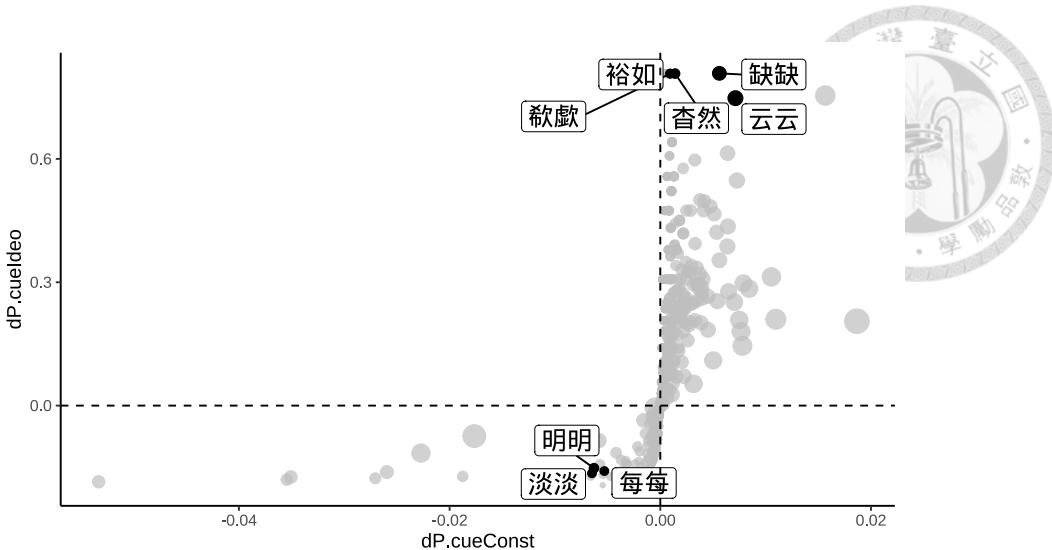


Figure 7.5: $\Delta P_{construction \rightarrow videophone}$ in the BARE IDEOPHONE_{PRED} construction

If we then look at the $\Delta P_{ideophone \rightarrow construction}$, with extremes highlighted in Figure 7.5, we can observe items like the compositional ideophone *yù-rú* 裕如 ‘with ease’, which belongs to an idiom *yìngfù yù-rú* deal.with with.easeVIDEO ‘be able to deal with’, *yǎo-rán* 杳然 ‘quiet, without a trace’, or *quē~quē* 缺缺 ‘lacking’. If one is given these items, there is a high chance that they will occur in a predicative construction. This is the opposite case of the ideophones in the bottom of Figure 7.5, e.g., *dàn~dàn* 淡淡 ‘indifferent, bland’. We thus have two ways of analyzing every plot. These show the tendencies of how different constructions interact with ideophones. In the next section we take a closer look at adverbial constructions.

7.3.2 Adverbial constructions

The second set of constructions we look at are adverbial usage of ideophones. Using the examples in Meng (2012) above (95-96) as an illustration, there seem to be two main constructions we will need to investigate. The

first includes that of adverbial adjuncts, the other adverbial complements.

The former is normally marked by the form *de* 地 and the latter by *de* 得.

In our experience, the latter is respected in most usage contexts, but the former often is interchanged with the “attributive DE”, i.e., *de* 的. However, as she mentions regularly, these are tendencies.

7.3.2.1 IDEOPHONE DE_{ADV} construction The adverbial adjunct is operationalized by the regular expression "`({ideos})_\\w+ (地 | 的)_DE \\w+_V`". This means the ideophones followed by a tag ("`{ideos}_\\w+`"), a space, and the two relevant versions of the DE marker. However, we need to take out at least those ideophones that were identified in Section 7.3.1.1, since this regular expression creates an overlap. For this case study, we will illustrate the method with *shēn~shēn* 深深 ‘deep’, which seems an intuitive good representative of this construction and ideophone, see (106).

We will follow the same steps as above.

(106) a. (ASBC n° 2052)

我 深深地 感謝 天主，
wǒ shēn~shēn=de gǎnxiè tiān-zhǔ
1SG deeply.IDEO=DE thank sky-master
“I deeply thank the Lord.”

b. (ASBC n° 2120)

但 我 深深的 以為，
 dàn wǒ shēn~shēn=de yǐwéi
 but 1SG deeply.IDEO=DE wrongly.think
 “But I was really convinced, ...”



Table 7.8: The contingency table for *shēn~shēn* 深深 in the IDEOPHONE DE_{ADV} construction

	construction	\neg construction	row totals
<i>shēn~shēn</i>	$a = 92$	$b = 291$	383
\neg <i>shēn~shēn</i>	$c = 1775$	$d = 20049$	21824
column totals	1867	20340	22207

The contingency table is illustrated in Table 7.8. After programmatically listing this for all items that occur in the construction, all $\Delta P_{ideophone \rightarrow construction}$ and $\Delta P_{construction \rightarrow ideophone}$ scores are calculated.

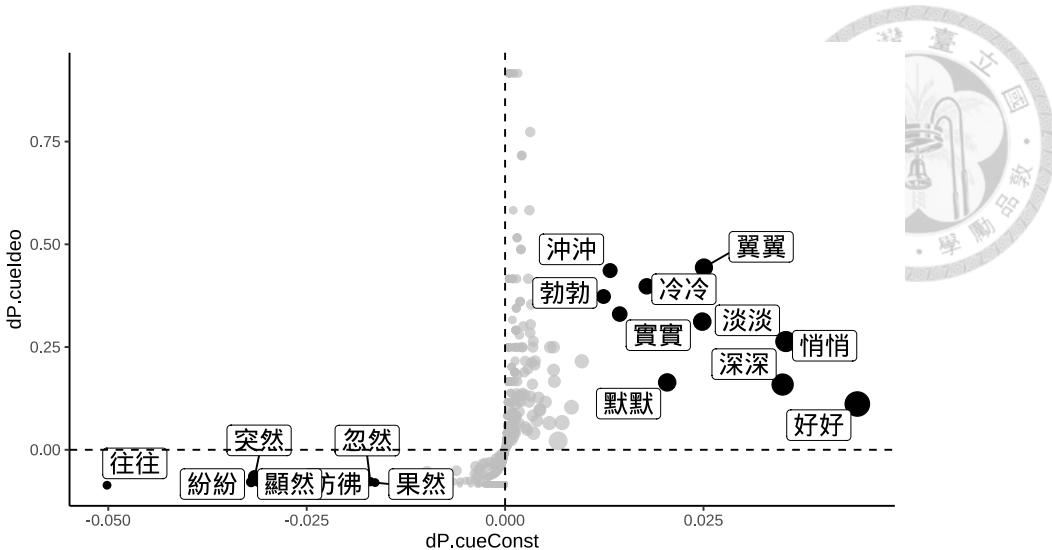


Figure 7.6: $\Delta P_{construction \rightarrow ideophone}$ in the IDEOPHONE DE_{ADV} construction

Let us first inspect the $\Delta P_{construction \rightarrow ideophone}$, shown in Figure 7.6. It can be seen that *shēn~shēn* 深深 has the third highest value for $\Delta P_{construction \rightarrow ideophone}$, meaning that if given this construction, *shēn~shēn* 深深 has the third highest chance of appearing, statistically speaking. But other ideophone items can also occur in the adverbial adjunct construction, e.g., *qiāo~qiāo* 悄悄, or *yì~yì*, which only occurs in the data embedded in the idiom *xiǎo.xīn-yìyì* 小心-翼翼 ‘with utmost caution’, see (107).

- (107) a. (ASBC n° 101483)

悄悄地 走進 山洞。
 qiāo~qiāo=de zǒu-jìn shān-dòng
 sneaky.IDEO=DE go-enter mountain-cave
 “Sneakily enter the cave.”



b. (ASBC n° 107090)

小心-翼翼 = 的 降落下來；
 xiǎo.xīn-yì~yì=de jiàng.luò-xià-lái
 watch.out-careful.IDEO=DE descend-down-come
 “Come down very carefully.”

In the other direction, we can look at $\Delta P_{ideophone \rightarrow construction}$. As Figure 7.7 shows, given the items on top, they are highly likely to co-occur with de_{adv} .

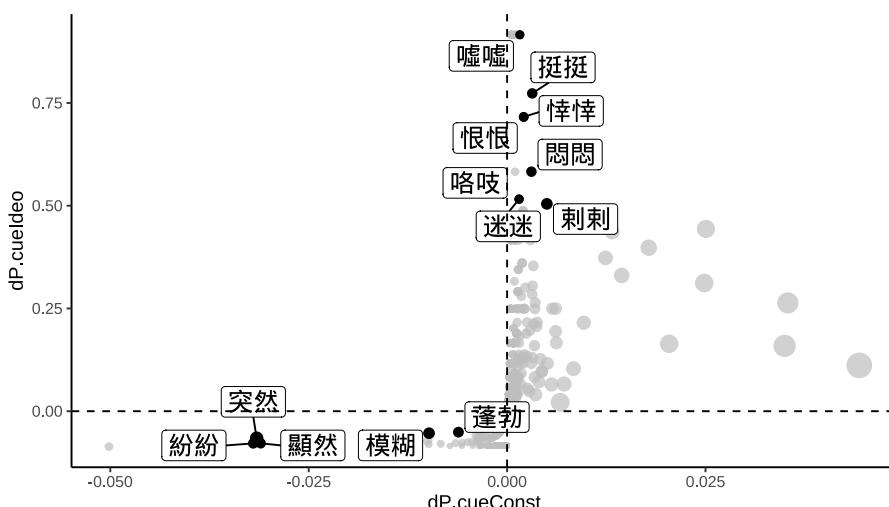


Figure 7.7: $\Delta P_{ideophone \rightarrow construction}$ in the IDEOPHONE DE_{ADV} construction

And once again we can also use the information from Figure 7.6 and Figure 7.7 to see which ideophones never occur in this kind of adverbial construction, e.g., móhū 模糊 or wáng~wáng. This intuitively seems correct because we have seen that the former occurred with a very high $\Delta P_{construction \rightarrow ideophone}$ in the IDEOPHONE_{PRED} construction; and the latter functions as a bare adverbial. Let us then inspect in the next section bare ideophone adverbials.

7.3.2.2 BARE IDEOPHONE_{ADV} construction It is well-known that adverbs in Chinese do not always co-occur with an overt marker like *de 地*. As we have seen above when discussing negatives with ideophones, there were only a few items that occurred in those environments, such as *hǎohǎo 好好* in example (92). This NEG IDEOPHONE schema was in fact an example of bare adverbial usage, often embedded in a somewhat idiomatic usage. It should then not come as a surprise that *hǎohǎo 好好* will jump out in the plots below⁹³.

But let us first go through the protocol with another item, namely *fēn~fēn 紛紛*, illustrated as occurring in this construction in (108). Its contingency table is shown in Table 7.9. As can be seen, cell *a* = 603 tokens, meaning that we are looking at a quite frequent ideophone, compared to the examples in the preceding sections. It should be remembered that the usage of cues as the association measures is contingency-based and thus will take this larger number into consideration, something that other measures would not necessarily provide. On the downside of being so frequent is of course that an item may be more entrenched and less likely to be judged as marked and consequently as ideophonic.

(108) a. (ASBC n° 2222)

都	紛紛	抬頭。
dōu	fēn~fēn	tái-tóu
all	in.succession.IDEO	raise-head
“They all looked up in succession.”		

⁹³Although once again, the ideophonehood of this item is questionable.



b. (ASBC nº 100622)

大家	紛紛	猜測，
dàjiā	fēn~fēn	cāi.cè
everybody in.succession.IDEO guess		

“One after another, everybody was guessing.”

Table 7.9: The contingency table for *fēn~fēn* 紛紛 in the BARE IDEOPHONE_{ADV} construction

	construction	¬construction	row totals
<i>fēn~fēn</i>	$a = 603$	$b = 119$	722
$\neg fēn~fēn$	$c = 7924$	$d = 13561$	21485
column totals	8527	13680	22207

After calculating the association measures, shown with highlights for the two different cues in Figure 7.8 and Figure 7.9, we once again can see that some ideophones are more likely to occur with a given construction, and that some are repulsed by it (right vs. left in Figure 7.8). And it can also be seen that this bare adverbial construction is likely to occur with a number of ideophones (top vs. bottom in Figure 7.9).

One of the nice outcomes of statistically approaching the data in this way is that we get a ranked score for all items. These are of potential help in future studies focusing on single items, because they provide perspective to a mere intuition-based approach. These often do not go beyond a listing of possible types, while foregoing their probability. But it seems probability is

tied to entrenchment and prototypicality in language in general (Geeraerts 2017), and presumably the same can be said for ideophones.

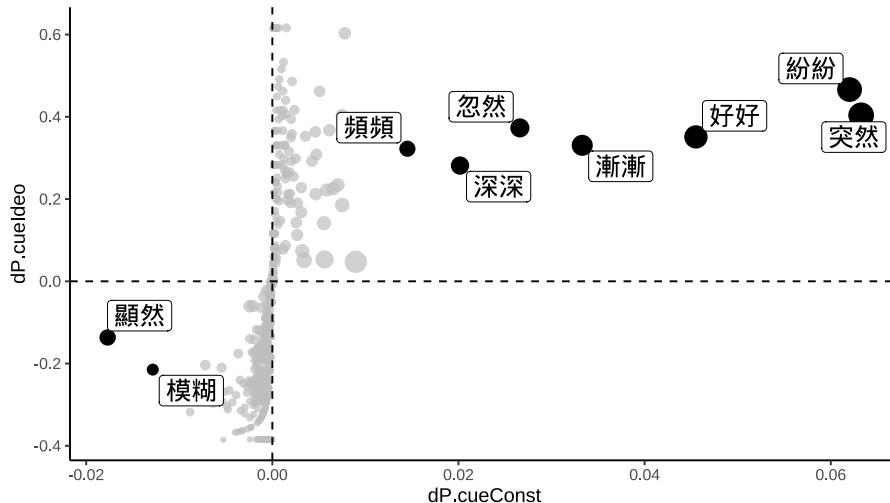


Figure 7.8: $\Delta P_{construction \rightarrow ideophone}$ in the BARE IDEOPHONE_{ADV} construction

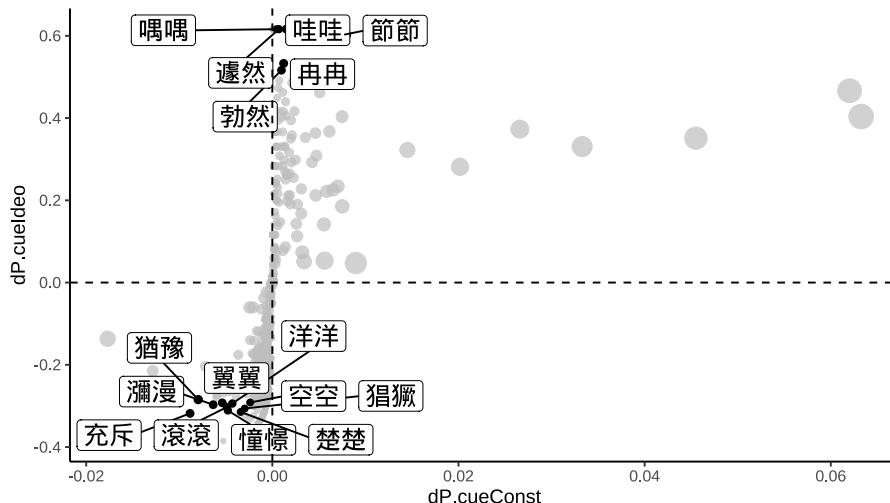
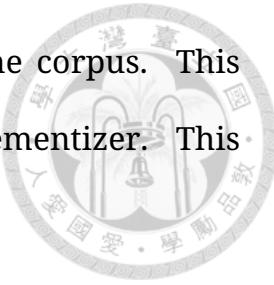


Figure 7.9: $\Delta P_{ideophone \rightarrow construction}$ in the BARE IDEOPHONE_{ADV} construction

7.3.2.3 DE_{COMP} IDEOPHONE construction The last adverbial construction that should be analyzed revolves around Meng's (2012) DE_{COMP} IDEOPHONE construction, as illustrated above in (96). The regular expression used to find these items is “得 _DE ({ideos})_ \\w+”. However, ideophones



followed by this DE_{COMPLEMENTIZER} occur only 21 times in the corpus. This suggests that ideophones are very rarely used as complementizer. This construction is illustrated in (109).

(109) a. (ASBC n° 107361)

郤 突兀得	令 人 難以 接受。
què tú~wù=de	líng rén nán.yǐ jiēshòu
but obtrusive.IDEO=DE CAUS people hard.to accept	
“It, however, was so obtrusively unacceptable.”	

b. (ASBC n° 107435)

她 感到 身上 被 月光 浸淫得	
tā gǎn.dào shēn-shàng bèi yuè-guāng qīn~yín=dé	
3SG feel body-on PASS moon-light soak.IDEO=DE	
寒怯難擋，	
hán.qiè.nán.dǎng	
timid.IDIOM	
“She felt timid with moonlight soaked in her body.”	

Surprisingly, there are not that many ideophones that fit this pattern. Rather than providing two small examples as before, I will show some “more frequent” collocations between verbs and ideophones in Table 7.10. If this construction had a higher frequency, such a listing could form the beginning for another application in the collostructional analysis method family, namely co-varying lexemes (Gries & Stefanowitsch 2004b).

Table 7.10: Collocates in the DE_{COMP} IDEOPHONE construction ($n > 3$)



verb	de	ideophone	n
活	得	好好	8
舉	得	高高	5
變	得	模糊	5
痛	得	哇哇	4

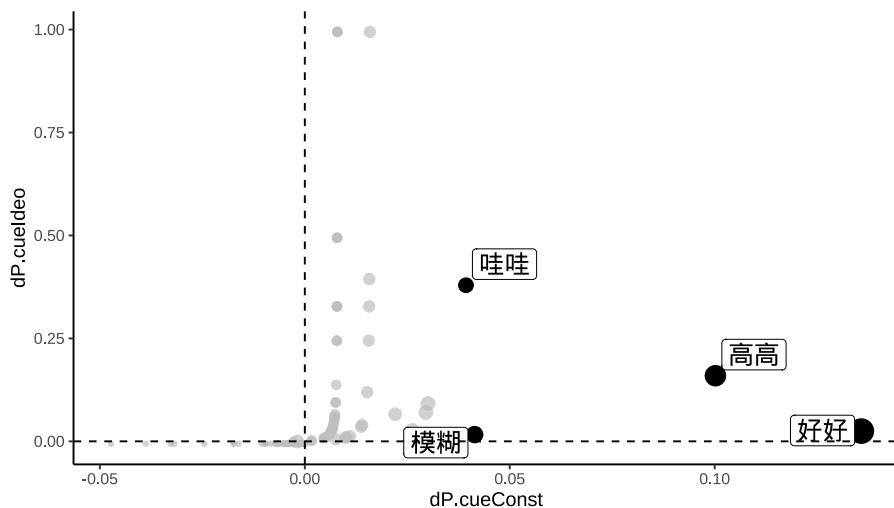
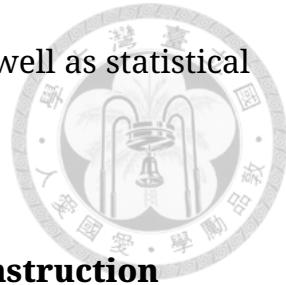


Figure 7.10: $\Delta P_{\text{construction} \rightarrow \text{ideophone}}$ in the DE_{COMP} IDEOPHONE construction

As such, we are still following the methodology we have used thus far. The results are shown in Figure 7.10. It stands out that, as one would expect, the items with the highest $\Delta P_{\text{construction} \rightarrow \text{ideophone}}$ values also occur in Table 7.10! By combining these two sets of information, one could start a lexical semantics project to study how verbs interact with ideophones. However, this matter will not be pursued here, for reasons similar to those raised by Gries (2019b:403), who excludes a number of data points from his analysis if they do not meet “certain frequency and dispersion thresholds”. While we would not go as far as completely omitting this construction type – intuitively it seems abundant – the corpus here does suggest to yield first to other constructions. Perhaps the last group of constructions we will dis-

cuss, attributive constructions, will provide more data as well as statistical solace.



7.3.3 Attributive constructions: IDEOPHONE DE_{ATT} construction

The last set of constructions to investigate is of the attributive type, as identified by Meng (2012), and illustrated in (98). These we schematize as the IDEOPHONE DE_{ATT} construction, and operationalize as “ideophone followed by DE followed by a nominal”.

The regular expression used to identify this construction is “(<{ideos})_\\w+._DE \\w+_N”, in which “\\w+_N” refers to any items tagged as nominals in the ASBC 4.0. In order to make sure there is no overlap with groups we already discussed above, we are taking out those items belonging to the IDEOPHONE_{PRED} DE construction (Section 7.3.1.1) and IDEOPHONE DE_{ADV} construction (Section 7.3.2.1), which also had a DE as overt marker.

Using *càn~làn* 燦爛 ‘bright’ as an example (110), we can construct once again a contingency table (Table 7.11).

(110) a. (ASBC nº 203313)

燦爛的 笑容，
càn~làn=de xiàoróng,
bright.IDEO=DE smile
“a bright smile”



b. (ASBC n° 106167)

燦爛的 陽光 在 茫茫的
 càn~làn=de yáng-guāng zài máng~máng=de lín-hǎi
 bright.IDEO=DE sun-light be.at vast.IDEO=DE forest-sea

耀起 無數的 亮點。
 yàoqǐ wú-shù=de liàng-diǎn

shine without-number=DE highlight-point

“The bright sunlight shone on countless points in the vast forest.”

Table 7.11: The contingency table for *càn~làn* 燦爛 in the IDEOPHONE DE_{ATT} construction

	construction	\neg construction	row totals
<i>càn~làn</i>	$a = 73$	$b = 68$	141
\neg <i>càn~làn</i>	$c = 2442$	$d = 19624$	22066
column totals	2515	19692	22207

Doing this for all items, allows us to visualize the $\Delta P_{construction \rightarrow ideophone}$ and $\Delta P_{ideophone \rightarrow construction}$ for all ideophone items in this construction. Once again, we have succeeded in bringing structure to our data: the ideophones with the highest $\Delta P_{construction \rightarrow ideophone}$ (Figure 7.11), e.g., *càn~làn* 燦爛 ‘bright’, *bīn~fēn* 繽紛 ‘in profusion (colors)’, *nóng~nóng* 濃濃 ‘dense, thick’ etc., are attracted by the construction. On the bottom left side, a number of adverbs are repulsed by it, as we have seen above. Ideophones which relatively depend more on the construction are shown on top in

Figure 7.12, while those that do not are shown in the bottom of this figure.

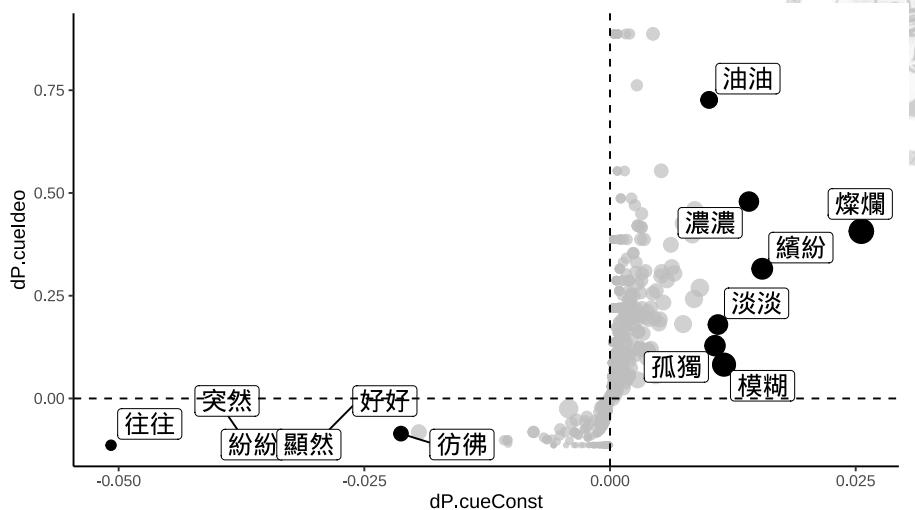


Figure 7.11: $\Delta P_{construction \rightarrow ideophone}$ in the IDEOPHONE DE_{ATT} construction

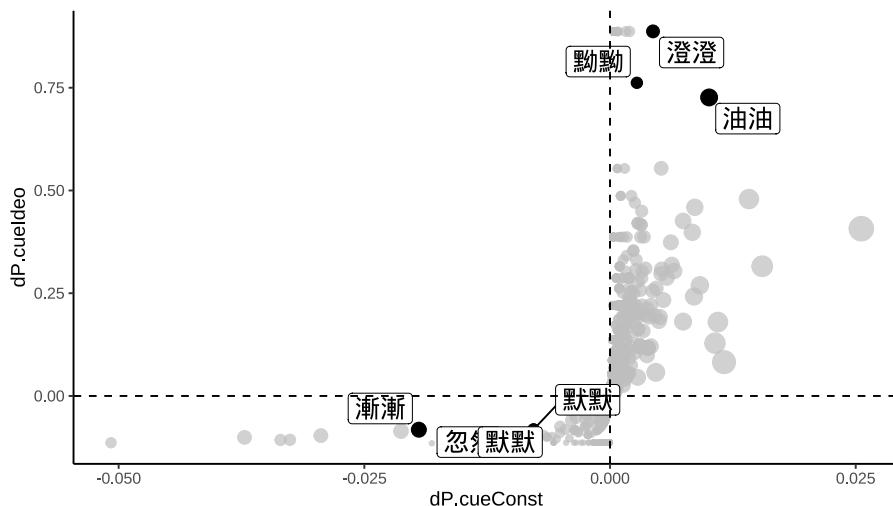
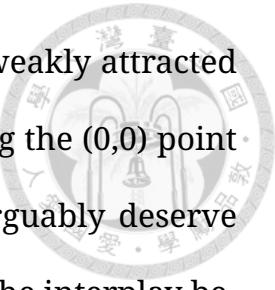


Figure 7.12: $\Delta P_{ideophone \rightarrow construction}$ in the IDEOPHONE DE_{ATT} construction

7.3.4 Collostructional analysis as a method for comparison

In the preceding sections, we have seen that not all ideophones are used equally in different constructions. This is unsurprising, since language use is economically structured. A well-known example is Zipf's law (Zipf 1949; Piantadosi 2014). For similar reasons our collostructional data all



seem to follow an S-curve. Most items are moderately or weakly attracted or repulsed by certain constructions, and are situated along the (0,0) point of the stretched S-curve, but some are outliers. These arguably deserve more attention than the moderate items, when describing the interplay between constructions and ideophones. As such, the highlights of each figure have focused on these outliers, which are situated along the dimensions of $\Delta P_{construction \rightarrow ideophone}$ (left-right) and $\Delta P_{ideophone \rightarrow construction}$ (top-bottom).

However, the moderate items can definitely be of interest as well. Let us say one is interested in one item in particular, e.g., *pái~huái* 徘徊 ‘go round and about, back and forth; shilly-shally, aimlessly irresolute, restless but hesitant, loiter and linger, dawdle and delay; also (med.) continuing ceaselessly, of sounds’. Looking at the definitions in the *Hànyǔ dà cídiǎn* 漢語大詞典, shown in Table 7.12, we can observe that the meanings of *pái~huái* 徘徊 are about movement, as well as a psychological state. It is also supposedly the same as *páng~huáng* 徨徨 ‘walk back and forth, shilly-shilly; vacillate, procrastinate; nervously pacing’, although the definitions in the *Hànyǔ dà cídiǎn* suggest that there is less polysemy in this item (Table 7.13).

Table 7.12: *Hànyǔ dà cídiǎn* definitions for *pái~huái* 徘徊

Meanings (English)	Meanings (Chinese)
go back and forth, circling	往返回旋；來回走動。
same as <i>páng~huáng</i>	猶彷徨。
linger on	流連；留戀。
walking slowly	安行貌；徐行貌。



Meanings (English)	Meanings (Chinese)
wind	猶回環。
flower	見“徘徊花”。

Table 7.13: *Hànyǔ dà cídiǎn* definitions for *páng~huáng* 徶徨

Meanings (English)	Meanings (Chinese)
waver back and forth, without peace of mind	來往走動、心神不寧貌。
go back and forth	回旋貌。

However, how similar are these items in terms of appearance in the constructions we have been discussing in this section? Now we can pick the fruits of our analyses. As Figure 7.13 and Figure 7.14 show, most of the constructions have the same interaction with these ideophones. They are both strong for the IDEOPHONE_{PRED} construction. However, there is a difference in frequency, with *pái~huái* being the more frequent one. Furthermore, *pái~huái* has a higher $\Delta P_{construction \rightarrow ideophone}$, suggesting that the cue of construction has a higher prototypical value for this item than for *páng~huáng*. The second big difference is in the IDEOPHONE DE_{ATT} construction, where it is *páng~huáng* that can be used. To perform a complete lexical semantic study of these two items, we should of course combine these findings with other informations, such as collocates, as we have done in previous chapters or will below in Section 7.4.

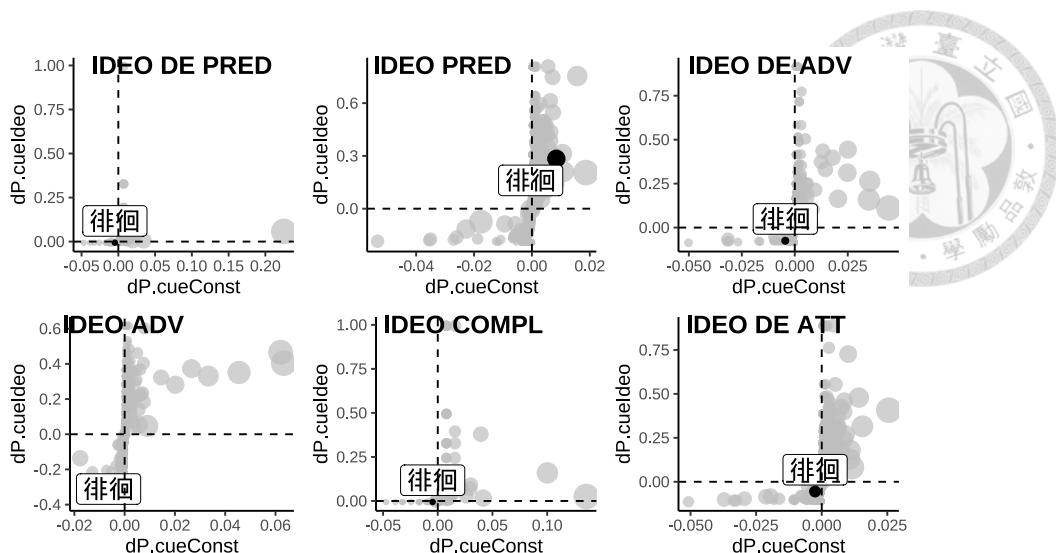


Figure 7.13: All constructions for *pái~huái* 徘徊 ‘waver, hesitate’

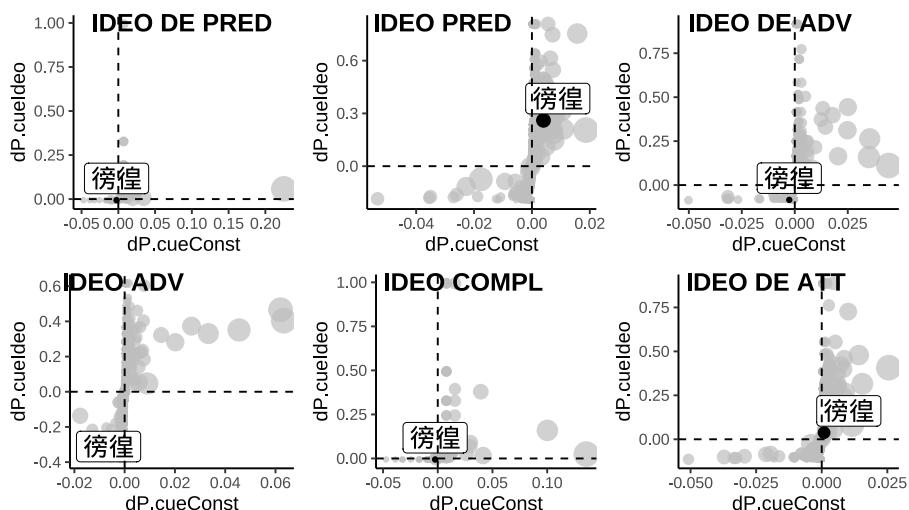


Figure 7.14: All constructions for *páng~huáng* 傍徨 ‘waver, hesitate’

The point here is not to do a full lexical semantic study, but rather to show again that ideophones as they occur in these constructions are also prototypically structured. In other words, while I agree with the brief descriptive overview by Meng (2012), her findings should be supplemented with data from real usage. To reiterate, if ideophones occur in a construction, the attraction is directional. Some ideophones rely on the construction to occur, i.e., when they have a high $\Delta P_{ideophone \rightarrow construction}$ but a low $\Delta P_{construction \rightarrow ideophone}$. For others, it is the construction that attracts

them (high $\Delta P_{construction \rightarrow ideophone}$). Of course, both can also occur, but typically no ideophones will occur evenly across all constructions.

This methodology thus seems to provide another piece of evidence to study the behavior and meaning of ideophones in Chinese. Further explorations could take the methodology outlined here as a starting point and perform detailed studies on specific items. Another issue is the *tupleization* that Gries (2019b) seems to advocate. In this study, dispersion (see Section 4.5.2) across the corpus has not been included, but we agree with him that this is a way to reduce skewedness towards one corpus part. For instance, it could be the case that one ideophonic item only occurs in one genre, or just so happens to occur a lot in one text but nowhere else, and thus is not very representative. These and related issues are well-known, and our solution has been to use the balanced corpus of ASBC 4.0, even though that is not as representative for e.g., internet language, or even spoken language⁹⁴.

That being said, we do think that the findings illustrated by the collexeme analysis shown in this section have been useful⁹⁵, in terms of attraction, but also repulsion. If it is true that repulsion is somehow correlated with ungrammaticality (Stefanowitsch 2006), collexeme analysis also provides a promising usage-based venue for future syntactic studies. It is for this reason we will use it once more in the next section, where the infamous ABB construction in Chinese is revisited in terms of association measures.

⁹⁴Another small issue is that the six constructions addressed in this section do not exhaust the data, i.e., there are many more (micro)constructions to be found and studied.

⁹⁵Especially as being a non-native speaker myself, it is useful to be able to demonstrate the interplay between items and constructions. This methodology can of course be used in many other constructions.



7.4 ABB constructions and ideophones

7.4.1 From ABB to COLLOCATE + IDEOPHONE

The literature on the ABB construction in Chinese is vast. One of the reasons is that this construction has been identified in a number of Sinitic languages, such as but not exclusive to Mandarin, Cantonese, and Taiwan Southern Min, as illustrated in (111).

- (111) a. Mandarin: *hēi-qī~qī* 黑漆漆 ‘pitch dark’
b. Cantonese: *hak1-maa4~maa4* 黑麻麻 ‘pitch dark’
c. Southern Min: *ó-sô~sô* 烏 sô-sô ‘pitch dark’

As T’sou (1978:67) states about the ABB template in Cantonese: “It was noted there that ABB is the basic form in which sound symbolism could be fruitfully explored”. This virtually accords prototype status to this construction. Subsequently, it has been studied mostly for its formal aspect, i.e., the BB patterns that co-occur with A, e.g., Bodomo (2006); Sew (2008); Lai (2015) for Cantonese; Chang (2009) for Southern Min; and a plethora of Chinese papers for Mandarin: Cáo (1995); Mok (2001); Zhāng (2005); Yáo (2006); Lǐ (2008); Sūn (2012); and Wang (2010); (2014).

The main perspective that these studies seem to take is that BB is an affix (*cízhùi* 詞綴) attached to a stem (*cígēn* 詞根). One will find a combination of the following opinions in the literature. Sometimes it is stated that the A in ABB can only be an adjective, but other times nouns and verbs are correctly identified in this A position. Some scholars maintain that BB has

no meaning, others do see some meaning in there. In a recent reference grammar (Huang, Jin & Shi 2016) the following is stated:

Affixation. Derivation by affixation predominantly involves monosyllabic adjectives. A common way to achieve this is to add to the root a suffix in the shape of a repeated syllable like the – 烘烘 *hong1hong1* [...]. Such a suffix typically does not have very clear semantic content but does make a distinctive contribution to the over-all meaning of the derived adjective. When – 烘烘 *hong1hong1* is suffixed to 臭 *chou4* ‘stinky,’ the combination contributes a special interpretation to the sentence that the kitchen waste is emitting a foul smell constantly and heavily. Similarly, the adjective derived from the suffixation of – 噴噴 *pen1pen1* to 香 *xiang1* ‘fragrant’ [...] has a reading related to 香 *xiang1* ‘fragrant’ with a more vivid flavor, namely, releasing a sweet scent continuously and extensively. The adjective 干巴巴 *gan1ba1ba1* ‘dreadfully dry’ [...] is derived from 干 *gan1* ‘dry’ but has the added flavor of awfulness and unattractiveness because of the suffix - 巴巴 *ba1ba1*.

Huang, Jin & Shi (2016:284)

It is clear that Huang, Jin & Shi (2016) are referring to ideophones, as can be seen by their usage of ‘more vivid’ as well as the sensory examples they provide on the semantic side, and the markedness on the formal side. Furthermore, their statement that this is an affixation process seems doubtful. After all, compared to some typical affixes, as they are presented by Packard

(2000) and illustrated in Table 7.14, these do not seem to conform.

Table 7.14: Affixes in Chinese (adapted from Packard 2000:174)

Wordforming prefix	Wordforming suffix	Grammatical affixes
<i>dì-</i> 第 ‘ordinalizer’	<i>-dù</i> 度 ‘degree’	<i>-men</i> 們 ‘human plural’
<i>fēi-</i> 非 ‘not’	<i>-huà</i> 化 ‘-ize/-ify’	<i>-le</i> 了 ‘perfective a.o.’
<i>fù-</i> 復 ‘again’	<i>-ér</i> 兒 ‘nominalizer’	<i>-zhe</i> 著 ‘durative’
<i>kě-</i> 可 ‘may’	<i>-rán</i> 然 ‘as’	<i>-guo</i> 過 ‘experiential’
<i>wú-</i> 無 ‘without’	<i>-tou</i> 頭 ‘nominalizer’	<i><bu></i> 不 ‘(negative) potential’
<i>wèi-</i> 未 ‘not yet’	<i>-xìng</i> 性 ‘nature’	<i><de></i> 得 ‘potential’
<i>zài-</i> 再 ‘again’	<i>-zhě</i> 者 ‘one who’	
<i>-zi</i> 子 ‘nominalizer’		

Since the items in Table 7.14 are affixes, they have a high frequency in the whole language. While they are not equally productive, they are placed on the grammatical, closed-class side of linguistic items (cf. Talmy 2000a:ch.1). When compared to this list in Table 7.14, any familiarity with the ABB construction in Chinese unequivocally leads one to the realization that BB does not behave like the elements in this list.

After all, what is an affix? According to Taylor (2003), (a) prototypical affixes are definitely bound, (b) usually cannot be stressed, (c) often somewhat integrate in the phonological shape of a word of which they are a part, (d) are highly selective to the items to which they attach, and (e) cannot be moved around independently of their stems. Criteria (a-b) and (d-e) intuitively seem true for most of the items in the Table 7.14. Prototypical words,

Table 7.15: Examples of ABB (adapted from Wang (2014:352))

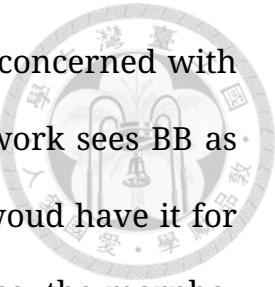
Adjective + BB	Noun + BB	Verb + BB
<i>ǎi-dūn~dūn</i> 矮墩墩 ‘short’	<i>shuǐ-líng~líng</i> 水靈靈 ‘charming’	<i>xiào-mī~mī</i> 笑咪咪 ‘smiley’
<i>là-sū~sū</i> 辣酥酥 ‘spicy’	<i>qì-gǔ~gǔ</i> 气鼓鼓 ‘angry’	<i>chàn-wéi~wéi</i> 颤巍巍 ‘shaky’
<i>lǎn-yáng~yáng</i> 懒洋洋 ‘lazy’	<i>lèi-wāng~wāng</i> 泪汪汪 ‘teary’	<i>chuǎn-xū~xū</i> 喘吁吁 ‘breathless’
<i>kōng-dàng~dàng</i> 空荡荡 ‘empty’	<i>hàn-jīn~jīn</i> 汗津津 ‘sweaty’	<i>zuì-xūn~xūn</i> 醉醺醺 ‘drunk’

Taylor argues, take the opposite of these five criteria. Clitics, he says, are not very selective (d), but for the most follow the same criteria as affixes. So BB is not a clitic either then.

But that being said, there are a number of BB elements that are argued to be (a) bound, because they don't occur other than in ABB environments (Wang 2014). It (b) is possible to stress BB. They do not (c) integrate in the phonological shape (the “root” A); instead they retain their full pronunciation. They are (d) highly selective to the items “to which they attach”, or rather *co-occur*. They (e) sometimes can be moved around independently of their “stems”. These differences with the prototypical affix strongly suggest that BB is not an affix, as Huang, Jin & Shi (2016) claim.

But what is it then? Wang (2014) argues that they are mostly compounds, and sometimes compounds after affixation. We will first briefly sketch Wang's data and argument and then come back to the compound question.

In the table Wang gives, reproduced in Table 7.15, he states that “the word stems can be an adjective, a noun, or a verb, but the reduplicated part is always an adjective” (Wang 2014:352). Our issue is not with the first part of the statement, but rather that the reduplicated part is always an adjective. It is better to characterize this element as an ideophone, belonging to the morphological template of Base-Base (BB). This is dif-



ferent than BB used by Wang, because that one is purely concerned with syllable/characters, while the ideophonic template framework sees BB as just one pattern among many, as would Mok (2001) also woud have it for instance. Characterizing the reduplicated part as Base-Base, the morphological template also has the advantage of differentiating with what is sometimes also understood as an ABB construction, namely ideophones like *gū~lū~lū* 咕嚕嚕 or *hōng~lóng~lóng* 轩隆隆, which purely on the basis of the phonology or characters do conform to an ABB template. These were characterized as ARR in CHIDEOD.

Table 7.16: Examples of ABB (adapted from Wang 2014:353)

Type	Adjective	Structure
1: A + BB	<i>ǎi-dūn~dūn</i> 矮墩墩 ‘short’	矮 + 墩墩
2: AB + B	<i>chì-luǒ~luǒ</i> 赤裸裸 ‘naked; undisguised’	赤 + 裸 + 裸
3: BA + B	<i>xiāng-pēn~pēn</i> 香噴噴 ‘delicious’	香噴噴

To continue, for Wang there are basically three types of ABB reduplications in Chinese, shown in Table 7.16. As he explains this table: “In Type 1, BB as a whole is attached to the left constituent and in Type 2, the right constituent B is reduplicated first and then attached to the left side. Type 3 is actually a two-step reduplication: First step, it is the reduplication on the left morpheme—BBA. Second step, BBA switches positions, A goes to the left side and BB goes to the right” (Wang 2014:354). The evidence for especially the third type comes from two sources. The first is the existence of items like *pèn-xiāng* 噴香 ‘delicious’, but not **xiāng-pèn* 香噴 or **xiāng-pèn~pèn* 香



噴噴. In the original Wang does not provide tone marks, but it may be of importance that *xiāng-pēn~pēn* has *pēn* with first tone, while *pèn-xiāng* 噴香 has *pèn* with fourth tone. The second source is Wang's native Shanghai dialect which follows the BBA pattern, "which suggests that our proposal is correct" (Wang 2014:354).

Subsequently, Wang states that these are prime examples of treating them as compounding, rather than affixation. The only exception is made for the BB element in e.g., *gān-bā~bā* 乾巴巴 'dry' or *shī-hū~hū* 濕乎乎 'wet'. Here, it is argued, these BB elements have lost their lexical meaning and "do not contribute to the meanings of the whole words, so they should be treated as suffixes" (Wang 2014:354).

That is all good and well, if you neglect the dangers of the fallacies like the exclusionary fallacy and the rule/list fallacy (Langacker 1987b:28–29) and the process metaphor (Langacker 1987b:63). These state that particular statements (lists) are to be excised from the grammar if general statements (rules) can be established that subsume them. For example, a noun like *stapler* presents a dilemma to a processual treatment (as used by Wang 2014 in type 3 above) of forms: if the form *stapler* is derived by rule, it is impossible to account for its meaning being more than 'something that staples'; if it is simply listed in the lexicon, the productive *V + -er* derivational pattern, which certainly can be identified here, cannot subsume it (Langacker 1987b:28). Or in other words, as Wang states, if the BB form has no direct lexical meaning, it is a suffix and nothing else; if it has meaning, it must be following a processual rule that is the sum of its components, which then

comes in three types.

The solution to these dilemmas is to recognize that so-called complex forms have a status of their own, in which it sometimes is possible to view them as the sum of their parts, but not always. This is illustrated in Figure 7.15. Following the notations in Langacker (1987b) and Tuggy (1992), the hatched rectangles are so-called elaboration sites. These mark that on one level of description, the component is dependent on a more autonomous component⁹⁶. In the case of Figure 7.15.a, *bīng~bīng* is dependent on *lěng*, just as *nuǎn* is more autonomous than *hōng~hōng*, and *gān* elaborates *bā~bā*. This is actually not a surprising phrasing, as similar ideas were used in Chapter 5, where the meaning of ideophones was elaborated by their usage and their reference.

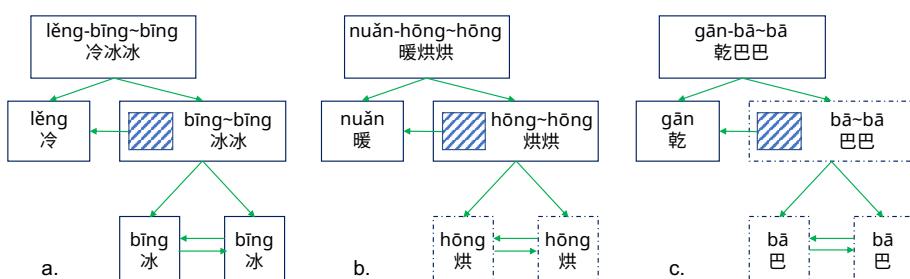


Figure 7.15: Three types of ABB compositional structures

A, b, and c in Figure 7.15 differ with respect to unit status on the different levels of composition. In Cognitive Grammar, a unit is a cognitive structure mastered by a language user to the point that it can be employed in a largely automatic fashion, without requiring attention to its individual parts or their arrangement. It is similar to phenomena like chunking or au-

⁹⁶Autonomy in Cognitive Grammar does not necessarily mean that a component can occur by itself. Related, dependence is not to be interpreted in the same way as dependency grammars often use it.

tomatization. In Figure 7.15.a items have reached unit status⁹⁷ on all levels.

For Figure 7.15.b, the situation is slightly different. There are a number of ABB forms that occur with *hōng~hōng* as its BB part, and they seem conventional enough to claim unit status. But *hōng* by itself seems much less accessible, at least in the meaning it has as a reduplication; as a verb ‘to dry; to bake etc.’ it of course can be considered a unit. In Figure 7.15.c, *bā~bā* is recognizable in a number of ABB patterns, and has achieved unit status, but much less than the other two cases. As a single character *bā* is even less available. These impressions are of course negated to a certain degree by Chinese relying on its semiotic folk model, which often recycles writing in many contexts.

To return to the issue at hand, it seems unnecessary to propose three different rules in two compounding / affixation + compounding paradigms, as Wang (2014) does. There are different levels of ENTRENCHMENT, related to the degree of unit status a component has reached. And sometimes we can see the composition clearly, but not always. And in any case, if it is compositionally transparent, the more complex structure usually attracts the strongest focus, and not its parts.

Is the compounding question resolved then? Chinese has been called a language of compound words (Arcodia 2007). Although the difference between compounds and phrases is not clear (Li & Thompson 1981; Huang 1984; Zhou, Ostrin & Tyler 1993; Bates et al. 1993), it seems reasonable to agree with Wang (2014) that we are dealing with a compounding process for the ABB-construction.

⁹⁷Unit status does not mean that a component can occur on its own.

A more important question pertains to the weight given to the ABB-construction. After all, the linguistic availability of this construction seems somewhat limited. Cáo (1995) reports identifying 338 types in his dictionary data. Wang (2014) twists these words and says: “According to CAO (1995), there are 338 ABB adjectives in Chinese” (Wang 2014:352), without referring to the nature of the data. Wang (2014) also makes use of dictionary material and identifies 336 types. And the studies referred to above make similar claims. What if we instead opened up the construction? If ABB is reanalyzed as a construction in which A is the COLLOCATE, we immediately can study more types. And if we, in a second move, not just focused on full reduplication, which in the examples given here, behave suspiciously similar to ideophones, in that they are marked words that depict sensory imagery which belong to an open lexical class⁹⁸, we can rephrase BB as IDEOPHONE. In sum, ABB should be rephrased as the COLLOCATE + IDEOPHONE construction. Given this more schematic construction, there certainly can be a special conceptual space for ABB as a special instance of that schema, but it should not be limited to that. The recategorization is diagrammed in Figure 7.16.

⁹⁸This is what makes the affix interpretation somewhat unavailable.

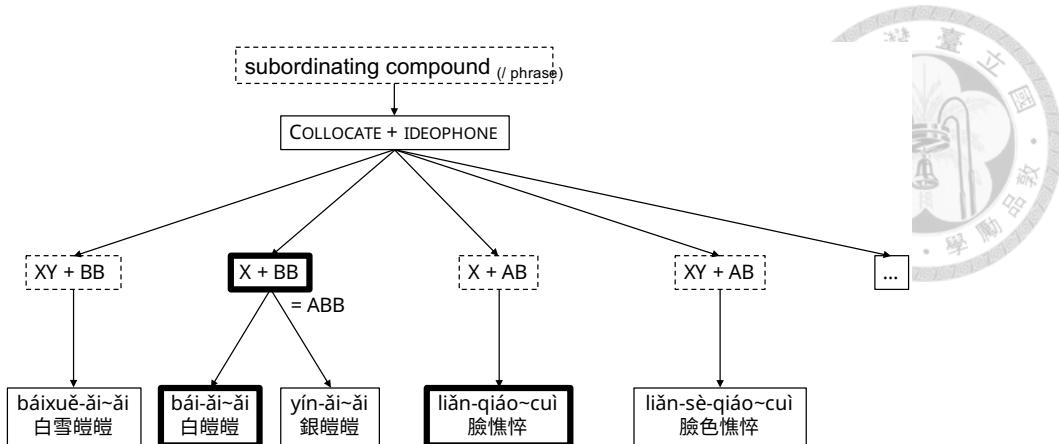


Figure 7.16: Simple schematic network of the COLLOCATE + IDEOPHONE construction

Figure 7.16 takes inspiration from Tuggy's insightful discussion on reduplication in Nahuatl (2003). The exemplars are shown on the bottom of the diagram, with their schemas on the mid-level (with ABB rephrased as XBB for clarity), and the most abstract schema on top.

For *ǎi~ǎi*, dictionary material (Zhāng 2005; Wang 2014) apparently only gives *bái* 白 ‘white’ as a possible A. However, a cursory search in the ASBC corpus yields also items like *yín* 銀 ‘silver’ or *bái-xuě* 白雪 ‘white snow’ as possible ’A’s. Since *bái* comes from the dictionary, it arguably is better entrenched and more prototypical, i.e., more likely to occur with *ǎi~ǎi* than the other two. But the danger lies in excluding the other two from consideration based on list data (dictionary) and intuition alone. Clearly, if the other two also occur with it, it tells us something about the semantics of *ǎi~ǎi*. In this case, a gloss like ‘pure white’ does seem to cover the semantics well enough, as both ‘silver’ and ‘white snow’ are located in that part of the color spectrum.

However, another issue is that *bái-xuě* (or maybe even *xuě* 雪 ‘snow’) would normally not even be considered in an ABB analysis, since it violates

that construction's arbitrary template. Along to the right side of the diagram, similar arguments in favor of opening up ABB to include other patterns can be made: these exemplars with *qiáo~cuì* would not even be considered under the traditional ABB analysis (which is why there were only some 330 items in previous studies).

Important to note, however, is that the schemas in Figure 7.16 are not necessarily equally entrenched or salient to speakers. What can be said in favor of ABB as a construction (once again: shown as XBB on the diagram), is that it has a somewhat prototypical position as a schema, and the others do not. But it *is* sanctioned by a more general schema, namely the COLLOCATE + IDEOPHONE construction, which in turn is sanctioned by the subordination type discussed above. Tuggy (2003) makes the same point for his description of reduplication in Nahuatl:

Although other models would posit only 1.b [the highest level schema], excising 1. c-d [midlevel schemas] because of it, the [Cognitive Grammar] model does not give us any reason to suppose that speakers' minds gravitate towards such higher level schemas as automatically as analysts' seem to; *ceteris paribus*, lower-level schemas are expected to be more salient.

Tuggy (2003:95)

In other words, we should not just describe these patterns as instances of subordination, but also chart the other patterns in data. The point is that ABB as a pattern, while salient to Chinese linguists as well as to many Chinese speakers through the SOUND-WRITING-MEANING semiotic folk model,

has a special place amidst other constructions. We should not leave those other constructions out of sight, and investigating the data with the higher level construction of COLLOCATE + IDEOPHONE provides good middle ground.

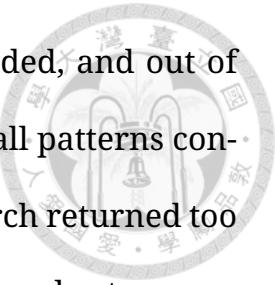
In the next section collexeme analysis will be used to show how such studies can be done.

7.4.2 Cues and the COLLOCATE + IDEOPHONE construction

Having argued in favor of a higher level construction has benefits for the investigation of ideophones and their relation to collocates. Not only will we find more tokens, and subsequently more data, but we can also chart the prototypicality of salient schemas such as the ABB schema. And we can also finetune or nuance statements like the autonomy-dependence characterization provided above. That is to say, some collocates may attract certain ideophones more than others (or even repulse them) and some ideophones may attract or repulse certain collocates as well.

As for data, the balanced ASBC corpus will be relied on once more. Starting from the same dataset as above, we are limiting the ideophones to two syllables again. This is a methodological choice to show how similar items like BB (from ABB), i.e., disyllabic ideophones of the morphological templates BB, RR, BR or RB can already greatly increase the scope of the data under investigation⁹⁹. Compositional ideophones, which are listed with suffixes like *-rán* 然 in CHIDEOD, e.g., *máng-rán* 茫然, are also excluded from the data. This is a first group of data.

⁹⁹ Including monosyllabic, trisyllabic and quatrissyllabic ideophones will increase this scope further, but it may be too far away from the ABB construction to make my point.



In order to make sure *all* the ABB candidates are included, and out of fear that some BB / ideophones are not listed in CHIDEOD, all patterns conforming to the ABB in the ASBC¹⁰⁰ are added. Since this search returned too many items, items with tags like noun, foreign word, numerals etc. were dropped. This is a second group of data.

A third group of data was selected according to the criteria of the first group, but now looked beyond the word border, with word defined in terms of the ASBC corpus tagging. That is to say, in group one the ASBC had segmented *huáng-dēng~dēng* 黃澄澄 ‘glistening yellow’ as one word, but now we are trying to get items segmented in two computational words, e.g., *bái-sè cāng~máng* 白色蒼茫 ‘indistinct white’. After all, these still conform to the larger COLLOCATE + IDEOPHONE construction, and it would not be all right to leave out data just because it was segmented in this way in this corpus.

These three groups are merged into one dataset where duplicate values have been left out. This gives us 5608 tokens to work with. In terms of types, it seems there are 1837 types of truly ABB types in this data source, a number that is vastly greater than the supposedly 338 ABB types found in dictionary data (Cáo 1995; Wang 2014). In terms of the schematic representation in Figure 7.16, there are 8 patterns in the data: the collocate ranges from X to XY to XYZ to XYZW; the ideophone is coded in terms of AB and BB¹⁰¹. As is shown in Table 7.17), full reduplication in the ideophone is indeed the most

¹⁰⁰I did this using the regular expression ""\\b[^](.)\\1(?= _\\w+)". This is to be read word boundary - non-space character - any character - repetition of any character - followed by a tag.

¹⁰¹Once again, I think it is better to characterize the ideophone in terms of the morphological template, but because the point is to broaden up the ABB construction, I resort to simple AB and BB notation.

Table 7.17: Patterns and frequencies in the COLLOCATE + IDEOPHONE construction

collocate_pattern	ideo_pattern	token_freq	type_freq	example
XY	BB	1856	755	文質彬彬
X	BB	1837	528	沈甸甸
XY	AB	1596	951	七彩繽紛
X	AB	179	127	霧瀰漫
XYZ	AB	80	74	小朋友躊躇
XYZ	BB	47	47	金銀色閃閃
XYZW	AB	8	8	千言萬語邏迤
XYZW	BB	5	5	天災人禍頻頻

frequent, in terms of token frequency. But somewhat surprisingly, the most common collocate is not of the X (or A) type! Instead it is XY, followed by an almost equal group of tokens containing X.

As for type frequency, XY AB ranks first, followed by XY BB. This is not that unexpected, since it is well-known that Chinese has a four character template, most familiar from various idioms. The pattern X AB is already less frequent, both in terms of token or type frequency. And below that there are even less tokens or types consisting in total of five or six characters.

It is then possible to see why exactly the trisyllabic construction of ABB (X-BB) has attracted this much discussion in Chinese linguistics: there is both a relatively high type and token frequency. But this is eclipsed by tetrasyllabic expressions that follow the XY-BB and XY-AB schema. Especially here the type frequency is of importance. As Bybee (2001) states:

“It appears that the productivity of a pattern, expressed in a schema, is largely, though not entirely, determined by its type frequency: the more items encompassed by a schema, the stronger it is, and the more available it is for application to new

items. Thus, the English Past Tense *-ed* applies to thousands of verbs and is much more productive than any of the irregular patterns, which are highly restricted in the number of verbs to which they apply (Bybee 1985, 1995; MacWhinney 1978)."

Bybee (2001:13)

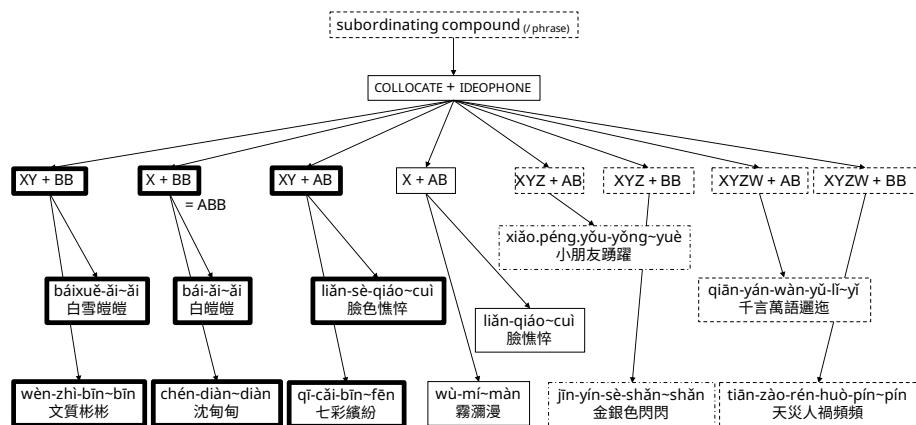
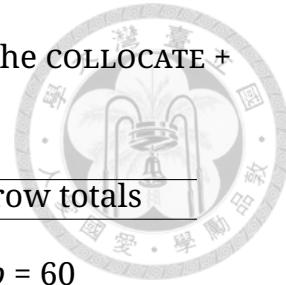


Figure 7.17: Revised schematic network of the COLLOCATE + IDEOPHONE constructions

We can capture the relative salience of the patterns in the revised diagram, shown in Figure 7.17. While such a diagram goes a great deal in conveying the message that ABB should be seen in context, or maybe competition, to other patterns, we can actually go beyond this. Since we have more data available, we can perform a co-varying collexeme analysis (Gries & Stefanowitsch 2004b). Given that our $\text{COLLOCATE}_{\text{SLOT } 1}$ $\text{IDEOPHONE}_{\text{SLOT } 2}$ has two slots, we can use the same contingency methods to explore the attraction of the collocate to the ideophone, and the attraction of the ideophone towards the collocate. Similar to the previous half of this chapter, I will thus use contingency based association measures, which are also directional, here respectively abbreviated as $\Delta P_{\text{collocate} \rightarrow \text{ideophone}}$ and $\Delta P_{\text{ideophone} \rightarrow \text{collocate}}$.

Table 7.18: The contingency table for *hēi-qī~qī* 黑漆漆 in the COLLOCATE + IDEOPHONE construction



collocate	ideophone	row totals
	漆漆 other ideophones	
黑	$a = 13$ $b = 47$	$a+b = 60$
other collocate	$c = 2$ $d = 5546$	$c+d = 5600$
column totals	$a+c = 15$ $b+d = 5593$	$a+b+c+d = 5608$

Let us revise, the first step is to construct a contingency table. However, instead of constructions ~ ideophone, we are now looking at collocate ~ ideophone. Illustrated with *hēi-qī~qī* 黑漆漆 in Table 7.18, it can be seen that the combination of *hēi-qī~qī* occurs 13 times in the dataset (cell *a*). *Qī~qī* occurs two more times with other ideophones (cell *c*). *Hēi* is found 47 more times with other ideophones (cell *b*). That makes that the value for the infamous cell *d* is 5,546, given a total of 5,608.

If we programmatically make the contingency table for all collocates and ideophones, we can also calculate their $\Delta P_{ideophone \rightarrow collocate}$ and $\Delta P_{collocate \rightarrow ideophone}$. These follow the same formulas as we used above. If the ideophone is the cue, $\Delta P_{ideophone \rightarrow collocate}$ can be calculated with the following formula: $\frac{a}{a+c} - \frac{b}{b+d}$. If, however, the collocate is the cue, $\Delta P_{collocate \rightarrow ideophone}$ follows the formula: $\frac{a}{a+b} - \frac{c}{c+d}$. For *hēi-qī~qī* 黑漆漆 this would mean that $\Delta P_{ideophone \rightarrow collocate} = 0.86$ and $\Delta P_{collocate \rightarrow ideophone} = 0.22$. Or visualized, Figure 7.18 shows that there is quite a bit of variation present in the data set.

In the bottom left quadrant of each plot we find values that are moderately attracted to the cue (in the left plot this is *hēi* and in the right plot it is *qī~qī*). It is maybe not very clear on these two plots but *hēi-qī~qī* has the

highest token frequency. But this usage-based perspective shows that if one is given *hēi*, it is not all too certain that the corresponding ideophone will be *qī~qī*; there seem to be many competitors¹⁰².

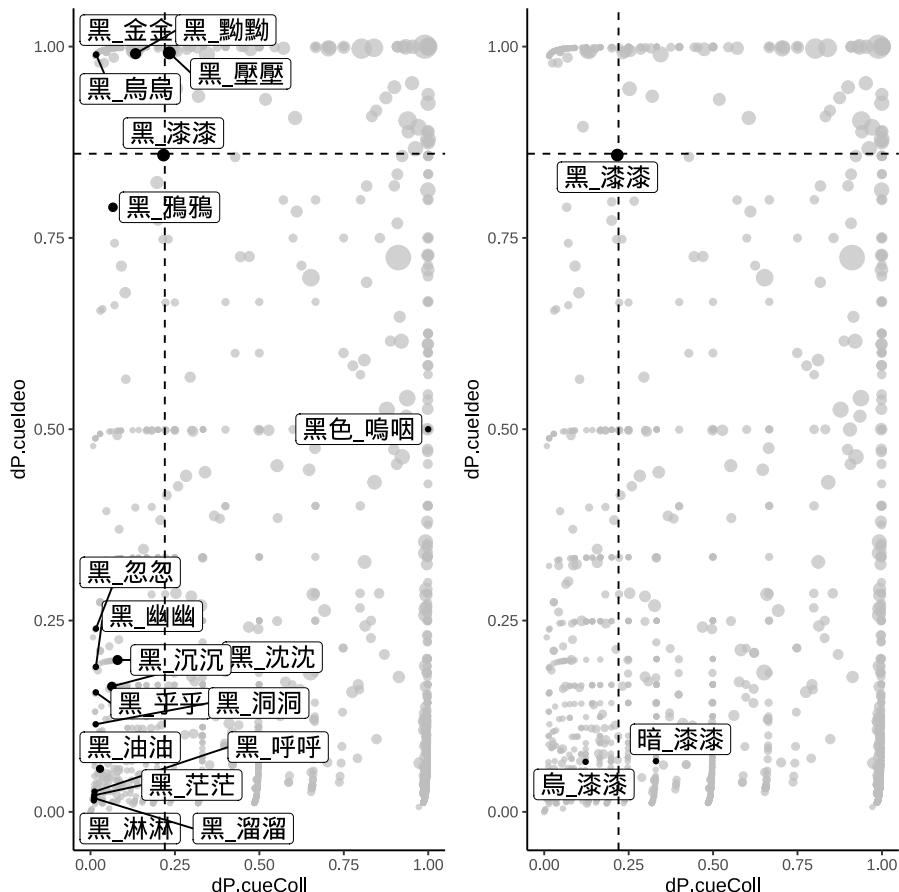


Figure 7.18: $\Delta P_{ideophone \rightarrow collocate}$ and $\Delta P_{collocate \rightarrow ideophone}$ visualized from the perspective of *hēi* 黑 and from *qī~qī* 漆漆

Let us inspect the plot for highlights for *qī~qī* as ideophone and its corresponding collocates. *Hēi* is in the same position, but here we find two other alternatives, namely *àn-qī~qī* 暗漆漆 ‘pitch dark’ and *wū-qī~qī* 暗漆漆 ‘pitch black’. This is surprising, because Wang (2014:353) explicitly agrees with Zhāng (2005) in stating that “漆漆 *qīqī* ‘paint’ only combine[s] with 黑

¹⁰² Although I must admit that when I did an informal inquiry to my fellow students, many of the other competitors seemed strange to them. Perhaps in the right context and with priming these could rise in naturalness. This is something that could be tested in future work.

hei ‘dark, black’ ” (emphasis mine). Two comments must be made. The first is that *hēi* seems the prototypical collocate for *qī~qī*, and in that sense the two scholars are not entirely wrong, but they are also not right – a consequence of taking dictionary material as the basis for generalizations. The second comment is that perhaps there is some transfer from Taiwanese into Mandarin. But even then, it does not seem strange that these two collocates appear here, since both are semantically close to the meaning of *hēi*.

Knowing this, it won’t be surprising what can be found for *bái* 白 and *ǎi~ǎi* 噼 噪, another pair for which there supposedly is only one collocate (*ibid.*). As we can see in Figure 7.19, *bái* has many construction types in the data set, and depending on what other word it modifies the following ideo-phone will be different. Looking from the perspective of *ǎi~ǎi*, it is clear that *bái-ǎi~ǎi* 白 噪 噪 is not the most prototypical value – that honor goes to *bái-xuě-ǎi~ǎi* 白 雪 噪 噪 ‘the white snow pure white’. We hope that the value of opening up the ABB construction to include more schematic patterns is demonstrated once again. It could be argued, of course, that “we should treat *bái-xuě* as A, so the ABB construction still stands”, but that is unfortunately not how the literature treats ABB. ABB is treated purely on the basis of characters, and that is simply not enough to do justice to the construction at large.

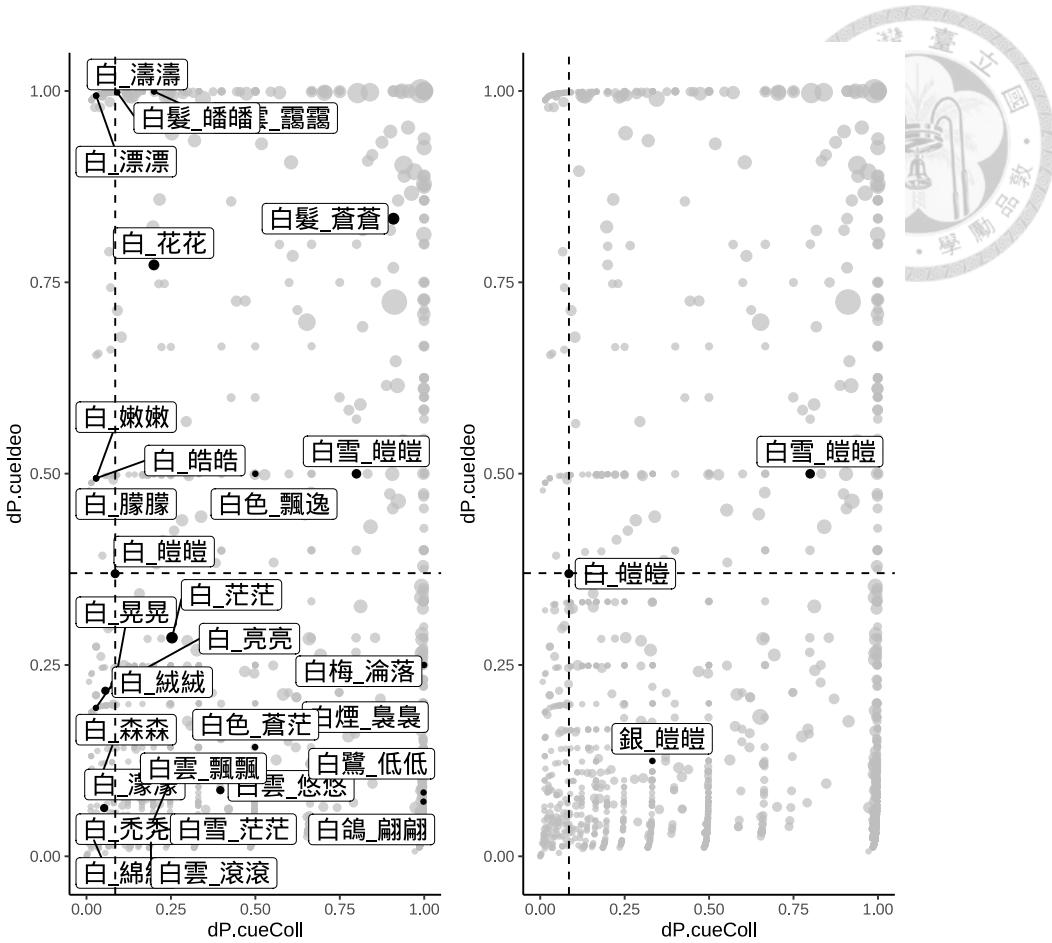


Figure 7.19: $\Delta P_{ideophone \rightarrow collocate}$ and $\Delta P_{collocate \rightarrow ideophone}$ visualized from the perspective of *bái* 白 and from *ǎi~ǎi* 眇眇

Let us examine two more examples. The first is *huó-shēng~shēng* 活生生 ‘lively’, which occurs quite often in the corpus¹⁰³. It occurs in written data, and has a very literary feeling to it, as shown in (112a).

¹⁰³Surprisingly not (yet) found in CHIDEOD.



(112) a. ASBC (nº 202255)

文化 是 一 條 活生生的、
 wénhuà shì yì tiáo huó-shēng~shēng=de
 culture COP one CLF living=DE

浩浩蕩蕩的 大江 大河，
 hào~hào-dàng~dàng=de dà-jiāng dà-hé
 vast.mighty.IDEO=DE big-stream big-river

“Culture is a living and mighty stream.”

b. ASBC (nº 107478)

他 也 曾經 碰過 活生生的 例子，
 tā yě céngjīng pèng-guò huó-shēng~shēng=de lìzi
 3SG also before touch-EXP lively=DE example

“He also encountered living examples.”

Its cell *a* value is 114, so it has a pretty high token frequency. In fact, 114 is the highest value for cell *a* in the dataset. Given the collocate *huó* as the cue, it seems *shēng~shēng* indeed is the most likely candidate, as can be seen in Figure 7.20, but far from the only one. From the perspective of *shēng~shēng*, we get collocates like *huó*, but also *yìng* 硬 and a few other alternatives. The more we look at these examples, the more it seems that the role of idiomticity cannot be overstated.

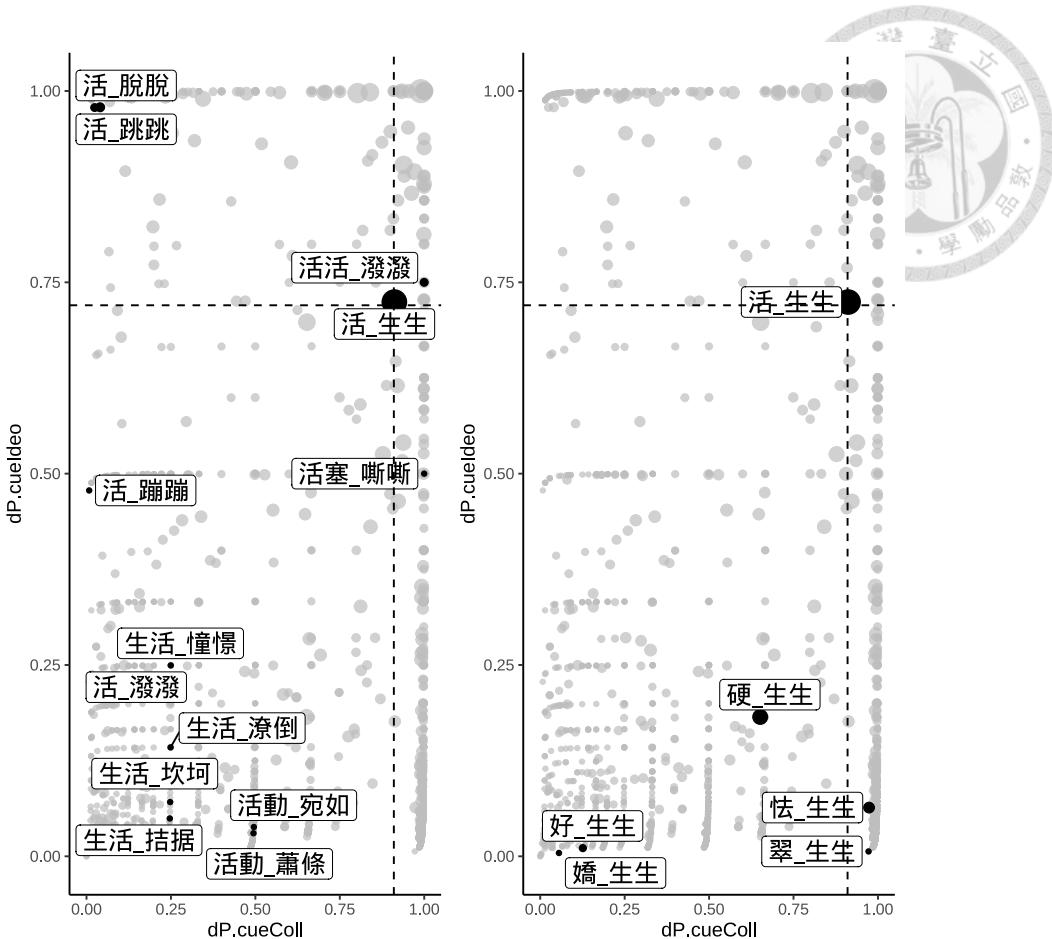


Figure 7.20: $\Delta P_{ideophone \rightarrow collocate}$ and $\Delta P_{collocate \rightarrow ideophone}$ visualized from the perspective of *huó* 活 and from *shēng*~*shēng* 生生

The idiomacity idea is also present in the item with the second highest cell a ($a = 97$), namely *xiǎo-xīn-yì~yì* 小心翼翼 ‘with utmost care’. Let us first start with the right panel in Figure 7.21. Here it can be seen that the $\Delta P_{ideophone \rightarrow collocate}$ and $\Delta P_{collocate \rightarrow ideophone}$ is nearing 1 for both measures. This means that they have a very strong attraction to each other from both sides. Given the collocate, the chance is extremely likely that this ideophone will follow, and vice versa. The left panel also shows an interesting observation. We see a very infrequent (cell $a = 1$) variant of *yì~yì* 翼翼, namely *yì~yì* 奕奕.

This seems to be a typo, but it could of course also be the beginning of

a conflation. In the case of the latter, we would need more current data to study this emergence. Note that *yì~yì 奕奕* does occur in two other exemplars: *shén-cǎi-yì~yì 神彩奕奕* ‘glowing with health and radiating vigour’ and *jīng-shén-yì~yì 精神奕奕* ‘vigorous spirit’.

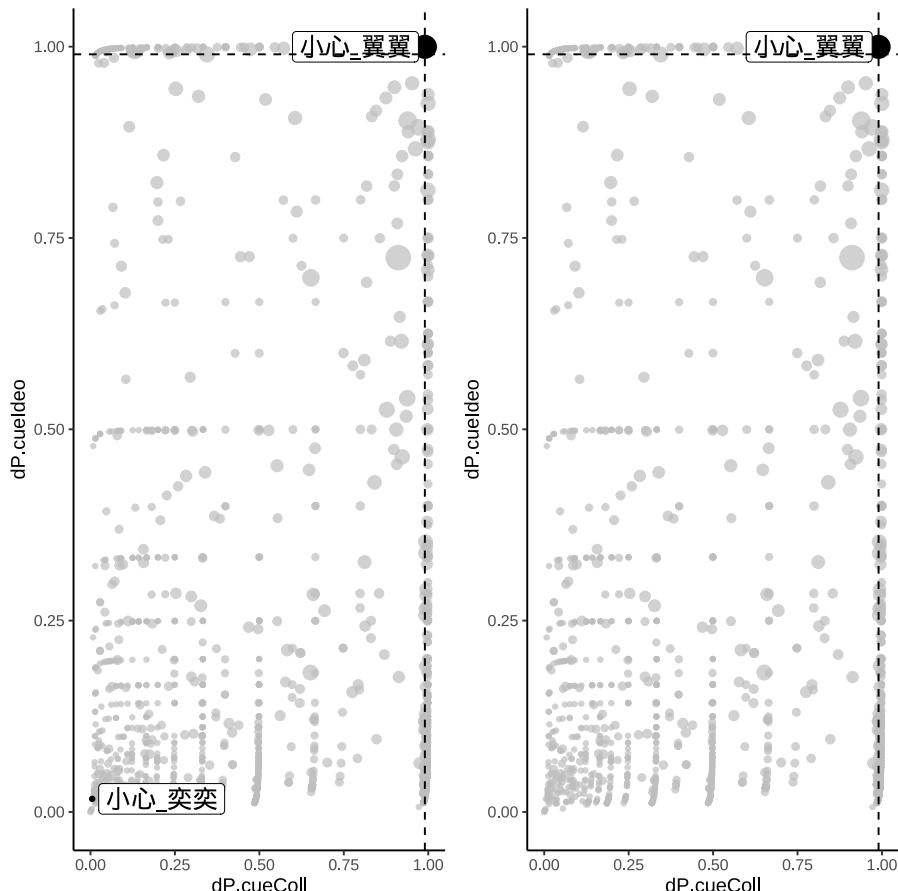


Figure 7.21: $\Delta P_{ideophone \rightarrow collocate}$ and $\Delta P_{collocate \rightarrow ideophone}$ visualized from the perspective of *xiǎo-xīn 小心* and from *yì~yì 翼翼*

In any case, the idiomacity of ‘the correct *yì~yì*’ is an important note to finish this section on. It seems that one cannot exist without the other in this case, i.e., the collocate and the ideophone. While there are presumably many other examples of (almost) hapax legomena in the data set, there is also a wide range of possibilities for the degree of attraction the ideophone exerts and the pull of the collocates.

Additionally, perhaps the examples provided above are not as ‘surprising’ to a native speaker of Chinese. However, the main contribution of the methodology in this section has consisted of usage-based approaches that illuminate the salience of native speakers, if indeed corpus data can represent this. Since it is impossible to verify which combinations of collocate and ideophone are somewhat out of the ordinary to a given reader, we have built a supplementary application¹⁰⁴ which allows the user to explore different combinations of collocates and ideophones. The constellation of the plot in the app is the same as the ones discussed here, but the user can enter collocates and ideophones to see how they relate. Another benefit of this app is that it can be used to quickly further explore the salience of these two structural groups, i.e., collocate and ideophone. This may result in future research that probes the psychological reality of ideophones, their iconicity value or ideophonicity (with ratings for example).

7.5 Conclusion

Throughout this chapter constructions pertaining to ideophones in Standard Chinese have been explored. This has been done on two different analytical levels, namely a macro and meso level¹⁰⁵. First, attention was devoted to the macro level. Based on previous discussions of complex adjectives and ideophone constructions six patterns of interest were identified, depending on predicative, adverbial or attributive usage of

¹⁰⁴The app is available here: https://simazhi.shinyapps.io/ABB_app/

¹⁰⁵The micro level, or ideophone internal level, has effectively been explored in Chapters 3 and 4.

ideophones. The main point of this chapter is that for the category of ideophones all these constructions can be identified, but not all ideophones occur in them equally, a notion rarely touched upon in previous research.

This notion was operationalized by using collexeme analysis, more specifically, collostructional analysis. The association measures involved contingency and directionality. That is to say, as Ellis & Ferreira-Junior (2009) convincingly argue, attraction and repulsion depends on relative frequency, and can be studied by giving a construction or an ideophone as a cue. High scoring elements in either or both measures $\Delta P_{ideophone \rightarrow construction}$ and $\Delta P_{construction \rightarrow ideophone}$ provide clues as to the prototypicality of a construction or ideophone in that construction. The resulting visualizations were explored. It could be seen that they followed a spread out S-curve, with most items being moderately attracted or repulsed by the construction (and vice versa). However, the outliers provided ample discussion points.

In the meso-level, we explored the notorious ABB construction in Chinese. First we looked at problems with this construction from a number of theoretical perspectives, such as the wrong characterization of it as an affix, followed by the probable characterization as a compound or possibly phrase – they are not clearly demarcated (in Chinese). Next, we argued for opening up both A and BB to allow for larger collocates and more ideophones. This larger scope enabled us to find more data, and it was found that ABB is but one of three patterns that are high in type frequency.

After this, another method from the family of collostructional analysis,

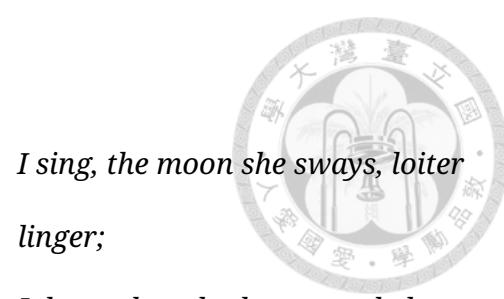
namely co-varying collexeme analysis was adopted. This allowed us to explore the attraction and repulsion between collocate and ideophone. It was demonstrated that moving beyond ABB to a COLLOCATE IDEOPHONE construction is a crucial step forward for future lexical semantic descriptions of particular ideophones and their collocates, which elaborate them.

Of course, this chapter is also not devoid of limitations. Methodologically, if we take the warning of Gries (2019b) seriously, we should have projected the dispersion on a separate axis. However, we find his 3D models hard to interpret in print, so we have foregone this part of his tuples.

We have also focused on the statistical methods that can be used to study ideophones in these constructions, rather than provide extremely detailed lexical semantic sketches of ideophones. We leave this for future research, where collostructional analysis hopefully will become a more ingrained tool in usage-based studies, because it has a tremendous potential for charting variation and prototypicality.



8 Conclusions



*I sing, the moon she sways, loiter
linger;*

*I dance, her shadows run, helter
skelter*

我歌月徘徊，我舞影凌亂。

Lǐ Báí 李白

This dissertation explored the variability and prototypicality of Chinese ideophones, within a Cognitive Linguistics framework. It adopted a quantified usage-based approach to study ideophones both synchronic and diachronic.

The main finding is that, indeed, ideophones are a heterogeneous category in Chinese. My significant original contributions to knowledge include (1) the creation of an open-source database of Chinese ideophones which contains data on parameters like the phonology, morphology, meaning, etc. of ideophones, and which can be updated with future work; (2) and four methodological perspectives that show how the variation of and in this category is structured, i.e., with Chinese ideophones as a category; from ideophone to meaning; from meaning to ideophone; between concepts; and in relation to constructions. We contribute to the field of Chinese linguistics more evidence for not relegating ideophones to the margins of language study. For typological approaches, it is a reminder that Chinese has a sizable inventory of ideophonic items, with a long history. And in relation to iconicity and motivation, it strongly suggests that interactions with writing

are not overlooked. Below we will summarize the argument of this dissertation and discuss its limitations as well as its extendability.



8.1 Summary

Let us reconstruct the argument that has been made in this thesis. The main assumption was that expressions like “the ideophonic lexicon” or “the mimetic lexicon” as opposed to “the prosaic lexicon” are too simple. Just like “the prosaic lexicon” contains a plethora of different clusters and sub-clusters, so too is “the ideophonic lexicon”. It is not monolithic. However, to arrive at a nuanced understanding of the different variability that can exist in such an ideophonic lexicon, a few steps needed to be taken.

The first step that was made was to delineate the scope of ideophones as a category in Chinese. As Chapter 2 showed, the scope is highly correlated with choice of terminology. From a mostly formal and structural point of view, one could be focusing on reduplication phenomena. One can take into account full reduplication and partial reduplication, or the number of syllables the result of the reduplication process has, with a special reference for disyllabic reduplicated words (binomes). In this case, one will treat related meanings as outcomes of morphology (Sun 1999), viz. reduplication and how it leads to vividness, but also to plant and animal names, or mark attenuation (Li & Ponsford 2018). After all, under the traditional understanding of the term ideophone, they are vivid (Doke 1935).

One can also start from semantics, in which case onomatopoeia – the depiction of SOUND – are the default point of departure. Onomatopoeia re-

search has a relatively long research history, but has often focused either on the structures onomatopoetic words can take (cf. Mok 2001), or simply list which sound words are used for which referents (Li 2007), in a fashion that reminds one of the dictionary.

A third perspective then has been to combine these two criteria, and adopt Dingemanse's definition Dingemanse (2012); (2019) for ideophones: *marked words that depict sensory imagery and belong to an open lexical class*. Such a characterization combines structural and functional elements. However, it was only the starting point for the investigation of this dissertation.

As detailed in Chapter 3, the first step of assessing variability and prototypicality phenomena in “the ideophonic lexicon” of Chinese was the construction of the Chinese Ideophone Database (CHIDEOD), whose aim is to go beyond traditional dictionaries. In conjunction with corpora, CHIDEOD provided the usage-based foundations of this dissertation.

Next, the research questions could be tackled. In Chapter 4, we first discussed items that would definitely fall outside the boundary of “the ideophonic lexicon”, on theoretical and typological grounds. Next, the idea of “the ideophonic lexicon” was operationalized by investigating the items in CHIDEOD, and those that were present in the synchronic Academia Sinica Balanced Corpus of Chinese 4.0. The parameters investigated were mostly situated along morpho-phonological markedness (reduplication patterns), orthographic markedness (repetition of semantic radicals) and depiction of meaning. The methodology used to probe these was Multiple Correspondence Analysis. It was found that the items in both case studies are divided

in two clusters that have overlapping boundaries – suggesting that SOUND is quite distinct from NON-SOUND ideophones, but that they have more in common than divides them. But this basic division was also nuanced: some clear preferential correspondences between different factors were discovered.

This fuzzy schism essentially is related to the issue of scope and terminology. There is a certain motivation in choosing to just study onomatopoeia or (non-sound) binomes. However, more can be gained by incorporating them under the same umbrella term, and explain the difference between the two clusters as a difference of type of iconicity relation. SOUND ideophones seem highly motivated through imagic iconicity, so-called “real iconicity” between form and referent. NON-SOUND ideophones, on the other hand, make use of diagrammatic iconicity, which can highlight Gestalts and relations, but are not mutually exclusive. If iconicity is present, one could term it a coerced kind of iconicity (Dingemanse 2011b).

Having observed the heterogeneity of ideophones as a category, we turned to the somewhat small lexical field of LIGHT ideophones in Chapters 5 and 6. In the first of these two chapters, three previous semantic frameworks for ideophones were compared. These were image schemas (Nuckolls 1996; 2017), Ideal Cognitive Models (Lu 2006), and frame semantics adaption (Akita 2012b; 2013b; Kiyama & Akita 2015). Through an analogy with the advance in the field of conceptual metaphors (Kövecses 2017), these three frameworks could be reclassified into a fourfold stratified network, with at the top the most abstract level of image schemas, below

that domains – explained in a manner that allowed for an interpretation of the term as ICM –, frames and on the bottom mental spaces. The latter relates to how each linguistic move is made. The semantic field of LIGHT ideophones provided an opportunity to explore Diachronic Prototype Semantics, because on the lowest level it could be apprehended as a quantified study of mental spaces, thereby completing the four levels of the stratified framework.

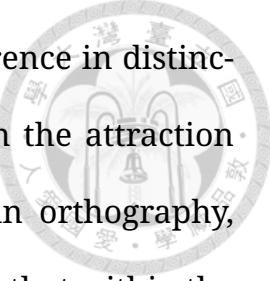
With a sample of LIGHT ideophones chosen on phonological grounds tokens were manually collected from the Scripta Sinica corpus. Having categorized these items, the relations between different collocates of ideophones in a dynamic network were visualized. We saw that meanings were prototypically structured, with some meanings taking up higher frequencies at a given point, and persistence as well as change throughout time. Token frequency effects were found as some meanings, or bundle of meanings, crystallized over time; type frequency effects could be seen in relation to the productivity of semantic clusters. A curious by-product of focusing on VISUAL ideophones, which depend on diagrammatic iconicity rather than imagic iconicity, is that there are orthographic variants that are often interpreted as synonyms. The usage-based approach showed that their behavior can differ quite remarkably, but that there is also conflation at various points, when one collocate hops to another variant, where it may prosper or dwindle.

Abstracting from the sample, we schematized the collocates that occurred more often into salient frames, e.g., LIGHT, SUN, GOLD, LIGHTNING,

STARS, but also SHADES OF RED. These could be further abstracted to domains-ICMs, depending on the nature of the light source. Finally, an image schema sanctioning all these depictions of marked light words was proposed, in which the normal folk idea of seeing is something active, but here the depiction reverses that perspective, in the sense that the speakers want to depict how the quality of the light impressed them.

In terms of statistics this was the least complex chapter, because it merely depended on token and type frequencies. In Chapter 6, we adopted more complex statistics. First, all ideophone types found in CHIDEOD in the Scripta Sinica corpus were gathered, in order to construct the Diachronic Chinese Ideophone Corpus (DIACHIC). Next, semantic vector spaces were calculated for every period in DIACHIC. These were used to study, once again, the heterogeneity within the category of ideophones. That is to say, the effects from the manual study in Chapter 5 were placed in a taxonomy of lexical salience phenomena.

As a case study we limited the study to LIGHT ideophones, in order to first compare whether distributional relational semantics could even come close to the findings from a traditional manual lexical semantic study. First, when studying semasiological salience effects, the results were comparable to those identified before, with the nuance that conceptual distance could now be quantified. Second, onomasiological salience allowed for a relatively easy reversal of the perspectives between meaning and naming: instead of investigating an ideophone and related collocates, we could see which ideophones were closest to a given collocate. Third, structural



salience was demonstrated in two ways: there was a difference in distinctive weight between different frames of collocates, and in the attraction and repulsion of these frames and structural elements in orthography, viz. semantic radicals. These three types of salience showed that within the semantic field of LIGHT, Chinese ideophones are not a homogeneous block, but they have different features and elements standing out, depending on one's perspective. Clearly, such an observation can be extended to other types of ideophones and should inform the generalizations we make about ideophones as a whole.

In Chapter 7 the diachronic studies were left behind as the synchronic grammatical behavior of ideophones was studied. Previous studies had presented a number of generalizations about the tendencies of ideophones in relation to certain constructions. Armed with association measures obtained through collostructional analysis, the heterogeneity of the “ideophonic lexicon” was further probed. It became clear that some items distinctly are attracted or repulsed by certain constructions, and that some items also depend on these constructions to even occur. Another demonstrated point is that some constructions, with ABB in particular because it had attracted a fair amount of research, should be seen as prototypical cases of a wider schematic construction. By first abstracting and then studying the inventory, it became clear that the association measures also showed that attraction and repulsion could be observed between collocate and ideophone.

The relevance of these four methodological chapters is that we now have

demonstrated four corpus based ways of approaching “the ideophonic lexicon”. The notion that ideophones should be conceived as a prototypical cross-linguistic concept has been advanced many times before (Childs 1994; Gabas & van der Auwera 2004; Akita 2009; Dingemanse 2011a; 2017). However, in the case of Chinese, if ideophones are even recognized and studies were based on data beyond the dictionary (Meng 2012) or intuition (Paul 2006), the tendencies were still presented with generalizations like “more likely” or “always”. One of the main outcomes of this dissertation is that while such statements may be true for the majority (or the prototype) of items, it hides the variation and unevenness behind them. And it is exactly this variation that we have attempted to expose: as items within the language particular category of ideophones, within a semantic field in terms of semasiological, onomasiological and structural salience, and interacting with other constructions.

8.2 Limitations and future extendability

As with any serious study, some (non-fuzzy) boundaries needed to be drawn. The first of these is related to the nature of the data and research orientation. We wanted to study ideophones in Chinese, both historically and presently. However, in doing so, we had to select variants. Because of the ubiquity of materials, we chose Standard Chinese (Modern Mandarin) as the variant for the modern language, while of course there are many other variants to choose from, e.g., Southern Min, Hakka, Cantonese, Jin, Gan, Wu, Huizhou, Ping, Xiang, Bai, etc. Taking into account the dialectal

level, one could get an even more fractured, albeit more precise descriptive dataset. Some cross-linguistic studies have been performed on the behavior of (mostly) SOUND ideophones across these main varieties, e.g., Arthur Lewis Thompson (2019a). One of the main reasons for choosing Mandarin was practical in that this dissertation has been completed at National Taiwan University, where a wealth of material is available. An example of this material was also the resources made by Academia Sinica, which provided a solid corpus for studying historical usage.

Since ideophones and related phenomena (cf. Chapter 2) have often been studied for their formal structures, such as Mok's (2001) comprehensive phonological treatment, we wanted to explore more the meaning side of ideophonic items, and in particular NON-SOUND ideophones, such as can be seen in the case studies on LIGHT ideophones. A first obvious extension for future work, then, is to study other groups of ideophones, in order to further lay bare the salient elements. However, future work should not just look at ideophones as they appear in a corpus of texts. After all, many current studies approach them from discourse or experimental settings, arguing that ideophones are mostly spoken and performed (Dingemanse 2019). They are also often accompanied by or supported by gestures (e.g., Haiman 2018 among others). Not surprisingly, the 12th International Symposium on Iconicity in Language and Literature (ILL-12) (2019) had as its theme “iconicity in cognition and across semiotic systems”, where gesture and sign languages played a huge part, e.g., Edwards's (2019) case study of the iconicity in Ghanese Sign Language, and iconic gestures identified in a multimodal

TV corpus (Woodin, Winter & Littlemore 2019). It stands to reason, that this dissertation can also benefit from multi-modal studies in the future. A seminal work for Chinese, based on the Beijing lect can be found in Sam-Sin (2008). My own experience is that there are a number of ideophones used in speech with which ideophones often co-occur, or for which an ideophone conjures up a vivid image in the mind's eye, like *xiū* 嘿 in example (2) in Chapter 1. For others, I do not necessarily get such a vivid feeling, but I am not a native speaker. The point is that this deserves further exploration.

This exploration can happen in four main ways (Tummers, Heylen & Geeraerts 2005). The first is introspection, which will always have a place in linguistics (Geeraerts 2010a), but will need to be supplemented with other more empirical approaches, under the guise of converging evidence, even if what counts as evidence can differ radically among linguists (Penke & Rosenbach 2007). The study of (Chinese) ideophones, in a sense, will always rely on the researcher's gut feeling about the topic.

The second way is surveys, which can be seen as a collective way of gathering introspection data (Tummers, Heylen & Geeraerts 2005:229). In terms of ideophone research, these have included open interviews, e.g., Nuckolls's research on Pastaza Quechua or Childs's work on Bantu ideophones. Norm ratings, such as is currently quite trendy in sensory linguistics (Winter 2019) are another form, and could almost readily be extended to ideophones – provided that there is an updated theory on the categories of senses that ideophones can express and which goes beyond the characterization in Dingemanse (2012). Related to this issue, future work should collect ratings for the

Chinese Ideophone Database, so we can see how “ideophonic” or “iconic” the items in it are. This will allow us to revisit previous studies and lexicographical works and instigate a discussion on which items belong to the core from that perspective. However, we should retain a degree of caution as to the validity of such ratings, as there can be a number of confounding factors, which has been demonstrated with Chinese (Chen et al. 2019) and Japanese (Thompson, Akita & Do). Another set of ratings that should be gathered in the near future pertains to the familiarity of the items in CHIDEOD. If what we want to study is how the current speakers of (Mandarin) Chinese use ideophones, it is of utmost importance that the familiarity of reported items is assessed. This can then be followed up on in the next group.

The third group is experimental settings, which in recent years have been confirming that the features of ideophones in some languages can be picked up by speakers of other unrelated languages. For example, Lockwood, Dingemanse & Hagoort (2016) found in a learning task that Dutch speakers were sensitive to the form-meaning mappings of Japanese ideophones. It would be interesting to perform this kind of study with Chinese ideophones, which are not as frequent as Japanese in daily usage, and that do not display the structural variance in the syllables like e.g., Korean and Japanese ideophones do (Kulemeka 1995). Another avenue for further research is psycholinguistic in nature. Relatively recent work has focused on the processing of Japanese mimetics (Kanero et al. 2014; Lockwood 2017), but also acquisition has been put to the test (Imai et al. 2008; Imai & Kita 2014). We hope that the construction of CHIDEOD can help the design of fu-

ture experimental studies; this is the main reason why we included a number of psycholinguistic measures present in the Chinese Lexical Database (Sun et al. 2018) into the Chinese Ideophone Database. For example, following up on the familiarity ratings raised in the previous group of methods, it will become possible then to perform a learning experiment on the least familiar items, as well as assess the transparency of orthography versus phonology.

The fourth main approach to data gathering concerns the corpus, typically a non-elicited collection of texts that are the product of language usage (rather than the process, or on-line). The current dissertation has followed on this usage-based approach to ask questions of ideophones in Chinese, but that does not mean that this is the end. As lightly touched upon above, multimodal corpora will provide us with a more comprehensive understanding of ideophones and the heterogeneity of the category. Apart from focusing on spoken forms and accompanying gestures, we would like to extend our understanding by including referential information as well, for instance in the form of social media depictions. The foundation of such an approach has already been laid out by Van Hoey & Hsu (2020). It mainly consists of replicating the methodology of a seminal study which compared the CLOTHING related vocabularies of Belgian Dutch and Netherlandic Dutch (Geeraerts, Grondelaers & Bakema 1994). These kinds of corpora allow for a fine-grained studies that can do justice to salience phenomena as they were explored in Chapter 6: it will enable the tracking of statistical patterns in the usage of ideophones in these multimodal settings, but also provide space for

novel and creative extensions of default usages.

To sum up, this dissertation does not contain the final word on Chinese ideophones, but it has shown that the data deserves to be taken seriously, both for typological studies, as well as in Chinese linguistics. We have shown that there is such a thing as the ideophonic lexicon. However, its boundaries are fuzzy and the status of the items in terms of psychological reality are still not fully understood. On the other hand, what is clear, is that the items are structured with different degrees of salience, depending on the methodology used to investigate the ideophonic lexicon. We look forward to extending these ideas in future research.



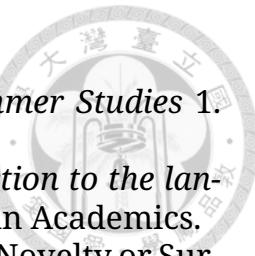


References

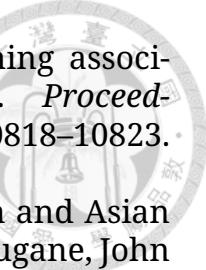


- Abbot-Smith, Kirsten & Michael Tomasello. 2006. Exemplar-learning and schematization in a usage-based account of syntactic acquisition. *The Linguistic Review* 23(3). doi:10.1515/TLR.2006.011.
- Academia Sinica. 1990a. Zhōngyāng yánjiūyuàn Shànggǔ Hànyǔ biāo jì yǔliàokù 中央研究院上古漢語標記語料庫 (Academia Sinica tagged corpus of Old Chinese). <http://lingcorpus.iis.sinica.edu.tw/ancient/>.
- Academia Sinica. 1990b. Zhōngyāng yánjiūyuàn Zhōnggǔ Hànyǔ biāo jì yǔliàokù 中央研究院中古漢語標記語料庫 (Academia Sinica tagged corpus of Middle Chinese). <http://lingcorpus.iis.sinica.edu.tw/middle/>.
- Academia Sinica. 1990c. Zhōngyāng yánjiūyuàn Jìndài Hànyǔ biāo jì yǔliàokù 中央研究院近代漢語標記語料庫 (Academia Sinica tagged corpus of Early Mandarin Chinese). <http://lingcorpus.iis.sinica.edu.tw/early/>.
- Academia Sinica 中央研究院. 2015. Scripta Sinica Database (Hanji quanwen ziliaoku jihua 漢籍全文資料庫). Database. *Scripta Sinica Database*. <http://hanchi.ihp.sinica.edu.tw/> (26 June, 2016).
- Adamska-Sałaciak, Arleta. 2015. Dictionary definitions: Problems and solutions. *Studia Linguistica Universitatis Iagellonicae Cracoviensis* 129(4). [unknown2].
- Ahlner, Felix & Jordan Zlatev. 2010. Cross-modal iconicity: A cognitive semiotic approach to sound symbolism. *Sign Systems Studies* 38(1/4). 298. doi:10.12697/SSS.2010.38.1-4.11.
- 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. 2012a. Multimedia Encyclopedia of Japanese Mimetics (ME-JaM). Database. <https://sites.google.com/site/jpmimeticthesaurus/>.
- Akita, Kimi. 2012b. Toward a frame-semantic definition of sound-symbolic words: A collocational analysis of Japanese mimetics. *Cognitive Linguistics* 23(1). 67–90.
- Akita, Kimi. 2013a. Constraints on the semantic extension of onomatopoeia. *Public Journal of Semiotics* 13(1). 21–37.
- Akita, Kimi. 2013b. The lexical iconicity hierarchy and its grammatical correlates. In Lars Elleström, Olga Fischer & Christina Ljungberg (eds.), *Iconic investigations*, 331–349. (Iconicity in Language and Literature volume 12). Amsterdam ; Philadelphia: John Benjamins Publishing Company.
- Akita, Kimi. 2015. Sound symbolism. In Jan-Ola Östman & Jef Verschueren (eds.), *Handbook of Pragmatics*, 1–24. Amsterdam: John Benjamins. 10.1075/%20hop.19.sou1.
- Akita, Kimi. 2016. A multimedia encyclopedia of Japanese mimetics: A frame-semantic approach to L2 sound-symbolic words. In Kaori Kabata & Kiyoko Toratani (eds.), *Cognitive-functional approaches to the study of Japanese as a second language*, 139–167. (Studies on Lan-

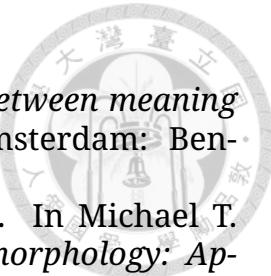
- guage Acquisition volume 46). Boston: De Gruyter Mouton.
- Akita, Kimi. 2017a. Grammatical and functional properties of mimetics in Japanese. In Noriko Iwasaki, Peter Sells & Kimi Akita (eds.), *The grammar of Japanese mimetics: Perspectives from structure, acquisition and translation*, 20–34. (Routledge Studies in East Asian Linguistics). Milton Park, Abingdon, Oxon ; New York, NY: Routledge.
- Akita, Kimi. 2017b. Decomposing the lexical iconicity hierarchy for ideophones. Paper presented at the Akita, Kimi. 2017. Decomposing the lexical iconicity hierarchy for ideophones. In CLS-MPI Iconicity Focus Group Workshop: Types of Iconicity in Language Use, Development and Processing. Nijmegen: Max Planck Institute for Psycholinguistics., Nijmegen: Max Planck Institute for Psycholinguistics.
- Akita, Kimi & Mark Dingemanse. 2019. Ideophones (Mimetics, Expressives). In Mark Aronoff (ed.), *Oxford research encyclopedia of linguistics*. Oxford: Oxford University Press.
- Akita, Kimi & Takeshi Usuki. 2016. A constructional account of the “optional” quotative marking on Japanese mimetics. *Journal of Linguistics* 52(02). 245–275. doi:10.1017/S002226715000171.
- Akita, Kimi, Jingyi Zhang & Katsuo Tamaoka. 2020. Systematic side of sound symbolism: The case of suffixed ideophones in Japanese. In, *KLS selected papers*. Nagoya: Kansai Linguistic Society.
- Allan, Kathryn. 2012. Using OED data as evidence for researching semantic change. In Kathryn Allan & Justyna A. Robinson (eds.), *Current methods in historical semantics*, 17–39. (Topics in English Linguistics 73). Berlin: De Gruyter Mouton.
- Ameka, Felix K. 1992. Interjections: The universal yet neglected part of speech. *Journal of Pragmatics* 18(2-3). 101–118.
- Appelhans, Tim, Florian Detsch, Christoph Reudenbach & Stefan Woellauer. 2019. *Mapview: Interactive viewing of spatial data in r*. <https://CRAN.R-project.org/package=mapview>.
- Arcodia, Giorgio Francesco. 2007. Chinese: A language of compound words? In Fabio Montermini, Gilles Boyé & Nabil Hathout (eds.), *Selected proceedings of the 5th Décembrettes: Morphology in Toulouse*, 79–90. Somerville, MA: Cascadilla Proceedings Project.
- Arcodia, Giorgio Francesco & Bianca Basciano. 2018. The Construction Morphology analysis of Chinese word formation. In Geert Booij (ed.), *The construction of words: Advances in Construction Morphology*, 219–253. New York, NY: Springer Berlin Heidelberg.
- Armoskaite, Solveiga & Päivi Koskinen. 2017. Structuring sensory imagery: Ideophones across languages and cultures. *Canadian Journal of Linguistics/Revue canadienne de linguistique*. 1–5. doi:10.1017/cnj.2017.12.
- Aston, W. G. 1894. Japanese onomatopes and the origin of language. *The Journal of the Anthropological Institute of Great Britain and Ireland* 23. 332–362. doi:10.2307/2842085.
- Bache, Stefan Milton & Hadley Wickham. 2014. *Magrittr: A forward-pipe operator for r*. <https://CRAN.R-project.org/package=magrittr>.
- Baldinger, Kurt. 1980. *Semantic theory: Towards a modern semantics*. Oxford: Clarendon Press.



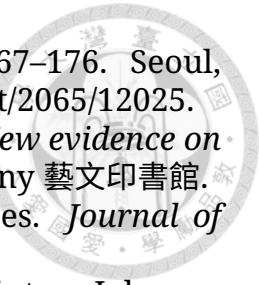
- ford: Blackwell.
- Banker, Elizabeth M. 1964. Bahnar reduplication. *Mon-Khmer Studies* 1. 119–134.
- Barnes, Archie B. 2007. *Chinese through poetry: An introduction to the language and imagery of traditional verse*. Durham: Alcuin Academics.
- Barto, Andrew, Marco Mirolli & Gianluca Baldassarre. 2013. Novelty or Surprise? *Frontiers in Psychology* 4(907). doi:10.3389/fpsyg.2013.00907.
- Bates, Elizabeth, Sylvia Chen, Ping Li & Ovid Tzeng. 1993. Where is the boundary between compounds and phrases in Chinese? A reply to Zhou et al. *Brain and Language* 45. 94–107.
- Baxter, William H. & Laurent Sagart. 2017. Old Chinese reconstruction: A response to Schuessler. *Diachronica* 34(4). 559–576. doi:10.1075/dia.17003.sag.
- Baxter, William Hubbard. 1992. *A handbook of Old Chinese phonology*. (Trends in Linguistics 64). Berlin ; New York: Mouton de Gruyter.
- Baxter, William Hubbard & Laurent Sagart. 1998. Word formation in Old Chinese. In Jerome Lee Packard (ed.), *New approaches to Chinese word formation: Morphology, phonology and the lexicon in modern and ancient Chinese*, 35–76. Berlin ; New York: Mouton de Gruyter.
- Baxter, William Hubbard & Laurent Sagart. 2014. *Old Chinese: A new reconstruction*. Oxford ; New York: Oxford University Press.
- Baxter, William Hubbard & Laurent Sagart. 2015. Baxter-Sagart Old Chinese reconstruction (v. 13 october 2015). <http://ocbaxtersagart.lsait.lsa.umich.edu/BaxterSagartOC2015-10-13.xlsx>.
- Benzécri, Jean-Paul. 1973. *L'analyse des données*, 2. *L'analyse des correspondances*. Paris: Dunod.
- Benzécri, Jean-Paul. 1984. *Analyse des correspondances, exposé élémentaire*. 2nd edn. Paris: Dunod.
- Berlin, Brent. 1976. The concept of rank in ethnobiological classification: Some evidence from Aguarana folk botany. *American Ethnologist* 3. 381–400.
- Berlin, Brent. 1978. Ethnobiological classification. In Eleanor Rosch & Barbara B. Lloyd (eds.), *Cognition and categorization*, 1978. Hillsdale: NJ: Erlbaum.
- Berlin, Brent, Dennis E. Breedlove & Peter H. Raven. 1973. General principles of classification and nomenclature in folk biology. *American Anthropologist* 75. 214–242.
- Berlin, Brent, Dennis E. Breedlove & Peter H. Raven. 1974. *Principles of Tzeltal plant classification: An introduction to the botanical ethnography of a Mayan-speaking people of Highland Chiapas*. New York: Academic Press.
- Bickel, Balthasar, Bernard Comrie & Martin Haspelmath. 2008. The Leipzig glossing rules: Conventions for interlinear morpheme by morpheme glosses (Revised version of February 2008). <https://www.eva.mpg.de/lingua/pdf/LGR08.02.05.pdf> (28 May, 2015).
- Bisang, Walter. 2008. Precategoriality and syntax-based parts of speech: The case of Late Archaic Chinese. *Studies in Language* 32(3). 568–589.
- Blasi, Damián E., Søren Wichmann, Harald Hammarström, Peter F.



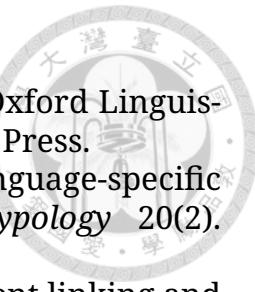
- Stadler & Morten H. Christiansen. 2016. Sound-meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences* 113(39). 10818–10823. doi:10.1073/pnas.1605782113.
- Bodomo, Adams. 2006. The structure of ideophones in African and Asian languages: The case of Dagaare and Cantonese. In John Mugane, John P. Hutchison & Dee A. Worman (eds.), *Selected Proceedings of the 35th Annual Conference on African Linguistics*, 203–213. Somerville, MA: Cascadilla Proceedings Project.
- Boleda, Gemma. 2020. Distributional semantics and linguistic theory. *Annual Review of Linguistics* 6(1). 213–234. doi:10.1146/annurev-linguistics-011619-030303.
- Bolinger, Dwight L. 1985. The inherent iconism of intonation. In John Haiman (ed.), *Iconicity in syntax: Proceedings of a Symposium on Iconicity in Syntax, Stanford, June 24 - [2]6, 1983*, 97–108. (Typological Studies in Language 6). Amsterdam: Benjamins.
- Booij, Geert. 2005. Compounding and derivation: Evidence for Construction Morphology. In Wolfgang U. Dressler, Dieter Kastovsky, Oskar E. Pfeiffer & Franz Rainer (eds.), *Morphology and its demarcations: Selected papers from the 11th Morphology Meeting, Vienna, February 2004*, 109–132. (Amsterdam Studies in the Theory and History of Linguistic Science v. 264). Amsterdam ; Philadelphia: John Benjamins Pub. Co.
- Booij, Geert. 2007. Construction Morphology and the lexicon. In Fabio Montemini, Gilles Boyé & Nabil Hathout (eds.), *Selected proceedings of the 5th Décembrettes: Morphology in Toulouse*, 34–44. Somerville, MA: Cascadilla Proceedings Project.
- Booij, Geert (ed.). 2018. *The construction of words: Advances in Construction Morphology*. New York, NY: Springer Berlin Heidelberg.
- Booij, Geert. 2010. Construction Morphology: Construction Morphology. *Language and Linguistics Compass* 4(7). 543–555. doi:10.1111/j.1749-818X.2010.00213.x.
- Booij, Geert & Jenny Audring. 2017. Construction Morphology and the Parallel Architecture of Grammar. *Cognitive Science* 41. 277–302. doi:10.1111/cogs.12323.
- Boussidan, Armelle, Eyal Sagi & Sabine Ploux. 2009. Phonaesthemic and etymological effects on the distribution of senses in statistical models of semantics. In Yves Peirsman, Yannick Versley & Tim Van de Cruys (eds.), *DiSCo 2009 [Distributional Semantics beyond Concrete Concepts] from CogSci 2009 Workshop*, 35–40.
- Brown, Dunstan, Marina Chumakina & Greville G. Corbett (eds.). 2013. *Canonical morphology and syntax*. First edition. Oxford: Oxford University Press.
- Burenhult, Niclas & Asifa Majid. 2011. Olfaction in Aslian Ideology and Language. *The Senses and Society* 6(1). 19–29.
- Bühler, Karl. 1934. *Sprachtheorie: Die Darstellungsfunktion der Sprache*. Jena: G. Fischer.
- Bybee, Joan. 2010. *Language, usage and cognition*. Cambridge ; New York:



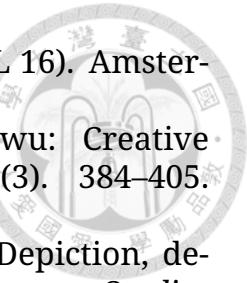
- Cambridge University Press.
- Bybee, Joan L. 1985. *Morphology: A study of the relation between meaning and form*. (Typological Studies in Language 9). Amsterdam: Benjamins.
- Bybee, Joan L. 1988. Morphology as lexical organization. In Michael T. Hammond & Michael P. Noonan (eds.), *Theoretical morphology: Approaches in modern linguistics*, 119–141. San Diego, CA: Academic Press.
- Bybee, Joan L. 2001. *Phonology and language use*. (Cambridge Studies in Linguistics). Cambridge, [England] ; New York: Cambridge University Press.
- Bybee, Joan L. & Paul J. Hopper. 2001a. Introduction to frequency and the emergence of linguistic structure. In Joan L. Bybee & Paul J. Hopper (eds.), *Frequency and the emergence of linguistic structure*, 1–24. (Typological Studies in Language 45). Amsterdam: Benjamins.
- Bybee, Joan L. & Paul J. Hopper (eds.). 2001b. *Frequency and the emergence of linguistic structure*. (Typological Studies in Language 45). Amsterdam: Benjamins.
- Carr, Denzel. 1966. Homorganicity in Malay/Indonesian in expressives and quasi expressives. *Language* 42(2). 370–377. doi:10.2307/411697.
- Cavnar, William B. & John M. Trenkle. 1994. N-gram-based-text categorization. *Ann Arbor MI* 48113(2). 161–175.
- Cáo, Ruifāng 曹瑞芳. 1995. Pǔtōnghuà ABB shì xíngróngcí de dìngliàng fēnxī 普通话 ABB 式形容词的定量分析 [quantitative analysis of Mandarin Chinese ABB adjectives]. *Yǔwén yánjiū* 语文学研究 3. 22–25.
- Chan, Marjorie K. M. 1996. Some thoughts on the typology of sound symbolism and the Chinese language. In Chin-chuan Cheng, Jerome Lee Packard, James Yoon & Yu-ling You (eds.), *Proceedings of the Eighth North American Conference on Chinese Linguistics*. Vol. 2., 1–15. Los Angeles: GSIL Publications.
- Chang, Yufen. 2009. The phonology of ABB reduplication in Taiwanese. In Yun Xiao (ed.), *Proceedings of the 21st North American Conference on Chinese Linguistics (NACCL-21)*: Vol. 1, 28–41. Smithfield, Rhode Island: Bryant University.
- Chao, Yuen Ren. 1968. *A grammar of spoken Chinese*. Berkely and Los Angeles: Univ. of California Press.
- Chen, Ching-Yu, Shu-Fen Tseng, Chu-Ren Huang & Keh-Jiann Chen. 1993. Some distributional properties of Mandarin Chinese – A study based on the Academia Sinica Corpus. In Chu-Ren Huang & Claire Hsun-Hui Chang (eds.), *Proceedings of Pacific Asia Conference on Formal and Computational Linguistics [PACFoCoL] 1*. Taipei: Computational Linguistics Society of R.O.C.
- Chen, I-Hsuan, Qingqing Zhao, Yunfei Long, Qin Lu & Chu-Ren Huang. 2019. Mandarin Chinese modality exclusivity norms. (Ed.) Zhiqiang Cai. *PLOS ONE* 14(2). e0211336. doi:10.1371/journal.pone.0211336.
- Chen, Keh-Jiann, Chu-Ren Huang, Li-Ping Chang & Hui-Li Hsu. 1996. SINICA CORPUS : Design methodology for balanced corpora. In Byung-Soo Park & Jong-Bok Kim (eds.), *Proceedings of the 11th pacific asia con-*



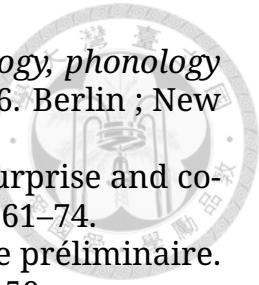
- ference on language, information and computation, 167–176. Seoul, Korea: Kyung Hee University. doi:<http://hdl.handle.net/2065/12025>.
- Chi, Hsiu-Sheng 季旭昇. 2014. *Shuōwén xīnzhèng* 說文新證 [New evidence on the Shuōwén jiězì]. Taipei: Yee Wen Publishing Company 藝文印書館.
- Childs, G. Tucker. 1988. The phonology of Kisi ideophones. *Journal of African Languages and Linguistics* 10(2). 165–190.
- Childs, G. Tucker. 1994. African ideophones. In Leanne Hinton, Johanna Nichols & John J. Ohala (eds.), *Sound symbolism*, 178–204. Cambridge [England]: Cambridge University Press.
- Childs, G. Tucker. 1996. Where have all the ideophones gone? The death of a word category in Zulu. *Toronto Working Papers in Linguistics* 15. 81–103.
- Chou, Ya-Min 周亞民. 2005. Hantology: The knowledge structure of Chinese writing system and its applications (Hànzì zhīshì běntǐ: Yǐ zì wéi běn de zhīshì jiàgòu yǔ qí yìngyòng shìlì) 漢字知識本體——以字為本的知識架構與其應用示例. Taipei: National Taiwan University PhD dissertation.
- Church, Kenneth W. & Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16(1). 22–29.
- CKIP group. 2019. *Ckiptagger 0.1.0*. <https://pypi.org/project/ckiptagger/>.
- CKIP group & Academia Sinica. 2013. Academia Sinica Balanced Corpus of Modern Chinese (ASBC 4.0) (Zhōngyāng yánjiūyuàn Xiàndài Hànyǔ biāojì yúliàokù 中央研究院現代漢語標記語料庫). <https://ckip.iis.sinica.edu.tw/CKIP/engversion/20corpus.htm>.
- Clark, Herbert H. 1997. Dogmas of understanding. *Discourse Processes* 23(3). Taylor & Francis. 567–598.
- Clark, Herbert H. 2016. Depicting as a method of communication. *Psychological Review* 123(3). 324–347. doi:10.1037/rev0000026.
- Connell, Louise, Dermot Lynott & Briony Banks. 2018. Interoception: The forgotten modality in perceptual grounding of abstract and concrete concepts. *Philosophical Transactions of the Royal Society B: Biological Sciences* 373(1752). 20170143. doi:10.1098/rstb.2017.0143.
- Corbett, Greville G. 2007. Canonical Typology, suppletion, and possible words. *Language* 83(1). 8–42. doi:10.1353/lan.2007.0006.
- Corbett, Greville G. 2008. Determining morphosyntactic feature values: The case of case. In Greville G. Corbett & Michael P. Noonan (eds.), *Case and grammatical relations: Papers in honor of Bernard Comrie*, 1–34. Amsterdam: Benjamins.
- Corbett, Greville G. 2011. Implicational hierarchies. In Jae Jung Song (ed.), *The Oxford Handbook of Language Typology*. Oxford: Oxford University Press.
- Corbett, Greville G. 2015. Morphosyntactic complexity: A typology of lexical splits. *Language* 91(1). 145–193. doi:10.1353/lan.2015.0003.
- Croft, William. 1991. *Syntactic categories and grammatical relations: The cognitive organization of information*. Chicago: University of Chicago Press.
- Croft, William. 2001. *Radical construction grammar: Syntactic theory in typology*.



- pological perspective. Oxford: Oxford Univ. Press.
- Croft, William. 2012. *Verbs: Aspect and causal structure*. (Oxford Linguistics). Oxford [England] ; New York: Oxford University Press.
- Croft, William. 2016. Comparative concepts and language-specific categories: Theory and practice. *Linguistic Typology* 20(2). doi:10.1515/lingty-2016-0012.
- Croft, William, Chiaki Taoka & Esther J. Wood. 2001. Argument linking and the commercial transaction frame in English, Russian and Japanese. *Language Sciences* 23. 579–602.
- Dai, John Xiang-Ling. 1998. Word formation in Old Chinese. In Jerome Lee Packard (ed.), *New approaches to Chinese word formation: Morphology, phonology and the lexicon in modern and ancient Chinese*, 103–134. Berlin ; New York: Mouton de Gruyter.
- Deerwester, Scott C., Susan T. Dumais, Thomas K. Landauer & Richard A. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41(6). 391–407.
- De Pascale, Stefano. 2019. Token-based vector space models as semantic control in lexical lexicometry. Leuven: KU Leuven PhD dissertation.
- Deshors, Sandra C. 2017. Zooming in on verbs in the progressive: A Collostructional and Correspondence Analysis approach. *Journal of English Linguistics* 45(3). 260–290. doi:10.1177/0075424217717589.
- Dewell, Robert B. 1994. Over again: Image-schema transformations in semantic analysis. *Cognitive Linguistics* 5(4). 351–380.
- Diatka, Vojtěch. 2014. Hindi ideophones. Prague: Karlova Univerzita Master thesis.
- Diffloth, Gérard. 1972. Notes on expressive meaning. *Chicago Linguistic Society* 8. 440–447.
- Diffloth, Gérard. 1976. Expressives in Semai. In Philip N. Jenner, Laurence C. Thompson & Stanley Starosta (eds.), *Austroasiatic studies*, vol. 1, 249–264. (Oceanic Linguistics Special Publications 13). Honolulu: Univ. of Hawaii Press.
- Diffloth, Gérard. 1979. Expressive phonology and prosaic phonology in Mon-Khmer. In Theraphan L. Thongkum (ed.), *Studies in Mon-Khmer and Thai phonology and phonetics in honor of E. Henderson*, 49–59. Bangkok: Chulalongkorn University Press.
- Dingemanse, Mark. 2011a. The meaning and use of ideophones in Siwu. Nijmegen: Radboud University Nijmegen dissertation.
- Dingemanse, Mark. 2011b. Ezra Pound among the Mawu: Ideophones and iconicity in Siwu. In Pascal Michelucci, Olga Fischer & Christina Ljungberg (eds.), *Semblance and signification*, 39–54. (Iconicity in Language and Literature 10). Amsterdam ; Philadelphia: John Benjamins Pub. Co.
- Dingemanse, Mark. 2012. Advances in the cross-linguistic study of ideophones. *Language and Linguistics Compass* 6(10). 654–672.
- Dingemanse, Mark. 2013. Ideophones and gesture in everyday speech. *Gesture* 13(2). 143–165. doi:10.1075/gest.13.2.02din.
- Dingemanse, Mark. 2019. 'Ideophone' as a comparative concept. In Kimi Akita & Prashant Pardeshi (eds.), *Ideophones, mimetics and expressiveness*.



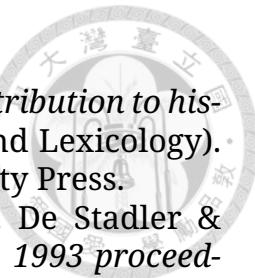
- sives, 13–33. (*Iconicity in Language and Literature*, ILL 16). Amsterdam : Philadelphia: John Benjamins Pub. Co.
- Dingemanse, Mark. 2014. Making new ideophones in Siwu: Creative depiction in conversation. *Pragmatics and Society* 5(3). 384–405. doi:10.1075/ps.5.3.04din.
- Dingemanse, Mark. 2015. Ideophones and reduplication: Depiction, description, and the interpretation of repeated talk in discourse. *Studies in Language* 39(4). 946–970. doi:10.1075/sl.39.4.05din.
- Dingemanse, Mark. 2016. Ideophones and sensory language in social interaction. In NINJAL (ed.), *Mimetics in Japanese and other languages in the world* (日本語と世界諸言語のオノマトペ). Tachikawa: NINJAL.
- Dingemanse, Mark. 2017. Expressiveness and system integration: On the typology of ideophones, with special reference to Siwu. *STUF - Language Typology and Universals* 70(2). 363–384. doi:10.1515/stuf-2017-0018.
- Dingemanse, Mark. 2018. Redrawing the margins of language: Lessons from research on ideophones. *Glossa: a journal of general linguistics* 3(1). 1–30. doi:10.5334/gjgl.444.
- Dingemanse, Mark & Kimi Akita. 2016. An inverse relation between expressiveness and grammatical integration: On the morphosyntactic typology of ideophones, with special reference to Japanese. *Journal of Linguistics*. 1–32. doi:10.1017/S002222671600030X.
- Dingemanse, Mark, Damián E. Blasi, Gary Lupyan, Morten H. Christiansen & Padraic Monaghan. 2015. Arbitrariness, iconicity and systematicity in language. *Trends in Cognitive Sciences* 19(10). 603–615.
- Dingemanse, Mark & Asifa Majid. 2012. The semantic structure of sensory vocabulary in an African language. In Naomi Miyake, David Peebles & Richard P Cooper (eds.), *Proceedings of the 34th Annual meeting of the Cognitive Science Society 2012 (CogSci 2012): Building bridges across cognitive sciences around the world : Sapporo, Japan, 1-4 August 2012*, 300–305. Red Hook, NY: Curran Associates, Inc.
- Dingemanse, Mark, Marcus Perlman & Pamela Perniss. 2020. Construals of iconicity: Experimental approaches to form-meaning resemblances in language. *Language and Cognition* 12. Cambridge University Press. 1–14. doi:10.1017/langcog.2019.48.
- Dingemanse, Mark, Will Schuerman, Eva Reinisch, Sylvia Tufvesson & Holger Mitterer. 2016. What sound symbolism can and cannot do: Testing the iconicity of ideophones from five languages. *Language* 92(2). e117–e133. doi:10.1353/lan.2016.0034.
- Dingemanse, Mark & Bill Thompson. 2020. Playful iconicity: Structural markedness underlies the relation between funniness and iconicity. *Language and Cognition*. 1–22. doi:10.1017/langcog.2019.49.
- Dixon, R. M. W. & Alexandra Y. Aikhenvald (eds.). 2004. *Adjective classes: A cross-linguistic typology*. (Explorations in Linguistic Typology 1). Oxford ; New York: Oxford University Press.
- Doke, Clement M. 1935. *Bantu linguistic terminology*. London ; New York: Longmans, Green.
- Duanmu, San. 1998. Wordhood in Chinese. In Jerome Lee Packard (ed.),



- New approaches to Chinese word formation: Morphology, phonology and the lexicon in modern and ancient Chinese, 135–196. Berlin ; New York: Mouton de Gruyter.
- Dunning, Ted. 1993. Accurate methods for the statistics of surprise and co-incidence. *Computational linguistics* 19(1). MIT Press. 61–74.
- Durand, Maurice. 1961. Les impressifs en vietnamien: Étude préliminaire. *Bulletin de la Société des Études Indochinoises* 36(1). 5–50.
- Edward, Mary. 2019. Patterned iconicity for different semantic categories: Evidence from sign languages and gestures. Oral presentation. Paper presented at the ILL 12 [International Symposium on Iconicity in Language and Literature], Lund: Lund University. <https://konferens.ht.lu.se/en/ill-12/>.
- Eifring, Halvor. 2019. Language contact across time: Classical Chinese on modern public signs in Taiwan. *Journal of Chinese Linguistics* 47(2). 562–614. doi:10.1353/jcl.2019.0023.
- Ellis, Nick C. 2006. Language Acquisition as Rational Contingency Learning. *Applied Linguistics* 27(1). 1–24. doi:10.1093/applin/ami038.
- Ellis, Nick C. & Fernando Ferreira-Junior. 2009. Constructions and their acquisition: Islands and the distinctiveness of their occupancy. *Annual Review of Cognitive Linguistics* 7. 188–221. doi:10.1075/arcl.7.08ell.
- Emeneau, M. B. 1969. Onomatopoetics in the Indian linguistic area. *Language* 45(2). 274–299. doi:10.2307/411660.
- Fauconnier, Gilles. 1994. *Mental spaces: Aspects of meaning construction in natural language*. Cambridge ; New York, NY, USA: Cambridge University Press.
- Fauconnier, Gilles & Eve Sweetser (eds.). 1996. *Spaces, worlds, and grammar*. (Cognitive Theory of Language and Culture). Chicago: University of Chicago Press.
- Fellbaum, Christiane (ed.). 1998. *WordNet: An electronic lexical database*. (Language, Speech, and Communication). Cambridge, Mass: MIT Press.
- Feng, Shengli. 1998. Prosodic structures and compound words in Classical Chinese. In Jerome Lee Packard (ed.), *New approaches to Chinese word formation: Morphology, phonology and the lexicon in modern and ancient Chinese*, 197–260. (Trends in Linguistics: Studies and Monographs 105). Berlin ; New York: Mouton de Gruyter.
- Ferrara, Lindsay & Gabrielle Hodge. 2018. Language as description, indication, and depiction. *Frontiers in Psychology* 9. 716. doi:10.3389/fpsyg.2018.00716.
- Féng, Jìng 冯静. 2016. *Shījīng zhōng de nǐhuìcí yánjiū* 《诗经》中的拟绘词研究 [Study of ideophones in the *Book of Odes*]. Harbin: Heilongjiang University Master thesis.
- Féng, Shènglì 冯胜利. 2010. Lún yùtǐ de jīzhì jí qí yǔfǎ zhǔxìng 论语体的机制及其语法属性 (The mechanism of register and its grammatical properties). *Zhōngguó yǔwén* 中国语文 5. 400–412.
- Fillmore, Charles J. 1977. Toppics in lexical semantics. In Roger Cole (ed.), *Current issues in linguistic theory*, 76–138. Bloomington: Indiana University Press.



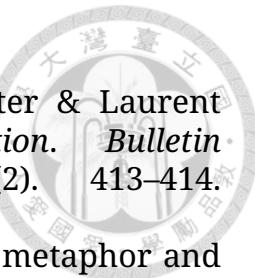
- Fillmore, Charles J. 2003. FrameNet and Frame Semantics. *International Journal of Lexicography* 16. 231–366.
- Fillmore, Charles J., Paul Kay & Mary Catherine O'Connor. 1988. Regularity and idiomticity in grammatical constructions: The case of Let Alone. *Language* 64(3). 501–538.
- Firke, Sam. 2019. *Janitor: Simple tools for examining and cleaning dirty data*. <https://CRAN.R-project.org/package=janitor>.
- Firth, John R. 1957. A synopsis of linguistic theory. In John R. Firth (ed.), *Studies in linguistic analysis*, 1–32. Oxford: Philological Society.
- Fischer-Jørgensen, Eli. 1978. On the universal character of phonetic symbolism with special reference to vowels. *Studia Linguistica* 32(1-2). 80–90. doi:10.1111/j.1467-9582.1978.tb00329.x.
- Flaksman, Maria. 2017. Iconic treadmill hypothesis: The reasons behind continuous onomatopoeic coinage. In Angelika Zirker, Matthias Bauer, Olga Fischer & Christina Ljungberg (eds.), *Iconicity in Language and Literature*, vol. 15. Amsterdam: John Benjamins Publishing Company. doi:10.1075/ill.15.02fla.
- Foolen, Ad. 1997. The expressive function of language: Towards a cognitive semantic approach. In Susanne Niemeier & René Dirven (eds.), *The Language of emotions: Conceptualization, expression, and theoretical foundation*. Amsterdam: John Benjamins.
- Fordyce, James Forrest. 1988. Studies in sound symbolism with special reference to English. Los Angeles: University of California, Los Angeles PhD dissertation.
- Forkel, Robert, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping & Russell D. Gray. 2018. Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data* 5. 180205. doi:10.1038/sdata.2018.205.
- Forker, Diana. 2016. Conceptualization in current approaches of language typology. *Acta Linguistica Hafniensia* 48(1). 70–84. doi:10.1080/03740463.2016.1176372.
- Fortune, G. 1962. *Ideophones in Shona: An inaugural lecture given in the University College of Rhodesia and Nyasaland on 28 April 1961*. London ; New York, N.Y: Oxford University Press.
- Francis, W. Nelson & Henry Kučera. 1964. *Manual of information to accompany a standard corpus of present-day edited American English for use with digital computers. Report to the U. S. Office of Education on Cooperative Research Project E-007*. Providence, RI: Brown University.
- Frawley, William. 1981. In defense of the dictionary: A response to Haiman. *Lingua* 55. 53–61.
- Fudge, Erik. 1970. Phonological structure and 'expressiveness'. *Journal of Linguistics* 6(2). 161–188.
- Gabas, Nilson Jr. & Johan van der Auwera. 2004. Ideophones in Karo. In Michel Achard & Suzanne Kemmer (eds.), *Language, culture and mind*, 397–414. Stanford, Calif: CSLI Publications.
- Geeraerts, Dirk. 1983. Prototype theory and diachronic semantics: A case



- study. *Indogermanische Forschungen* 88. 1–32.
- 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.
- Geeraerts, Dirk. 1999. Beer and semantics. In Leon G. De Stadler & Christoph Eyrich (eds.), *Issues in cognitive linguistics: 1993 proceedings of the International Cognitive Linguistics Conference*, 35–55. (Cognitive Linguistics Research 12). Berlin ; New York: Mouton de Gruyter.
- Geeraerts, Dirk. 2000. Salience phenomena in the lexicon: A typology. In Liliana Albertazzi (ed.), *Meaning and cognition: A multidisciplinary approach*, 125–136. (Converging Evidence in Language and Communication Research 2). Amsterdam: Benjamins.
- Geeraerts, Dirk. 2001. The definitional practice of dictionaries and the Cognitive Semantic conception of polysemy. *Lexicographica* 17. 6–21.
- Geeraerts, Dirk. 2003. Meaning and definition. In P. G. J. van Sterkenburg (ed.), *A practical guide to lexicography: Edited by Piet van Sterkenburg*, 83–93. (Terminology and Lexicography Research and Practice v. 6). Amsterdam ; Philadelphia: John Benjamins Pub.
- Geeraerts, Dirk. 2006a. Salience phenomena in the lexicon. A typology. In Dirk Geeraerts (ed.), *Words and other wonders: Papers on lexical and semantic topics*, 74–96. (Cognitive Linguistics Research 33). Berlin ; New York: Mouton de Gruyter.
- Geeraerts, Dirk. 2006b. Beer and Semantics. In Dirk Geeraerts (ed.), *Words and other wonders: Papers on lexical and semantic topics*, 252–271. (Cognitive Linguistics Research 33). Berlin ; New York: Mouton de Gruyter.
- Geeraerts, Dirk. 2006c. Vagueness's puzzles, polysemy's vagaries. In Dirk Geeraerts (ed.), *Words and other wonders: Papers on lexical and semantic topics*, 99–148. (Cognitive Linguistics Research 33). Berlin ; New York: Mouton de Gruyter.
- Geeraerts, Dirk. 2006d. *Words and other wonders: Papers on lexical and semantic topics*. (Cognitive Linguistics Research 33). Berlin: Mouton de Gruyter.
- Geeraerts, Dirk. 2010a. The doctor and the semantician. In Dylan Glynn & Kerstin Fischer (eds.), *Quantitative methods in cognitive semantics: Corpus-driven approaches*, 63–78. (Cognitive Linguistics Research 46). Berlin ; New York: De Gruyter Mouton.
- Geeraerts, Dirk. 2010b. *Theories of Lexical Semantics*. Oxford: Oxford Univ. Press.
- Geeraerts, Dirk. 2017. Entrenchment as onomasiological salience. In Hans-Jörg Schmid (ed.), *Entrenchment and the psychology of language learning: How we reorganize and adapt linguistic knowledge*, 153–174. (Language and the Human Lifespan). Washington, DC : Berlin: American Psychological Association ; De Gruyter Mouton.
- Geeraerts, Dirk, Stefan Grondelaers & Peter Bakema. 1994. *The structure of lexical variation: Meaning, naming, and context*. (Cognitive Linguistics Research 5). Berlin: Mouton de Gruyter.

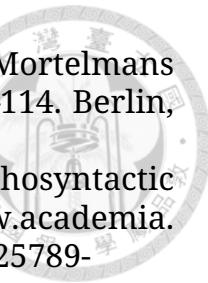


- Giora, Rachel. 2003. *On our mind: Salience, context, and figurative language*. Oxford: Oxford University Press.
- Gipper, Helmut. 1959. Sessel oder Stuhl? Ein Beitrag zur Bestimmung von Wortinhalten im Bereich der Sachkultur. In Helmut Gipper (ed.), *Sprache, Schlüssel zur Welt: Festschrift für Leo Weisgerber*, 271–292. Düsseldorf: Schwann.
- Glur, Christoph. 2019. *Data.Tree: General purpose hierarchical data structure*. <https://CRAN.R-project.org/package=data.tree>.
- 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*, 307–341. (Human Cognitive Processing volume 43). Amsterdam ; Philadelphia: John Benjamins Publishing Company.
- Glynn, Dylan. 2015. The conceptual profile of the lexeme home: A multifactorial diachronic analysis. In Javier E. Díaz Vera (ed.), *Metaphor and metonymy across time and cultures: Perspectives on the sociohistorical linguistics of figurative language*, 265–293. (Cognitive Linguistics Research volume 52). Berlin ; Boston: De Gruyter Mouton.
- Glynn, Dylan. 2016. Quantifying polysemy: Corpus methodology for prototype theory. *Folia Linguistica* 50(2). doi:10.1515/flin-2016-0016.
- Goddard, Cliff. 2018. *Ten lectures on Natural Semantic Metalanguage: Exploring language, thought and culture using simple, translatable words*. (Distinguished Lectures in Cognitive Linguistics 21). Leiden ; Boston: Brill.
- Goddard, Cliff & Anna Wierzbicka. 1994. *Semantic and lexical universals: Theory and empirical findings..* Vol. 25. Amsterdam: John Benjamins Publishing.
- Goddard, Cliff & Anna Wierzbicka. 2002. *Meaning and universal grammar: Theory and empirical findings..* Vol. 1. Amsterdam: John Benjamins Publishing.
- Goddard, Cliff & Anna Wierzbicka. 2014a. *Words and meanings: Lexical semantics across domains, languages, and cultures*. (Oxford Linguistics). Oxford: Oxford University Press.
- Goddard, Cliff & Anna Wierzbicka. 2014b. Semantic fieldwork and lexical universals. *Studies in Language* 38(1). 80–126.
- Goldberg, Adele E. 1995. *Constructions: A construction grammar approach to argument structure*. (Cognitive Theory of Language and Culture). Chicago: University of Chicago Press.
- Goldberg, Adele E. 2006. *Constructions at work: The nature of generalization in language*. (Oxford Linguistics). Oxford ; New York: Oxford University Press.
- Goldberg, Adele E. 2009. The nature of generalization in language. *Cognitive Linguistics* 20(1). 93–127.
- Goldberg, Adele E. 2013. Constructionist approaches. In Thomas Hoffmann & Graeme Trousdale (eds.), *The Oxford handbook of construction grammar*, 15–31. Oxford ; New York: Oxford University Press.
- Goldberg, Yoav & Omer Levy. 2014. Word2vec explained: Deriving Mikolov et al.'s negative-sampling word-embedding method. <http://arxiv.org/>



- abs/1402.3722.
- Goldstein, David M. 2015. Review of: William H. Baxter & Laurent Sagart. 2014. *Old Chinese: A New Reconstruction*. *Bulletin of the School of Oriental and African Studies* 78(2). 413–414. doi:10.1017/S0041977X15000361.
- Goossens, Louis. 1990. Metaphtonymy: The interaction of metaphor and metonymy in expressions for linguistic action. *Cognitive Linguistics* 1(3). 323–340.
- Gómez, Gale Goodwin. 2009. Reduplication, ideophones, and onomatopoeic repetition in the Yanomami languages. *Grazer Linguistische Studien* 71. 1–18.
- Gōng, Liángyù 龔良玉 (ed.). 1991. *Xiāngshēngcí cídiǎn* 象声词词典 [Dictionary of onomatopoeia]. Guìyáng: Guīzhōu jiàoyù chūbǎnshè.
- Grammont, Maurice. 1901. Onomatopées et mots expressifs. *Trentenaire de la Société pour l'Étude des Langues Romanes*. 261–322.
- Greenacre, Michael. 2006. From simple to multiple correspondence analysis. In Michael Greenacre & Jorg Blasius (eds.), *Multiple correspondence analysis and related methods*, 41–76. London: Chapman & Hall.
- Greenacre, Michael. 2007. *Correspondence analysis in practice*. 2nd edn. Boca Raton, FL: Chapman & Hall.
- Gries, Stefan Th. 2008. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics* 13(4). 403–437. doi:10.1075/ijcl.13.4.02gri.
- Gries, Stefan Th. 2012. Frequencies, probabilities, and association measures in usage-/exemplar-based linguistics. *Studies in Language* 11(3). 477–510.
- Gries, Stefan Th. 2019a. 15 years of collostructions: Some long overdue additions/corrections (to/of actually all sorts of corpus-linguistics measures). *International Journal of Corpus Linguistics* 24(3). 385–412. doi:10.1075/ijcl.00011.gri.
- Gries, Stefan Th. 2015. More (old and new) misunderstandings of collostructional analysis: On Schmid and Küchenhoff (2013). *Cognitive Linguistics* 26(3). 505–536. doi:10.1515/cog-2014-0092.
- Gries, Stefan Th. 2019b. On classification trees and random forests in corpus linguistics: Some words of caution and suggestions for improvement. *Corpus Linguistics and Linguistic Theory* 15(1). doi:10.1515/cllt-2018-0078.
- Gries, Stefan Th. & Andrea L. Berez. 2017. Linguistic annotation in/for corpus linguistics. In Nancy Ide & James Pustejovsky (eds.), *Handbook of linguistic annotation*, 379–409. Dordrecht: Springer Netherlands. doi:10.1007/978-94-024-0881-2.
- Gries, Stefan Th. & Anatol Stefanowitsch. 2004a. Extending collostructional analysis: A corpus-based perspective on “alternations”. *International Journal of Corpus Linguistics* 9(1). 97–129. doi:10.1075/ijcl.9.1.06gri.
- Gries, Stefan Th. & Anatol Stefanowitsch. 2004b. Covarying collexemes in the *Into-Causative*. In Michel Achard & Suzanne Kemmer (eds.), *Language, culture and mind*, 225–236. Stanford, Calif: CSLI Publications.
- Grondelaers, Stefan & Dirk Geeraerts. 2003. Towards a pragmatic model of

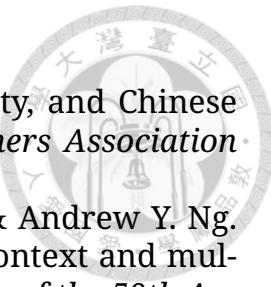
- cognitive onomasiology. In Hubert Cuyckens, René Dirven & John R. Taylor (eds.), *Cognitive approaches to lexical semantics*, 67–92. (Cognitive Linguistics Research 23). Berlin: Mouton de Gruyter.
- Guthrie, David, Ben Allison, Wei Liu, Louise Guthrie & Yorick Wilks. 2006. A closer look at skip-gram modelling. In Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odijk & Daniel Tapias (eds.), *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, 1222–1225. Genoa: European Language Resources Association.
- Güldemann, Tom. 2008. *Quotative indexes in African languages: A synchronic and diachronic survey*. (Empirical Approaches to Language Typology 34). Berlin ; New York: Mouton de Gruyter.
- Haiman, John. 1980. The iconicity of grammar: Isomorphism and motivation. *Language* 56(3). 515–540.
- Haiman, John. 1982. Dictionaries and encyclopedias again. *Lingua* 56. 353–355.
- Haiman, John (ed.). 1985. *Iconicity in syntax: Proceedings of a Symposium on Iconicity in Syntax, Stanford, June 24 - [2]6, 1983*. (Typological Studies in Language 6). Amsterdam: Benjamins.
- Haiman, John. 2011. *Cambodian: Khmer*. (London Oriental and African Language Library v. 16). Amsterdam ; Philadelphia: John Benjamins Pub. Co.
- Haiman, John. 2018. *Ideophones and the evolution of language*. Cambridge University Press. doi:10.1017/9781107706897.
- Hamano, Shoko. 1998. *The sound-symbolic system of Japanese*. (Studies in Japanese Linguistics). Stanford, Calif. ; Tokyo: CSLI Publications ; Kurosio.
- Hamilton, William L., Jure Leskovec & Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. <http://arxiv.org/abs/1605.09096>.
- Hampe, Beate & Joseph E. Grady (eds.). 2005. *From perception to meaning: Image schemas in cognitive linguistics*. (Cognitive Linguistics Research 29). Berlin ; New York: Mouton de Gruyter.
- Harbsmeier, Christoph. 2016. Irrefutable Conjectures. A Review of William H. Baxter and Laurent Sagart. *Old Chinese: A New Reconstruction. Monumenta Serica* 64(2). 445–504. doi:10.1080/02549948.2016.1259882.
- Harris, Zellig S. 1954. Distributional structure. *Word* 10. 146–62.
- Harris, Zellig S. 1970. *Papers in structural and transformational linguistics*. Dordrecht: Reidel.
- Hasada, Rie. 1994. The semantic aspects of onomatopoeia: Focusing on Japanese psychocomimes. Canberra: The Australian National University Master thesis.
- Haspelmath, Martin. 2010. Comparative concepts and descriptive categories in crosslinguistic studies. *Language* 86(3). 663–687. doi:10.1353/lan.2010.0021.
- Haspelmath, Martin. 2011. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica* 45(1). 31–80.
- Haspelmath, Martin. 2018. How comparative concepts and descriptive lin-



- guistic categories are different. In Daniël Olmen, Tanja Mortelmans & Frank Brisard (eds.), *Aspects of Linguistic Variation*, 83–114. Berlin, Boston: De Gruyter. doi:10.1515/9783110607963-004.
- Haspelmath, Martin. Towards standardization of morphosyntactic terminology for general linguistics. In. <http://www.academia.edu/download/61066836/Standardization06.1520191030-25789-15moenr.pdf>.
- Hàndiǎn 漢典. 2004–2018. Hàndiǎn 漢典 [Chinese dictionary]. <http://www.zdic.net/> (24 May, 2018).
- Heath, Jeffrey. 2014. Review Article on Langacker 2009: Online Supplement. *Language* 90(1). s1–s3. doi:10.1353/lan.2014.0021.
- Heath, Jeffrey. 2019. The dance of expressive adverbials (“ideophones”) in Jamsay (Dogon). *Folia Linguistica* 53(1). 1–24. doi:10.1515/flin-2019-2002.
- Henderson, Eugénie J. A. 1965. Final -k in Khasi: A secondary phonological pattern. *Lingua* 14. 459–466.
- Herlofsky, William J. 2019. Re-vision: Are there ideophones in signed languages? Oral presentation. Paper presented at the ILL 12 [International Symposium on Iconicity in Language and Literature], Lund: Lund University. <https://konferens.ht.lu.se/en/ill-12/>.
- Hester, Jim. 2019. *Glue: Interpreted string literals*. <https://CRAN.R-project.org/package=glue>.
- Hester, Jim & Hadley Wickham. 2019. *Fs: Cross-platform file system operations based on 'libuv'*. <https://CRAN.R-project.org/package=fs>.
- 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.
- Hill, Nathan W. 2012. The six vowel hypothesis of Old Chinese in comparative context. *Bulletin of Chinese Linguistics* 6(2). 1–69.
- Hill, Nathan W. 2017. Review of William H. Baxter & Laurent Sagart. *Old Chinese: A New Reconstruction*. Oxford University Press, 2014. *Archiv Orientální* 85. 135–140.
- Hinton, Leanne, Johanna Nichols & John J. Ohala (eds.). 1994. *Sound symbolism*. Cambridge [England]: Cambridge UP.
- Hiraga, Masako. 2005. *Metaphor and iconicity: A cognitive approach to analysing texts*. Hounds mills, Basingstoke, Hampshire ; New York: Palgrave Macmillan.
- Hiraga, Masako, William J. Herlofsky, Kazuko Shinohara & Kimi Akita (eds.). 2015. *Iconicity: East meets West*. (Iconicity in Language and Literature 14). Amsterdam ; Philadelphia: John Benjamins Publishing Company.
- Ho, Dah-an. 2016. Such errors could have been avoided: Review of *Old Chinese: A New Reconstruction*. *The Journal of Chinese Linguistics* 44(1).

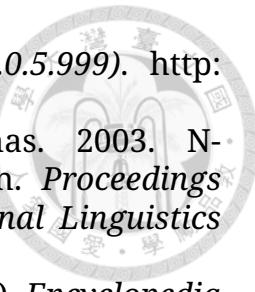


- 175–230.
- Hoffkann, Karl. 1952. “Wiederholende” Onomatopoetika im Altindischen. *Indogermanische Forschungen* 60. 254–264.
- Hoffmann, Thomas & Graeme Trousdale (eds.). 2013. *The Oxford handbook of construction grammar*. Oxford ; New York: Oxford University Press.
- Hong, Jia-Fei & Chu-Ren Huang. 2013. A hanzi radical ontology based approach towards teaching Chinese characters. In Donghong Ji & Guozheng Xiao (eds.), *Chinese Lexical Semantics: 13th workshop, CLSW 2012, Wuhan, China, July 6-8, 2012: Revised selected papers*, vol. 7717, 745–755. (Lecture Notes in Computer Science). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-36337-5.
- Hoon, Randall Yao Tong & Chiew Pheng Phua. 2009. Metaphorical conceptions of MIND in Pre-Qin Confucian mind-nature discourses. In Siva Sivakumar, Wenny Puspadewi & Khai Ge Luah (eds.), *Proceedings of the URECA*. Singapore: Nanyang Technological University.
- Hoshi, Hideyuki, Nahyun Kwon, Kimi Akita & Jan Auracher. 2019. Semantic associations dominate over perceptual associations in vowel-size iconicity. *i-Perception* 10(4). 204166951986198. doi:10.1177/2041669519861981.
- Hosmer, David W. & Stanley Lemeshow. 2000. *Applied logistic regression*. Hoboken, NJ, USA: John Wiley & Sons, Inc. doi:10.1002/0471722146.
- Hsieh, Feng-Fan. 2017. Reduplication. In Rint Sybesma, Wolfgang Behr, Yueguo Gu, Zev Handel, C.-T. James Huang & James Myers (eds.), *Encyclopedia of Chinese language and linguistics*, vol. III, 548–555. Leiden: Brill.
- 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.
- Hsieh, Shu-Kai & Chu-Ren Huang. 2009–2019. Chinese WordNet (Zhongwen cihui wanglu 中文詞彙網路). <http://lope.linguistics.ntu.edu.tw/cwn/> (9 January, 2016).
- Huang, Chu-Ren. 2000. From quantitative to qualitative studies: Developments in Chinese computational and corpus linguistics. *Chinese Studies 漢學研究* 18. 473–509.
- Huang, Chu-Ren & Keh-jiann Chen. 1992. A Chinese Corpus for Linguistics Research. In Christian Boitet (ed.), *Proceedings of the 1992 International Conference on Computational Linguistics (COLING-92)*, 1214–1217. Nantes.
- Huang, Chu-Ren, Shu-Kai Hsieh & Keh-jiann Chen (eds.). 2017. *Mandarin Chinese words and parts of speech: A corpus-based study*. (Routledge Studies in Chinese Linguistics). London ; New York: Routledge, Taylor & Francis Group.
- Huang, Chu-Ren, Ya-Jun Yang & Sheng-Yi Chen. 2008. An ontology of Chinese radicals: Concept derivation and knowledge representation based on the semantic symbols of the four hoofed-mammals. In, *Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation*, 189–196. The University of the Philippines Visayas Cebu Col-



- lege, Cebu: De La Salle University.
- Huang, C-T James. 1984. Phrase structure, lexical integrity, and Chinese compounds. *Journal of the Chinese Language Teachers Association* 19(2). 53–78.
- Huang, Eric H., Richard Socher, Christopher D. Manning & Andrew Y. Ng. 2012. Improving word representations via Global Context and multiple word prototypes. In Haizou Li (ed.), *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, 873–882. Jeju Island, Korea: ACL.
- Huang, Shi-Zhe, Jing Jin & Dingxu Shi. 2016. Adjectives and adjective phrases. In Chu-Ren Huang & Dingxu Shi (eds.), *A reference grammar of Chinese*, 276–296. Cambridge: Cambridge University Press.
- Hung, Jen-Jou, Marcus Bingenheimer & Simon Wiles. 2010. Quantitative evidence for a hypothesis regarding the attribution of early Buddhist translations. *Literary and Linguistic Computing* 25(1). 119–134. doi:10.1093/llc/fqp036.
- Hurch, Bernhard (ed.). 2005. *Studies on reduplication*. (Empirical Approaches to Language Typology 28). Berlin ; New York: Mouton de Gruyter.
- Huyssteen, Gerhard B. van. 2004. Motivating the composition of Afrikaans reduplications: A cognitive grammar analysis. In Günter Radden & Klaus-Uwe Panther (eds.), *Studies in linguistic motivation*, 269–292. (Cognitive Linguistics Research 28). Berlin ; New York: Mouton de Gruyter.
- Iannone, Richard. 2019. *DiagrammeR: Graph/Network visualization*. <https://CRAN.R-project.org/package=DiagrammeR>.
- Ibarretxe-Antuñano, Iraide. 2006. Estudio lexicológico de las onomatopeyas vascas: El Euskal Onomatopeien Hiztegia: Euskara-Ingelesera-Gaztelania. *Fontes Linguae Vasconum* 101. 145–159.
- Ibarretxe-Antuñano, Iraide. 2017. Basque ideophones from a typological perspective. *Canadian Journal of Linguistics/Revue canadienne de linguistique* 62(02). 196–220. doi:10.1017/cnj.2017.8.
- Ide, Nancy & James Pustejovsky (eds.). 2017. *Handbook of linguistic annotation*. Dordrecht: Springer Netherlands. doi:10.1007/978-94-024-0881-2_16.
- Imai, Mutsumi & Sotaro Kita. 2014. The sound symbolism bootstrapping hypothesis for language acquisition and language evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences* 369(1651). 1–13. doi:10.1098/rstb.2013.0298.
- Imai, Mutsumi, Sotaro Kita, Miho Nagumo & Hiroyuki Okada. 2008. Sound symbolism facilitates early verb learning. *Cognition* 109(1). 54–65. doi:10.1016/j.cognition.2008.07.015.
- Iwasaki, Noriko. 2017. Use of mimetics in Motion event descriptions by English and Korean learners of L2 Japanese: Does language typology make a difference? In Noriko Iwasaki, Peter Sells & Kimi Akita (eds.), *The grammar of Japanese mimetics: Perspectives from structure, acquisition and translation*, 193–218. (Routledge Studies in East Asian Linguistics). Milton Park, Abingdon, Oxon ; New York, NY: Routledge.

- Iwasaki, Noriko, Peter Sells & Kimi Akita (eds.). 2017. *The grammar of Japanese mimetics: Perspectives from structure, acquisition and translation*. (Routledge Studies in East Asian Linguistics). Milton Park, Abingdon, Oxon ; New York, NY: Routledge.
- Iwasaki, Shoichi. 2015. A multiple-grammar model of speakers' linguistic knowledge. *Cognitive Linguistics* 26(2). doi:10.1515/cog-2014-0101.
- Jacques, Guillaume. 2013. Ideophones in Japhug (Rgyalrong). *Anthropological Linguistics* 55(3). 256–287.
- Jacques, Guillaume. 2017. La réception du système de Baxter-Sagart: Premier bilan. *Panchronica*.
- Jakobson, Roman. 1960. Why "mama" and "papa". In Bernard Kaplan & Seymour Wapner (eds.), *Perspectives in psychological theory: Essays in honor of Heinz Werner*, 124–134. New York: International Universities Press.
- Jakobson, Roman & Linda R. Waugh. 1979. *The sound shape of language*. Bloomington: Indiana University Press.
- Jendraschek, Gerd. 2002. *Semantische Eigenschaften von Ideophonen im Türkischen*. (ed. Linguistik 30). München: LINCOM Europa.
- Jespersen, Otto. 1922. *Language: Its nature, development and origin*. London: Allen & Unwin.
- Johnson, Marion R. 1976. Toward a definition of the ideophone in Bantu. *OSU WPL* 21. 240–253.
- Johnson, Mark. 2005. The philosophical significance of image schemas. In Beate Hampe & Joseph E. Grady (eds.), *From perception to meaning: Image schemas in cognitive linguistics*, 15–33. (Cognitive Linguistics Research 29). Berlin ; New York: Mouton de Gruyter.
- Joo, Ian. 2018. Spoken language iconicity: An articulatory-based analysis of 66 languages. Hsinchu: National Chiao Tung University Master thesis.
- Joo, Ian. 2019. Phonosemantic biases found in Leipzig-Jakarta lists of 66 languages. *Linguistic Typology*(ahead of print). doi:10.1515/lingty-2019-0030.
- Jorden, Eleanor Harz & Mari Noda. 1987. *Japanese, the spoken language*. (Yale Language Series). New Haven: Yale University Press.
- Joseph, Brian D. 1997. On the linguistics of marginality: The centrality of the periphery. *Chicago Linguistic Society* 33. 197–213.
- Junod, Henri A. 1896. *Grammaire Ronga*. Lausanne: Imprimerie Georges Bridel & Cie.
- Kanero, Junko, Mutsumi Imai, Jiro Okuda, Hiroyuki Okada & Tetsuya Matsuda. 2014. How sound symbolism is processed in the brain: A study on Japanese mimetic words. (Ed.) Hisao Nishijo. *PLoS ONE* 9(5). e97905. doi:10.1371/journal.pone.0097905.
- Kassambara, Alboukadel. 2019. *Ggpubr: 'Ggplot2' based publication ready plots*. <https://CRAN.R-project.org/package=ggpubr>.
- Kassambara, Alboukadel & Fabian Mundt. 2019. *Factoextra: Extract and visualize the results of multivariate data analyses*. <https://CRAN.R-project.org/package=factoextra>.
- Kassambara, Alboukadel & Fabian Mundt. 2017. *Factoextra: Extract and vi-*



- sualize the results of multivariate data analyses (v. 1.0.5.999). <http://www.sthda.com/english/rpkgs/factoextra>.
- Kešelj, Vlado, Fuchun Peng, Nick Cercone & Calvin Thomas. 2003. N-gram-based author profiles for authorship attribution. *Proceedings of the conference Pacific Association for Computational Linguistics [PACLING]* 3. 255–264.
- Kilian-Hatz, Christa. 2006. Ideophones. In Keith Brown (ed.), *Encyclopedia of language & linguistics*, 508–512. Oxford: Elsevier.
- Kim-Renaud, Young-Key. 1978. Semantic features in phonology: Evidence from vowel harmony in Korean. *Korean Linguistics* 1(1). 1–18. doi:10.1075/kl.1.01ykk.
- Kita, Sotaro. 1993. Language and thought interface: A study of spontaneous gestures and Japanese mimetics. Chicago: University of Chicago PhD dissertation.
- Kita, Sotaro. 1997. Two-dimensional semantic analysis of Japanese mimetics. *Linguistics* 35. 379–415.
- Kiyama, Naoki & Kimi Akita. 2015. Gradability and mimetic verbs in Japanese: A frame-semantic account. *Proceedings of the Annual Meeting of the Berkeley Linguistics Society* 41. 245–265.
- Kizu, Mika & Naomi Cross. 2017. Translating into Japanese mimetics: Grammatical class-shifts and historical development. In Noriko Iwasaki, Peter Sells & Kimi Akita (eds.), *The grammar of Japanese mimetics: Perspectives from structure, acquisition and translation*, 221–237. (Routledge Studies in East Asian Linguistics). Milton Park, Abingdon, Oxon ; New York, NY: Routledge.
- Kockelman, Paul. 2003. The meanings of interjections in Q'eqchi' Maya: From emotive reaction to social and discursive action. *Current Anthropology* 44(4). 467–497.
- Koelle, Sigismund Wilhelm. 1854. *Grammar of the Bórnū or Kānurī language*. London: Church Missionary House.
- Köhler, Wolfgang. 1929. *Gestalt psychology*. New York: Liveright.
- Kövecses, Zoltán. 2017. Levels of metaphor. *Cognitive Linguistics* 28(2). doi:10.1515/cog-2016-0052.
- Kǒng, Lìlì 孔丽丽. 2017. Hán Tài nǐshēngcí bìjiào yánjiū 汉泰拟声词比较研究 (A comparative study of Chinese and Thai onomatopoeia). Tianjin: Tianjin University Master thesis.
- Kroll, Paul W. 2015. *A student's dictionary of Classical and Medieval Chinese*. (Handbook of Oriental Studies: Section 4 China 30). Leiden: Brill.
- Kroll, Paul W. 2017. *A student's dictionary of Classical and Medieval Chinese: Revised edition*. (Handbook of Oriental Studies. Section 4 China 30). Leiden ; Boston, MA: Brill.
- Kulemeka, A. T. 1995. Sound symbolic and grammatical frameworks: A typology of ideophones in Asian and African languages. *South African Journal of African Languages* 15(2). 73–84.
- Kunene, Daniel P. 1965. The ideophone in Southern Sotho. *Journal of African Languages* 4. 19–39.
- Künstler, Mieczysław Jerzy. 1967. *Les formations adverbiales à quasi-suffixe en chinois arachaique et dans la langue de l'époque Han*. (Prace Orient.



- talistyczne 17). Warszawa: Państwowe Wydawnictwo Naukowe.
- Kwok, Bit-Chee 郭必之. 2012. Cóng Nánníng Yuèyǔ de zhuàngmàocí kàn Hànyǔ fāngyán yǔ mínzú yǔyán de jiēchù 从南宁粤语的状貌词看汉语方言与民族语言的接触 [Studying language contact between Chinese dialects and folk languages through investigating ideophones in Nanning Yue]. *Mínzú yǔwén* 民族语文 3. 16–24.
- Kwon, Nahyun. 2015. The natural motivation of sound symbolism. Brisbane: University of Queensland PhD dissertation.
- Kwon, Nahyun. 2017. Total reduplication in Japanese ideophones: An exercise in Localized Canonical Typology. *Glossa: a journal of general linguistics* 2(1). 40. doi:10.5334/gjgl.267.
- Kwon, Nahyun & Keiko Masuda. 2019. On the ordering of elements in ideophonic echo-words versus prosaic dvandva compounds, with special reference to Korean and Japanese. *Journal of East Asian Linguistics* 28(1). 29–53. doi:10.1007/s10831-019-09189-1.
- Kwon, Nahyun & Erich R. Round. 2015. Phonaesthemes in morphological theory. *Morphology* 25(1). 1–27. doi:10.1007/s11525-014-9250-z.
- Labov, William. 1973. The boundaries of words and their meanings. In Charles-James Bailey & Roger W. Shuy (eds.), *New ways of analysing variation in English*. Washington, DC: Georgetown University Press.
- Lahaussois, Aimée. 2017. Ideophones in Khaling Rai. *Linguistics of the Tibeto-Burman Area* 40(2). 179–201. doi:hal-01895513.
- Lai, Yik-Po 黎奕葆. 2015. Xiānggǎng yuèyǔ shuāng yīnjié zhuàng mào hòuzhuì de yīnyùn tèsè 香港粤语雙音節狀貌後綴的音韻特色 (Phonology of disyllabic ideophonic suffix in Hong Kong Cantonese). *Language and Linguistics* 16(5). 691–729.
- Lakoff, George. 1987. *Women, fire, and dangerous things: What categories reveal about the mind*. Chicago: The Univ. of Chicago Press.
- Lakoff, George & Mark Johnson. 1980. *Metaphors we live by*. Chicago: University of Chicago Press.
- Lakoff, George & Mark Johnson. 1999. *Philosophy in the flesh: The embodied mind and its challenge to Western thought*. New York: Basic Books.
- Landauer, Thomas K. & Susan T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review* 104(2). 211–240.
- Langacker, Ronald W. 1987a. *Foundations of Cognitive Grammar 1: Theoretical prerequisites*. Stanford, California: Stanford University Press.
- Langacker, Ronald W. 1987b. Nouns and Verbs. *Language* 63(1). 53–94.
- Langacker, Ronald W. 1991. *Foundations of Cognitive Grammar 2: Descriptive application*. Stanford, California: Stanford University Press.
- Langacker, Ronald W. 2005. Construction Grammars: Cognitive, radical, and less so. In Francisco José Ruiz de Mendoza Ibáñez & M. Sandra Peña Cervel (eds.), 101–159. (Cognitive Linguistics Research 32). Berlin ; New York: Mouton de Gruyter.
- Langacker, Ronald W. 2008a. *Cognitive grammar: A basic introduction*. Oxford ; New York: Oxford University Press.
- Langacker, Ronald W. 2008b. Sequential and summary scanning: A reply.

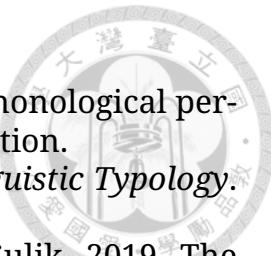
- Cognitive Linguistics* 19(4). doi:10.1515/COGL.2008.022.
- LaPolla, Randy J. 1994. An experimental investigation into phonetic symbolism as it relates to Mandarin Chinese. In, *Sound symbolism*, 130–147. Cambridge [England]: Cambridge University Press.
- Larson, Richard K. 1991. Some issues in verb serialization. In Claire Lefebvre (ed.), *Serial verbs: Grammatical, comparative and cognitive approaches*, 185–211. Amsterdam: John Benjamins.
- Lawler, John. 1988. Time is money: The anatomy of a metaphor. Ann Arbor: University of Michigan, ms.
- Le, Minh Thanh 黎明清. 2017. Yuè Hán nǐshēngcí jí qí hù yì yánjiū 越汉拟声词及其互译研究 (The study of contrast and translation between Vietnamese and Chinese onomatopoetic words). Wuhan: Wuhan University PhD dissertation.
- Lee, John. 2012. A Classical Chinese corpus with nested Part-of-Speech Tags. *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. (Associaton for Computational Linguistics). 75–84.
- Lenci, Alessandro. 2018. Distributional Models of Word Meaning. *Annual Review of Linguistics* 4(1). 151–171. doi:10.1146/annurev-linguistics-030514-125254.
- Leskien, A. 1902. Schallnachahmungen und Schallverba im Litauischen. *Indogermanische Forschungen* 13. 165–212.
- Levin, Beth. 1993. *English verb classes and alternations: A preliminary investigation*. Chicago: University of Chicago Press.
- Levshina, Natalia. 2015. *How to do linguistics with R: Data exploration and statistical analysis*. Amsterdam ; Philadelphia: John Benjamins Publishing Company.
- Levshina, Natalia. 2019. *Rling: R for linguistics*.
- Levy, Omer & Yoav Goldberg. 2014. Linguistic regularities in sparse and explicit word representations. *Proceedings of the 18th Conference on Computational Language Learning*. 171–180.
- Levy, Omer, Yoav Goldberg & Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics* 3. 211–225.
- Lê, Sébastien, Julie Josse & François Husson. 2008. FactoMineR: An R package for multivariate analysis. *Journal of Statistical Software* 25(1). 1–18.
- Li, Charles N. & Sandra A. Thompson. 1981. *Mandarin Chinese: A functional reference grammar*. 1. paperback print, repr. Berkeley, Calif.: Univ. of California Press.
- Li, Jian. 2013. The rise of disyllables in Old Chinese: The role of lianmian words. New York, N.Y.: The City University of New York PhD dissertation.
- Li, Peng-Hsuan, Tsu-Jui Fu & Wei-Yun Ma. 2019. Why Attention? Analyze BiLSTM Deficiency and Its Remedies in the Case of NER. <http://arxiv.org/abs/1908.11046> (15 January, 2020).
- Li, Yueyuan. 2015. Verb reduplication: A cross-linguistic survey with special focus on Mandarin Chinese. Lancaster: Lancaster University PhD



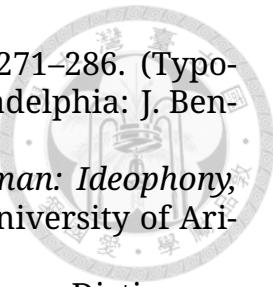
- dissertation.
- Liberman, Mark. 1975. The intonational system of English. Cambridge, MA: MIT PhD dissertation.
- Lieven, Elena & Michael Tomasello. 2008. Children's first language acquisition from a usage-based perspective. In Peter Jake Robinson & Nick C. Ellis (eds.), *Handbook of cognitive linguistics and second language acquisition*, 168–196. New York: Routledge.
- Lijffijt, Jeffrey & Stefan Th. Gries. 2012. Correction to Stefan Th. Gries' "Dispersions and adjusted frequencies in corpora", International Journal of Corpus Linguistics. *International Journal of Corpus Linguistics* 17(1). 147–149. doi:10.1075/ijcl.17.1.08lij.
- List, Johann-Mattis. 2018. *Sinopy: A Python library for quantitative tasks in Chinese historical linguistics*. Jena: Max Planck Institute for the Science of Human History. <https://github.com/lingpy/sinopy>.
- List, Johann-Mattis, Nathan W. Hill & Christopher J. Foster. 2019. Towards a standardized annotation of rhyme judgments in Chinese historical phonology (and beyond). *Journal of Language Relationship Bonposы языкового родства* 17(1-2). 26–43.
- List, Johann-Mattis, Jananan Sylvestre Pathmanathan, Nathan W. Hill, Eric Baptiste & Philippe Lopez. 2017. Vowel purity and rhyme evidence in Old Chinese reconstruction. *Lingua Sinica* 3(1). doi:10.1186/s40655-017-0021-8.
- Liú, Dānqīng 刘丹青. 2009. Shíci de nǐshēnghuà chóngdié jí qí xiāngguān gòushì 实词的拟声化重叠及其相关构式 (Ideophonic reduplication of content words and the related constructions in Mandarin Chinese). *Zhōngguó yǔwén* 中国语文 1. 22–31.
- Lǐ, Jīng'ér 李镜儿. 2007. *Xiàndài Hànyǔ nǐshēngcí yánjiū* 现代汉语拟声词研究 (*Onomatopoeics in Modern Chinese*). Shànghǎi: Xuélín chūbǎnshè.
- Lǐ, Jìnróng 李劲荣. 2008. ABB xíngshì róng cí de gòuchéng fāngshì ABB 形式容词的构成方式 (The structure type of adjectives in the form ABB). *Journal of Gannan Normal University* 赣南师范学院学报 1. 87–91.
- Lǐ, Shuǐ 李水. 2018. Jìn shí nián Xiàndài Hànyǔ nǐshēngcí yánjiū de xīn dòngxiàng (2008–2018) 近十年现代汉语拟声词研究的新动向 (2008~2018) [A summary of the new studies on Chinese onomatopoeic words in the past ten years]. *Journal of Yunnan Normal University* 云南师范大学学报 (对外汉语教学与研究版) 16(3). 55–60.
- Lǐ, Yuèyuán & Dan Ponsford. 2018. Predicative reduplication: Functions, their relationships and iconicities. *Linguistic Typology* 22(1). 51–117. doi:10.1515/lingty-2018-0003.
- Lǐ, Yún'bīng 李云兵. 2006. Miáo yǔ chóngdiéshì de gòuchéng xíngshì, yǔyì hé jùfǎ jiégòu tèzhēng 苗语重叠式的构成形式、语义和句法结构特征 (On the forms, semantics and syntactic [sic] of characteristics of reduplication in Miao language). *Yǔyán kēxué* 语言科学 5(2). 85–103.
- Lockwood, Gwilym. 2017. Talking sense: The behavioural and neural correlates of sound symbolism. Nijmegen: Radboud University Nijmegen PhD dissertation.
- Lockwood, Gwilym & Mark Dingemanse. 2015a. Iconicity in the lab: A review of behavioral, developmental, and neuroimaging research

- into sound-symbolism. *Frontiers in Psychology* 6(1246). 1–14. doi:10.3389/fpsyg.2015.01246.
- Lockwood, Gwilym & Mark Dingemanse. 2015b. Corrigendum: Iconicity in the lab: A review of behavioral, developmental, and neuroimaging research into sound-symbolism. *Frontiers in Psychology* 6(1624). 1–2. doi:10.3389/fpsyg.2015.01624.
- Lockwood, Gwilym, Mark Dingemanse & Peter Hagoort. 2016. Sound-Symbolism Boosts Novel Word Learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. doi:10.1037/xlm0000235.
- Loewe, Michael (ed.). 1993. *Early Chinese texts: a bibliographical guide*. (Early China Special Monograph Series no. 2). Berkeley, Calif.: Society for the Study of Early China : Institute of East Asian Studies, University of California, Berkeley.
- Lu, Chiarung 呂佳蓉. 2006. *Giongo, gitai 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.
- Luo, Wen, Fumito Masui & Michal Ptaszynski. 2014. Explaining Japanese onomatopoeia in Chinese using translated paraphrases. In, *Proceedings of the International Workshop on Modern Science and Technology*. Wuhan University of Science and Technology.
- Luó, Zhúfēng 罗竹风 (ed.). 1993. *Hànyǔ dà cídiǎn* 漢語大詞典 (*Comprehensive Dictionary of Chinese*). Shànghǎi: Shanghai Lexicographical Publishing House.
- Lǚ, Shūxiāng 吕叔湘 (ed.). 2005. *Xiàndài Hànyǔ cídiǎn* 现代汉语词典 [*Dictionary of Modern Chinese*]. 5th ed. Beijing: Commercial Press.
- Lynott, Dermot & Louise Connell. 2009. Modality exclusivity norms for 423 object properties. *Behavior Research Methods* 41(2). 558–564. doi:10.3758/BRM.41.2.558.
- Lynott, Dermot & Louise Connell. 2013. Modality exclusivity norms for 400 nouns: The relationship between perceptual experience and surface word form. *Behavior Research Methods* 45(2). 516–526. doi:10.3758/s13428-012-0267-0.
- MacWhinney, Brian. 2000. *The CHILDES project: Tools for analyzing talk*. 3rd ed. Mahwah, NJ: Lawrence Erlbaum.
- Magnus, Margaret. 2001. What's in a word? Studies in phonosemantics. Trondheim: Norwegian University of Science and Technology (NTNU) PhD dissertation.
- Marneffe, Marie-Catherine de & Christopher Potts. 2017. Developing Linguistic Theories Using Annotated Corpora. In Nancy Ide & James Pustejovsky (eds.), *Handbook of Linguistic Annotation*, 411–438. Dordrecht: Springer Netherlands. doi:10.1007/978-94-024-0881-2_16.
- Matthews, Danielle, Elena Lieven, Anna Theakston & Michael Tomasello. 2005. The role of frequency in the acquisition of English word order. *Cognitive Development* 20(1). 121–136. doi:10.1016/j.cogdev.2004.08.001.
- Mǎ, Kūn 马坤. 2017. Shǐ bījiào xià de Shànggǔ Hán yǔ gòunǐ — Bái Yípíng, Shā Jiā’ěr (2014) tǐxì shùpíng 史比较下的上古汉语构拟——白一平、沙佳儿 (2014) 体系述评

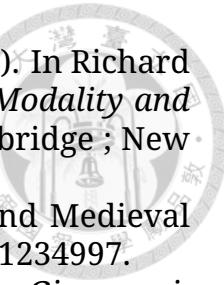
- 加尔 (2014) 体系述评 [A systematic review of the reconstruction of Old Chinese from a historical-comparative perspective — William H. Baxter and Laurent Sagart 2014]. *Zhongguo yuwen* 中国语文 4. 496–509.
- Mǎn, Fāng 滿芳. 2009. *Shījīng zhōng de AABB shì zǔhé* 《诗经》中的 AABB 式组合 (The AABB structure in the Shijing). *Shāndōng jiàoyù xuéyuàn xuébào* 山东教育学院学报 24(1). 42–44.
- McCarthy, John J. 1983. Phonological features and morphological structure. In John F. Richardson, Mitchell Marks & Amy Chuckerman (eds.), *Papers from the Parasession on the Interplay of Phonology, Morphology and Syntax*, 135–161. Chicago, IL: Chicago Linguistic Society.
- McCawley, James. 1992. Justifying part-of-speech assignment in Mandarin Chinese. *Journal of Chinese Linguistics* 20(2). 211–245.
- McLaren, James. 1886. *An introductory Kafir grammar with progressive exercises*. Lovedale.
- McLean, Bonnie. 2019. One form, many meanings: Iconicity in phonological and semantic development. Canberra: The Australian National University Honours thesis.
- McNally, Louise. 1991. Multiplanar reduplication: Evidence from Sesotho. In Aaron Halpern (ed.), *Proceedings of WCCFL 9*, 331–346.
- Meinhof, Carl. 1906. *Grundzüge einer vergleichenden Grammatik der Bantusprachen*. Berlin: Eckhardt & Messtorff.
- Mel'čuk, Igor A. 1988. *Dependency syntax: Theory and practice*. (SUNY Series in Linguistics). Albany (New York): State University of New York Press.
- Meng, Chenxi. 2012. A description of ideophonic words in Mandarin Chinese. Leiden: Leiden University Research Master in Linguistics.
- Mervis, Carolyn B. & Eleanor Rosch. 1981. Categorization of natural objects. *Annual Review of Psychology* 32. 89–115.
- Mester, R. Armin & Junko Itô. 1989. Feature predictability and underspecification: Palatal prosody in Japanese mimetics. *Language* 65(2). 258–293. doi:10.2307/415333.
- Michel, Jean-Baptiste, Yuan Kui Shen, Aviva P. Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, et al. 2011. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science* 331(6014). 176–182. doi:10.1126/science.1199644.
- Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. Distributed representation of words and phrases and their compositionality. In Christopher J. C. Burges, Léon Bottou, Max Welling, Zoubin Ghahramani & Kilian Q. Weinberger (eds.), *Advances in neural information processing systems (Proceedings of Neural Information Processing Systems [NIPS 26])*, 3111–3119.
- 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.
- Mithun, Marianne. 1982. The synchronic and diachronic behavior of plops, squeaks, croaks, sighs, and moans. *International Journal of American Linguistics*.



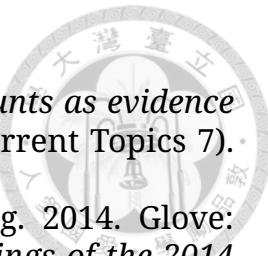
- linguistics* 48(1). 49–58.
- Mok, Waiching Enid. 2001. Chinese sound symbolism: A phonological perspective. Hawai'i: University of Hawai'i PhD dissertation.
- Moroz, George. 2017. *Lingttypology: Easy mapping for Linguistic Typology*. <https://CRAN.R-project.org/package=lingtypology>.
- Motamed, Yasamin, Hannah Little, Alan Nielsen & Justin Sulik. 2019. The iconicity toolbox: Empirical approaches to measuring iconicity. *Language and Cognition* 11(02). 188–207. doi:10.1017/langcog.2019.14.
- Müller, Kirill. 2017. *Here: A simpler way to find your files*. <https://CRAN.R-project.org/package=here>.
- Müller, Max. 1861. *Lectures on the science of language* 1. London: Longmans, Green.
- Nännny, Max & Olga Fischer (eds.). 1999. *Form miming meaning*. (Iconicity in Language and Literature 1). Amsterdam: Benjamins.
- Nenadic, Oleg & Michael Greenacre. 2007. Correspondence Analysis in R, with two- and three-dimensional graphics: The ca package. *Journal of Statistical Software* 20(3). 1–13.
- Newman, Paul. 1968. Ideophones from a syntactic point of view. *Journal of West African Languages* 5. 107–117.
- Newmeyer, Frederick J. 1992. Iconicity and generative grammar. *Language*. 756–796.
- Nielsen, Alan KS & Mark Dingemanse. 2020. Iconicity in word learning and beyond: A critical review. *Language and Speech*. 1–21. doi:10.1177/0023830920914339.
- NINJAL. 2016. Lago Word Profiler for the Balanced Corpus of Contemporary Written Japanese (NINJAL-LWP for BCCWJ). Database. <http://nlb.ninjal.ac.jp/search/> (25 March, 2019).
- Norman, Jerry. 1988. *Chinese*. (Cambridge Language Surveys). Cambridge [Cambridgeshire] ; New York: Cambridge University Press.
- Noss, Philip A. 1986. The ideophone in Gbaya syntax. In Gerrit J. Dimmenaal (ed.), *Current approaches to African linguistics*. Dordrecht: Foris.
- Noss, Philip A. 1989. The ideophone poems of Dogobadomo. *Croscurrent (Special Issue: The Ancestor's Beads)* 23(4). 33–43.
- Noss, Philip A. 1999. The ideophone: A dilemma for translation and translation theory. In Paul F. A. Kotey (ed.), *New dimensions in African linguistics and languages*, 261–272. Trenton, NJ: Africa World Press.
- Noss, Philip A. 2001. Ideas, phones and Gbaya verbal art. In Erhard Friedrich Karl Voeltz & Christa Kilian-Hatz (eds.), *Ideophones*, 259–270. (Typological Studies in Language v. 44). Amsterdam ; Philadelphia: J. Benjamins.
- Nuckolls, Janis B. 1995. Quechua texts of perception. *Semiotica* 103(1/2). 145–169.
- Nuckolls, Janis B. 1996. *Sounds like life: Sound-symbolic grammar, performance, and cognition in Pastaza Quechua*. (Oxford Studies in Anthropological Linguistics 2). New York: Oxford University Press.
- Nuckolls, Janis B. 1999. The case for sound symbolism. *Annual Review of Anthropology*. 225–252.
- Nuckolls, Janis B. 2001. Ideophones in Pastaza Quechua. In Erhard Friedrich



- Karl Voeltz & Christa Kilian-Hatz (eds.), *Ideophones*, 271–286. (Typological Studies in Language v. 44). Amsterdam ; Philadelphia: J. Benjamins.
- Nuckolls, Janis B. 2010. *Lessons from a Quechua strongwoman: Ideophony, dialogue, and perspective*. (First Peoples). Tucson: University of Arizona Press.
- Nuckolls, Janis B. 2016a-. The Quechua Ideophonic Dictionary. Dictionary. *Quechua realwords: An audiovisual ANTI-dictionary of expressive Quechua ideophones*. <http://nongrat.us/quechua/> (9 January, 2017).
- Nuckolls, Janis B. 2019. The sensori-semantic clustering of ideophonic meaning in Pastaza Quichua. In Kimi Akita & Prashant Pardeshi (eds.), *Ideophones, mimetics and expressives*, 167–198. (Iconicity in Language and Literature, ILL 16). Amsterdam : Philadelphia: John Benjamins Pub. Co.
- Nuckolls, Janis B. 2014. Ideophones' challenges for typological linguistics: The case of Pastaza Quichua. *Pragmatics and Society* 5(3). 355–383. doi:10.1075/ps.5.3.03nuc.
- Nuckolls, Janis B. 2016b. Rethinking mono-sensory implicational approaches to ideophones in Pastaza Quichua. In NINJAL (ed.), *Mimetics in Japanese and other languages in the world* (日本語と世界諸言語のオノマトペ). Tachikawa: NINJAL.
- Nuckolls, Janis B., Joseph A. Stanley, Elizabeth Nielsen & Roseanna Hopper. 2016. The systematic stretching and contracting of ideophonic phonology in Pastaza Quichua. *International Journal of American linguistics* 82(1). 95–116.
- Nuckolls, Janis B. & Tod D. Swanson. 2019. Quechua Real Words: An audiovisual ANTI-dictioanry of expressive Quechua ideophones. <http://quechuarealwords-dev.byu.edu/index.php> (12 March, 2019).
- Nuckolls, Janis B., Tod D. Swanson, Diana Shelton, Alexander Rice & Sarah Hatton. 2017. Lexicography in-your-face: The active semantics of Pastaza Quichua ideophones. *Canadian Journal of Linguistics/Revue canadienne de linguistique* 62(02). 154–172. doi:10.1017/cnj.2017.9.
- Oakley, Todd. 2007. Image schemas. In Dirk Geeraerts & H. Cuyckens (eds.), *The Oxford handbook of cognitive linguistics*, 214–235. (Oxford Handbooks). Oxford ; New York: Oxford University Press.
- Occhino, Corrine, Benjamin Anible, Erin Wilkinson & Jill P. Morford. 2017. Iconicity is in the eye of the beholder: How language experience affects perceived iconicity. *Gesture* 16(1). 100–126. doi:10.1075/gest.16.1.04occ.
- Ogden, Charles K. & Ivor A. Richards. 1923. *The meaning of meaning: A study of the influence of language upon thought and of the science of symbolism*. Magdalene College: Univ. of Cambridge.
- Ohala, John J. 1994. The frequency code underlies the sound-symbolic use of voice pitch. In Leanne Hinton, Johanna Nichols & John J. Ohala (eds.), *Sound symbolism*, 325–347. Cambridge [England]: Cambridge University Press.
- Okrent, Arika. 2002. A modality-free notion of gesture and how it can help us with the morpheme vs. gesture question in sign language



- linguistics (Or at least give us some criteria to work with). In Richard P. Meier, Kearsy Cormier & David Quinto-Pozos (eds.), *Modality and structure in signed and spoken languages*, 175–198. Cambridge ; New York: Cambridge University Press.
- O'Neill, Timothy. 2016. A Student's Dictionary of Classical and Medieval Chinese. *Tang Studies*. 1–11. doi:10.1080/07375034.2016.1234997.
- Ono, Masahiro 小野正弘 (ed.). 2007. *Nihongo onomatope jiten: Giongo, gitango 4500* 日本語オノマトペ辞典：擬音語、擬態語 4500. Tōkyō: Shōgakkan.
- Osswald, Rainer & Robert D. Jr. Van Valin. 2014. FrameNet, frame structure, and the syntax-semantics interface. In Thomas Gamerschlag, Doris Gerland, Rainer Osswald & Wiebke Petersen (eds.), *Frames and concept types: Applications in language and philosophy*, vol. 94, 125–156. Cham: Springer International Publishing. doi:10.1007/978-3-319-01541-5_6.
- Ou, Hsiu-Hui 歐秀慧. 1992. *Shījīng nǐ shēng cí yánjiū* 詩經擬聲詞研究 [A Study of Onomatopoeia in the *Shijing*]. Chiayi: National Chung Cheng University 國立中正大學 Master thesis.
- Packard, Jerome Lee (ed.). 1998. *New approaches to Chinese word formation: Morphology, phonology and the lexicon in modern and ancient Chinese*. (Trends in Linguistics: Studies and Monographs 105). Berlin ; New York: Mouton de Gruyter.
- Packard, Jerome Lee. 2000. *The morphology of Chinese: A linguistic and cognitive approach*. Cambridge, UK ; New York, NY, USA: Cambridge University Press.
- Packard, Jerome Lee. 2001. *The morphology of Chinese: A linguistic and cognitive approach* (Hànyǔ xíngtài xué: Yǔyán rènzhī yánjiù fǎ 汉语形态学: 语言认知研究法). (Ed.) Shí Dìngxǔ 石定栩. Beijing: Wàiyǔ jiāoxué yǔ yánjiū chǔbǎnshè.
- Packard, Jerome Lee. 2016. Lexical word formation. In Chu-Ren Huang & Dingxu Shi (eds.), *A reference grammar of Chinese*, 67–80. Cambridge: Cambridge University Press.
- Paul, Waltraud. 2006. Zhu Dexi's two classes of adjectives revisited. In Christoph Anderl & Halvor Eifring (eds.), *Studies in Chinese language and culture: Festschrift in honour of Christoph Harbsmeier on the occasion of his 60th birthday*, 303–315. Oslo: Hermes Academic Publishing.
- Paul, Waltraud. 2015a. *New perspectives on Chinese syntax*. (Trends in Linguistics. Studies and Monographs 271). Berlin ; Boston: De Gruyter.
- Paul, Waltraud. 2015b. The syntax and semantics of the sentence periphery (part I) - what the topic is (not) about. In, *New perspectives on Chinese syntax*, 193–248. (Trends in Linguistics. Studies and Monographs 271). Berlin ; Boston: Mouton de Gruyter.
- Peeters, Bert. 2000. Setting the scene: Some recent milestones in the lexicon-encyclopedia debate. In Bert Peeters (ed.), *The lexicon-encyclopedia interface*, 1–52. Amsterdam ; New York: Citeseer.
- 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.
- Penke, Martina & Anette Rosenbach (eds.). 2007. *What counts as evidence in linguistics: The case of innateness*. (Benjamins Current Topics 7). Amsterdam ; Philadelphia: J. Benjamins Pub. Co.
- Pennington, Jeffrey, Richard Socher & Christopher Manning. 2014. Glove: Global vectors for word representation. In, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing [EMNLP]*, 1532–1543. Doha: Association for Computational Linguistics.
- Perlman, Marcus, Hannah Little, Bill Thompson & Robin L. Thompson. 2018. Iconicity in signed and spoken vocabulary: A comparison between American Sign Language, British Sign Language, English, and Spanish. *Frontiers in Psychology* 9. 1433. doi:10.3389/fpsyg.2018.01433.
- Perniss, Pamela & Gabriella Vigliocco. 2014. The bridge of iconicity: From a world of experience to the experience of language. *Philosophical transactions of The Royal Society* 369. 1–13.
- Perry, Lynn K., Marcus Perlman & Gary Lupyan. 2015. Iconicity in English and Spanish and its relation to lexical category and age of acquisition. (Ed.) Johan J Bolhuis. *PLOS ONE* 10(9). e0137147. doi:10.1371/journal.pone.0137147.
- Piantadosi, Steven T. 2014. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review* 21(5). 1112–1130. doi:10.3758/s13423-014-0585-6.
- Posner, Michael. 1986. Empirical studies of prototypes. In Colette Craig (ed.), *Noun classes and categorization*, 53–61. Amsterdam: Benjamins.
- Potts, Christopher. 2007. The expressive dimension. *Theoretical Linguistics* 33(2). 165–198.
- Priestly, Tom M. S. 1994. On levels of analysis of sound symbolism in poetry, with an application to Russian poetry. In Leanne Hinton, Johanna Nichols & John J. Ohala (eds.), *Sound symbolism*, 237–248. Cambridge [England]: Cambridge University Press.
- Pulleyblank, Edwin G. 1973. Some new hypotheses concerning word families in Chinese. *Journal of Chinese Linguistics* 1. 111–125.
- Pulleyblank, Edwin G. 1995. *Outline of classical Chinese grammar*. Vancouver, BC: UBC Press.
- Pustejovsky, James. 1995. *The generative lexicon*. Cambridge, Mass: MIT Press.
- Qin, Wenfeng & Yanyi Wu. 2019. *jiebaR: Chinese Text Segmentation* (v. 0.10.99). <https://CRAN.R-project.org/package=jiebaR>.
- Qiu, Yixuan. 2019. *Showtext: Using fonts more easily in r graphs*. <https://CRAN.R-project.org/package=showtext>.
- Radden, Günter & Klaus-Uwe Panther (eds.). 2004. *Studies in linguistic motivation*. (Cognitive Linguistics Research 28). Berlin ; New York: Mouton de Gruyter.
- Ramachandran, Vilayanur S & Edward M Hubbard. 2001. Synaesthesia: A window into perception, thought and language. *Journal of Consciousness Studies* 8(12). Imprint Academic. 3–34.
- Rácz, Péter. 2013. *Salience in sociolinguistics: A quantitative approach..* Vol.

84. Berlin: De Gruyter Mouton.
- R Core Team. 2019. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Reddy, Michael J. 1979. The conduit metaphor: A case of frame conflict in our language about language. In Andrew Ortony (ed.), *Metaphor and thought*, 284–324. Cambridge: Cambridge Univ. Press.
- Reid, David. 1967. *Sound symbolism*. Edinburgh: T. & A. Constable.
- Reisinger, Joseph & Raymond J. Mooney. 2010. Multi-prototype vector-space models of word meaning. In Ron Kaplan, Jill Burstein, Mary Harper & Gerald Penn (eds.), *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL 2010)*. Los Angeles, CA: ACL.
- Rosch, Eleanor. 1975. Cognitive representations of semantic categories. *Journal of Experimental Psychology: General* 104(3). 192–233.
- Rosch, Eleanor. 1978. Principles of categorization. In Eleanor Rosch & Barbara B. Lloyd (eds.), *Cognition and categorization*, 27–48. Hillsdale, N.J.: Erlbaum.
- Rosch, Eleanor. 1988. Coherences and categorization: A historical view. In Frank S. Kessel (ed.), *The development of language and language researchers: Essays in honor of Roger Brown*, 373–392. Hillsdale, N.J.: Erlbaum.
- Rosch, Eleanor & Carolyn B. Mervis. 1975. Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology* 7(4). 573–605.
- Rossum, Guido van & Fred L. Drake. 2009. *Python 3 Reference Manual*. (Python Documentation Manual). Scotts Valley, CA: CreateSpace.
- Ruette, Tom. 2012. Aggregating lexical variation towards large-scale lexical lectometry. Leuven: KU Leuven PhD dissertation.
- Ruiz Martínez, Daniel. 2019. Dictionaries of Japanese mimetic words: Defining the mimetic category by the selection of lexical items. *Lexicography* 6(1). 1–19. doi:10.1007/s40607-019-00055-9.
- Sadowski, Piotr. 2001. The sound as an echo to the sense: The iconicity of English *G1*-words. In Olga Fischer & Max Nänny (eds.), *The motivated sign*, 69–88. (Iconicity in Language and Literature 2). Amsterdam: Benjamins.
- Sagart, Laurent. 1999. *The roots of old Chinese*. (Amsterdam Studies in the Theory and History of Linguistic Science Series 4, Current Issues in Linguistic Theory 184). Amsterdam: Benjamins.
- 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.
- Samarin, William J. 1970a. Inventory and choice in expressive language. *Word* 26(2). 153–169. doi:10.1080/00437956.1970.11435590.
- Samarin, William J. 1970b. Field procedures in ideophone research. *Journal of African Languages* 9(1). 27–30.
- Samarin, William J. 1971. Survey of Bantu ideophones. *African Language*



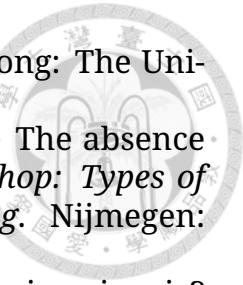
- Studies* 12. 130–168.
- Sam-Sin, Fresco. 2008. The ideophone in Peking Patois: Sounds & shapes. Leiden: Leiden University Master thesis.
- Sanchez, Gaston. 2012. 5 functions to do Multiple Correspondence Analysis in R. *Visually enforced*. <http://www.gastonsanchez.com/visually-enforced/how-to/2012/10/13/MCA-in-R/> (27 April, 2020).
- Sapir, Edward. 1921. *Language: An introduction to the study of speech*. New York: Harcourt Brace & Co.
- Sasamoto, Ryoko. 2019. *Onomatopoeia and relevance: Communication of impressions via sound*. (Palgrave Studies in Sound). Cham: Palgrave Macmillan.
- Sasamoto, Ryoko & Rebecca Jackson. 2016. Onomatopoeia – showing-word or saying-word? Relevance Theory, lexis, and the communication of impressions. *Lingua* 175–176. 36–53. doi:10.1016/j.lingua.2015.11.003.
- Saussure, Ferdinand de. 2005. *Cours de linguistique générale*. Éd. critique, [Nachdr. der Ausg. 1916]. (Grande bibliothèque Payot). Paris: Payot.
- Schmid, Hans-Jörg. 2000. *English abstract nouns as conceptual shells: From corpus to cognition*. (Topics in English Linguistics 34). Berlin ; New York: Mouton de Gruyter.
- Schmid, Hans-Jörg. 2017. A framework for understanding linguistic entrenchment and its psychological foundations. In Hans-Jörg Schmid (ed.), *Entrenchment and the psychology of language learning: How we reorganize and adapt linguistic knowledge*, 9–35. (Language and the Human Lifespan). Washington, DC : Berlin: American Psychological Association ; De Gruyter Mouton.
- Schmid, Hans-Jörg & Franziska Günther. 2016. Toward a Unified Socio-Cognitive Framework for Salience in Language. *Frontiers in Psychology* 7. doi:10.3389/fpsyg.2016.01110.
- Schmid, Hans-Jörg & Helmut Küchenhoff. 2013. Collostructional analysis and other ways of measuring lexicogrammatical attraction: Theoretical premises, practical problems and cognitive underpinnings. *Cognitive Linguistics* 24(3). 531–577. doi:10.1515/cog-2013-0018.
- Schryver, Gilles-Maurice de & Tshwane Dje. 2009. The lexicographic treatment of ideophones in Zulu. *Lexicos* 19. (AFRILEX). 34–54.
- Schuessler, Axel. 2015. New Old Chinese. *Diachronica* 32(4). 571–598. doi:10.1075/dia.32.4.04sch.
- Sells, Peter. 2017. The significance of the grammatical study of Japanese mimetics. In Noriko Iwasaki, Peter Sells & Kimi Akita (eds.), *The grammar of Japanese mimetics: Perspectives from structure, acquisition and translation*, 7–19. (Routledge Studies in East Asian Linguistics). Milton Park, Abingdon, Oxon ; New York, NY: Routledge.
- Sew, Jyh Wee. 2008. From South Asian echo formation to Cantonese phonetic repetition. *The International Journal of Language, Society and Culture* 24. 72–83.
- Sidhu, David M. & Penny M. Pexman. 2017. Five mechanisms of sound symbolic association. *Psychonomic Bulletin & Review*. doi:10.3758/s13423-017-1361-1.
- Silge, Julia. 2017a. Word vectors with tidy data principles. [2017-30-10].



- https://juliasilge.com/blog/tidy-word-vectors/ (4 November, 2018).
- Silge, Julia. 2017b. Tidy word vectors, take 2! [2017-11-27]. https://juliasilge.com/blog/word-vectors-take-two/ (4 November, 2018).
- Silge, Julia & David Robinson. 2016. Tidytext: Text mining and analysis using tidy data principles in r. *JOSS* 1(3). The Open Journal. doi:10.21105/joss.00037.
- Simone, Raffaele. 1995. Iconic aspects of syntax: A pragmatic approach. In Raffaele Simone (ed.), *Iconicity in language*, 153–169. (Amsterdam Studies in the Theory and History of Linguistic Science 110). Amsterdam ; Philadelphia: J. Benjamins.
- Simpson, J. A. & E. S. C. Wiener (eds.). 1989. *Oxford English Dictionary*. Oxford: Clarendon Press.
- Slobin, Dan I. 2004. The many ways to search for a frog. In Sven Strömqvist & Ludo Verhoeven (eds.), *Relating events in narrative, volume 2: Typological and contextual perspectives*, 219–257. Mahwah, NJ: L. Erlbaum Associates.
- Smith, Jonathan. 2015. Sound Symbolism in the reduplicative vocabulary of the *Shijing*. *Journal of Chinese Literature and Culture* 2(2). 258–285. doi:10.1215/23290048-3324236.
- Soares Costa, Patrício, Nadine Correia Santos, Pedro Cunha, Jorge Cotter & Nuno Sousa. 2013. The use of Multiple Correspondence Analysis to explore associations between categories of qualitative variables in healthy ageing. *Journal of Aging Research* 2013. 1–12. doi:10.1155/2013/302163.
- Sohn, Ho-Min. 2001. *The Korean language*. (Cambridge Language Surveys). Cambridge: Cambridge Univ. Press.
- Staden, Paul M. S. von. 1977. Some remarks on ideophones in Zulu. *African Studies* 36(2). 195–224. doi:10.1080/00020187708707502.
- Stamatatos, Efstatios, Nikos Fakotakis & George Kokkinakis. 2000. Automatic text categorization in terms of genre and author. *Computational Linguistics* 26(4). 471–495.
- Starostin, George. 2015. Review of W. Baxter and L. Sagart, Old Chinese. A New Reconstruction. *Journal of Language Relationship* Bonposы языкового родства 13(4). 383–389.
- Stefanowitsch, Anatol. 2006. Negative evidence and the raw frequency fallacy. *Corpus Linguistics and Linguistic Theory* 2(1). doi:10.1515/CLLT.2006.003.
- Stefanowitsch, Anatol & Stefan Th. Gries. 2003. Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8(2). 209–243. doi:10.1075/ijcl.8.2.03ste.
- Strik Lievers, Francesca & Bodo Winter. 2018. Sensory language across lexical categories. *Lingua* 204. 45–61. doi:10.1016/j.lingua.2017.11.002.
- Sun, Ching Chu, Peter Hendrix, Jianqiang Ma & Rolf Harald Baayen. 2018. Chinese lexical database (CLD): A large-scale lexical database for simplified Mandarin Chinese. *Behavior Research Methods* 50(6). 2606–2629. doi:10.3758/s13428-018-1038-3.
- Sun, Jackson T.S. 孙天心 & Danluo 石丹罗 Shi. 2004. Cǎodèng Jiāróngyǔ de zhuàngmàocí 草登嘉戎语的狀貌词 [Ideophones in rGyalrong Tshob-



- dun]. *Mínzú yǔwén* 民族语文 5. 1–11.
- Sun, Jingtao. 1999. Reduplication in Old Chinese. Vancouver: University of British Columbia PhD dissertation.
- Sūn, Jǐngtáo 孙景涛. 2008. *Gǔ Hanyǔ chóngdié gòucífǎ yánjiū* 古汉语重叠构词法研究 (*A study of reduplicative morphology in Old Chinese*). (Zhōngguó Dāngdài Yǔyán Xué Cóngshū 中国当代语言学丛书). Shanghai: Shanghai jiaoyu chubanshe.
- Sūn, Wàn mí 孙万蜜. 2012. Guānyú ABB shì xíngróngcí zhōng BB de dùyīn wèntí 关于 ABB 式形容词中 BB 的读音问题 (The pronunciation problems of the BB in the ABB construction adjectives). *Yǔyán yánjiū* 语言研究 5. 117–120.
- Sweetser, Eve. 1990. *From etymology to pragmatics: Metaphorical and cultural aspects of semantic structure*. Transferred to digital print. (Cambridge Studies in Linguistics 54). Cambridge: Cambridge Univ. Press.
- Tadmor, Uri. 2009. Loanwords in the world's languages: Findings and results. In Martin Haspelmath & Uri Tadmor (eds.), *Loanwords in the world's languages: A comparative handbook*, 55–75. Berlin, Germany: De Gruyter Mouton.
- Talmy, Leonard. 2000a. *Toward a cognitive semantics: Volume I: Concept structuring systems*. (Language, Speech, and Communication). Cambridge, Mass: MIT Press.
- Talmy, Leonard. 2000b. *Toward a cognitive semantics: Volume II: Typology and process in concept structuring*. (Language, Speech, and Communication). Cambridge, Mass: MIT Press.
- Tamori, Ikuhiro 田守育啓 & Lawrence Schourup. 1999. *Onomatope: Keitai-to imi オノマトペ: 形態と意味* [*Onomatopoeia: Form and meaning*]. Tōkyō: Kuroshio Publ.
- Tang, Ting-Chih 湯廷池. 1985. Guóyǔ xíngróngcí de chóngdié guīlǜ 國語形容詞的重疊規律 [Reduplication rules for adjectives in Mandarin Chinese]. In Tang Ting-Chih 湯廷池 (ed.), *Hanyǔ cífǎ jùfǎ lùnjí* 漢語語法論集 [*Studies on Chinese morphology and syntax*], 29–57. Taipei: Student Book Company.
- Taoka, Chiaki. 2000. Aspect and argument structure in Japanese. Manchester: University of Manchester PhD dissertation.
- Taub, Sarah F. 2001. *Language from the body: Iconicity and metaphor in American Sign Language*. Cambridge, UK New York: Cambridge University Press.
- Taylor, John R. 1988. Contrasting prepositional categories: English and Italian. In Brygida Rudzka-Ostyn (ed.), *Topics in cognitive linguistics*, 299–326. (Amsterdam Studies in the Theory and History of Linguistic Science v. 50). Amsterdam ; Philadelphia: J. Benjamins.
- Taylor, John R. 2003. *Linguistic categorization*. 3rd ed. (Oxford Textbooks in Linguistics). New York: Oxford University Press.
- Taylor, John R. 2004. The ecology of constructions. In Günter Radden & Klaus-Uwe Panther (eds.), *Studies in linguistic motivation*, 49–73. (Cognitive Linguistics Research 28). Berlin ; New York: Mouton de Gruyter.
- Thompson, Arthur Lewis. 2019a. Movement as meaning: An articulatory



- investigation into the iconicity of ideophones. Hong Kong: The University of Hong Kong PhD dissertation.
- Thompson, Arthur Lewis. 2017. Debunking phonaesthemes: The absence of iconicity. In, *CLS-MPI Iconicity Focus Group Workshop: Types of Iconicity in Language Use, Development and Processing*. Nijmegen: Max Planck Institute for Psycholinguistics.
- Thompson, Arthur Lewis. 2018. Are tones in the expressive lexicon iconic? Evidence from three Chinese languages. *PLOS ONE* 13(12). e0204270. doi:10.1371/journal.pone.0204270.
- Thompson, Arthur Lewis. 2019b. Unconventional iconicity can be conventional: Evidence from demonstrations following quotatives in American English. Poster. Paper presented at the ILL 12 [International Symposium on Iconicity in Language and Literature], Lund: Lund University. <https://konferens.ht.lu.se/en/ill-12/>.
- Thompson, Arthur Lewis, Kimi Akita & Youngah Do. Iconicity ratings across the Japanese lexicon: A comparative study with English. In press. *Linguistics Vanguard*.
- Thompson, Arthur Lewis & Youngah Do. 2019. Defining iconicity: An articulation-based methodology for explaining the phonological structure of ideophones. *Glossa: a journal of general linguistics* 4(1). 72. doi:10.5334/gjgl.872.
- Tian, Fei, Hanjun Dai, Jiang Bian, Bin Gao, Rui Zhang, Enhong Chen & Tie-Yan Liu. 2014. A probabilistic model for learning multi-prototype word embeddings. In Junichi Tsujii & Jan Hajic (eds.), *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 151–160. Dublin: Dublin City University and Association for Computational Linguistics.
- Tolskaya, Maria. 2011. Ideophones as positive polarity items. Cambridge (Mass.): Harvard University Master thesis.
- Tomasello, Michael. 2003. *Constructing a language: A usage-based theory of language acquisition*. Cambridge, Mass: Harvard University Press.
- Toratani, Kiyoko. 2005. A cognitive approach to mimetic aspect in Japanese. *Annual Meeting of the Berkeley Linguistics Society* 31(1). 335–346.
- Tran, Truong Huynh Le 陈张黎. 2017. Hanyǔ hé Yuènányǔ nǐshēngcí duìbǐ yánjiū 汉语和越南语拟声词对比研究 [A comparative study of onomatopoeias in Modern Chinese and Vietnamese]. Central China Normal University PhD dissertation.
- Traunmüller, Hartmut. 1994. Sound symbolism in deictic words. In Hans Auli & Peter af Trampe (eds.), *Tongues and texts unlimited. Studies in honour of Tore Jansson on the occasion of his sixtieth anniversary*, 213–234. Stockholm: Dept. of Classical languages, Stockholm University.
- T'sou, B. K. 1978. Sounds symbolism and some socio- and historical linguistic implications of linguistic diversity in Sino-Tibetan languages. *Cahiers de linguistique - Asie orientale* 3(1). 67–76. doi:10.3406/clao.1978.1039.
- Tsujimura, Natsuko & Masanori Deguchi. 2007. Semantic integration of mimesis in Japanese. *Chicago Linguistic Society* 39(1). 339–353.
- Tsur, Reuven. 1992. *What makes sound patterns expressive? The poetic mode of speech perception*. Durham & London: Duke University Press.

- Tuggy, David. 1992. The affix-stem distinction: A Cognitive Grammar analysis of data from Orizaba Nahuatl. *Cognitive Linguistics* 3(3). 237–300.
- Tuggy, David. 2003. Reduplication in Nahuatl: Iconicities and paradoxes. In Eugene H. Casad & Gary B. Palmer (eds.), *Cognitive linguistics and non-Indo-European languages*, 91–133. (Cognitive Linguistics Research 18). Berlin ; New York: Mouton de Gruyter.
- Tuggy, David. 1993. Ambiguity, polysemy, and vagueness. *Cognitive Linguistics* 4(3). 273–290. doi:10.1515/cogl.1993.4.3.273.
- Tummers, Jose, Kris Heylen & Dirk Geeraerts. 2005. Usage-based approaches in Cognitive Linguistics: A technical state of the art. *Corpus Linguistics and Linguistic Theory* 1(2). doi:10.1515/cllt.2005.1.2.225.
- Tyler, Andrea & Vyvyan Evans. 2003. *The semantics of English prepositions: Spatial scenes, embodied meaning, and cognition*. Cambridge ; New York: Cambridge University Press.
- Uhlenbeck, E. M. 1952. The study of wordclasses in Javanese. *Lingua* 3. 322–354. doi:10.1016/0024-3841(52)90028-4.
- Ullmann, Stephen. 1957. *The principles of semantics*. 2nd ed. Glasgow: Jackson, Son & Co.
- Urtel, Herman. 1919. Zur baskischen Onomatopoeis. *Sitzungsberichte der Akademie der Wissenschaften Berlin XIII*.
- Van der Auwera, Johan & Kalyanamalini Sahoo. 2015. On comparative concepts and descriptive categories, *Such as they are*. *Acta Linguistica Hafniensia* 47(2). 136–173. doi:10.1080/03740463.2015.1115636.
- Van Hoey, Thomas. 2014. De *Gǔwén Yùndòng* 古文运动 (Klassieke Prozabeweging) tijdens de Tang en Song: Ideologische én stijlistische vernieuwing? [唐宋古文运动——意识形态以及文体风格的创新？] [The *Gǔwén Yùndòng* 古文运动 (Classical Prose Movement) during Tang and Song: Ideological as well as stylistic innovation?]. Leuven: KU Leuven Master thesis.
- 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. 2018a. Does the thunder roll? Mandarin Chinese meteorological expressions and their iconicity. *Cognitive Semantics* 4(2). 230–259.
- Van Hoey, Thomas. 2020. *CHIDEOD: Data package containing the chinese ideophone database*.
- Van Hoey, Thomas. 2016a. Ideophones in Old Chinese: The case of the *Shījīng* 詩經. In, *ISACG 9 [International Symposium on Ancient Chinese Grammar]*. Berlin: Humboldt University.
- Van Hoey, Thomas. 2016b. Ideophones in Premodern Chinese: Revisiting Dingemanse's implicational hierarchy (poster). In NINJAL (ed.), *Mimetics in Japanese and other languages in the world* (日本語と世界諸言語のオノマトペ). Tachikawa: NINJAL.
- Van Hoey, Thomas. 2017. The thunder rolls: Iconicity and ideophones in Chinese meteorological expressions. Poster. Paper presented at the CLS-MPI Iconicity Focus Group Workshop: Types of Iconicity in Language Use, Development and Processing, Nijmegen: Max Planck Institute for Psycholinguistics.

- Van Hoey, Thomas. 2018b. The blending of bending: Worldbuilding in Avatar: The Last Airbender and The Legend of Korra. Poster. Paper presented at the RaAM (Association for Researching and Applying Metaphor) 12: Metaphor Across Contexts and Domains: From Descriptions to Applications. 26-30 June 2018, Hong Kong: Hong Kong Polytechnic University.
- Van Hoey, Thomas. 2019a. Defining Chinese ideophones: A family of constructions. In, *International Workshop on Mimesis (Ideophones, Expressives) III: Crucibles of Mimesis*. Nagoya: Nanzan University.
- Van Hoey, Thomas. 2019b. Radiant suns, burning fires and brilliant flowers: The onomasiology and radical support of Chinese literary LIGHT ideophones. Oral presentation. Paper presented at the ICLC 15 [International Cognitive Linguistics Conference], Nishinomiya: Kwansei Gakuin University. <https://iclc2019.site/>.
- Van Hoey, Thomas & Iju Hsu. 2020. Chinese ideophones in advertisements and social media: Motivated by multimodal metaphors. Paper presented at the RaAM (Researching and Applying Metaphor) 13: Metaphorical creativity in a multilingual world. 18-21 June 2020, Hamar: Inland Norway University of Applied Sciences (INN).
- Van Hoey, Thomas & Chiarung Lu. 2019a. Lexical variation of ideophones in Chinese classics: Their implications in embodiment and migration. In Janice Fon (ed.), *Dimensions of diffusion and diversity*, 195–226. (Cognitive Linguistics Research 63). Berlin; Boston: De Gruyter Mouton.
- Van Hoey, Thomas & Chiarung Lu. 2018. All that glitters is not gold: Prototypical semantic change in shiny Literary Chinese ideophones. Oral presentation. Paper 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 & Chiarung Lu. 2019b. Reduplication as a trigger of intersubjectivity: Mandarin Chinese ideophones and reduplication in the CHILDES corpora. Oral presentation. Paper presented at the ICLC 15 [International Cognitive Linguistics Conference], Nishinomiya: Kwansei Gakuin University. <https://iclc2019.site/>.
- Van Hoey, Thomas & Arthur Lewis Thompson. 2019. Bridging phonology, meaning, and written form across time: Introducing a database of Chinese literary ideophones. Oral presentation. Paper presented at the ILL 12 [International Symposium on Iconicity in Language and Literature], Lund: Lund University. <https://konferens.ht.lu.se/en/ill-12/>.
- Van Hoey, Thomas & Arthur Lewis Thompson. The Chinese Ideophone-Database (CHIDEOD). In press. *Cahiers de linguistique - Asie orientale*.
- Van Valin, Robert D. & Randy J. LaPolla. 1997. *Syntax: Structure, meaning, and function*. (Cambridge Textbooks in Linguistics). Cambridge, U.K. ; New York, NY: Cambridge University Press.
- Vendler, Zeno. 1957. Verbs and times. *The Philosophical Review* 66(2). 143–160.
- Vidal, Owen Emeric. 1852. Introductory remarks. In Samuel Ajayi Crowther

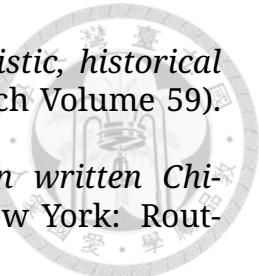
- (ed.), *A vocabulary of the Yoruba language*, 1–38. London: Seeleys.
- Vigliocco, Gabriella & Sotaro Kita. 2006. Language-specific properties of the lexicon: Implications for learning and processing. *Language and Cognitive Processes* 21(7-8). 790–816. doi:10.1080/016909600824070.
- Voeltz, Erhard Friedrich Karl & Christa Kilian-Hatz (eds.). 2001. *Ideophones*. (Typological Studies in Language v. 44). Amsterdam ; Philadelphia: J. Benjamins.
- Wang, Zhijun. 2010. The head of the Chinese adjectives and ABB reduplication. In Lauren Ebyd Clemens & Chi-Ming Louis Liu (eds.), *Proceedings of the 22nd North American Conference on Chinese Linguistics (NACCL-22) & the 18th International Conference on Chinese Linguistics (IACL-18)*: Vol. 1, 232–245. Cambridge (Mass.): Harvard Univ. Press.
- Wang, Zhijun. 2014. The head of the Chinese adjectives and ABB reduplication. *US-China Foreign Language* 12(5). 349–359.
- Watson, Burton. 2003. *Zhuangzi: Basic writings*. (Translations from the Asian Classics). New York: Columbia University Press.
- Watson, Richard L. 1966. Reduplication in Pacoh. *Hartford Studies in Linguistics* 21. 1–138.
- Waugh, Linda R. 1992. Presidential address: Let's take the con out of iconicity. *American Journal of Semiotics* 9(1). 7–47.
- Wáng, Lì 王力. 1984. *Zhōngguó yǔfǎ lǐlùn (1944-1945)* 中国语法理论 (1944-1945) [*Theories of Chinese grammar (1944-1945)*]. (Wáng Lì Wénjí 1). Jīnán: Shāndōng jiàoyǔ chūbǎnshè.
- Wáng, Lì 王力. 1985. *Zhōngguó xiàndài yǔfǎ (1943)* 中国现代语法 (1943) [*A modern grammar of Chinese (1943)*]. (Wáng Lì Wénjí 2). Jīnán: Shāndōng jiàoyǔ chūbǎnshè.
- Wáng, Wànréν 王万仁 (ed.). 1987. *Xiàngshēngcí lì shì 象声词例释* [Examples and explanations of onomatopoeia]. Nánníng: Guǎngxī jiàoyù chūbǎnshè : Guǎngxī xīnhuá shūdiàn fāxíng.
- Wälchli, Bernhard. 2015. *Istiktukai “eventives” — The Baltic precursors of ideophones and why they remain unknown in typology*. In Peter Arkadiev, Axel Holvoet & Björn Wiemer (eds.), *Contemporary approaches to Baltic linguistics*, 491–521. (Trends in Linguistics Studies and Monographs volume 276). Berlin ; Boston: De Gruyter Mouton.
- Webster, Anthony K. 2008. “To Give an Imagination to the Listener”: The neglected poetics of Navajo ideophony. *Semiotica* 171. 343–365. doi:10.1515/SEMI.2008.081.
- Webster, Anthony K. 2017. “So it's got three meanings dil dil:” Seductive ideophony and the sounds of Navajo poetry. *Canadian Journal of Linguistics/Revue canadienne de linguistique* 62(2). 173–195. doi:10.1017/cnj.2017.11.
- Webster, Anthony K. 2014. Rex Lee Jim's 'Na'asts'qosí': On iconicity, interwoven-ness, and ideophones. *Pragmatics and Society* 5(3). 431–444. doi:10.1075/ps.5.3.07web.
- Westermann, Diedrich Hermann. 1905. *Wörterbuch der Ewe-Sprache I. Teil: Ewe-Deutsches Wörterbuch*. Berlin: Dietrich Reimer (Ernst Vohsen).
- Westermann, Diedrich Hermann. 1907. *Grammatik der Ewe-Sprache*. Berlin: Dietrich Reimer. 10.1515/9783111694191.



- Westermann, Diedrich Hermann. 1937. Laut und Sinn in einigen westafrikanischen Sprachen. *Archiv für Vergleichende Phonetik* 1. 154–172, 193–211.
- Whitehead, John. 1899. *Grammar and dictionary of the Bobangi language*. London: Kegan Paul, Trench, Trübner and Co.
- Wickham, Hadley. 2014. Tidy data. *Journal of Statistical Software* 59(10). 1–23. doi:10.18637/jss.v059.i10.
- Wickham, Hadley. 2017. *Tidyverse: Easily install and load the 'Tidyverse'* (v. 1.2.1). <https://CRAN.R-project.org/package=tidyverse>.
- Wielfaert, Thomas, Kris Heylen & Dirk Speelman. 2013. Interactive visualization of semantic vector spaces for lexicological analysis. *TALN-RÉCITAL 2013*. 154–166.
- Wierzbicka, Anna. 1972. *Semantic primitives*. Frankfurt: Athenaeum.
- Wierzbicka, Anna. 1992. Defining emotion concepts. *Cognitive science* 16. 539–581.
- Wilkinson, Endymion Porter. 2015. *Chinese history: A new manual: Fourth edition*. Fourth edition. (Harvard-Yenching Institute Monograph Series 100). Cambridge, Massachusetts: Harvard University Asia Center, for the Harvard-Yenching Institute.
- Williams, Joseph M. 1976. Synaesthetic adjectives: A possible law of semantic change. *Language* 52(2). 461–478.
- Winkler-Breslau, Heinrich. 1907. Elamisch und Kaukasisch. *Orientalistische Literaturzeitung* 10(1-6). 565–573. doi:10.1524/olzg.1907.10.16.287.
- Winter, Bodo. 2019. *Sensory Linguistics: Language, perception and metaphor*. (Converging Evidence in Language and Communication Research 20). Amsterdam: John Benjamins Publishing Company. doi:10.1075/celcr.20.
- Wnuk, Ewelina & Asifa Majid. 2014. Revisiting the limits of language: The odor lexicon of Maniq. *Cognition* 131(1). 125–138. doi:10.1016/j.cognition.2013.12.008.
- Woodin, Greg, Bodo Winter & Jeannette Littlemore. 2019. Degrees of metaphoricity: A large-scale, quantitative analysis of iconic gestures in the TV News Archive. Oral presentation. Paper presented at the ILL 12 [International Symposium on Iconicity in Language and Literature], Lund: Lund University. <https://konferens.ht.lu.se/en/ill-12/>.
- Wu, Mengqi. 2014. The structure of ideophones in Southern Sinitic. Hong Kong: University of Hong Kong Master thesis.
- Xie, Yihui. 2014. Knitr: A comprehensive tool for reproducible research in R. In Victoria Stodden, Friedrich Leisch & Roger D. Peng (eds.), *Implementing reproducible computational research*. Chapman and Hall/CRC. <http://www.crcpress.com/product/isbn/9781466561595>.
- Xie, Yihui. 2015. *Dynamic documents with R and knitr*. Boca Raton, Florida: Chapman and Hall/CRC. <https://yihui.org/knitr/>.
- Xie, Yihui. 2016. *Bookdown: Authoring books and technical documents with R markdown*. Boca Raton, Florida: Chapman and Hall/CRC. <https://github.com/rstudio/bookdown>.
- Xie, Yihui. 2019. *Bookdown: Authoring books and technical documents with r markdown*. <https://github.com/rstudio/bookdown>.



- Xie, Yihui. 2020. *Knitr: A general-purpose package for dynamic report generation in r*. <https://yihui.org/knitr/>.
- Xíng, Fúyì 邢福义. 2004. Nǐyīncí nèibù de yízhìxìng 拟音词内部的一致性 (Internal consistency of onomatopes). *Zhōngguó yǔwén* 中国语文 5(302). 417–429.
- Xue, Shiqi. 1982. Chinese Lexicography Past and Present. *Dictionaries: Journal of the Dictionary Society of North America* 4(1). 151–169. doi:10.1353/dic.1982.0009.
- Xue, Shiqi. 2003. Chinese lexicography past and present. In R. R. K. Hartmann (ed.), *Lexicography: Critical concepts*, 158–173. London ; New York: Routledge.
- Xú, Tiānyún 徐天云. 2000. Liánmiáncí yánjiū de lìshǐ guān yǔ fēi lìshǐ guān 联绵词研究的历史观与非历史观 (Conception and Non-conception of History of Researches on Binomes). *Gǔhànyǔ yánjiū* 古汉语研究 47(2). 14–18.
- Xú, Zhènbāng 徐振邦 (ed.). 2013. *Liánmiáncí dà cídiǎn* 聯綿詞大詞典 [Great dictionary of binomes]. Beijing: Commercial Press.
- Xǔ, Zhōngshū 徐中舒 (ed.). 1995. *Hànyǔ dà zìdiǎn* 漢語大字典 (Great Compendium of Chinese Characters).. 3 vols. Wuhan: Sichuan cishu chubanshe.
- Yang, Yike, Chu-Ren Huang, Sicong Dong & Si Chen. 2018. Semantic transparency of radicals in Chinese characters: An ontological perspective. In, *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*. Hong Kong: Association for Computational Linguistics. <https://www.aclweb.org/anthology/Y18-1091>.
- Yasuoka, Kōichi 安岡孝一. 2018. Kanbun no izonbunpō kaiseki to kaeriten no kankei nitsuite 漢文の依存文法解析と返り点の関係について [On the relationship between dependency grammar analysis of Kanbun]. In, *Proceedings of the 1st Research Conference of the Kanji Society of Japan* 日本漢字学会第1回研究大会予稿集, 33–48. Kyōto: Kyōto University.
- Yáng, Shùsēn 杨树森. 2006. Lùn xiàngshēngcí yǔ tàn cí de chàyxìng 论象声词与叹词的差异性 (On the difference between onomatopoeias and exclamations). *Zhōngguó yǔwén* 中国语文 3(312). 206–215.
- Yáo, Zhànlóng 姚占龙. 2006. Tónggēn ABB shì zhuàngtài xíngróngcí jí qí liàng jí kǎochá 同根 ABB 式状态形容词及其量级考察 (An investigation of statal adjectives of the ABB pattern and their gradability). *Shijè Hànyǔ jiàoxué* 世界汉语教学 3. 72–79.
- Yu, Hongyan 于红岩. 2004. *Yuan qu xuān zhuāngtai xíngróngcí yanjiu* 《元曲选》状态形容词研究 [State words in *Selected Yuan dramas*]. Shanghai: Fudan University PhD dissertation.
- Yu, Weilun 游韋倫. 2015. Nichū ryō gengo ni okeru giongo no imi to imi kakuchō: Furēmu imi-ron no kanten kara no apurōchi 日中両言語における擬音語の意味と意味拡張 — フレーム意味論の観点からのアプローチ — [The meaning and semantic extensions of onomatopoeia in Japanese and Chinese: A Frame Semantics approach]. Kobe: Kobe University PhD dissertation.
- Yutani, Hiroaki. 2018. *Gghighlight: Highlight Lines and Points in “ggplot2”*. <https://CRAN.R-project.org/package=gghighlight>.



- Zhang, Weiwei. 2016. *Variation in metonymy: Cross-linguistic, historical and lectal perspectives*. (Cognitive Linguistics Research Volume 59). Berlin ; Boston: De Gruyter Mouton.
- Zhang, Zheng-sheng. 2017. *Dimensions of variation in written Chinese*. (Routledge Studies in Chinese Linguistics). New York: Routledge/Taylor & Francis Group.
- Zhào, Àiwǔ 赵爱武. 2005. Xiàngshēngcí: Cóng *Shījīng* dào *Yuánqǔ* 象声词：从《诗经》到《元曲》(Mimetic word: From *The Book of Songs* to *Yuan Drama*). *Hénán Kēji Dàxué xuébào* 河南科技大学学报 (*Shèkē bǎn* 社科版) 23(2). 47–50.
- Zhào, Àiwǔ 赵爱武. 2008. Jìn 20 nián Hànyǔ xiàngshēngcí yánjiū zōngshù 近 20 年汉语象声词研究综述 (Chinese onomatopoeia study in recent 20 years: Summary). *Wǔhàn dàxué xuébào* 武汉大学学报 (*rénwén kēxué bǎn* 人文科学版) 61(2). 180–185.
- Zhāng, Dān 张丹. 2005. Hànyǔ zhòng ABB xíng zhuàngtài xíngróngcí de gòuchéng fènxī 汉语中 ABB 型状态形容词的构成分析 [An analysis of the formation of ABB type state adjectives in Mandarin Chinese]. *Journal of Shenyang Agricultural University (Social Sciences Ed.)* 沈阳农业大学学报 (! 社会科学版) 7(1). 124–126.
- Zhou, Xiaolin, Ruth K. Ostrin & Lorraine K. Tyler. 1993. The noun-verb problem and Chinese aphasia: Comments on Bates et al. (1991). *Brain and Language* 45. 86–93.
- Zhu, Hao. 2019. *kableExtra: Construct complex table with 'kable' and pipe syntax*. <https://CRAN.R-project.org/package=kableExtra>.
- Zhū, Déxī 朱德熙. 1956. Xiàndài hànyǔ xíngróngcí yánjiū 现代汉语形容词研究 [A Study of Adjectives in Modern Chinese]. *Yǔyán yánjiū* 语言研究 1.
- Zhū, Déxī 朱德熙. 1961. Shuō *de* 说 “的” [On *De*]. *Zhōngguó yǔwén* 中国语文 12. 1–15.
- Zhū, Déxī 朱德熙 & Shūxiāng 吕叔湘 Lǚ. 1951. *Yǔfǎ xiūcí jiǎnghuà* 语法修辞 讲话. Beijing: Kāimíng shūdiàn 開明書店.
- Zipf, George Kingsley. 1949. *Human behavior and the principle of least effort: An introduction to human ecology*. Cambridge (Mass.): Addison-Wesley Press.
- Zwicky, Arnold M. & Geoffrey K. Pullum. 1987. Plain morphology and expressive morphology. In John Aske, Natasha Beery, Laura Michaelis & Hana Filip (eds.), *Proceedings of the Thirteenth Annual Meeting February 14-16, 1987: General Session and Parasession on Grammar and Cognition*, vol. VII, 330–340. Berkeley, Calif.: Berkely Linguistics Society.

Appendix 1: Data and code used in this dissertation



In this appendix an overview of links to the data and code used in this dissertation will be given.

The Chinese Ideophone Database (CHIDEOD) (Chapter 3)

- URL: <https://osf.io/kpwgf/>
- doi: 10.17605/OSF.IO/KPWGF

App version of CHIDEOD (Chapter 3)

- https://simazhi.shinyapps.io/chideod_appversion/

The Diachronic Chinese Ideophone Corpus (DIACHIC) (Chapter 3)

- URL: https://osf.io/dc3uj/?view_only=c1c25056f1084bb4ab22c9e5c1cd182c

Cue validity for COLLOCATE-IDEOPHONE constructions (Chapter 7)

- https://simazhi.shinyapps.io/ABB_app/

Code and data files of this dissertation

- Available upon request

PDF version of this dissertation

- <https://www.dropbox.com/sh/2pmkq2qc03p06wi/AADHLzfOtb37igDQbaT3pgOY?dl=0>

R packages

In the last version of this dissertation, R 4.0.2 (“Taking Off Again”) was used. A list of packages used in this dissertation with references is shown in Table 8.1.

Table 8.1: R packages used in this thesis

R packages used	Reference to authors
bookdown	Xie (2019); (2016)
ca	Nenadic & Greenacre (2007)
CHIDEOD	Van Hoey (2020)
data.tree	Glur (2019)



R packages used	Reference to authors
DiagrammeR	Iannone (2019)
factoextra	Kassambara & Mundt (2019)
FactoMineR	Lê, Josse & Husson (2008)
fs	Hester & Wickham (2019)
gghighlight	Yutani (2018)
ggpubr	Kassambara (2019)
glue	Hester (2019)
here	Müller (2017)
janitor	Firke (2019)
kableExtra	Zhu (2019)
knitr	Xie (2014); (2015); (2020)
lingtypology	Moroz (2017)
magrittr	Bache & Wickham (2014)
mapview	Appelhans et al. (2019)
Rling	Levshina (2019)
showtext	Qiu (2019)
tidytext	Silge & Robinson (2016)
tidyverse	Wickham (2017)

python libraries

For some parts of the analysis, Python version 3.6.9 was used. The used libraries include the following.

R packages used	Reference
selenium	https://pypi.org/project/selenium/
ckiptagger	CKIP group (2019)
tensorflow	https://www.tensorflow.org/



Appendix 2: Typological data concerning terminology



The map presented in Figure 2.1 in Chapter 2 is based on 62 languages for which ideophones, mimetics, expressives etc. have been described. The data comes from literature surveys in Voeltz & Kilian-Hatz (2001); Dingemanse (2011a; 2018), and Kwon (2015). It is not exhaustive but shows that in Western linguistics, the term IDEOPHONE has become dominant, although there of course often exist specialized words for these phenomena in local languages.

Table 8.3: Coverage and absolute frequency of items in CHIDEOD

language	terminology	source
Adangme	other	Christaller (1888) ; descriptive adverbs
Asheninka Perene	ideophone	Mihas (2012)
Bahnar	expressive	Diffloth (1994)
Baka (Cameroon)	ideophone	Kilian-Hatz (2001)
Bangi	onomatopoeia	Whitehead (1899) ; indeclinable adjectives
Basque	ideophone	Urtel (1917); Ibarretxe-Antuñano (2017); Schuchardt (1919) ; onomatopoeia; ideophones; sound words (Schallwörter)
Central Dagaare	ideophone	Bodomo (2006)
Didinga	ideophone	de Jong (2001)
Elamite	onomatopoeia	Winkler-Breslau (1907) ; Sound figures (Klangfiguren)
Emai-Iuleha-Ora	ideophone	Egbokhare (2001); Schaefer (2001)
English	onomatopoeia	Sapir (1929); Bolinger (1950); Bloomfield (1953); Marchand (1983); Magnus (2001); etc.
Estonian	ideophone	Mikone (2001)

Ewe	other	Westermann (1905: 1907); Schlegel (1857); Ameka (2001) ; sound pictures (Lautbilder); intensity and frequency adverbs (Intensitäts- und Frequenzadverbien)
Finnish	expressive	Jarva (2001); Mikone (2001)
French	expressive	Grammont (1901) ; expressive (mots expressifs)
Gan Chinese	ideophone	Wu (2015) ; 象聲詞 [xiangshengci] “onomatope” ; 擬聲詞 · 擬態詞 · 擬請詞 [nishengci; nitaici; niqingci] “phonomime; phenomime; psychomime”
Gbaya-Bossangoa	ideophone	Samarin (1965); Noss (2001); Roulon-Doko 2001)
Gooniyandi	ideophone	McGregor (2001)
Hakka Chinese	ideophone	Mok (2001); Bodomo (2006); de Sousa (2008); Wu (2015) ; 象聲詞 [xiangshengci] “onomatope” ; 擬聲詞 · 擬態詞 · 擬請詞 [nishengci; nitaici; niqingci] “phonomime; phenomime; psychomime”
Hausa	ideophone	Newman (1968) ; specific intensifying adverbs (Spezifische Verstärkungsadverbien) (Prietze 1908)
Ila	other	Smith (1920) ; echoisms
Iloko	onomatopoeia	Rubino (2001) ; onomatopoetics
Indonesian	expressive	Carr (1966)
Jaminjung	ideophone	Schultze-Berndt (2001)
Japanese	mimetic	Kita (1993; 1997); Lu (2006); Akita (2009); Kita (1997); Mester & Itō (1989); Rodrigues (1604) ; オノマトペ · 擬音語 · 擬態語 · 擬情語 [onomatope; giongo; gitago; gjōgo] “onomatopeia; phonomime; phenomime; psychomime”



Kambera	ideophone	Klamer (1998; 2000)
Kamu	ideophone	Svantesson (1983)
Kanembu	other	(Prietze 1908) ; specific intensifying adverbs (Spezifische Verstärkungsadverbien)
Kedah Malay	expressive	Collins (1979)
Khasi	other	Henderson (1965) ; phonaesthetic words
Kisi	ideophone	Childs (1988)
Korean	ideophone	You (1991); Lee (1992) ; 의성어 [uiseong-eo]; 의태 [uitae]; 의정어 [uijeong-eo]
Kota (India)	onomatopoeia	Emeneau (1969) ; onomatopoetics
Kwini	ideophone	McGregor (2001)
Khoe	ideophone	Kilian-Hatz (2001)
Lao	expressive	Crisfield (1983); Waylang (1996)
Lithuanian	onomatopoeia	Leskien (1902) ; sound imitations (Schallnachahmungen)
Luba-Katanga	ideophone	Kabuta (2001)
Mandarin Chinese	ideophone	Mok (2001); Zhao (2008); Zhang (1999); Yao (2004); Lu (2006); Li (2007); Meng (2012); Van Hoey (2015; 2017) ; 象聲詞 [xiangshengci] “onomatope” ; 擬聲詞·擬態詞·擬請詞 [nishengci; nitaici; niqingci] “phonomime; phenomime; psychomime”
Mundang	ideophone	Elders (2001)
Myene	onomatopoeia	Wilson (1847) ; (onomatopoeic) interjections
Northern Pastaza Quichua	ideophone	Nuckolls (1992; 1996; 2004; 2017)
Nyanja	ideophone	Kulemeka (1994) (lang: Chichewa); 1996; 1997)
Pacoh	ideophone	Watson (1966)
Semai	expressive	Diffloth (1976); (Tufvesson 2011)
Shona	ideophone	Fortune (1962); Klassen (1999)
Siwu	ideophone	Dingemanse (2011)
Somali	ideophone	Dhoorre & Tosco (1998)



Southern Sotho	ideophone	Kunene (1965)
Tetela	ideophone	Tassa (2001)
Tswana	ideophone	Creissels (2001)
Upper Necaxa Totonac	ideophone	Beck (2008)
Vietnamese	impressif	Durand (1961)
Warrwa	ideophone	McGregor (2001)
Wolaytta	ideophone	Amha (2001)
Wu Chinese	ideophone	Wu (2015) ; 象聲詞 [xiangshengci] “onomatope” ; 擬聲詞·擬態詞·擬請 詞 [nishengci; nitaici; niqingci] “phonomime; phenomime; psychomime”
Xhosa	onomatopoeia	McLaren (1906) ; indeclinable verbal particles
Xiang Chinese	ideophone	Wu (2015) ; 象聲詞 [xiangshengci] “onomatope” ; 擬聲詞·擬態詞·擬請 詞 [nishengci; nitaici; niqingci] “phonomime; phenomime; psychomime”
Yir-Yoront	ideophone	Alpher (1994; 2001)
Yoruba	ideophone	Rowlands (1970) ; specific adverbs (Vidal 1852)
Yue Chinese	ideophone	Mok (2001); Bodomo (2006); de Sousa (2008) ; 象聲詞 [xiangshengci] “onomatope” ; 擬 聲詞·擬態詞·擬請詞 [nishengci; nitaici; niqingci] “phonomime; phenomime; psychomime”
Zulu	ideophone	Doke (1935) ; Msimang & Poulos (2001) ; radical descriptives (Doke 1927)

Appendix 3: Recategorizing semantic radicals

Table 8.4 shows all recategorized semantic radicals, as used in Sections 4.4 and Section 4.5. As explained in Section 3.2.3.2, these radicals are based on the simplified form of the character because that is how they were stored in the Chinese Lexical Database (Sun et al. 2018). In Table 8.4, the variable ‘radical_support’ refers to #radical_support in CHIDEOD. Consequently, ‘radical_support_typefreq’ shows the type frequency of these patterns. The variable ‘radical’ represents the recategorized variable, keeping those values where the following condition was met: radical_support_typefreq > 25. As can be seen, the biggest group in ‘radical’ is “norad”, meaning that there is no radical support. Next we find a number of radicals like “mouth”, “water”, etc. And finally, when the condition is not met, they are categorized as “otherrad”.

Table 8.4: Distribution of the radicals participating in the MCA of CHIDEOD

radical	radical_typefreq	radical_support	radical_support_typefreq
norad	2379	NA	2379
mouth	926	口	926
water	358	氵	346
grass	123	艹	123
heart	130	忄	98
mountain	82	山	82
body	278	亼	80
body	278	足	69
body	278	扌	64
wood	61	木	61
woman	51	女	51
silk	46	糸	46
jade	40	王	40
body	278	辵	40
metal	40	金	40
fire	63	火	38
stone	37	石	37
speak	36	言	36
sun	35	日	35
heart	130	心	32



moon	32	月	32
eye	31	目	31
feather	27	羽	27
body	278	彳	25
fire	63	灬	25
otherrad	707	虫	25
otherrad	707	門	21
otherrad	707	阝	21
otherrad	707	車	20
otherrad	707	雨	19
otherrad	707	馬	19
otherrad	707	旳	18
otherrad	707	土	18
otherrad	707	穴	18
otherrad	707	竹	17
otherrad	707	宀	14
otherrad	707	田	14
otherrad	707	衣	14
otherrad	707	匚	13
otherrad	707	巾	12
water	358	水	12
otherrad	707	酉	12
otherrad	707	大	11
otherrad	707	欠	11
otherrad	707	白	11
otherrad	707	黑	11
otherrad	707	丶	9
otherrad	707	彑	9
otherrad	707	戈	8
otherrad	707	方	8
otherrad	707	夊	8
otherrad	707	禾	8
otherrad	707	耳	8
otherrad	707	隹	8
otherrad	707	貞	8
otherrad	707	儿	7
otherrad	707	刀	7



otherrad	707	犬	7
otherrad	707	走	7
otherrad	707	力	6
otherrad	707	匚	6
otherrad	707	厂	6
otherrad	707	弓	6
otherrad	707	手	6
otherrad	707	爻	6
otherrad	707	文	6
otherrad	707	糸	6
otherrad	707	齒	6
otherrad	707	一	5
otherrad	707	又	5
otherrad	707	子	5
otherrad	707	尸	5
otherrad	707	广	5
otherrad	707	聿	5
otherrad	707	钅	5
otherrad	707	鹿	5
otherrad	707	一	4
otherrad	707	二	4
otherrad	707	ト	4
otherrad	707	毛	4
otherrad	707	讠	4
otherrad	707	革	4
otherrad	707	页	4
otherrad	707	克	3
otherrad	707	口	3
otherrad	707	厂	3
otherrad	707	十	3
otherrad	707	口	3
otherrad	707	小	3
otherrad	707	戶	3
otherrad	707	止	3
otherrad	707	歹	3
otherrad	707	气	3
otherrad	707	广	3



otherrad	707	皿	3
otherrad	707	𠔿	3
otherrad	707	舌	3
otherrad	707	谷	3
otherrad	707	阜	3
otherrad	707	雲	3
otherrad	707	香	3
otherrad	707	彫	3
otherrad	707	鳥	3
otherrad	707	丿	2
otherrad	707	匚	2
otherrad	707	匚	2
otherrad	707	寸	2
otherrad	707	平	2
otherrad	707	幺	2
otherrad	707	升	2
otherrad	707	斤	2
otherrad	707	殳	2
otherrad	707	爪	2
otherrad	707	牛	2
otherrad	707	𠂇	2
otherrad	707	矢	2
otherrad	707	立	2
otherrad	707	至	2
otherrad	707	虍	2
otherrad	707	行	2
otherrad	707	角	2
otherrad	707	貝	2
otherrad	707	贝	2
otherrad	707	赤	2
otherrad	707	邑	2
otherrad	707	食	2
otherrad	707	马	2
otherrad	707	丨	1
otherrad	707	乙	1
otherrad	707	云	1



otherrad	707	人	1
otherrad	707	八	1
otherrad	707	几	1
otherrad	707	匚	1
otherrad	707	匚	1
otherrad	707	厄	1
otherrad	707	士	1
otherrad	707	夭	1
otherrad	707	委	1
otherrad	707	尺	1
otherrad	707	屮	1
otherrad	707	《《	1
otherrad	707	爻	1
otherrad	707	彑	1
otherrad	707	支	1
otherrad	707	曰	1
otherrad	707	母	1
otherrad	707	爻	1
otherrad	707	牙	1
otherrad	707	玄	1
otherrad	707	玉	1
otherrad	707	瓦	1
otherrad	707	生	1
otherrad	707	用	1
otherrad	707	由	1
otherrad	707	疋	1
otherrad	707	矛	1
otherrad	707	示	1
otherrad	707	米	1
otherrad	707	缶	1
otherrad	707	羊	1
otherrad	707	𦥑	1
otherrad	707	肅	1
otherrad	707	肉	1
otherrad	707	西	1
otherrad	707	见	1
otherrad	707	车	1



otherrad	707	采	1
otherrad	707	门	1
otherrad	707	青	1
otherrad	707	非	1
otherrad	707	面	1
otherrad	707	风	1
otherrad	707	飞	1
otherrad	707	骨	1
otherrad	707	高	1
otherrad	707	鬼	1
otherrad	707	魚	1
otherrad	707	鱼	1
otherrad	707	鸟	1
otherrad	707	黃	1
otherrad	707	鼎	1
otherrad	707	鼻	1

Appendix 4: Reproduction of selected figures



Variable categories - MCA

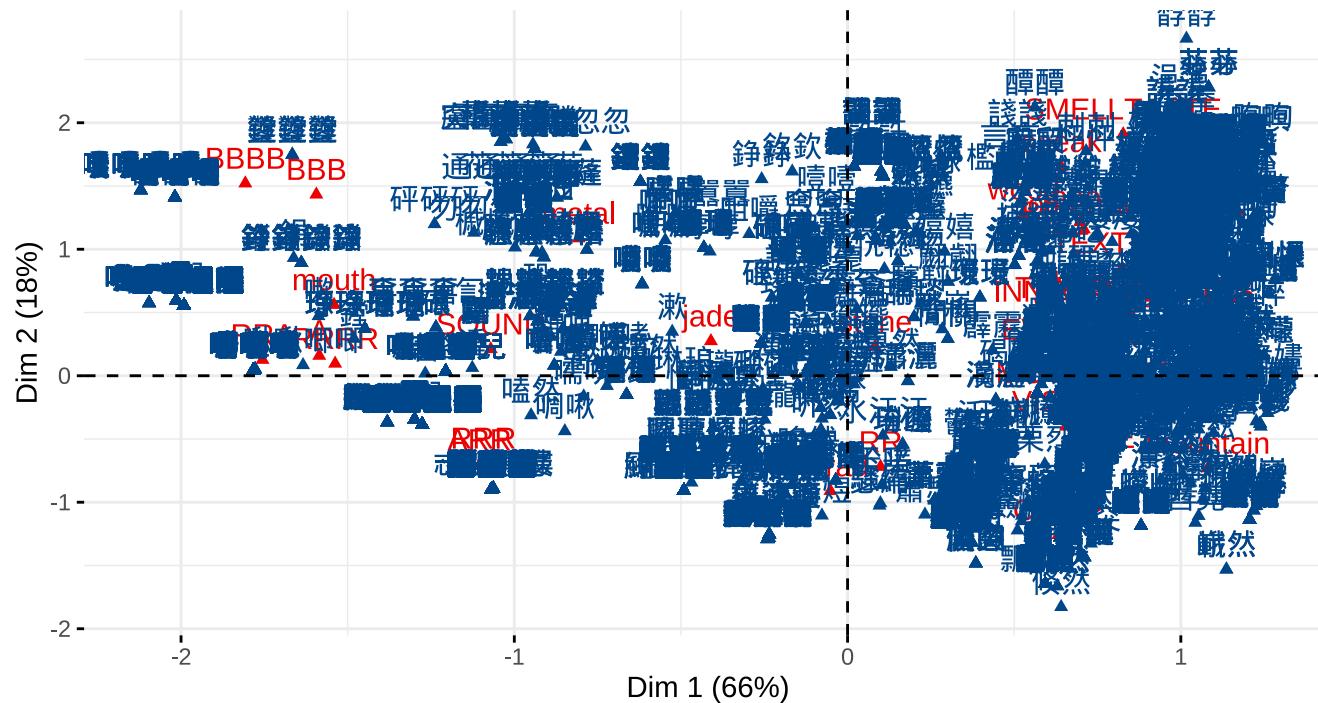


Figure 8.1: MCA plot of the barycenters of the **CHIDEOD** data with the supplementary values (the ideophones themselves) showing. Reproduction of Figure 4.9.

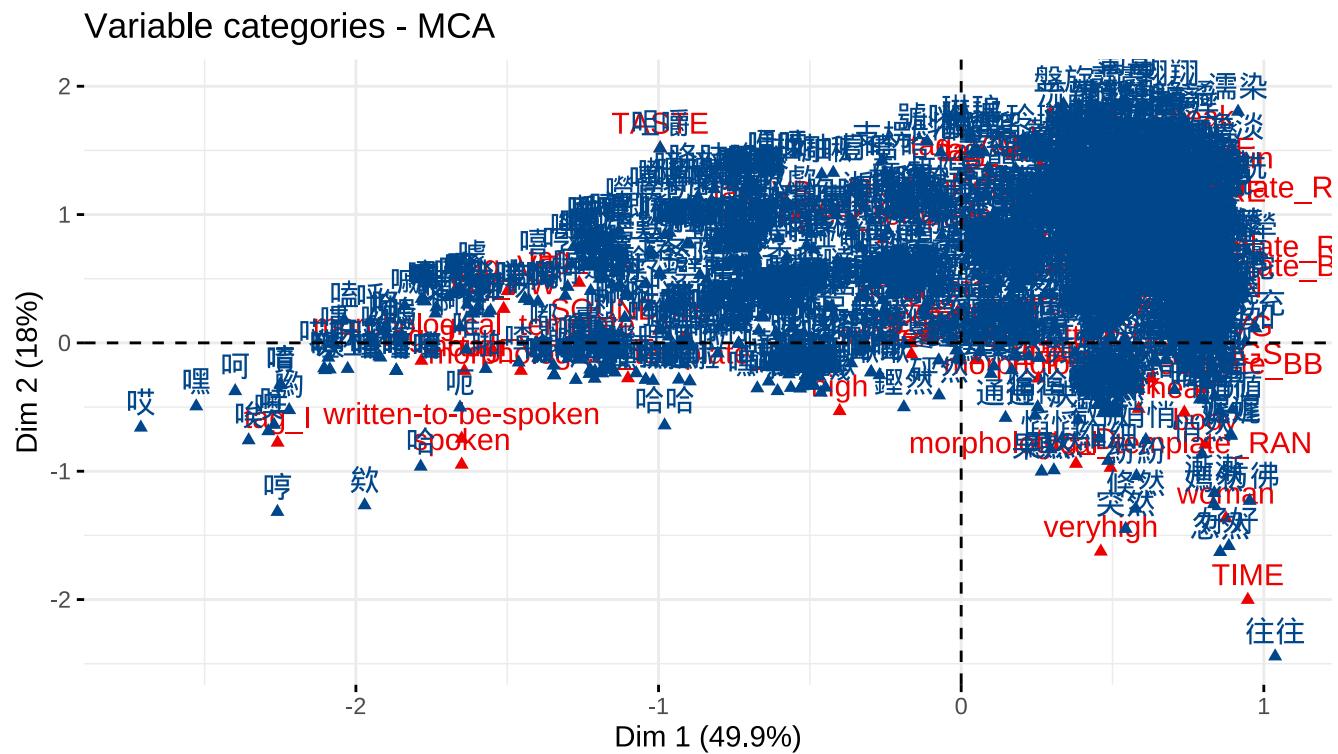


Figure 8.2: MCA plot of the barycenters of the **CHIDEOD-ASBC data** with the supplementary values (the ideophones themselves) showing. Reproduction of Figure 4.18.