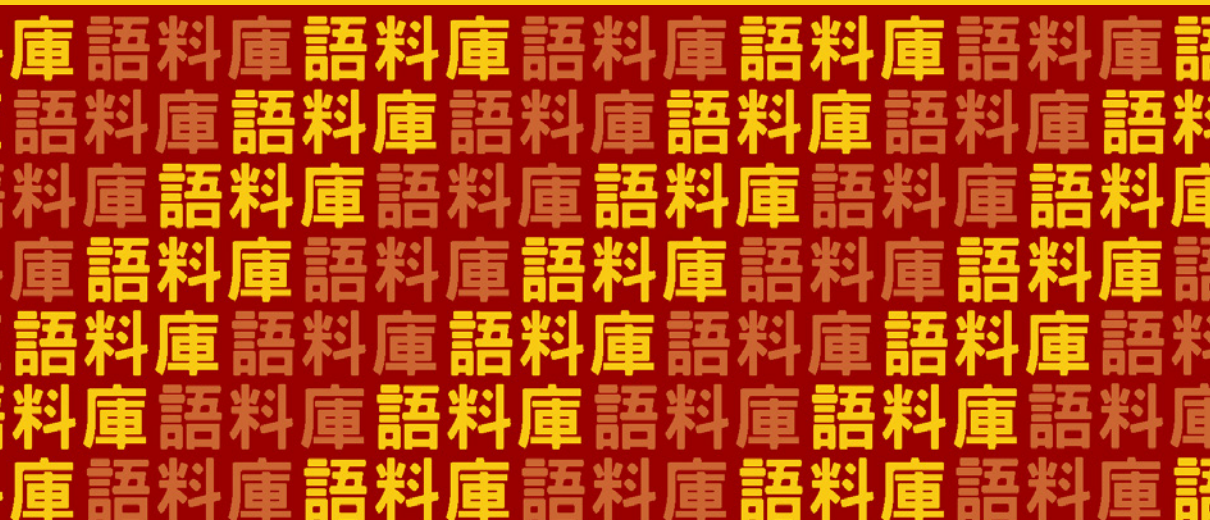
Sinica venetiana 6

e-ISSN 2610-9042
ISSN 2610-9654

Corpus-Based Research on Chinese Language and Linguistics

edited by
Bianca Basciano, Franco Gatti, Anna Morbiato

Edizioni
Ca'Foscari



Corpus-Based Research on Chinese Language and Linguistics

Sinica venetiana

Serie diretta da
Tiziana Lippiello e Chen Xiaoming

6

Edizioni
Ca'Foscari

Sinica venetiana

Direzione e redazione | Head office
Dipartimento di Studi sull'Asia e sull'Africa Mediterranea
Università Ca' Foscari Venezia
Palazzo Vendramin dei Carmini
Dorsoduro 3462
30123 Venezia
Italia

e-ISSN 2610-9042
ISSN 2610-9654

URL https://edizionicafoscari.unive.it/it/edizioni/collane/sinica-venetiana/

Corpus-Based Research on Chinese Language and Linguistics

edited by
Bianca Basciano, Franco Gatti, Anna Morbiato

Venezia
Edizioni Ca' Foscari - Digital Publishing
2020

Corpus-Based Research on Chinese Language and Linguistics
Bianca Basciano, Franco Gatti, Anna Morbiato (edited by)

Edizioni Ca' Foscari - Digital Publishing
Fondazione Università Ca' Foscari Venezia | Dorsoduro 3246 | 30123 Venezia
http://edizionicafoscari.unive.it | ecf@unive.it

1st edition December 2020
ISBN 978-88-6969-406-6 [ebook]
ISBN 978-88-6969-407-3 [print]

Certificazione scientifica delle Opere pubblicate da Edizioni Ca' Foscari - Digital Publishing: tutti i saggi pubblicati hanno ottenuto il parere favorevole da parte di valutatori esperti della materia, attraverso un processo di revisione anonima sotto la responsabilità del Comitato scientifico della collana. La valutazione è stata condotta in aderenza ai criteri scientifici ed editoriali di Edizioni Ca' Foscari.

Scientific certification of the works published by Edizioni Ca' Foscari - Digital Publishing: all essays published in this volume have received a favourable opinion by subject-matter experts, through an anonymous peer review process under the responsibility of the Scientific Committee of the series. The evaluations were conducted in adherence to the scientific and editorial criteria established by Edizioni Ca' Foscari.

Corpus-Based Research on Chinese Language and Linguistics / Bianca Basciano, Franco Gatti, Anna Morbiato (edited by) — 1. ed. — Venezia: Edizioni Ca' Foscari - Digital Publishing, 2020. — 366 pp; 23 cm. — (Sinica venetiana; 6). — ISBN 978-88-6969-407-3

URL https://edizionicafoscari.unive.it/en/edizioni/libri/978-88-6969-407-3/
DOI http://doi.org/10.30687/978-88-6969-406-6

Corpus-Based Research on Chinese Language and Linguistics
edited by Bianca Basciano, Franco Gatti, Anna Morbiato

Table of Contents

MORPHOLOGY AND THE LEXICON

SOCIOLINGUISTICS

CORPUS AND DATABASE BUILDING

Corpus-Based Research on Chinese Language and Linguistics
edited by Bianca Basciano, Franco Gatti, Anna Morbiato

Introduction

Bianca Basciano
Università Ca' Foscari Venezia, Italia

Franco Gatti
Università Ca' Foscari Venezia, Italia

Anna Morbiato
Università Ca' Foscari Venezia, Italia; The University of Sydney, Australia

In the past decades, corpus-based research has been gaining momentum in contemporary linguistics. While corpora, intended as large collections of naturally occurring texts, have always existed, rapid advances in computation and technology have provided tools for faster and more effective corpus construction and consultation. Chinese makes no exception: corpus data are now considered among the main resource for many linguists, while large-scale surveys are beginning to be taken as an important tool for linguistic investigation.

Among the reasons beyond the increasing number of corpus-based studies is the availability of "a myriad of large and publicly available Chinese corpora" (Xu 2015, 219), which include general purpose corpora, such as the CCL (Centre for Chinese Linguistics, Peking University) corpus or the BCC (Beijing Languages and Cultures University) corpus, interlanguage corpora, such as the BLCU International Corpus of Learner Chinese, and specialised corpora, such as the ZHTenTen simplified Chinese corpus mounted at Sketch Engine, the LDC (Linguistic Data Consortium at UPenn) or the ELRA (European Language Resources Association). Smaller, genre- or domain-specific corpora, such as e.g. the Leiden Weibo Corpus or the *Renmin Ribao* 'People's Daily' database, are also growing in number. Other resources include multilingual corpora and databases – e.g. a number

of English-Chinese parallel corpora, translational Chinese corpora, Chinese dialects databases and corpora of ethnic languages in China.

The great advantage of corpora lies in the fact that they offer access to large amounts of authentic, naturally occurring linguistic data produced by a variety of speakers or writers, thus providing more robust, statistically significant foundations for linguistic accounts and analyses. There is now considerable emphasis on the reliability of linguistic materials: several scholars stress the need for a shift to a more empirical mode of investigation, as rigorous theoretical advances need to be "grounded in solid empirical data" (Jing-Schmidt 2013, 1). A further advantage is that corpus queries may also reveal the statistical relevance of a specific linguistic phenomenon, e.g. a lexical item or a grammatical pattern, as well as possible changes or developments of its behaviours over time. Moreover, corpus queries may also allow searching for significant interactions between domain variables (Wallis, Nelson 2001). Finally, these tools may help reveal new words or patterns that were previously unobservable or, else, regarded as non-existent or marginal. In short, corpora allow qualitative and quantitative, synchronic and diachronic investigations of the language, providing factual, frequency, and interaction evidence for linguistic analyses (Wallis 2019). They not only offer new insights within the core subfields of linguistics – including syntax, semantics and lexicography, pragmatics and language use, information structure – but also provide precious material for disciplines such as language acquisition, with the analysis of learners' corpora and interlanguage development, or sociolinguistics, with synchronic and diachronic studies on language and society, socio-linguistic comparison, as well as the development of buzzwords in social media and the Internet.

The past decade has seen the rapid development of corpus-based research in many aspects of Chinese language and linguistics. One of the most popular types of research is the compilation of frequency character/word lists (Xu 2015): after Li Jinxi's *A Statistical Analysis of Basic Chinese Vocabulary* (1922), lexical studies received increasing interest, with many scholars applying corpus tools to all aspects of lexicography, including selecting words to be included in a dictionary on a statistical basis, identifying word senses, ordering of polysemous and homograph items, as well as determining word classes and singling out illustrative examples of words' uses (see McEnery, Xiao 2016, 442). Among the most recent lexical frequency and word list projects, there are the latest national Chinese character list, i.e. the 通用规范汉字表 *Tōngyòng Guīfàn Hànzì Biǎo* (A General Service List of Chinese Characters), released in 2013, and Xiao, Rayson and McEnery's (2009) *A Frequency Dictionary of Mandarin Chinese* (see McEnery, Xiao 2016 for a review). Corpus-based researches on second language acquisition and interlanguage development have also been increasing over the last couple of decades, with early projects at

BLCU now developed into the BLCU International Corpus of Learner Chinese, followed by other studies (Tao 2008, 2009; Xiao 2007; Zou, Smith, Hoey 2016, *inter alia*; for a review, see Xu 2015; McEnery, Xiao 2016; Zhang, Tao 2018). On the other hand, scholars agree that corpus-based sentential/grammatical level research is practically negligible if compared with lexical studies, although it is now receiving increasing attention with the introduction of more sophisticated query tools. For example, there have been some innovative corpus studies on morphological aspects of Chinese, e.g. on compounds and affixes (Sproat, Shih 1996; Nishimoto 2003; Arcodia, Basciano 2012) and on 离合词 *líhécí* 'separable words' (Siewierska, Xu, Xiao 2010; Wang C. 2001, Wang H. 2011). With respect to syntax, remarkable insights have been gained by scholars using corpora on syntactic patterns and behaviours of, e.g. adjectives (Thompson, Tao 2010), adverbial clauses (Wang 2006), and verbal coercion (Tao 2000). Interesting work has also been done on discourse/pragmatics (Jing-Schmidt, Kapatsinsky 2012). Contrastive studies also constitute a promising line of research, with main works done on the differences between English and Chinese (Xiao, McEnery 2008, 2010). Other significant areas of inquiry include corpus and database construction (Zhan 2019) and historical linguistics (Halliday 1959; Cook 2011; Ji 2010); for an overview, see Xu (2015). However, apart from these notable exceptions, Chinese corpus-based theoretical linguistics studies are scarce and by no means the mainstream (Xu 2015), partly due to the technological and methodological limitations connected with corpus interrogation. McEnery and Xiao (2016) also hold that research in corpus-based descriptive grammar in Chinese is rather sporadic and fragmentary, and has focused on specific linguistic features of interest to individual researchers.

This volume wants to contribute to filling this gap and stems from the idea that a lot can still be done: issues that have not received a commonly accepted account may benefit from corpus-based investigation conducted from a different angle, qualitative and/or quantitative; second, corpora may reveal linguistic phenomena, patterns and constructions that have not yet been investigated, thus enriching our knowledge of grammar; finally, new corpora or corpus-tagging methods that allow more precise analyses in specific research fields, ranging from diachronic linguistics to sociolinguistics, syntax and pragmatics, can be identified and suggested for future lines of research.

Studies presented in this volume are both quantitative and qualitative, as well as synchronic and diachronic, and are grounded in the tenet that corpora provide a more robust, statistically significant foundation for linguistic analyses. As corpus linguistics is not a monolithic, consensually agreed set of methods and procedures (McEnery, Hardie 2011), differences inevitably exist regarding approaches and methodologies in the different contributions, which may be both

discipline-specific and also due to the different aim and focus of each study. The contributions provide different insights not only into the potential of using corpora as tools allowing access to authentic language material, but also into the challenges involved in corpus interrogation, analysis, and building. All in all, they contribute to answering three fundamental questions: how can corpora improve current theoretical accounts of Chinese grammar in general? What do corpora reveal about the statistical relevance of linguistic phenomena and constructions? What are the limitations and the drawbacks of using corpora to investigate Chinese languages?

As reflected in the five sections of the volume, the contributions cover different fields of linguistics, including syntax and pragmatics, semantics, morphology and the lexicon, sociolinguistics, and corpus building.

The first section explores issues in Chinese syntax and pragmatics. Tao, Jin and Zhang's paper proposes an investigation of manner and state complement constructions combining corpus-based and corpus-driven methods, based on a corpus of written Chinese, offering both a theoretical account and an exploration of the implications for Chinese L2 learning. The study highlights preferred forms and functions of Manner/State Complement Constructions: monosyllabic verbs, basic action verbs, or psychological state verbs tend to co-occur with complements of adjectival, clausal, or idiomatic expressions. The authors conclude that Manner/State Complement Constructions are an assessment device indexing speaker evaluative stance, and that the loaded affective meanings account for the larger and more complex forms than their standard counterparts.

Morbiato provides quantitative and qualitative evidence of the existence of indefinite NPs in the sentence-initial and preverbal position, thus ruling out strict associations between definiteness, givenness, and the sentence-initial position and related restrictions often referred to in the literature. She examines big-size, generalised corpora, such as the PKU CCL corpus (Peking University), the BCC corpus (Beijing Language and Culture University), and the ZHTenTen (Stanford Tagger) corpus mounted at Sketch Engine. Her statistical data show that this phenomenon is neither rare nor marginal. Furthermore, they reveal that animate indefinites are significantly more likely to occur sentence-initially, while locatability and partitivity are frequent traits of inanimate SIIs. Finally, it singles out and discusses a new pattern featuring a proper noun introduced by the indefinite marker '一 *yī* + CLASSIFIER', thus confirming that corpora indeed contribute towards a more complete understanding of a language system by allowing to single out new, previously underdescribed linguistic patterns and phenomena.

Tantucci and Wang explore the V-过 *guo* construction by examining its evidential *versus* experiential usages in two comparable writ-

ten corpora, i.e. the Lancaster Corpus of Mandarin Chinese and the UCLA corpus of written Mandarin. The results of this study shed light on the relationship between the formal and functional categories of the V-过 *guo* construction and the textual environment in which it occurs, showing that specific genres and textual environments favour the evidential usage of 过 *guo* and that evidentiality is an important grammatical category of documentary, factual and academic prose. This study also shows that the categorial separation between evidential and experiential usages of the construction is a result of features underpinning form, usage and 'contextual situatedness'. The authors conclude that evidentiality emerges from specific intersections among these three dimensions and from distinctive illocutional concurrences of conventionalized behaviour.

The second section is devoted to semantic studies. Shi, Liu and Jing-Schmidt present a usage-based, quantitative and qualitative corpus investigation of action metaphors involving manual object manipulation. Two transitive constructions, [抓紧 *zhuājǐn* 'grab tightly, clutch' NP] and [把住 *bǎzhù* 'grasp firmly' NP], and a causative construction, [把 *bǎ* NP 捧 *pěng* COMPL] 'lift NP with deliberation' (with a metaphoric sense), are examined: results reveal that the former systematically imply a keen sense of urgency and/or importance, while the latter involves over-promotion of an undeserving entity. The study highlights the methodological importance of quantitative studies in establishing the conventionality, productivity, and semantic subclassification of metaphors encoded in syntactic patterns. It has both implications for theoretical hypotheses regarding the embodiment of conceptualisation and for language learning and teaching.

The contribution by Sparvoli focuses on modality, in particular on the factuality reading triggered by Chinese modals in past contexts. Through a corpus-based investigation, conducted in the English Chinese Parallel Concordancer, published by the Hong Kong Institute of Education, the author tests the hypothesis that deontic modals trigger counterfactual inference, while anankastic/goal-oriented modals either trigger an actuality entailment effect or a generic non-factual reading. The results of her investigation confirm the crucial role played by the deontic vs. anankastic contrast in the marking of factuality in Chinese, showing a gradient cline, from anankastic/goal-oriented modals to deontic modals, along which the factuality value decreases. The two extreme poles of the cline get a unique reading, i.e. past counterfactual for pure deontic modals and factual for strong anankastic modals. Finally, some pedagogical implications are discussed.

Boaretto and Castello propose a corpus-based study of Chinese modality by comparing the English and Chinese versions of Pope Francis' second encyclical *Laudato Si'*, focusing on different areas of modality, i.e. prediction/volition/intention, lack of possibility/abil-

ity/permission, and obligation. Meaningful translation correspondences are investigated to define their semantic space and detect possible cases of explicitation. While corpus data confirm predictable parallel expressions such as *will* and 会 *huì*, *cannot* and 不能 *bù néng*, they also reveal new correspondences, such as no overt modal expression in English and 会 *huì*, or *cannot* and 无法 *wúfǎ*. Overall, the study highlights how the translation of highly grammaticalised items undergoes a process of interpretation and adaptation: some translation choices are due to the translator's attempt to make the text explicit and to adapt it to the target culture. The corpus-based approach adopted reveals a network of semantically connected modal expressions and helps to identify the linguistic choices made by the writer and the translator to convey the intended semantic meanings. The authors point out that, while parallel concordancing software could help speed up this type of analysis, human scrutiny and judgement are still needed.

The third section proposes research into the lexicon and morphology of Chinese. Specifically, Dosedlová and Lu propose a corpus-based study on near-synonymy of classifiers: in Chinese there are many classifiers which are near-synonymous and interchangeable in some contexts. In particular, the study investigates two near-synonymous classifiers, i.e. 棵 *kē* and 株 *zhū*, based on co-varying collexeme analysis, which belongs to collostrucional methods (i.e. corpus-based quantitative methods which measure mutual attraction between lexemes and constructions), and on Euclidean distance. Such an approach allows to obtain a clearer picture on the co-occurrence of certain classifiers with certain nouns and on different usages. However, the authors suggest that it is highly recommendable to combine different methodological approaches for the analysis of near synonymy, in order to obtain a more comprehensive picture, able to reveal different aspects of the phenomenon.

The contribution by Basciano and Bareato focuses on word-formation, specifically, on new word-formation patterns emerged in the last few decades under the influence of foreign languages and netspeak. The authors present a corpus-based investigation on three emerging suffixes, i.e. 族 *zú*, 党 *dǎng*, and 客 *kè*, all forming nouns indicating persons with certain characteristics or behaviour, or doing a certain activity, examining neologisms drawn from the following sources: the 新世纪新词语大词典 *Xin shiji xinciyu da cidian* (New Century Comprehensive Dictionary of Neologisms), the Leiden Weibo Corpus, and the Buzzwords section of the *Shanghai Daily*. After describing the three word-formation patterns, the paper describes their evolution over time, and the semantic shift and meaning generalisation characterising their grammaticalization path. The study also proposes an analysis of productivity measures for the three word-formation patterns and discusses their diffusion in Chinese.

The fourth section explores applications of corpus tools to the investigation of sociolinguistic aspects. Specifically, Chin proposes a novel use of the Corpus of Mid-20th Century Hong Kong Cantonese, i.e. as a window on Hong Kong society, and specifically its family structure and marital life. It consists of a corpus-based sociolinguistic investigation of kinship terms and terms related to marriage, which reveals significant differences in family structure as compared to contemporary Hong Kong society.

The fifth section tackles issues on corpus and database construction. Zhan et al. present their work in progress and the challenges encountered in the creation of a Chinese constructicon provisionally named CCL-CxnBank. The project has been carried out since 2015 by the Center for Chinese Linguistics of Peking University and, at the moment, the constructicon includes more than 1,000 constructions and records their syntactic, semantic, and pragmatic information, as well as synonymy, antonymy, and hyponymy/hypernymy relations. In addition, the project includes the annotation of a corpus collecting instances of various usages of the constructions in real contexts: the corpus annotates the internal structure and the subjective attitude meaning of each construct, in order to provide a comprehensive description of the actual usages of the constructions.

Lastly, Anderl presents some reflections on the Database of Medieval Chinese Texts, an international and collaborative project, drawing on the expertise of specialists in various fields, the main partners being Ghent University and Dharma Drum Institute of Liberal Arts (Taiwan). The database collects manuscript texts, with a focus on the period between ca. 700 and 1000 CE. While there is a variety of digital databases for premodern Chinese texts, specialised databases on non-canonical manuscripts are still very rare and provide rather limited information. Therefore, this project is very valuable, since it aims at providing high-quality digital editions of Late Medieval Chinese key texts, which are of great importance for research on early colloquial grammatical markers and syntactic constructions, also developing an analytical apparatus. The paper presents the technical framework, the reference data collections, the process of digitalisation of the texts, the various modules of the database, and proposes some reflections. The paper also discusses the importance of the database as a pedagogical tool.

We would like to thank all the anonymous reviewers for their precious help. We would also like to express our heartfelt gratitude to Magda Abbiati and Federico Greselin for their generous support. Lastly, we wish to express our gratitude to the editorial staff of Edizioni Ca' Foscari.

Bibliography

Arcodia, G. F.; Basciano, B. (2012). "On the Productivity of the Chinese Affixes –兒 *–r*, –化 *–huà* and –頭 *–tou*". *Taiwan Journal of Linguistics*, 10, 89-118. http://dx.doi.org/10.6519/TJL.2012.10(2).3.

Cook, A. (2011). "Recent Developments in the Use of the Plural Marker *Men* in Modern Standard Chinese in Taiwan". *Chinese Language and Discourse*, 2(1), 80-98. https://doi.org/10.1075/cld.2.1.04coo.

Ji, M. (2010). "A Corpus-Based Study of Lexical Periodization in Historical Chinese". *Literary and Linguistic Computing*, 25(2), 199-213. https://doi.org/10.1093/llc/fqq002.

Jing-Schmidt, Z. (2013). *Increased Empiricism: Recent Advances in Chinese Linguistics*. Amsterdam: John Benjamins.

Jing-Schmidt, Z.; Kapatsinsky, V. (2012). "The Apprehensive: Fear as Endophoric Evidence and Its Pragmatics in English, Mandarin, and Russian". *Journal of Pragmatics*, 44, 346-73. https://doi.org/10.1016/j.pragma.2012.01.009.

Halliday, M. (1959). *The Language of the Chinese "Secret History of the Mongols"*. Oxford: Basil Blackwell.

Li J. 黎锦熙 (1922). "Guoyu zhong jiben yuci de tongji yanjiu" 国语中基本语词的统计研究 (Statistical Considerations of Basic Vocabulary in Chinese). *Guowen xuehui congkan*, 1(1), 81-4.

McEnery, T.; Hardie, A. (2011). "What Is Corpus Linguistics?". McEnery, T.; Hardie, A. (eds), *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press, 1-24.

McEnery, T.; Xiao, R. (2016). "Corpus-Based Study of Chinese". Chan, S. (ed.), *The Routledge Encyclopedia of the Chinese Language*. New York: Routledge, 438-51.

Nishimoto, E. (2003). "Measuring and Comparing the Productivity of Mandarin Chinese Suffixes". *Computational Linguistics and Chinese Language Processing*, 8(1), 49-76.

Siewierska, A; Xu, J.; Xiao, R. (2010). "Bang-le Yi Ge Da Mang (Offered a Big Helping Hand): A Corpus Study of the Splittable Compounds in Spoken and Witten Chinese". *Language Sciences*, 32(4), 464-87. https://doi.org/10.1016/j.langsci.2009.08.002.

Sproat, R.; Shih, C. (1996). "A Corpus-Based Analysis of Mandarin Nominal Root Compounds". *Journal of East Asian Linguistics*, 5, 49-71. https://doi.org/10.1007/BF00129805.

Tao H. 陶红印 (2000). "Cong 'Chi' Kan Dongci Lunyuan Jiegou de Dongtai Tezheng" 从"吃"看动词论元结构的动态特征 ('Eating' and Emergent Argument Structure). *Yuyan Yanjiu*, 20(3), 21-38.

Tao, H. (2008). "The Role of Corpora in Chinese Language Teaching and Teacher Education". Duff, P.; Lester, P. (eds), *Issues in Chinese Language Education and Teacher Development*. Vancouver: Centre for Research in Chinese Language and Literacy Education, University of British Columbia, 90-102.

Tao, H. (2009). "Core Vocabulary in Spoken Mandarin and the Integration of Corpus- Based Findings into Language Pedagogy". Xiao Y. (ed.), *Proceedings of the 21st North American Conference on Chinese Linguistics*. Smithfield (Rhode Island): Bryant University, 13-27.

Thompson, S.; Tao, H. (2010). "Conversation, Grammar, and Fixedness: Adjectives in Mandarin Revisited". *Chinese Language and Discourse*, 1(1), 3-30.

Wallis, S. (2019). "Grammar and Corpus Methodology". Aarts, B.; Bowie, J.; Popova, G. (eds), *The Oxford Handbook of English Grammar*. Oxford; New York: Oxford University Press, 59-83.

Wallis, S.; Nelson, G. (2001). "Knowledge Discovery in Grammatically Analysed Corpora". *Data Mining and Knowledge Discovery*, 5(4), 305-35.

Wang C. 王春霞 (2001). "Jiyu yuliaoku de liheci yanjiu" 基于语料库的离合词研究 (A Corpus-Based Study of Splittable Compounds) [MA dissertation]. Beijing: Beijing Language and Culture University.

Wang H. 王海峰 (2011). *Xiandai Hanyu liheci lixi xingshi gongneng yanjiu* 现代汉语离合词离析形式功能研究 (A Functional Study of the Split Forms of Separable Words in Modern Chinese). Beijing: Peking University Press.

Wang, Y. (2006). "The Information Structure of Adverbial Clauses in Chinese Discourse". *Taiwan Journal of Linguistics*, 4(1), 49-88.

Xiao, R. (2007). "What Can SLA Learn From Contrastive Corpus Linguistics? The Case of Passive Constructions in Chinese Learner English". *Indonesian Journal of English Language Teaching*, 3(2), 1-19.

Xiao, R.; McEnery, T. (2008). "Negation in Chinese: A Corpus-Based Study". *Journal of Chinese Linguistics*, 36(2), 274-330. https://www.jstor.org/stable/23756111.

Xiao, R.; McEnery, T. (2010). *Corpus-based Contrastive Studies of English and Chinese*. London/New York: Routledge.

Xiao, R.; Rayson, P.; McEnery, T. (2009). *A Frequency Dictionary of Mandarin Chinese: Core Vocabulary for Learners*. London: Routledge.

Xu, J. (2015). "Corpus-Based Chinese Studies: A Historical Review From the 1920s to the Present". *Chinese Language and Discourse*, 6(2), 218-44. https://doi.org/10.1075/cld.6.2.06xu.

Zhan W. 詹卫东 (2019). "Beijing Daxue CCL Yuliaoku de Yanzhi" 北京大学CCL语料库的研制. *Yuliaoku yuyanxue*, 6(1), 71-86.

Zhang, J.; Tao H. (eds) (2018). *Corpus-Based Research in Chinese as a Second Language*. London: Routledge.

Zou, B.; Smith, S.; Hoey M. (eds) (2016). *Corpus linguistics in Chinese contexts*. New York: Palgrave Macmillan.

Syntax and Pragmatics

Corpus-Based Research on Chinese Language and Linguistics
edited by Bianca Basciano, Franco Gatti, Anna Morbiato

A Corpus-Based Investigation of Manner/State Complement Constructions in Mandarin Chinese

Hongyin Tao
UCLA, USA

Hong Gang Jin
University of Macau, China

Jie Zhang
University of Oklahoma, USA

Abstract This study is an investigation of the complement constructions of manner and state (CM/S, e.g. 他的字写得好 *tā de zì xiě de hǎo* 'he writes characters well') based on a corpus of written Chinese. We find that CM/S have preferred forms and functions. Formally speaking, a monosyllabic verb, preferably 变 *biàn* 'change, become', basic action verbs, or psychological state verbs tend to co-occur with complements of adjectival, clausal, or idiomatic expressions. CM/S are argued to be an assessment device indexing speaker evaluative stances. The loaded affective meanings, we contend, account for the larger and more complex forms than their standard assessment counterparts. The implications of these findings on Chinese syntactic research and on L2 learning are explored.

Keywords Chinese Complement Construction. Complement of Manner. Complement of State. Assessment. Evaluative Stance. Construction Grammar. Iconicity.

Edizioni Ca'Foscari

Sinica venetiana 6
e-ISSN 2610-9042 | ISSN 2610-9654
ISBN [ebook] 978-88-6969-406-6 | ISBN [print] 978-88-6969-407-3

Peer review | Open access **19**
Submitted 2020-06-30 | Accepted 2020-10-14 | Published 2020-12-21

DOI 10.30687/978-88-6969-406-6/001

1 Introduction

Mandarin Chinese is known to have a variety of complement constructions (CC) that are highly productive, constituting some of the most unique features of its syntactic system (Shen 2003). These complement constructions exhibit a diverse range of syntactic, semantic, and pragmatic functions, indicating, e.g. result, degree, manner, possibility, direction, among others, and have been the subject of intense research from diverse linguistic theoretical persuasions (Chao 1968; Lü 1979; Li, Thompson 1981; Chu 1983; Cheung et al. 1994; Shen 2003, *inter alia*).

The current study restricts itself to just one type of CC, which we call complements of manner or state (CM/S, 情态 *qíngtài*/状态 *zhuàngtài*/方式 *fāngshì*). CM/S constructions typically consist of three components: the verb predicate (VP), the complementiser *de* (得), and complements of different syntactic structures. CM/S indicate either the manner in which the action named by the verbal predicate is executed or evaluated or a state toward which the action is carried out. Two quick examples illustrating these patterns can be found in (1) and (2).[1]

1. 父亲的围棋下得很好。(G48)
   fùqīn *de* *wéiqí* *xià* *de* *hěn* *hǎo*
   father ATT go play DE very well
   'Father plays *go* very well'.

2. 让摇滚乐变得更主流。(A33)
   ràng *yáogǔnyuè* *biàn* *de* *gèng* *zhǔliú*
   make rock.roll.music become DE even.more mainstream
   'Make Rock N Roll music even more mainstream'.

In (1) the complement 很好 *hěn hǎo* 'very well' can be seen as an evaluation ('how well') of the verbal predicate 下 *xià* 'play'. In (2), on the other hand, the complement 更主流 *gèng zhǔliú* 'even more mainstream' can be understood to be the state toward which the action of 变 *biàn* 'change, become' is to be carried out.[2]

A review of the literature shows that structural approaches to CM/S, and CC in general, which are dominant, have tended to focus on a few areas. First, syntactic configurations, especially the struc-

1 The glosses follow the general guidelines of the Leipzig Glossing Rules. Additional glosses include: ATT = 'attributive'; BA = 'disposal marker *bǎ*'; BEI = 'passive marker *bèi*'; BI = 'comparative marker *bǐ*'; DE = 'complementizer *de*'; JIANG = 'disposal marker *jiāng*'; MOD = 'modifier'; NONG = 'delexical verb *nòng*'; PRT = 'utterance final particle'.

2 More discussion on the identification of CC subtypes can be found in § 2.

ture of the complement, have been described as ranging from simple adjectival phrases (e.g. 快 *kuài* 'fast'; 非常好 *fēicháng hǎo* 'very good'; 十分客气 *shífēn kèqì* 'quite courteous'), to larger phrasal (and often idiomatic) units, such as 哭得像个泪人 *kū de xiàng gè lèi rén* 'cry with tears welled up', and all the way to complex clausal units, e.g. 弄得人人皆以绅士为流氓 *nòng de rénrén jiē yǐ shēnshì wéi liúmáng* 'make everyone treat gentlemen as hooligans' (Li 1963; Nie 1992, *inter alia*).

A great deal of work has concentrated on the second area: semantic features. Here, three types of meaning-related issues have been explored: the verb predicate, the complement, and the semantic focus of the structure. Verb predicates that are commonly brought into discussion include single or disyllabic action verbs indicating completed or ongoing actions. Complement types are reportedly to vary, and sometimes the same surface structure is shown to indicate different meanings (e.g. state *vs* result with the same adjective). Complements are also said to exhibit two types of semantic focus (Lü 1979; Lu 1993; Fan 1992; Wu 2002; Jiang 2005). The first type is said to be focusing on the action itself, where the complement describes and evaluates how the action itself is carried out, as illustrated in extract (1). In this regard, most researchers agree that action-focused CM/S are the most prototypical type with a high frequency of usage (Fan 1992; Zhang 2002; Wu 2002). The second type of semantic focus is said to be on the non-action elements of the construction: either the agent, the patient, the overall causality expressed in a CM/S, or some combinations thereof. Causality is also said to be achieved in conjunction with a disposal 把 *bǎ* construction, a 被 *bèi* passive construction, or a causative 让 *ràng*, 使 *shǐ*, or 将 *jiāng* construction. Thus, in extract (2) discussed earlier, a 让 *ràng* 'cause/causal' construction is observed and the focus can be said to be on the argument 'Rock N Roll music', which exhibits features of a pivotal entity – being both the causee of the causative verb 让 *ràng* and the agent of the following predicate of change of state ('becoming more mainstream'). The frequency of this type is believed to be lower than the action-focused type (Fan 1992; Zhang 2002; Wu 2002).

Finally, with regard to the pragmatics of CC, it has been claimed that CC are fundamentally a topic-comment structure (Chao 1968; Lu 1992; Liu 2005; Lu, Ying, Zhang 2015). Under this view, the subject and predicate of CC together function as the topic, signalling the old or known information, while the complement represents the comment, carrying the new or primary information. Because of its pragmatic nature, Li (1963, 1980) and researchers following him (e.g. Lu, Ying, Zhang 2015) claim that the complement of CC, at least with some of them, is the natural focus and the most salient part of the construction. Recent studies, however, have disputed this claim with a host of syntactic diagnostics, and a wide variety of proposals have been made (see Shen 2003 for a comprehensive review). Our data will

show that while this is an interesting angle from which to approach CC, there are actually more critical issues to be explored, which have received scant attention thus far.

In short, existing studies have approached CC from multiple structural perspectives, highlighting the fact that this is an important and unique feature of the syntax of Chinese. However, a number of shortcomings can be identified for most structural studies. First, most of these studies are based on intuition, as exemplified by the data samples used in the analysis, which are for the most part constructed sentences; and if actual usage samples are used, they typically involve a small quantity from individual collections. To be sure, there have been several corpus-based studies of CC in recent years; however, these studies tend to use either mixed genres (Li 1994; Wang 2011; Ma, Chen 2014) or single genres such as fiction (Wang 2001; Chen 2013) or school texts (Wu 2018), limiting to various degrees the validity of such studies. Second, most studies deal with resultative (动结式 *dòng jié shì*) and motion (动趋式 *dòng qū shì*) complements, as they are believed to be the most prolific types of all CC. While this may be a reasonable choice, we would like to show that other CC types may have their own characteristics and communicative utility and are thus equally worthy of our attention. Finally, most studies have tended to focus on individual components or isolated classes of elements (e.g. verbs, adjectives etc.) in the CC, and not from the perspective of meaning-form pairing (Fillmore, Kay, O'Connor 1988; Goldberg 1995, 2003) or co-occurrence/contingency patterns (Gries, Ellis 2015). One consequence of such an approach is that while it may enable us to see some of the admissible elements in a CC when individual features are focused, we know surprisingly little about the functional motivations of these constructions as opposed to others and how specific components/forms and meanings pair up and why.

In this study, we intend to address the shortcomings of the existing studies by using a modest sized corpus, the million-word UCLA Corpus of Written Chinese (more on this in § 2) and analyse CM/S constructions exhaustively. We also intend to pursue CM/S from the perspective of usage-based linguistics, paying particular attention to (type/token) frequency information (Bybee, Thompson 2000) and the notions of construction grammar (CxG; Fillmore, Kay, O'Connor 1988; Goldberg 1995, 2003). According to CxG, syntactic structures can be viewed as constructed on specific building blocks that are unique in their own ways, resulting in specific form-meaning mappings and conventionalised configurations whose meanings may not be readily deducted from the meanings of individual components. These pairings can be regarded as entrenched language knowledge for production and comprehension. Thus, in the case of CC, and CM/S in particular, we would expect that different types of CC or CM/S attract different types of component elements, resulting in different syntac-

tic configurations and unique meanings and functions. We will also attempt to apply functional linguistic principles such as the iconicity principle (Haiman 1983) and the prototypicality principle (Rosch 1973; Rosch, Mervis 1975; Hopper, Thompson 1984) to account for patterns revealed from corpora. These patterns, we hope, will not only deepen our knowledge about CC and CM/S, but also raise questions about a number of important theoretical issues, including L1 and L2 knowledge and how best to approach Chinese syntax in general.

In what follows, we will first describe the corpus and key concepts used for this study. We will then report the corpus-based findings before discussing the results from the point of view of usage-based functional linguistics and CxG. In the conclusion section, some generalisations about methodology and implications for other fields, such as L1 and L2 studies, will be provided.

2 Data and Methodology

2.1 The Corpus

The UCLA Corpus of Written Chinese (Tao, Xiao 2007-20) used for this study is designed as a Chinese counterpart for the FLOB and Frown corpora of British and American English for contrastive research, as well as an update of the Lancaster Corpus of Modern Chinese (LCMC; McEnery, Xiao 2004). The samples in the corpus are all collected from written modern Chinese available from the internet, during the periods of 2000-05 and 2005-12, with fifteen genres such as news, fiction, academic prose, and essays.[3] The data were word-segmented and tagged for parts-of-speech (POS) information by the software program ICTALCS (Zhang et al. 2002; Xiao, Rayson, McEnery 2009, 3-4), which uses algorithms based on statistical models. There are over one million tokens and near 60,000 types in the corpus.

3 Text genres and file numbers from the UCLA Corpus are indicated at the end of each extract (e.g. A05 for genre A, file no. 05). The 15 genres in the corpus are labelled as follows. A: Press reportage; B: Press editorials; C: Press reviews; D: Religion; E: Skills, trades and hobbies; F: Popular lore; G: Essays and biographies; H: Misc. (reports and official documents); J: Academic prose; K: General fiction; L: Mystery and detective stories; M: Science fiction; N: Adventure stories; P: Romantic fiction; R: Humor.

2.2 Inclusion of CM/S

While CM/S structures discussed here indicate manners or states,[4] as a whole they can appear as either the main clause (as in (1) and (2)) or part of a larger structure. For example, in (3), a CM/S is part of a copula 是 *shì* clause, specifically, being at the end of an equative structure and embedded in a 把 *bǎ* construction.

3. 交流感情, 起码的要求是把字写得规范、整洁、清楚 […](F32)
   jiāoliú gǎnqíng qǐmǎ de yāoqiú shì bǎ
   exchange affection minimal ATT requirement COP BA
   zì xiě de guīfàn zhěngjié qīngchu
   character write DE standard neat legible
   'To communicate your affection effectively, one needs minimally to write standard scripts, and write neatly and legibly […]'

In the next example, the CM/S is part of a relative clause modifying the head noun 现在 *xiànzài* 'nowadays':

4. 对食物的成分已了解得较为透彻的现在 […] (J98)
   duì shíwù de chéngfèn yǐ liǎojiě
   about food ATT composition already understand
   de jiàowéi tòuchè de xiànzài
   DE relatively thorough REL nowadays
   'In this day and age when we know a great deal about the ingredients of food […]'

In this study, both independent and embedded CM/S structures are included.

2.3 Corpus Approaches

In corpus linguistics, a broad distinction is made between a corpus-based and corpus-driven approach. In general, corpus-based research relies on established linguistic forms and theory to conduct investigations, while a corpus-driven approach relies more on corpus data itself in delineating features and the scope of a linguistic investigation (Biber 2009). Thus in our case, while we employ constructs such as CC and CM/S as they have been subject to intense previous research, making this project more of a corpus-based type, the spe-

4 As is well known, complement types may not always be clear-cut and borderline cases do exist. The selection of the tokens in this paper represents the best judgement of the three authors. We thank an anonymous reviewer for emphasising this point.

cific types of constructs – including their components and subcategories – will emerge mainly from the corpus itself. In this sense our study uses mixed methods of corpus-based and corpus-driven.

Looking at the key components in the CM/S structure, we will start our investigation with the following: 1) verb classes in the main predicate (here, instead of looking at individual verbs alone, we will examine classes of verbs and their frequency distribution in the corpus); 2) complement types (although complement types have been subject to intense study in the literature, in this study we will rely on frequency information of the corpus data to define the types of complements to focus on); 3) co-occurrence patterns of the verb classes and complement types. Once verb classes and complements types are identified, we will look into their correlation, via Correspondence Analysis (Glynn 2014) and other methods, as a window into the overall constructions they form and special meanings they may convey.

2.4 Macro and Micro Analyses

Finally, we will combine the macro level analysis with case studies, especially the high frequency items and the constructions that they help form. Case studies will be provided in § 5 after report of general corpus findings and discussions in §§ 3 and 4 respectively.

3 Corpus Findings

Our corpus investigation produced the following results, which will be reported in terms of verb classes, complement types, and co-occurrence patterns.

3.1 Verb Classes

In the UCLA Corpus, 769 tokens and 251 types of verb predicates in CM/S structures are found. Among them, 173 types (665 tokens) are monosyllabic, while 78 types (104 tokens) are disyllabic, showing a preference for monosyllabic verbs. The top 31 types, appearing at least five times and being all monosyllabic, are listed below (more items can be found in the Appendix B).

Table 1 Frequency of occurrences of top predicate verbs in the corpus

Ranking	Freq	Token	Ranking	Freq	Token	Ranking	Freq	Token
1	132	变	11	11	哭	21	6	卖
2	23	弄	12	11	活	22	6	急
3	22	过	13	10	做	23	6	放
4	20	笑	14	10	听	24	6	玩
5	18	说	15	9	忙	25	6	睡
6	17	吓	16	9	长	26	6	聊
7	16	写	17	8	穿	27	5	想
8	16	打	18	8	跑	28	5	搞
9	15	吃	19	7	看	29	5	来
10	14	走	20	6	冻	30	5	羞
						31	5	羞

This list of top verbs shows some interesting tendencies.

First, 变 ***biàn*** 'change, become'[5] stands out as the most frequent token, with an overwhelmingly high frequency of 132, accounting for 17% of all CM/S tokens found in the data.

Second, some of the top verbs are of the empty/**delexical** type. These verbs include 弄 *nòng*, 打 *dǎ*, 做 *zuò*, 搞 *gǎo*, 干 *gàn*, 进行 *jìnxíng*, and 办 *bàn*, akin to the English delexical verbs such as *do*, *make*, *take*, *get* etc. (Sinclair 1990, 147).[6]

Third, another group of verbs can also be identified as lexically less concrete, i.e. **general**, yet their referential meaning is somewhere between delexical verbs and common action verbs (to be described below). Examples of this kind include 过 *guò* 'live', 活 *huó* 'live', 放 *fàng* 'arrange', 玩 *wán* 'play', and 想 *xiǎng* 'think, desire'.

The next prominent group of verbs depicts various everyday **basic actions** (看 *kàn* 'look'; 卖 *mài* 'sell'; 睡 *shuì* 'sleep'), sometimes with opposite meanings: 笑 *xiào* 'laugh' / 哭 *kū* 'cry', 说 *shuō* 'speak'| 讲 *jiǎng* 'talk' | 聊 *liáo* 'chat'; 听 *tīng* 'listen', 写 *xiě* 'write', 走 *zǒu* 'walk' / 跑 *pǎo* 'run' / 来 *lái* 'come', 吃 *chī* 'eat', 穿 *chuān* 'wear'.

Finally, the last group consists of verbs that can be either transitive or intransitive, with some of them indicating a **psychological** state as the result of some impactful actions. Top examples in this category include 吓 *xià* 'scare/frightened', 急 *jí* 'anxious', 羞 *xiū* 'shy'.[7]

By applying the classification of high frequency verbs in this way, and having **others** as a separate category for all those that do not belong to any of the above semantic categories, we found the distribution of verb types and tokens in the corpus as follows.

5 Biber et al. call verbs such as *change, become* etc. "occurrence verbs" (1999, 364).

6 Of course *do* in English is also a widely used auxiliary verb.

7 We note that these verbs can be used with complement of degree. However, for this study, all degree complements are excluded.

Table 2 Frequency of occurrences of predicate verb types in the CM/S

Verb Types and Sample Tokens	Type	Token	%
A. * 变 *biàn* 'change, become'	1	132	17
B. * Delexical verbs: 弄 *nòng* 'do'; 打 *dǎ* 'hit'; 做 *zuò* 'make'; 搞 *gǎo* 'do'; 干 *gàn* 'do'; 进行 *jìnxíng* 'engage'; 办 *bàn* 'process'	8	65	8
C. General: 过 *guò* 'pass'; 活 *huó* 'live'; 放 *fàng* 'arrange'; 玩 *wán* 'play'; 想 *xiǎng* 'plan'	36	111	14
D. Basic actions: 睡 *shuì* 'sleep'; 写 *xiě* 'write'; 吃 *chī* 'eat'; 哭 *kū* 'cry'; 笑 *xiào* 'laugh'; 洗 *xǐ* 'wash'; 看 *kàn* 'look'; 卖 *mài* 'sell'	81	254	33
E. Psychological states: 急 *jí* 'be anxious/ worried'; 爱 *ài* 'love'; 疼 *téng* 'ache'; 病 *bìng* 'sick'; 累 *lèi* 'be tired'; 羞 *xiū* 'be shamed'; 感动 *gǎndòng* 'be moved'	35	89	12
F. Others: 升 *shēng* 'lift'; 围 *wéi* 'surround'; 定 *dìng* 'determine'; 掩饰 *yǎnshì* 'cover up'; 提 *tí* 'lift'; 销售 *xiāoshòu* 'sell'; 折磨 *zhémó* 'torment'	90	118	15
Total	**251**	**769**	**100**

*Exhaustive listing.

A frequency-based ranking list is given in (5):

5. Basic action > *Biàn* > Others > General > Psychological > Delexical

Overall the tendency seems to be from concrete everyday actions to more abstract (including mental) activities.

3.2 Complement Types

For complements, four general patterns emerge from the data. They are: 1) adjectival units of various kind. For example, a simple adjective such as 好 *hǎo* 'well' in 自己过得好 *zìjǐ guò de hǎo* '(doing) well', or an adjective with a modifier, as 这么漂亮 *zhème piàoliang* 'so pretty' in 谁让你长得这么漂亮 *shéi ràng nǐ zhǎng de zhème piàoliang* 'it doesn't help that you look so pretty'; 2) clausal units, where a complement contains a verbal predicate with or without a subject, e.g.:

6. 高烧未退，烧得她昏迷不醒。(G41)
gāoshāo wèi tuì shāo de tā hūnmí bù xǐng
high.fever NEG recede heat DE 3SG in.coma NEG wake
'High fever persists, keeping her in a state of deep coma'.

7. 我听得入了神。(P32)
   wǒ *tīng* *de* *rùleshén*
   1SG listen DE captivated
   'I was captivated by listening to it'.

In (6) there is a subject and a verb predicate in the complement, whereas in (7) the subject in the complement is implicit as it shares with the subject of the main clause 我 *wǒ* 'I'.

3) Formulaic expressions. By this we mean expressions that have paired, parallel, or contrastive elements, which are often similar in form, to highlight some quality in the expressed meanings. Typical examples may include the following.

8. 不管我把母亲写得多形象多生动多具体 […] (K36)
   bùguǎn *wǒ* *bǎ* *mǔqīn* *xiě* *de* *duō* *xíngxiàng*
   regardless 1SG BA mother depict DE how life-like
   duō *shēngdòng* *duō* *jùtǐ*
   how vivid how detailed
   'No matter how life-like, vivid, and detailed I depict mother […]'

9. 杀得越来越起劲。(L06)
   shā *de* *yuèláiyuè* *qǐjìn*
   kill DE more.more strong
   'Kill with increasing intensity'.

10. 吃得香美而不奢靡。(F34)
    chī *de* *xiāngměi* *ér* *bù* *shēmí*
    eat DE splendidly but NEG extravagant
    'Eat splendidly yet not extravagantly'.

In (8) three adjectives with modifiers are placed in tandem. In (9) the formula 越来越 *yuèláiyuè* is used; and finally, in (10) a positive adjective (with two coordinated morphemes) is used in contrast with a negative one, constituting a contrastive structure. Although these instances may be seen as subcategories of adjectival expressions, their special structural formations make them stand out as a unique feature to make a case for a separate category.

Finally, 4) idiomatic expressions. Typically in the form of 成语 *chéngyǔ* 'fixed (four-character) expressions', a large number of idioms appears as the main component of the complement. Some, as 前仰后翻 *qiányǎng-hòufān* 'rolling back and forth' in (11), can be seen as more fixed, while others, e.g. 稀稀拉拉 *xīxi-lālā* 'scattered around' in (12), may not be as fixed.

11. 直把我们笑得前仰后翻的。 (N37)
zhí *bǎ* *wǒmen* *xiào* *de* *qiányǎng-hòufān* *de*
finally BA 1PL laugh DE rolling.back.forth PRT
'Made us laugh so hard, almost rolling back and forth'.

12. 能坐下上百人的会议室里听众坐得稀稀拉拉。 (A42)
néng *zuòxià* *shàngbǎi* *rén* *de*
capable.of sit hundread.plus person REL
huìyìshì-lǐ *tīngzhòng* *zuò* *de* *xīxi-lālā*
meeting.room-in audience sit DE scattered
'In a room capable of seating over a hundred people, only a few people scattered around'.

The distribution of the four complement types can be found in table 3.

Table 3 Distribution of CM/S complements in the corpus

Complement Type	N	%
Adjectival	397	51.6
Clausal	157	20.4
Idiomatic	150	19.5
Formulaic	65	8.5
Total	**769**	**100**

Notable results from the data include the following. First, at just over 50%, adjectival expressions are the dominant single category for the complements. This gives a more accurate picture of the makeup of complements, as most earlier studies simply rely on intuition and estimate that adjectives are the majority, at least for some of the complements (e.g. resultatives, Shen 2003, 21).

Second, CM/S with an idiomatic expression are as frequent as clausal units. Idiomatic expressions, especially those of the four-character type, typically indicate a strong affective stance on the part of the speaker/writer. This, along with the proliferation of adjectival expressions in general, suggests that CM/S constructions are affect-laden and highly subjective (more discussion on this in § 4).

Third, while formulaic expressions may not be as fixed as the idiomatic expressions, they are also a notable type, and their function is very close to the idiomatic ones, with the only difference perhaps lying in the degree of fixedness: looser in formulaic expressions and more conventionalised in the idiomatic ones. If we combine these two together, however, this would be a very notable phenomenon to be accounted for. Again we will divulge this more in § 4.

3.3 Verbal Predicate and Complement Co-Occurrence Patterns

Having culled data about the two key individual components, let us now examine how verb predicates and complements co-occur with each other. Our goal for this exercise is to find out the attested preferred configurations that these key components may form. Table 4 provides an overview of the data in this respect.

Table 4 Co-occurrence patterns of verbal predicates and complements in CM/S constructions[8]

X^2=113.46, df=15, p < .0001

VType/Comp	Adjectival	Clausal	Formulaic	Idiomatic	Total
A. *Biàn*	73	13	16	30	132
B. Delexical	27	17	5	16	65
C. General	77	8	9	17	111
D. Action	146	34	23	51	254
E. Psych	25	49	3	12	89
F. Other	49	36	9	24	118
Total	**397**	**157**	**65**	**150**	**769**

There are a number of ways to look at the data. We can examine the percentages of complements across verbal categories, and the result is shown in both table 5 and figure 1.

Table 5 Complements across verbal types in percentages

	Biàn		Delexical		General		Action		Psych		Other	
	N	**%**	**N**	**&**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
Adjectival	73	55.3	27	41.5	77	69.4	146	57.5	25	28.1	49	41.5
Clausal	13	9.8	17	26.2	8	7.2	34	13.4	49	55.1	36	30.5
Formulaic	16	12.1	5	7.7	9	8.1	23	9.1	3	3.4	9	7.6
Idiomatic	30	22.7	16	24.6	17	15.3	51	20.1	12	13.5	24	20.3
Total	**132**	**100**	**65**	**58.6**	**111**	**100**	**254**	**100**	**89**	**100**	**118**	**100**

A number of properties can be noted here. First, while adjectival complements can co-occur with most of the verbal categories (see also 3.2), psychological state verbs correlate most often with clausal complements. Some examples of the latter can be found in (13) and (14).

8 Specific configuration patterns can be found in Appendix C.

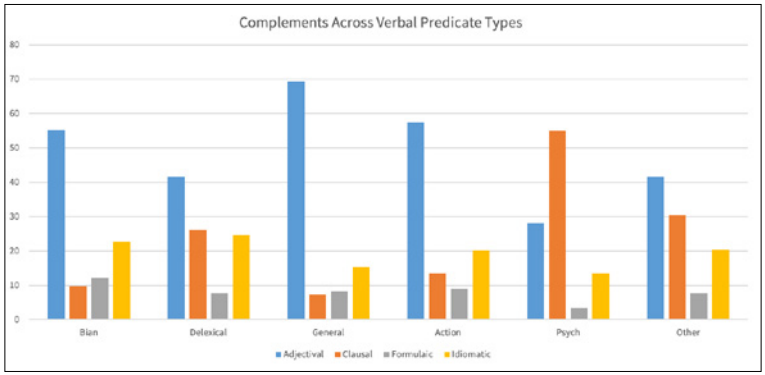


Figure 1 Complements across verbal types in percentages

13. 吓得倒抽一口冷气 […] (L31)
xià *de* *dàochōu* *yī* *kǒu* *lěngqì*
frightened DE inhale one mouth cold.air
'So frightened that they inhaled a mouthful of cold air […]'

14. 急得我天天上物价局打听去 […] (R15)
jí *de* *wǒ* *tiāntiān* *shàng* *wùjiàjú*
worried DE 1SG everyday go.to price.bureau
dǎtīng *qù*
inquire go
'I was so worried that I went to the consumer price bureau everyday to find out more information […]'

Second, although adjectival complements are generally common, they are even more dominant in three types of verbal predicates: 变 *biàn* (55.3%), general verbs (69.4%), and basic action verbs (57.5%), and this is especially the case of general verb constructions, where they make up the largest proportion.

A Correspondence Analysis,[9] which transforms the two dimensions from numerical information into a spatial display (Glynn 2014, Zhang 2017), shows similar patterns. Specifically, on the left sphere of the biplot graph, adjectival complements cluster with general verbs, basic action verbs, and 变 *biàn*, while clausal complements and psychological state verbs cluster on the extreme right.

9 Correspondence Analysis was performed through XLSTAT (Addinsoft 2020), a statistical and data analysis add-in for Excel.

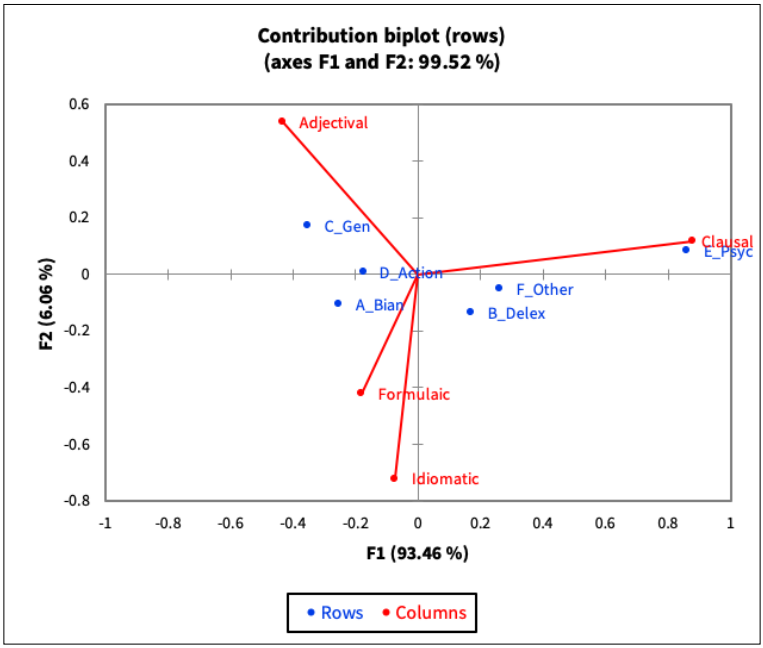


Figure 2 Correspondence analysis of the contingency data

Finally, we can also examine specific configuration patterns, making use of the ranked list of the observed combinations based on the frequency of the subtypes of each of the two major component categories. The result is shown in table 6.

Table 6 Construction patterns based on subtypes of the verbal predicates and the complements in CM/S constructions

	V	Comp	N	%
1	B. Action	Adjectival	146	19
2	General	Adjectival	77	10
3	*Biàn*	Adjectival	73	9.5
4	B. Action	Idiomatic	51	6.6
5	Psych	Clausal	49	6.4
6	Other	Adjectival	49	6.4
7	Other	Clausal	36	4.7
8	B. Action	Clausal	34	4.4
9	*Biàn*	Idiomatic	30	3.9
10	Delexical	Adjectival	27	3.5

11	Psych	Adjectival	25	3.3
12	Other	Idiomatic	24	3.1
13	B. Action	Formulaic	23	3
14	Delexical	Clausal	17	2.2
15	General	Idiomatic	17	2.2
16	*Biàn*	Formulaic	16	2.1
17	Delexical	Idiomatic	16	2.1
18	*Biàn*	Clausal	13	1.7
19	Psych	Idiomatic	12	1.6
20	General	Formulaic	9	1.2
21	Other	Formulaic	9	1.2
22	General	Clausal	8	1
23	Delexical	Formulaic	5	0.7
24	Psych	Formulaic	3	0.4
	Total		**769**	**100**

Not surprisingly, among the top six (all with a percentage of ≥ 6) syntactic configurations, the combinations of basic action verbs and an adjectival complement are dominant: two with action verb categories and four involving adjectival complements. The exceptional cases include one of the distinct patterns we have previously discussed: the mutual attraction of psychological state verbs and clausal complements, plus the other type of verbs with adjectival complements.

4 Summary and Discussion

4.1 Major Patterns

So far, our data extrapolation has yielded several notable patterns, which come from verbal predicates, complements, and their co-occurrences.

In terms of verbal predicates, 1) monosyllabic verbs dominate; 2) 变 *biàn* 'change, become' is the single most frequent verb among all verb tokens; and 3) overall verb frequency has the following hierarchical relations: Basic action > *Biàn* > Others > General > Psychological > Delexical, seemingly reflecting a larger hierarchy of concrete everyday actions over mental and abstract activities. These patterns, as will be elaborated in the next section, help us understand some of the important constructional features associated with these CM/S.

In terms of complements, the frequency hierarchy is: Adjectival > Clausal > Idiomatic > Formulaic.

Finally, in terms of verb predicate and complement co-occurrences, there are three notable patterns: 1) the top configurations are [Basic Action + Adjectival] > [General Verb + Adjectival] > [*Biàn* + Ad-

jectival] > [Basic Action + Idiomatic] > [Psychological State Verbs + Clausal / Other + Adjectival]; 2) psychological state verbs and clausal complements are mutually attractive. Finally, 3) verbal predicates with 变 *biàn* are robust in three of the four complement types (formulaic, idiomatic, and adjectival), except for clausal.

4.2 Some Generalisations

Given all these patterns, what underlying principles might there be that hold them all together? We would like to think of these underlying principles as construction functions. In this respect, the following generalisations may be proposed.

a. CM/S may be formed with any combinations of verb predicates and complements, yet they exhibit preferred syntactic structures, which involve a monosyllabic verb, as typified by 变 *biàn* 'change, become', or others denoting basic actions or psychological states, plus a complement of the adjectival, clausal, or idiomatic types.
b. CM/S constructions are an assessment device indexing speaker evaluative stances.
c. CM/S differ from other assessment devices with additional features, including assessing the process of the state of affairs and with strong affective qualities. These additional features result in, iconically, longer and more complex constructions than many simple assessment forms.

We now explicate these generalisations in turn.

4.2.1 Formal Preferences

Most syntactic studies have assumed that CM/S may be formed by any combination of verb predicates and complements, a claim that our data can be said to support if one just looks at the admissible items found in the corpus, which are highly varied. Others have speculated about frequency differences between action-focused and non-action focused CC, as we have seen in the Introduction section earlier. Yet our corpus results point to notable preferred syntactic structures – and new angles – for contemplation, involving some combinations of the key elements: a monosyllabic verb, preferably 变 *biàn* 'change, become', basic action verbs, or psychological state verbs, while the complement tends to be an adjectival, idiomatic, or clausal expression. These combinations can be schematised in figure 3.

{	*Biàn* 'change, become'		Adjectival complement
	Basic action verb	+	Idiomatic complement
	Psychological verb		Clausal complement
			}

Figure 3 Preferred syntactic constructions in CM/S

Although the schematisation allows the free combination of any of the items on the left and right columns, we reiterate a couple of strong tendencies: 1) 变 *biàn* 'change, become', by virtue of its sheer token frequency, should be recognised as a prototypical CM/S construction by itself; 2) there is a divide between action verbs and psychological state verbs: while the former can be combined with many complement types, the latter strongly attracts clausal complements.

4.2.2 CM/S as an Assessment Device

As stated earlier, CM/S constructions as a whole can be taken to be an assessment device through which speakers index their evaluative stance. In a natural conversation-based study, Thompson and Tao (2010) find that although adjectives in Mandarin Chinese can function either attributively (as a modifier) or predicatively (as a predicate), 80% of the adjectives in their conversational data are found to be of the predicative type, a result similar to what have been reported for both English (Thompson 1988; Englebretson 1997) and Japanese (Ono, Thompson 2009). In explaining this discourse preference, Thompson and Tao assert that predicate adjectives in conversation are deployed by speakers to "assess the world around them, and that assessments, including reactive tokens, are a primary way for people to negotiate stance, alignment, and perspective" (2010, 22). The fact that adjectives are pervasively used in the complements suggests that they are a primary device to reflect the speaker's subjectivity and in negotiating identity through assessing activities (Du Bois 2007; Englebretson 2007).

4.2.3 CM/S Differ from Other Assessment Devices and Iconicity

As one of the basic human conversational activities, assessment has been shown to be accomplished through a variety of syntactic configurations (Pomerantz 1984; Goodwin, Goodwin 1987, 1992; Thompson, Fox, Couper-Kuhlen 2015). It is widely believed that the most basic form of assessment involves "an assessable item + a copula + an assessment term", as illustrated by the English utterance 'It was so good' (Goodwin, Goodwin 1987). In Chinese, research has also shown that assessments can be done in a variety of ways, including structures similar to

the English copula construction (Fang 2018). Given this, how can CM/S be seen as different from basic assessment forms and what more can such a longer and more complicated form accomplish in language use?

We believe that CM/S have more expansive uses over other assessment devices due to their built-in features, which can be explained with the functional principle of iconicity as proposed in Haiman (1983). To be specific, we contend that CM/S differ from basic assessment forms in the following ways.

First, CM/S not only provide a simple assessment, they also assess the process of the state of affairs. This is best represented in 变 *biàn*-centred CM/S (e.g. (15)) but also in many other CM/S constructions. For example,

15. (政策)使巴以和平前景变得更加黯淡。 (A05)
(*zhèngcè*) *shǐ* *Bā* *Yǐ* *hépíng* *qiánjǐng*
(policy) cause Palestine Israel peace outlook
biàn *de* *gèngjiā* *àndàn*
become DE even.more bleak
'Implementation of such policies made the Palestine and Israel peace outlook even more glum'.

16. 小女孩托着下巴, 听得入了迷。(M21)
xiǎo *nǚhái* *tuō-zhe* *xiàbā* *tīng* *de* *rùlemí*
little girl hold-DUR chin listen DE mesmerise
'The little girl holds her chin, mesmerised by listening to it'.

17. 成绩不但没受影响, 而且比以前学得还好, 学得还主动。(C19)
chéngjī *bùdàn* *méi* *shòu* *yǐngxiǎng* *érqiě* *bǐ*
grade not.only NEG receive impact but BI
yǐqián *xué* *de* *hái* *hǎo* *xué* *de*
past study DE even better study DE
hái *zhǔdòng*
even.more motivate
'The grade is not only not negatively impacted, it's getting even better, and (the student) has even stronger learning motivations'.

In (15), the verb 变 *biàn* 'change, become' indicates that the glum outlook is the state that has been reached after the implementation of certain policy, which by definition involves a process. In (16), the verb 听 *tīng* 'listen', which implies a process, together with other elements in the utterance, such as 托着下巴 *tuōzhe xiàbā* 'holds (her) chin', which indicate a duration, reinforce the notion of a process. In (17), the notion of process is expressed with the comparative structure 比以前 *bǐ yǐqián* 'compared to before'.

Second, CM/S convey strong affective qualities. Although we cannot say categorically that simple assessment statements such as cop-

ular structures always carry a weak force, CM/S accomplish strong assessment power through a variety of linguistic features, such as idioms and multiple items of various formation in formulaic structures. The pervasive use of idioms and to some extent of formulaic structures, both of which cluster on the biplot graph in figure 2, are particularly noteworthy. Many discourse linguists have shown that idioms, broadly defined, serve evaluative functions in narratives and other discourse contexts (McCarthy 1998). As such, idioms are also said to carry high emotional or affective loads, such that in conversational discourse it is claimed that "a high degree of intimacy and in-group membership is projected by such idiomatic usage" (O'Keeffe, McCarthy, Carter 2007, 91). Our data corroborate these claims. Thus in the following set of expressions involving the delexical verb 弄 *nòng* 'do, get, make' (Tao, Hu 2019), different forms can be argued to display varying degrees of affective load: (18), which has no complements, is for information seeking and can be said to carry the least amount of affective load; (19), by contrast, has a simple (negative) complement, 不清 *bù qīng* 'NEG figure-out', which carries a slightly higher affective load than (18); and finally, (20) has a pair of idioms with strong judgmental and emotional slants, carrying arguably the highest degree of negative affective load, as it expresses the author's strong dislike of the protagonist Mo Huairen, a negative character portrayed in the story.

18. 你说他弄凉粉儿, 他弄两瓶酱油? (R15) (No complement)
nǐ *shuō* *tā* *nòng* *liángfěnr* *tā* *nòng* *liǎng*
2SG say 3SG get jelly 3SG get two
píng *jiàngyóu*
bottle soy.sauce
'Did you say that he got some jelly, and he got two bottles of soy sauce?'

19. 也弄不清它背后到底在搞些什么。 (D26) (Simple complement)
yě *nòng* *bù* *qīng* *tā* *bèihòu* *dàodǐ*
anyway figure.out NEG clear 3SG behind after.all
zài *gǎo* *xiē* *shénme*
PROG do some what
'Can't figure out exactly what is going on behind all this'.

20. (莫怀仁对歌), 又被刘三姐等弄得丑态百出, 大败而归。
(F16) (Double idiom-formed CM/S)
(*Mò Huáirén* *duì* *gē*) *yòu* *bèi* *Liúsānjiě* *děng*
Mo Huairen compete song again bei Liusanjie others
nòng *de* *chǒutài-bǎichū* *dàbài-érguī*
make DE display.all.ugliness end.in.total.defeat
'In a singing competition, Mo Huairen was once again defeated badly by Liusanjie and her friends and withdraw in total disgrace'.

Given the complexity of the meanings of CM/S, it is not surprising to see that CM/S structures are in general larger and more complex – being extensible as they often are to multiple clausal units in the complements – than the standard assessment forms such as copula constructions or simple statements such as 我喜欢 *wǒ xǐhuān* 'I like (it)' (Fang 2018). Here we find Haiman's (1983) iconicity principle highly relevant in explaining the differences. According to this functional principle, longer and more complicated forms tend to correspond to higher degrees of conceptual complication, such as longer processes, and more intense social meanings. In this case, the iconicity principle seems able to explain well both the *process* connotation and the more *loaded affective meanings* encoded in CM/S constructions that we have tried to elucidate, and these key ingredients may not necessarily be found in simple, standard assessment forms.

5 Cases Studies

Having provided an overall account of the major tendencies of CM/S constructions, we now turn to a few selected patterns and examine them in some more detail.

5.1 *Biàn* 'Change, Become'

The distribution of 变 *biàn* across complement types is given in table 7.

Table 7 *Biàn* and its complement types in the corpus

Adjectival	Clausal	Formulaic	Idiomatic	Total
73	13	1316	30	132

As shown above, 变 *biàn* has two prototypical use patterns: adjectival and idiomatic complements. In the case of adjectival complements, many constructions indicate a state that has been reached (perfective), as in (21), or one that starts to change (inchoative), as in (22).

21. 就要注册结婚了,远却变得陌生了。(G34)
jiùyào zhùcè jiéhūn le Yuǎn què biàn de
nearly register marry PRT Yuan however become DE
mòshēng le
strange PRT
'While they are about to register and get married, Yuan somehow becomes detached'.

22. 上升到政治高度, 马上就变得严肃起来。 (B04)
shàngshēng dào zhèngzhì gāodù mǎshàng jiù
elevate reach politics elevation soon then
biàn de yánsù qǐlái
become DE serious upward
'As soon as one politicises it, (things) suddenly become serious'.

Since idioms have been argued to play a special role in language, carrying particularly high emotional or affective loads as well as serving to index the evaluative stance of the speaker/writer, we now examine some specific instances of '变 *biàn* + idiom' combinations to demonstrate this property.

Many of the '变 *biàn* + idiom' combinations are used for the speaker/writer to depict an object or event in the outside world through an affective, hence subjective, lens. For example, in (23) the reporter uses a highly metaphorical idiom, 扑朔迷离 *pūshuò-mílí* (lit. 'hard to tell who is who between a jumping bunny couple'), to characterise the uncertainties surrounding a major political event.

23. 备受拖累, 两会行情的预期也由此变得扑朔迷离。 (A27)
bèi shòu tuōlèi liǎng-huì hángqíng de
severely get drag.down two-assembly prospect ATT
yùqí yě yóucǐ biàn de pūshuò-mílí
forecast also thus become DE bunny.couple.jumping
'This dragged down everything, making the prediction of the outcome of the two congressional sessions anyone's guess'.

Such a characterisation dramatises the political environment of the reported event and makes the report more emotional in comparison with a case like (15) that we saw earlier, repeated below. (15), as can be recalled, comes from another political event report; however, in this case, a relatively plain adjective form 黯淡 *àndàn* 'dark, glum' is used. In comparison with (23), considerably less emotional quality is expressed here, although the expression can still be argued to be metaphorical (using a dark colour describing a political prospect).

15. (政策)使巴以和平前景变得更加黯淡 。 (A05)
(*zhèngcè*) *shǐ Bā Yǐ hépíng qiánjǐng*
(policy) cause Palestine Israel peace outlook
biàn de gèngjiā àndàn
become DE even.more bleak
'Implementation of such policies made the Palestine and Israel peace outlook even more glum'.

Another comparison that we can make is to contrast the different types of idiom used to describe similar discourse objects. In (24) and

(25), for example, a common discourse entity, women, can be seen to be involved. In (24), soccer cheer-leader squads, typically consisting of young females, are associated with the sport event being described in the complement with the idiom 活色生香 *huósè-shēngxiāng* (lit. 'raising colours and spreading fragrance'). This metaphor, aided with the choice of 宝贝 *bǎobèi* 'babes' for the cheerleaders, of colour and scent applied to the female sex has a strong sexual connotation and indexes the way the writer projects their stance toward the role of the female cheerleader squads in the reported event (World Cup).

16. 有了足球宝贝, 世界杯变得更加活色生香。 (B29)
yǒu *le* *zúqiú* *bǎobèi* *Shìjiè* *Bēi* *biàn* *de*
have PFV soccer babe World Cup become DE
gèngjiā *huósè-shēngxiāng*
even.more raise.color.spread.fragrance
'With the soccer babes' presence, the World Cup becomes even more glitzy and attractive'.

By contrast, in (25), the author chooses to describe, with the idiomatic expression 风和日丽 *fēnghé-rìlì* (lit. 'calm wind and bright sunshine'), the environment (i.e. weather) where the female character is embedded. Here the overall imagery depicted is no less pleasant and uplifting than that of (24), yet it is free of any conceivable sexual biases.

17. (云)又突然全散了s天气又变得风和日丽, 织女也回到了家中 […] (F16)
(*yún*) *yòu* *túrán* *quán* *sàn* *le* *tiānqì*
(cloud) again suddenly totally dissipate PRT weather
yòu *biàn* *de* *fēnghé-rìlì* *Zhīnǚ*
again become DE caml.wind.pretty.sunshine Zhīnǚ
yě *huí* *dào* *le* *jiā-zhōng*
also return reach PVF home-in
'Once again all of a sudden the cloud dissipates completely. The weather then becomes sunny and bright with calming wind. Goddess Zhinü returns home as well […]'

These examples demonstrate clearly that choice of idiomatic complements over others is very much determined by the degree to which the speaker/writer projects their affective stance, and the different types of idioms chosen index divergent biases from which a stance is projected.

5.2 Delexical Verbs

Turning now to delexical verbs, the frequency distribution information for all eight of the identified verbs can be found in table 8.

Table 8 Delexical verbs and their complement types in the corpus

	Adjectival	Clausal	Formulaic	Idiomatic	Total
弄	3	10	2	8	23
打	8	4	2	2	16
做	9	0	0	1	10
搞	0	2	0	3	5
进行	4	0	0	0	4
办	0	0	1	2	3
干	2	1	0	0	3
作	1	0	0	0	1
Total	**27**	**17**	**5**	**16**	**65**

The most frequent token in this group is obviously 弄 *nòng* 'do, get, make', a prototypical delexical verb in Chinese (Tao, Hu 2019). Earlier through extracts (18)-(20), we have contrasted three utterances involving 弄 *nòng*, showing that with or without a complement and with different types of complement, the affective load can vary, again with idiomatic complements carrying the strongest load.

5.3 Psychological State Verb + Clausal Complement

Finally, let's take a look at some of the examples of verbs of psychological states and clausal complement constructions. The top five such tokens are given in table 9.

Table 9 Psychological state verbs and their complement types in the corpus

	Adjectival	Clausal	Formulaic	Idiomatic	Total
吓	3	13	0	1	17
忙	2	4	0	3	9
冻	1	1	0	4	6
急	1	5	0	0	6
羞	1	4	0	0	5

The most representative one is 吓 *xià* 'scare, frightened'. The patterns with 吓 *xià* constructions are of two types: in the first, the main agent and the agent of the complement clause are identical, as shown in (26) and (27).

18. 那么近，那么近。菁晓已吓得说不出话来。(L12)

nàme	*jìn*	*nàme*	*jìn*	*Jīngxiǎo*	*yǐ*	*xià*
that	close	that	close	Jingxiao	already	frightened

de *shuō* *bu* *chū* *huà* *lái*
DE speak NEG out word come
'It's so so close. Jingxiao is already too frightened to say anything'.

19. 一下便飞到了半空中。我吓得赶紧闭上了眼睛。 (L31)
yīxià *biàn* *fēi-dào* *le* *bànkōng-zhōng* *wǒ* *xià*
shortly already fly-reach PFV midair-in 1SG frighten
de *gǎnjǐn* *bì-shàng* *le* *yǎnjīng*
DE hurry close-up PFV eye
'It reached midair in no time. I was so frightened that I hurriedly closed my eyes'.

The second pattern involves an external agent causes a psychological state change of the subject in the complement clause. Thus in (28)-(30), the external forces of some naked person, the damaged poles and trees, and a sudden kiss cause 我 *wǒ* 'I', bike riders and passers-by, and Jianwen, respectively, to perform actions described in the complement clause in a panic manner.

20. 一位男人光着身体正站在我身旁买面包, 吓得我差点把两瓶果酱打翻在地。(N08)
yī *wèi* *nánrén* *guāng-zhe* *shēntǐ* *zhèng* *zhàn* *zài*
one CLF man naked-DUR body PROG stand by
wǒ *shēn* *páng* *mǎi* *miànbāo* *xià* *de* *wǒ*
1SG body next buy bread scare DE 1SG
chàdiǎn *bǎ* *liǎng* *píng* *guǒjiàng* *dǎfān* *zài* *dì*
nearly BA two jar jam throw.out to ground
'A naked man stood next to me checking out some bread, and this frightened me so much that I almost threw two jars of jam to the ground'.

21. 当场将电线杆和行道树砸断, 吓得骑车者、过路人惊慌奔逃 […] (A16)
dāngchǎng *jiāng* *diànxiàngān* *he* *xíngdàoshù* *zá*
on.spot JIANG utility.pole and street.tree crash
duàn *xià* *de* *qíchēzhě* *guòlùrén* *jīnghuāng* *bēntáo*
down frighten DE bike.rider passer.by panic run.away
'It crashed a utility pole and a street tree on the spot, scaring away bike riders and passers-by […]'

22. 这突如其来的一吻, 吓得健文一跳。(K40)
zhè *tūrúqílái* *de* *yī* *wěn* *xià* *de* *Jiànwén*
this unexpected ATT one kiss startle DE Jianwen
yī *tiào*
one jump
'The unexpected kiss startled Jianwen so much that he almost jumped'.

Given that these constructions tend to focus on a traumatising psychological effect and its ensuing consequences, a clausal complement serves the need nicely in being deployed to express the consequence component.

6 Conclusions

This study finds that CM/S constructions in a written Chinese corpus have preferred forms and functions. Formally speaking, a monosyllabic verb, preferably 变 *biàn* 'change, become', basic action verbs, or psychological state verbs tend to co-occur with complements of adjectival, clausal, or idiomatic expressions. CM/S are argued to be an assessment device indexing speaker evaluative and affective stances. The loaded affective meanings, we contend, account for the larger and more complex forms than their standard assessment counterparts.

We believe that these findings have important implications for a number of theoretical concerns. First, a corpus-based and corpus-driven mixed approach proves to be fruitful for investigating Chinese syntactic constructions. For example, while we began our study on the assumption of standard grammatical studies on CM/S forms, we let the corpus data drive us to the conclusion that key components (e.g. 变 *biàn* alone as a verbal predicate category or idiomatic expression as a complement category) and co-occurrence patterns (e.g. the mutual attraction of psychological state verbs and clausal complements) as stand-out attested categories must be recognised.

Second, with a usage-based approach and the view of construction grammar, investigation of syntactic structures can lead to new directions. While standard approaches to CC in Chinese have focused on issues such as semantic focus and what is called pragmatic meanings in topic-comment structure and information status, such views turn out to be rather limiting since constructional form-meaning pairing has shown that 1) different key components may display different tendencies in their co-occurrence with other constituents, and that 2) constructional meanings may differ from that of individual components (e.g. assessments of states and processes and affective loading may not be deducted from the complement or verbal predicate alone). In this regard, we believe that traditional concerns such as admissible elements and different kinds of focus in Chinse CC (action-centred *vs* other-than-action-centred) may be inadequate and need to be supplemented with the usage-based approach advocated here, which emphasises constructional meanings and functions of CM/S as an assessment device for affective stance marking, which in return explains their more complex forms and structural preferences.

Finally, given our own interest in comparing L1 and L2 language knowledge and acquisition processes, we believe that a realistic un-

derstanding of how CC, and CM/S in particular, work in the first language population provides a solid foundation as baseline data from which to evaluate L2 learning patterns and pedagogical practices: for example, how to prioritise teaching foci to reflect L1 constructional frequency information, including contingency information; how to focus the pedagogy on affective stance marking, and how to explain L2 developmental stages with CC and CM/S. We intend to explore those issues in a separate study (Jin, Zhang, Tao forthcoming).

Acknowledgements

We wish to thank the two anonymous reviewers for their careful reading of the paper and constructive suggestions and Fang Di for bibliographical assistance. The first author also acknowledges the support of a faculty research grant from the UCLA Academic Senate for 2019-20 and a University of Macau Distinguished Visiting Scholar award in July 2019, during which this project was initiated.

Appendix A: Verb Frequency Ranking List

Rank	V	Freq	Rank	V	Freq	Rank	V	Freq
1	变	132	42	唱	3	83	闷	2
2	弄	23	43	学	3	84	闹	2
3	过	22	44	干	3	85	飞	2
4	笑	20	45	惹	3	86	骂	2
5	说	18	46	改	3	87	骗	2
6	吓	17	47	离	3	88	下	1
7	写	16	48	累	3	89	传	1
8	打	16	49	跳	3	90	作	1
9	吃	15	50	争	2	91	刨	1
10	走	14	51	切	2	92	到	1
11	哭	11	52	升	2	93	刻	1
12	活	11	53	吹	2	94	剪	1
13	做	10	54	害	2	95	去	1
14	听	10	55	恨	2	96	吵	1
15	忙	9	56	找	2	97	呛	1
16	长	9	57	拍	2	98	喜	1
17	穿	8	58	挤	2	99	嚼	1
18	跑	8	59	握	2	100	围	1
19	看	7	60	搅	2	101	坠	1
20	冻	6	61	摔	2	102	堵	1
21	卖	6	62	撞	2	103	定	1
22	急	6	63	晒	2	104	当	1
23	放	6	64	杀	2	105	待	1
24	玩	6	65	泡	2	106	念	1
25	睡	6	66	洗	2	107	懂	1
26	聊	6	67	淋	2	108	扔	1
27	想	5	68	烤	2	109	托	1
28	搞	5	69	熬	2	110	扣	1
29	来	5	70	生	2	111	扩	1
30	羞	5	71	画	2	112	抓	1
31	讲	5	72	留	2	113	抢	1
32	坐	4	73	站	2	114	抱	1
33	开	4	74	考	2	115	拂	1
34	爱	4	75	肿	2	116	拉	1
35	谈	4	76	脱	2	117	拌	1
36	输	4	77	荡	2	118	拖	1
37	伤	3	78	记	2	119	招	1
38	刺	3	79	踢	2	120	挂	1
39	办	3	80	转	2	121	挖	1
40	压	3	81	逗	2	122	捆	1
41	叫	3	82	逼	2	123	捣	1

Rank	V	Freq
124	推	1
125	揉	1
126	提	1
127	摆	1
128	撇	1
129	撑	1
130	撕	1
131	撩	1
132	教	1
133	晃	1
134	栽	1
135	沾	1
136	涂	1
137	涨	1
138	混	1
139	滑	1
140	满	1
141	炒	1
142	炸	1
143	烧	1
144	烫	1
145	照	1
146	煮	1
147	理	1
148	用	1
149	疼	1
150	病	1
151	盯	1
152	砍	1
153	破	1
154	碎	1
155	等	1
156	给	1
157	耍	1
158	落	2
159	藏	1
160	蜇	1
161	补	1
162	读	1
163	败	1
164	起	1
165	蹲	1
166	躲	1
167	载	1
168	连	1
169	醉	1
170	铺	1
171	错	1
172	靠	1
173	驳	1
174	发展	4
175	感动	4
176	表现	4
177	进行	4
178	保持	3
179	折磨	3
180	控制	3
181	体现	2
182	兴奋	2
183	刻画	2
184	发挥	2
185	服侍	2
186	消失	2
187	考虑	2
188	装扮	2
189	了解	1
190	亲近	1
191	休息	1
192	出台	1
193	出落	1
194	剪裁	1
195	压抑	1
196	压迫	1
197	厌恶	1
198	厮杀	1
199	变性	1
200	叙说	1
201	吸引	1
202	回答	1
203	复习	1
204	完成	1
205	崇拜	1
206	工作	1
207	延续	1
208	建设	1
209	心疼	1
210	怀旧	1
211	愉快	1
212	感激	1
213	打扫	1
214	把握	1
215	挤压	1
216	掌握	1
217	掩映	1
218	掩饰	1
219	搅和	1
220	撩拨	1
221	收拾	1
222	暴露	1
223	树立	1
224	模拟	1
225	流行	1
226	消耗	1
227	消费	1
228	清洗	1
229	溶解	1
230	演奏	1
231	演绎	1
232	激励	1
233	照顾	1
234	生产	1
235	生活	1
236	知道	1
237	磨砺	1
238	纯洁	1
239	结合	1
240	编写	1
241	老练	1
242	衬托	1
243	装点	1
244	装饰	1
245	解释	1
246	认识	1
247	辉映	1
248	过渡	1
249	运用	1
250	销售	1
251	难过	1

Appendix B: V+Comp Distribution Patterns

V/C	Adjectival	Clausal	Formulaic	Idiomatic	Total
A	73	13	16	30	132
变	73	13	16	30	132
B	27	17	5	16	65
弄	3	10	2	8	23
打	8	4	2	2	16
做	9			1	10
搞		2		3	5
进行	4				4
办			1	2	3
干	2	1			3
作	1				1
C	77	8	9	17	111
过	17	3	1	1	22
活	5	1	1	4	11
长	7	1		1	9
放	6				6
玩	2		2	2	6
想	4		1		5
开	4				4
表现	2		1	1	4
输	2			2	4
学	3				3
控制	1		2		3
改		2		1	3
离	3				3
服侍				2	2
生	1		1		2
考虑	2				2
记	2				2
骗	1			1	2
了解	1				1
休息	1				1
到	1				1
完成	1				1
工作	1				1
当	1				1
懂	1				1
招		1			1
照顾	1				1
生产	1				1
生活	1				1

用	1				1
知道	1				1
结合				1	1
给	1				1
认识	1				1
运用				1	1
靠	1				1
D	146	34	23	51	254
笑	7	8	1	4	20
说	14	1		3	18
写	8		2	6	16
吃	10		2	3	15
走	10		4		14
哭	6	2		3	11
听	5	2		3	10
穿	5	1	1	1	8
跑	3	1		4	8
看	1	3		3	7
卖	4	1	1		6
睡	5	1			6
聊	5		1		6
来	4	1			5
讲	3		1	1	5
坐	3			1	4
谈	3			1	4
叫	1	1		1	3
唱	1		2		3
跳	1	1		1	3
争				2	2
切	1		1		2
吹		1		1	2
找	2				2
拍	2				2
挤				2	2
摔	1	1			2
晒	2				2
杀			1	1	2
洗	1			1	2
烤	2				2
熬	2				2
画	2				2
留	2				2
站	2				2
考	2				2

脱		2			2
装扮	1		1		2
踢	1		1		2
飞	2				2
骂		2			2
刨	1				1
剪	1				1
去	1				1
吵	1				1
嚼		1			1
回答		1			1
复习	1				1
抓		1			1
抱	1				1
拉			1		1
拖	1				1
挂	1				1
挖				1	1
捣	1				1
推	1				1
揉				1	1
摆				1	1
撕	1				1
收拾				1	1
教	1				1
暴露			1		1
栽	1				1
涂	1				1
消费			1		1
清洗			1		1
演奏	1				1
烧		1			1
烫	1				1
煮		1			1
盯	1				1
砍	1				1
编写				1	1
解释				1	1
读	1				1
起	1				1
蹲	1				1
躲	1				1
醉				1	1
铺				1	1

驳				1	1
E	25	49	3	12	89
吓	3	13		1	17
忙	2	4		3	9
冻	1	1		4	6
急	1	5			6
羞	1	4			5
感动		2		2	4
爱	2	1	1		4
伤	3				3
压		3			3
累	1	2			3
兴奋		2			2
恨		1		1	2
肿	2				2
闷	2				2
亲近			1		1
压抑	1				1
压迫			1		1
厌恶		1			1
喜		1			1
崇拜		1			1
心疼	1				1
怀旧		1			1
愉快		1			1
感激		1			1
涨	1				1
满		1			1
激励		1			1
疼	1				1
病	1				1
破		1			1
纯洁		1			1
老练	1				1
败				1	1
错	1				1
难过		1			1
F	49	36	9	24	118
发展	3	1			4
保持	3				3
刺		3			3
惹		3			3
折磨		2		1	3
体现	1			1	2

刻画				2	2
升	1		1		2
发挥	1			1	2
害		2			2
握	2				2
搅		2			2
撞		2			2
泡		1	1		2
消失	1			1	2
淋		2			2
荡		1	1		2
落				2	2
转	2				2
逗		2			2
逼		2			2
闹	1	1			2
下	1				1
传	1				1
出台	1				1
出落			1		1
刻				1	1
剪裁	1				1
厮杀				1	1
变性	1				1
叙说	1				1
吸引		1			1
呛		1			1
围	1				1
坠	1				1
堵				1	1
定	1				1
延续	1				1
建设	1				1
待		1			1
念		1			1
打扫				1	1
扔		1			1
托			1		1
扣			1		1
扩	1				1
把握	1				1
抢				1	1
拂			1		1
拌			1		1

挤压	1				1
捆		1			1
掌握	1				1
掩映	1				1
掩饰	1				1
提	1				1
搅和				1	1
撇				1	1
撑		1			1
撩	1				1
撩拨	1				1
晃		1			1
树立		1			1
模拟				1	1
沾		1			1
流行		1			1
消耗	1				1
混	1				1
溶解			1		1
滑	1				1
演绎				1	1
炒				1	1
炸	1				1
照				1	1
理				1	1
碎				1	1
磨砺	1				1
等				1	1
耍	1				1
藏	1				1
蜇				1	1
补	1				1
衬托				1	1
装点	1				1
装饰	1				1
载	1				1
辉映		1			1
过渡	1				1
连	1				1
销售	1				1
Total	397	157	65	150	769

Bibliography

Addinsoft (2020). *XLSTAT Statistical and Data Analysis Solution*. New York. https://www.xlstat.com.

Biber, D. (2009). "Corpus-Based and Corpus-Driven Analyses of Language Variation and Use". Heine, B.; Narrog, H. (eds), *The Oxford Handbook of Linguistic Analysis*. 1st ed. Oxford: Oxford University Press, 159-92. https://doi.org/10.1093/oxfordhb/9780199544004.013.0008.

Biber, D. et al. (1999). *Longman Grammar of Spoken and Written English*. Harlow: Pearson Education Limited.

Bybee, J.; Thompson, S.A. (2000). "Three Frequency Effects in Syntax". *Berkeley Linguistic Society*, 23(1), 65-85. https://doi.org/10.3765/bls.v23i1.1293.

Chao, Y.R. (1968). *A Grammar of Spoken Chinese*. Berkeley: University of California Press.

Chen X. 陈小曼 (2013). "Jiyu yuyi zhixiang fenxi de 'de' zi ju Yingyi yanjiu" 基于语义指向分析的"得"字句英译研究 (A Semantic Focus-Based Study of English Translations of *De* Constructions). *Waiguo Yuwen*, 29(5), 113-18.

Cheung, H.-N.S. et al. (1994). *A Practical Chinese Grammar*. Hong Kong: Chinese University Press.

Chu, C.C. (1983). *A Reference Grammar of Mandarin Chinese for English Speakers*. New York: Peter Lang.

Ding P. 丁萍 (2012). "Lun jieguo buyu 'hao' yu 'wan' dui shuyu dongci de xuanze jizhi" 论结果补语"好"与"完"对述语动词的选择机制 (Mechanisms of the Selection of Resultative *Hao* and *Wan* and Verbal Predicates). *Xibei Minzu Daxue Xuebao. Zhexue Shehui Kexue Ban*, 2, 99-104.

Du Bois, J.W. (2007). "The Stance Triangle". Englebretson, R. (ed.), *Stancetaking in Discourse. Subjectivity, Evaluation, Interaction*. Amsterdam; Philadelphia: John Benjamins, 139-82.

Englebretson, R. (1997). "Genre and Grammar. Predicative and Attributive Adjectives in Spoken English". *Berkeley Linguistic Society*, 23(1), 411-21. https://doi.org/10.3765/bls.v23i1.1272.

Englebretson, R. (ed.) (2007). *Stancetaking in Discourse. Subjectivity, Evaluation, Interaction*. Amsterdam; Philadelphia: John Benjamins.

Fan X. 范晓 (1992). "V de ju de 'de' hou chengfen" V 得句的"得"后成分 (Post-*De* Elements in the V *De* Construction). *Hanyu Xuexi*, 6, 5-8.

Fang D. 方迪 (2018). *Hanyu Kouyu Pingjia Biaoda Yanjiu. Jiyu Hudong Shijiao* 汉语口语评价表达研究——基于互动视角 (Expressions of Assessment in Spoken Chinese: An Interactional Approach) [PhD Dissertation]. Beijing: The Chinese Academy of Social Sciences.

Fillmore, C.; Kay, P.; O'Connor, C. (1988). "Regularity and Idiomaticity in Grammatical Constructions. The Case of *Let Alone*". *Language*, 64, 501-38. https://doi.org/10.2307/414531.

Glynn, D. (2014). "Correspondence Analysis. Exploring Data and Identifying Patterns". Glynn, D.; Robinson, J.A. (eds), *Corpus Methods for Semantics. Quantitative Studies in Polysemy and Synonymy*. Amsterdam: John Benjamins, 443-85.

Goldberg, A.E. (1995). *Constructions. A Construction Grammar Approach to Argument Structure*. Chicago; London: University of Chicago Press.

Goldberg, A.E. (2003). "Constructions. A New Theoretical Approach to Language". *Trends in Cognitive Sciences*, 7(5), 219-24. https://doi.org/10.1016/s1364-6613(03)00080-9.

Goodwin, C.; Goodwin, M.H. (1987). "Concurrent Operations on Talk. Notes on the Interactive Organization of Assessments". *IPRA Papers in Pragmatics*, 1(1), 1-54. https://doi.org/10.1075/iprapip.1.1.01goo.

Goodwin, C.; Goodwin, M.H. (1992). "Assessments and the Construction of Context". Goodwin, C.; Duranti, A. (eds), *Rethinking Context. Language as an Interactive Phenomenon*. Cambridge: Cambridge University Press, 147-89.

Gries, S.T.; Ellis, N.C. (2015). "Statistical Measures for Usage-Based Linguistics". *Language Learning*, 65(S1), 228-55. https://doi.org/10.1111/lang.12119.

Haiman, J. (1983). "Iconic and Economic Motivation". *Language*, 59(4), 781-819. https://doi.org/10.2307/413373.

Hopper, P.J.; Thompson, S.A. (1984). "The Discourse Basis for Lexical Categories in Universal Grammar". *Language*, 60(4), 703-52. https://doi.org/10.2307/413797.

Jiang, C. 姜春华 (2005). *Xiandai Hanyu zhuangtai buyu yu chengdu buyu yanjiu* 现代汉语状态补语与程度补语研究 (A Study in State and Degree Complements) [PhD Dissertation]. Shanghai: Shanghai Normal University.

Jin, H.G.; Zhang, J.; Tao, H. (forthcoming). "A Comparative Corpus Study on L1 and L2 Verb Complement Constructions of Manner (VCM)". Chen, H.; Mochizuki, K.; Tao, H. (eds), *Learner Corpora: Construction and Explorations in Chinese and Related Languages*. Singapore: Springer Nature.

Li, C.; Thompson, S.A. (1981). *Mandarin Chinese. A Functional Reference Grammar.* Berkeley: University of California Press.

Li L. 李临定 (1963). "Dai 'de' zi de buyuju" 带"得"字的补语句 (Complements with *De*). *Zhongguo Yuwen*,5, 396-410.

Li L. 李临定 (1980). "Dongbuge jushi" 动补格句式 (Complement Constructions). *Zhongguo Yuwen*, 2, 93-102.

Li X. 李小荣 (1994). "Dui shujieshi dai binyu gongneng de kaocha" 对述结式带宾语功能的考察 (An Investigation of Resultative Complements Taking an Object). *Hanyu Xuexi*, 1994(5), 32-8.

Liu D. 刘丹青 (2005). "Cong suowei 'buyu' tan gudai Hanyu yufaxue tixi de canzhao xi" 从所谓"补语"谈古代汉语语法学体系的参照系 (Baseline Reference Systems for Classical Chinese Grammar Based on the So-Called Complements). *Hanyu Shi Xuebao*, 5, 37-49.

Lu B. 陆丙甫; Ying X. 应学凤; Zhang G. 张国华 (2015). Zhuangtai buyu shi Hanyu de xianhe chengfen 状态补语是汉语的显赫成分 (State Complements as Salient Features of Chinese Grammar). *Zhongguo Yuwen*, 3, 195-205+287.

Lu J. 鲁健骥 (1992). "Zhuangtai buyu de yujing beijing ji qita" 状态补语的语境背景及其他 (Contextual Factors and Other Issues in State Complements). *Yuyan Jiaoxue yu Yanjiu*, 1, 32-42.

Lu J. 鲁健骥 (1993). "Zhuangtai buyu de jufa, yuyi, yuyong fenxi zai jiaoxue zhong de yingyong" 状态补语的句法、语义、语用分析在教学中的应用 (Syntactic, Semantic, and Pragmatic Analyses of State Complements and Teaching Applications). *Yuyan Jiaoxue yu Yanjiu*, 2, 22-31.

Lü S. 吕叔湘 (1979). *Hanyu yufa fenxi wenti* 汉语语法分析问题 (Issues in Analysing Chinese Grammar). Beijing: Commercial Press.

Ma T. 马婷婷; Chen B. 陈波 (2014). "Jieguo buyu 'dao' shiyong de yuyi tiaojian fenxi" 结果补语"到"使用的语义条件分析 (Semantic Conditions for the Use of Complements with *Dao*). *Linyi Daxue Xuebao*, 36(3), 55-8.

McCarthy, M.J. (1998). *Spoken Language and Applied Linguistics*. Cambridge: Cambridge University Press.

McCarthy, M.J.; Carter, R.A. (1997). "Grammar, Tails and Affect. Constructing Expressive Choices in Discourse". *Text*, 17(3), 405-29. https://doi.org/10.1515/text.1.1997.17.3.405.

McEnery, A.; Xiao, Z. (2004). "The Lancaster Corpus of Mandarin Chinese. A Corpus for Monolingual and Contrastive Language Study". *Proceedings of the Fourth International Conference on Language Resources and Evaluation* (*LREC*) *2004* (Lisbon, 24-30 May 2004), 1175-8.

Nie Z. 聂志平 (1992). "Youguan 'de' ziju de jige wenti" 有关"得"字句的几个问题 (Some Issues in *De* Constructions). *Liaoning Shifan Daxue Xuebao*, 3, 52-8.

O'Keeffe, A.; McCarthy, M.; Carter, R. (2007). *From Corpus to Classroom. Language Use and Language Teaching*. Cambridge: Cambridge University Press.

Ono T.; Thompson, S.A. (2009). "Fixedness in Japanese Adjectives in Conversation. Toward a New Understanding of a Lexical (Part-of-Speech) Category". Corrigan, R. et al. (eds), *Formulaic Language*. Amsterdam: Benjamins, 117-45.

Pomerantz, A.M. (1984). "Agreeing and Disagreeing with Assessment. Some Features of Preferred/Dispreferred Turn Shapes". Atkinson, J.M.; Heritage, J. (eds), *Structure of Social Action. Studies in Conversation Analysis*. Cambridge: Cambridge University Press, 57-101.

Rosch, E. (1973). "Natural Categories". *Cognitive Psychology*, 4(3), 328-50. https://doi.org/10.1016/0010-0285(73)90017-0.

Rosch, E.; Mervis, C. (1975). "Cognitive Representations of Semantic Categories". *Journal of Experimental Psychology. General*, 104(3), 192-233. https://doi.org/10.1037/0096-3445.104.3.192.

Shen J. 沈家煊 (2003). "Xiandai Hanyu dongbu jiegou de leixingxue kaocha" 现代汉语"动补结构"的类型学考察 (A Typological Investigation of Verb Complement Constructions in Modern Chinese). *Shijie Hanyu Jiaoxue*, 3, 17-23.

Sinclair, J. (ed.) (1990). *Collins Cobuild English Grammar.* Glasgow: HarperCollins.

Tao, H.; Hu, J. (2019). "Structural, Semantic, and Pragmatic Properties of *Nong* (弄) Constructions in Mandarin Discourse. Evidence from Corpora". *International Journal of Chinese Linguistics*, 6(1), 162-76. https://doi.org/10.1075/ijchl.18003.tao.

Tao, H.; Xiao, R. (2007-20). *The UCLA Corpus of Written Chinese.* Los Angeles; Lancaster: University Centre for Computer Corpus Research on Language.

Thompson, S.A. (1988). "A Discourse Approach to the Cross-Linguistic Category 'Adjective'". Hawkins, J. (ed.), *Explanations for Language Universals*. Oxford: Blackwell, 167-85.

Thompson, S.A.; Fox, B.A.; Couper-Kuhlen, E. (2015). *Grammar in Everyday Talk*. Cambridge: Cambridge University Press.

Thompson, S.A.; Tao, H. (2010). "Conversation, Grammar, and Fixedness. Adjectives in Mandarin Revisited". *Chinese Language and Discourse*, 1(1), 3-30. https://doi.org/10.1075/cld.1.1.01tho.

Wang H. 王红旗 (2001). "Dongjieshi shubu jiegou zai ba zi ju he chongdong ju zhong de fenbu" 动结式述补结构在把字句和重动句中的分布 (Distribution

of Verb Complements in *Ba* and Verb Reduplication Constructions). *Yuwen Yanjiu*, 1, 6-11.

Wang Y. 王媛 (2011). "Xiandai Hanyu dongjieshi de jinxingti" 现代汉语动结式的进行体 (The Progressive Aspect of Modern Chinese Resultative Complements). *Yuyan Kexue*, 10(1), 70-82.

Wu X. 吴笑双 (2018). "Xiaoxue yuwen kewen zhong de dongbu jiegou jiliang jiqi yuyi guanxi yanjiu" 小学语文课文中的动补结构计量及其语义关系研究 (A Qualitative Study of Verb Complement Structures in Primary School Chinese Textbooks) [MA Thesis]. Nanjing: Nanjing Normal University.

Wu Y. 吴颖 (2002). "Dongci + 'de' + buyu de fenlei he yuyi tezheng fenxi" 动词+"得"+补语的分类和语义特征分析 (Types of Verb + Complements and Their Semantic Features). *Suzhong Daxue Xuebao* (*Zhexue Shehui Kexue Ban*), 2, 91-4.

Xiao, R.; Rayson, P.; McEnery, T. (2009). *A Frequency Dictionary of Mandarin Chinese. Core Vocabulary for Learners*. *Routledge Frequency Dictionaries*. London; New York: Routledge.

Zhang, H. et al. (2002). "Automatic Recognition of Chinese Unknown Words Based on Role Tagging". *Proceedings of the 1st SIGHAN Workshop, COLING 2002* (Taipei, 24 August-1 September 2002), 71-7.

Zhang Y. 张豫峰 (2002). "'De'ziju buyu de yuyi zhixiang" "得"字句补语的语义指向 (Semantic Focuses of *De* Complements). *Shanxi Shida Xuebao*, 29(1), 116-19.

Zhang, Z. (2017). *Dimensions of Variation in Written Chinese*. London: Routledge.

Corpus-Based Research on Chinese Language and Linguistics
edited by Bianca Basciano, Franco Gatti, Anna Morbiato

Chinese Sentence-Initial Indefinites: What Corpora Reveal

Anna Morbiato
Università Ca' Foscari Venezia, Italia; The University of Sydney, Australia

Abstract While the sentence-initial position in Chinese is generally related to givenness/definiteness, instances of informationally new or indefinite sentence-initial NPs may be found in language in use. This paper systematically explores the phenomenon of sentence-initial indefinites (SIIs), their statistical relevance, and the interaction with features typically connected to linear order, such as animacy or locatability. Results of a quantitative and qualitative analysis conducted on three major big-size, generalised corpora show that SIIs in Chinese are not only possible, but also statistically relevant. Animacy and locatability are found to play a key role in increasing SIIs acceptability. Finally, data reveal a new pattern featuring SIIs with proper nouns.

Keywords Sentence-initial indefinites (SIIs). Chinese. Animacy. Information structure. Corpus study. Quantitative analysis. Qualitative analysis.

Edizioni Ca'Foscari

Sinica venetiana 6
e-ISSN 2610-9042 | ISSN 2610-9654
ISBN [ebook] 978-88-6969-406-6 | ISBN [print] 978-88-6969-407-3

Peer review | Open access
Submitted 2020-05-11 | Accepted 2020-08-18 | Published 2020-12-21

DOI 10.30687/978-88-6969-406-6/002

1 Introduction[1]

The sentence-initial position in Chinese is generally associated with, and often defined in terms of, a specific information status, i.e. that of givenness/identifiability and, consequently, definiteness. This association is widely accepted in the literature (Xu 1995) and is supported by the fact that bare nouns in Chinese receive a definite reading when preverbal (1a). Furthermore, it is often maintained that indefinite NPs cannot occur in the sentence-initial position (1b): to be first introduced, indefinites should be preceded by an existential or presentational verb, and then predicated upon, hence the construction in (1c) – all examples from Hole (2012, 61):

1. a. 外国人遇到了张三。
wàiguórén *yùdào-le* *Zhāngsān*
foreigner meet-PFV Zhangsan
'The foreigner met Zhangsan'.

   b. * 一个外国人遇到了张三。
yí *ge* *wàiguórén* *yùdào-le* *Zhāngsān*
one CLF foreigner meet-PFV Zhangsan
'A foreigner met Zhangsan'.

   c. 有一个外国人遇到了张三。
yǒu *yí* *ge* *wàiguórén* *yùdào-le* *Zhāngsān*
exist one CLF foreigner meet-PFV Zhangsan
'A foreigner met Zhangsan'.

In Li and Thompson's grammar, the sentence-initial position is the position for the topic, which "always refers either to something that the hearer already knows about – that is, it is definite – or to a class of entities – that is, it is generic" (1981, 85). Newly-introduced referents cannot be topics, hence they "must follow the main verb of the presentative sentence" (1981, 509), as in (1c). Most subsequent literature on topic-comment structures and word order makes similar observations (Chu 2006; Li 2005; Shyu 2016; Tsao 1977, 1989; Xu 1995; Xu, Liu 2007; Zhu 1982, among others); Ho (1993) holds that the fact that the sentence-initial position should be occupied by a definite el-

1 In this paper, I use the term 'Chinese' to refer to *Pǔtōnghuà*, the standard language of the PRC. Simplified Chinese characters and the *Pinyin* romanisation system have been used throughout the article. The glosses follow the general guidelines of the Leipzig Glossing Rules. Additional glosses include: BEI = 'Chinese 被 *bèi* marker'; COS = 'change of state'; EXP = 'experiential aspect'; MKR = 'marker'; NMLZ = 'nominalizer'; SFP = 'sentence-final particle'; SP = 'structural particle'. I am very grateful to the two anonymous reviewers for their constructive comments and suggestions.

ement "is so strictly adhered to that [...] Chinese has a last resort, which is to prefix a dummy verb 有 *you* [...] to postpone the indefinite NP in the initial position", as in (1c).

However, observations have been raised against the generalisations above. In particular, it has been noted that not all sentence-initial referents are informationally old, i.e. known both to the hearer and to the speaker (Paul 2015): they may be specific – i.e. non identifiable by the hearer – and even indefinite (Bisang 2016; Lu, Pan 2009; Morbiato 2018; Wu 1998). The possibility of indefinites to occur sentence-initially was also stressed by Fan (1985) and subsequent literature by Chinese scholars (Fang 2019; Fu 2013; Liu 2018; Liu, Zhang 2004; Lu, Pan 2009; Tang 2011; Wang 2003; Xu 1997, 1999; Zhang 2007; Zhou, Chen 2013, among others) on so-called 'indefinite-subject sentences' (无定主语句 *wúdìng zhǔyǔ jù*) (see § 2) and is borne out by corpus data:

2. 一位年轻助教谈起了他刚读过一本关于文物保护的著作 [...] (PKUcorpus)
   yí *wèi* *niánqīng* *zhùjiào* *tán-qǐ-le*
   one CLF young teaching.assistant tell-start-PFV
   tā *gāng* *dú-guo* *yì* *běn* *guānyú* *wénwù*
   3SG.M just read-EXP one CLF on cultural.relic
   bǎohù *de* *zhùzuò*
   protection SP work
   'A young teaching assistant started telling he had just read a book on cultural heritage protection'.

This challenges the widely accepted association of the sentence-initial position with topichood, givenness, and definiteness, as well as analyses that postulate a definiteness restriction on the sentence-initial position. However, several aspects of sentence-initial indefinites (henceforth SIIs) in Chinese have not yet been fully explored: how widespread is this phenomenon? How does it interact with other features typically connected to the sentence-initial position (such as animacy and locatability)? Crucially, corpus-based studies on the topic remain the minority and are usually conducted on relatively small, genre-specific corpora.

This paper adopts corpus methodologies and tools to investigate SIIs, with a particular focus on determining (i) the statistical relevance of SIIs of the type of '一 *yī* CLF N' in big-size corpora and (ii) its interaction with the semantic feature of animacy and, secondly, with the referential property of locatability. To this end, it proposes the results of a large-scale, quantitative and qualitative analysis conducted on three major big-size, generalised corpora, namely the PKU CCL corpus (Centre for Chinese Linguistics, Peking University, 470 million characters, henceforth PKU), the BCC corpus of Modern Chinese (Beijing Language and Culture University, 15 billion characters, henceforth BCC), and the Sketch Engine ZHTenTen (Stanford

Tagger) simplified Chinese corpus (13,5 billion characters, henceforth ZHTenTen (ST)). A corpus approach is chosen as it contributes to grounding the analysis on empirical, natural data: corpora allow adhering more to real language in use; moreover, they may help reveal new patterns or phenomena, thus contributing towards deeper and more complete linguistic descriptions even for languages that are over-described, like Chinese.

The rest of the article is organised as follows: § 2 provides an overview of the literature on Chinese SIIs and their characteristics. § 3 presents the study, its research questions, methodology, and linguistic data. §§ 4 and 5 discuss the findings of the quantitative and qualitative analyses, respectively. § 6 draws the conclusions and briefly discusses the implications of such findings on theoretical accounts of the sentence structure of Chinese and onto Chinese as a second/foreign language teaching.

2 (In)definiteness and the Sentence-Initial Position in the Literature

The term 'definiteness' denotes a grammatical category featuring a formal distinction that marks an NP as *identifiable*:[2] this formal distinction may consist of a variety of grammatical means, "including phonological, lexical, morphological, and word order" (Chen 2015, 408). Among the first linguists that associated definiteness with word order in Chinese is Chao, who claims that the encoding of definite/indefinite reference is not much connected to grammatical functions (subject/object): rather, it is the "position in an earlier or later part of the sentence that makes the difference" (1968, 76-7). Crucially, Chao himself proposes a counterexample of SII, of the type of a thetic judgement (3a), commenting that it is a less preferred pattern if compared to the definite>verb>indefinite pattern displayed by (3b):

3. a. 一个卖刷子的在门口呐。

yí	*ge*	*mài*	*shuāzi*	*de*	*zài*	*ménkǒu*	*na*
one	CLF	sell	brush	NMLZ	be.at	door	SFP

b. 门口有一个卖刷子的。

ménkǒu	*yǒu*	*yí*	*ge*	*mài*	*shuāzi*	*de*
door	exist	one	CLF	sell	brush	NMLZ

'A brush peddler is at the door'.

2 Identifiability is an addressee-oriented notion relating to the speaker's assumptions as to whether the addressee "is able to identify the particular entity in question among other entities of the same or different class in the context" (Chen 2015, 408).

Li and Thompson (1981, 167-8) also identify exceptions to their above-mentioned definiteness restriction to the preverbal position, which they illustrate with sentences in (4a)-(4d). All four sentences feature sentence-initial NPs of the type of '一 *yī* CLF N'; however, Li and Thompson hold that such exceptions are only apparent: all sentence-initial NPs in (4) are indeed formally indefinite, but according to them they all receive a definite reading. In (4a), 一 *yī* refers to a specific "absolute quantity" and is therefore definite; in (4b), 一 *yī* in fact means "each", hence, it is not indefinite; in (4c)-(4d), they maintain, 一 *yī* introduces "something that is part of an entity already known by the hearer" (i.e. the leg of a known person, the peasants of a known village) and "can therefore be considered a definite noun phrase":

4. a. 一个人就够了。
yí ***ge*** ***rén*** *jiù* *gòu* *le*
one CLF person then (be).enough PFV/COS
'One person will be enough'.

b. 一个人吃一口。
yí ***ge*** ***rén*** *chī* *yì* *kǒu*
one CLF person eat one mouth
'Each person gets one mouthful'.

c. 一条腿断了。
yì ***tiáo*** ***tuǐ*** *duàn-le*
one CLF leg break-PFV/COS
'One of its legs is broken'.

d. 一个农夫说,"我想出一个办法了"。
yí ***gè*** ***nóngfū*** *shuō* *wǒ* *xiǎng-chū* *yí* *ge* *bànfǎ* *le*
one CLF peasant say 1SG think-exit one CLF way COS
'A peasant said "I've thought of a way"'.

Indeed, the examples above show that not all sentence-initial NPs of the type of '一 *yī* CLF N' are true indefinites. They may emphasise the *quantity* (4a) or receive a *distributive* reading (4b) (see also Lu, Pan 2009). Other readings are possible, e.g. *generic* reference (to a specific class), as in (5) below:

5. 一个年轻人应当有志气。 (Lu, Pan 2009)
yí ***ge*** ***niánqīng*** ***rén*** *yīngdāng* *yǒu* *zhìqì*
one CLF young man should have ambition
'A young man / Young men should be ambitious'.

However, the underlined NPs in (4c)-(4d) can hardly be labelled as definite. In (4c), the implicit body-part (or possession/containment etc.)

relationship might enable the hearer to identify the referent the leg belongs to; however, which specific leg is broken (left/right?) is not identifiable. Similarly, in (4d), 一个农夫 *yí ge nóngfū* 'a peasant' might be assumed to be specific (known by the speaker) but can hardly be considered identifiable by the hearer, especially with no context. On the other hand, the context of these utterances may render the referent *locatable* (Morbiato 2018; Wu 1998), i.e. located within a given/identifiable set (i.e. the two legs) or setting (i.e. the village where the peasant lives; the notion of locatability will be discussed in more depth below). Moreover, none of Li and Thompson's explanations account for Chao's example in (3), a SII *tout court*.

Some scholars put forward a more nuanced view of the definiteness-preverbal position association: Chen (2015, 410), for example, talks about definiteness- and indefiniteness-inclined positions, holding that preverbal NPs are overwhelmingly, but not exclusively, definite. Hole (2012, 61-2), after commenting on (1) that "subject DPs in Chinese must be interpreted as definite", adds that indefinite subjects are barred from the sentence-initial position in *non-thetic* (i.e. all-focus, topicless) sentences, thus implying that SIIs may occur in thetic judgements. However, examples of thetic sentences he includes, such as 一张床睡三个人 *yì zhāng chuáng shuì sān ge rén* 'one bed accommodates three people', do not display an indefinite reading, but rather a distributive one. Lu, Zhang and Bisang (2015) and Bisang (2016) go one step further, arguing that subjects, unlike topics, may be indefinite (they see indefiniteness as a subjecthood test): in thetic sentences, they claim, "preverbal indefinite subjects are acceptable" (Bisang 2016, 356):

6. 一个杯子被我打碎了。[3] (Bisang 2016, 356)
   yí *ge* *bēizi* *bèi* *wǒ* *dǎ-suì-le*
   one CLF cup BEI 1SG hit-break-PFV/COS
   'A cup was broken by me'.

Major contributions to the literature on SIIs come from Chinese scholars. In his influential paper, Fan (1985) notes that SIIs are not only possible, but also rather common in some genres such as news reports: sentences with indefinite subject NPs, he claims, do constitute a sentence pattern in Chinese – they are neither uncommon nor peculiar. Since then, a number of studies have followed (Fang 2019; Fu 2013; Liu 2018; Liu, Zhang 2004; Lu, Pan 2009; Tang 2011; Wang 2003; Xu

3 Note, however, that such a string in Google obtains only 5 results, none of which are thetic sentences (they all have a topic beforehand). A similar string with a third person pronoun 他 *tā* 'he', as in 一个杯子被他打碎了 *yí ge bēizi bèi tā dǎ suì-le* 'a glass was broken by him' gives two occurrences, both of which in grammars that list the sentence as ungrammatical.

1997, 1999; Zhang 2007; Zhou, Chen 2013, among others), mostly focusing on the semantic and syntactic characteristics that license or increase the acceptability of SIIs. Generally, these regard: (i) the type of predicate – highly transitive, dynamic, and stage-level predicates are preferred over low-transitive, stative, and individual-level ones; (ii) the referential characteristics of the SII – the more information is provided that increases the referent's identifiability, the higher the SII's acceptability; and (iii) information structure – thetic sentences may host SIIs, especially when the referent is locatable in clear spatio-temporal frames. In what follows, main contributions will be briefly presented, with particular reference to corpus-based studies.

Several scholars focused on singling out properties and related licensing conditions to SIIs. Tang (2005) holds that SIIs are acceptable only in highly transitive sentences. Zhang (2007) concludes that SIIs occur in topicless (非主题判断 *fēi zhǔtí pànduàn*) – i.e. thetic – judgements, whereby the entire clause is a single unit conveying new information. Lu and Pan (2009) elaborate on this and claim that SIIs occur in (a) thetic sentences, where the whole predicate is projected into the core domain and is constrained by an existence operator, and (b) with stage-level predicates (expressing an event), but not with individual-level predicates (that express some judgement). Chen (2015) also remarks that SIIs are more acceptable with dynamic predicates but hardly occur as subject with stative ones (7):

7. *一个人很聪明。 (Chen 2015, 410)

yí	*ge*	*rén*	*hěn*	*cōngming*
one	CLF	person	very	smart

'One person is very smart'.

With reference to the above considerations, Wang (2003), Huang (2004), Wei and Chu (2007), and Lu and Pan (2009), among others, put forward a number of corollary licensing conditions to SIIs – e.g. SIIs cannot occur with modal verbs, negative adverbs, and tense. However, corpus studies found that most of these conditions are only tendencies, as counterexamples can be found for each parameter. Specifically, Zhou and Chen (2013) measured the descriptional accuracy of such licensing conditions with the method of parameter setting and measurement against a relatively small test corpus (i.e. a 1,000-sentence subcorpus of the PKU). From their analysis, it appears that all factors indeed contribute through a complex interplay to increasing SII's identifiability, and hence acceptability rate, but none constitutes an absolute restriction.

A widely accepted generalisation on SIIs is that the greater the amount of information on the referent (e.g. by means of longer nominal modifiers), the higher its degree of identifiability and, hence, its acceptability (Xu 1999). Wang (2003), for example, talks about degree

of (cognitive) *accessibility* (可及度 *kějídù*) and of *identifiability* (个体化程度 *gètǐhuà chéngdù*). Indeed, the acceptability difference between (8a) and (8b) lies in the long, informationally-rich (complex relative clause plus noun) modifier of the SII:

8. a. *一种方法最近问世。 (Zhou, Chen 2013, 373)
yì zhǒng fāngfǎ zuìjìn wènshì
one CLF method recently come.out
'A method was recently introduced'.

b. 一种取几根头发就可准确断定被检测者是不是吸毒者的检毒方法最近问世。
yì zhǒng qǔ jǐ gēn tóufǎ jiù kě zhǔnquè
one CLF pick some CLF hair then can accurately
duàndìng bèijiǎncèzhě shì-bú-shì xīdúzhě de
determine subject be-NEG-be drug.addict SP
jiǎndú fāngfǎ zuìjìn wènshì
detection method recently come.out
'A hair drug test for accurately determining whether a subject is a drug addict has recently come out'.

A very interesting perspective is provided by Fu's (2013) corpus-based, diachronic study, which reveals that SIIs very likely originated during the Song Dynasty (960-1279) and evolved from earlier constructions whereby an indefinite NP is the subject of the sentence following a perceptual verb, like 见 *jiàn* 'see'. Early instances of 'see' + indefinite NP patterns – e.g. (9) from *Zhuangzi* – also specify the scene witness (the <seer>, in this case King Wen). Later, the construction became impersonal, by means of markers that express the idea of 'seeing', such as 只见 *zhǐjiàn* and 则见 *zéjiàn*: sentences like (10) are interpreted as if the witness were an omniscient narrator. Later, these markers disappeared (11) (all examples are from Fu 2013):

9. 文王观于臧, 见一丈人钓 […] (*Zhuangzi*, *Tianzifang*)
Wén wáng guān yú Zāng ***jiàn*** *yí zhàngrén diào*
Wen king look SP Zang **see** one man fish
'King Wen was (once) looking about him at Zang, when he saw an old man fishing […]'[4]

4 Translation source: the *Chinese Text Project* (https://ctext.org).

10. 两边人犹未散, 只见一个庄客在东边墙角下叫道 […] (*Stories to Awaken the World*, 1627)

liǎng-biān	*rén*	*yóu*	*wèi*	*sàn*		
two-part	people	still	NEG	scatter		
zhǐjiàn	*yí*	*ge*	*zhuāngkè*	*zài*	*dōng-bian*	*qiángjiǎo*
MKR	one	CLF	farm.worker	at	east-side	corner
xià	*jiàodào*					
under	say					

'The people had not yet scattered; a farm worker at the east corner said […]'

11. 正说处, 一个小和尚点了灯来请洗澡。 (*Journey to the West*, § 62)

zhèng	*shuōchù*	*yí*	*ge*	*xiǎo*	*héshang*	*diǎn-le*	*dēng*
right	say.out	one	CLF	little	monk	light-PFV	lamp
lái	*qǐng*	*xǐzǎo*					
come	invite	shower					

'As they were talking, a young monk came in to light the lamp and invite Sanzang to take his bath'.[5]

Locatability. From the data in the literature analysed so far, an important feature of SIIs that scholars, however, never explicitly mention seems to be locatability, intended as identifiability of the referent's setting rather than identifiability of the referent itself. An example of non-identifiable, locatable referent is the sentence-initial NP in *a person in the airplane started shouting*: the hearer (and even the speaker) might not know who this person is, but they are definitely able to locate the referent within the group of people on that specific airplane. In other words, the referent itself is not identifiable: what can be identified is the scene/setting/set/frame where the referent is located. Locatability is typically granted by the presence of a phrase that expresses a temporal or spatial frame for the utterance, which is an inherent characteristic of Chinese topics (Chafe 1976; Her 1991; Morbiato 2018; Paul 2015) and is the property Li and Thompson tried to recall with respect to (4c)-(4d): the referents are not identifiable/definite, but rather locatable within a known set – one of two legs of an individual in (4c) – or a temporal/spatial setting – one of the peasants of a known village in (4d). This also suggests that locatability, rather than givenness and identifiability, is a more accurate restriction to the preverbal position in Chinese (see Morbiato 2018, 2020 for discussion). This is confirmed by Liu and Zhang's (2004) corpus investigation of eight novels and children stories: most (although no statistics are provided) of the SIIs they detected feature a temporal or spatial reference occurring before the indefinite NP. Such tem-

5 Translation from 'Internet archive' (`https://bit.ly/3pu33AZ`).

poral or spatial reference situates the referent within identifiable spatio-temporal coordinates. It may be either a phrase (12) or a sentence/clause (13). Other sentences may feature no explicit temporal reference, but according to Liu and Zhang (2004, 99) "从上下文中, 可以明显看出指的就是'正在此时'的意思" (the context allows the identification of the reference time as 'just now' [Author's translation]). In other words, they have an implicit *stage topic*.[6]

12. 1990年11月, 一份诉状递到了北京市西城区人民法院。
yījiǔjiǔlíng *nián* *shíyī* *yuè* (SPATIO-TEMPORAL FRAME)
1990 year 11 month
yí *fèn* *sùzhuàng* *dìdào-le* *Běijīng* *shì* *Xīchéng*
one CLF complaint submit-PFV Beijing city Xicheng
qū *Rénmín* *Fǎyuàn*
district People Court
'In November 1990, a complaint was submitted to the People's Court of Xicheng District, Beijing'.

13. 正在审问的时候, 一只大老虎跳进公堂 […]
zhèngzài *shěnwèn* *de* *shíhòu* (SPATIO-TEMPORAL FRAME)
PROG interrogate SP time
yì *zhǐ* *dà* *lǎohǔ* *tiáo-jìn* *gōng-táng*
one CLF big tiger jump-enter public-hall
'During the interrogation, a big tiger jumped into the public hall […]'

An account in terms of locatability also explains Xiong's (2008) claim that SIIs admissibility depends on the presence of a specific component that meets the topic's needs: what Xiong actually means is that some contextual element is needed that renders the topic referent locatable; such an element may be a temporal/locative phrase, even an implicit one (stage topic). It also sheds light on Liu's (2003) observation that the role of SIIs within the narration is to create a plot transition: in this case, the new topic also involves a shift of setting (for example, a new scene or a new time reference, with different spatio-temporal coordinates).

All the above studies highlight significant features of SIIs. However, they reveal little about their statistical relevance, as most corpus-based studies are qualitative and/or conducted on small-size corpora. Furthermore, little is said on another rather significant cross-linguis-

6 Given an utterance, stage topics are its implicit spatio-temporal coordinates that allow the assessment of its truth value. This captures the fact that a sentence like *it is snowing!* is true and informative only with reference to the temporal and spatial setting of its discourse. According to Erteschik-Shir, "thetic sentences are viewed as having implicit 'stage' topics indicating the spatio-temporal parameters of the sentence (here-and-now of the discourse). These are contextually defined" (2007, 16).

tic feature of the sentence-initial position, i.e. *animacy*: does this semantic trait interact at all with SIIs in Chinese?

3 The Study. What Corpora Tell on SIIs

As said earlier, this study adopts a corpus approach, with the aim to ground the analysis on empirical, natural data. Specifically, corpora contribute towards: (i) verifiability and reproducibility as monitoring mechanisms for a given analysis, as results can be checked by repeating the same query; and (ii) highlighting facts, data, or details that had not been observed before and have not yet been integrated in linguistic descriptions. Let us now turn to corpus data: a banal query with the string '。一位' (. *yí wèi*) in the PKU corpus gives 5,751 results; the first 5 occurrences are reported in table 1. The same query gives 1,466 results in the BCC corpus and 605,379 in the ZHTenTen (ST) corpus. On the other hand, the string '。一个' (. *yí ge*) occurs 13,399 times in the PKU corpus; the first 5 occurrences are shown in table 2.

Table 1 PKU corpus: first 5 occurrences of the string '。一位' (. *yí wèi*)

[…] 两位具有马克思主义传统的欧洲思想家	。一位	是意大利共产党领导人和理论家安东尼·葛兰西, 另一位是 […]
liǎng wèi jùyǒu Mǎkèsī-zhǔyì chuántǒng de Ōuzhōu sīxiǎngjiā	*. yí wèi*	*shì Yìdàlì Gòngchǎndǎng lǐngdǎorén hé lǐlùnjiā Āndōngní Gélánxī, lìng yí wèi shì*
[…] two European thinkers within the Marxist tradition	. One	is the leader and theoretician of the Italian Communist Party, Antonio Gramsci, the other is […]
当时有两位大史学家 […]	。一位	是黄梨洲, 他著了一部《明夷待访录》[…]
dāngshí yǒu liǎng wèi dà shǐxuéjiā	*. yí wèi*	*shì Huáng Lízhōu, tā zhù le yí bù Míngyí Dàifǎng Lù*
At that time, there were two great historians […]	. One	is Huang Lizhou, who wrote the *Mingyi Daifang Lu* […]
[…] 这就是哲学家康德和他的仆人拉普	。一位	传记家赞叹道: "康德的一生就像是一个最规则的动词 […]
zhè jiù shì zhéxuéjiā Kāngdé hé tā de púrén Lāpǔ	*. yí wèi*	*zhuànjìjiā zàntàn dào: "Kāngdé de yìshēng jiù xiàng shì yí ge zuì guīzé de dòngcí*
[…] these are the philosopher Kant and his manservant Lampe	. A	biographer said admiringly: "Kant's life is like a regular verb […]
这项研究已经成为社会学学术进展的一个很重要的组成部分	。一位	著名的美国社会学家就认为, 这方面的研究已经不是在与主流社会学 […]
zhè xiàng yánjiū yǐjīng chéngwéi shèhuìxué xuéshù jìnzhǎn de yí ge hěn zhòngyào de zǔchéng bùfèn	*. yí wèi*	*zhùmíng de Měiguó shèhuìxuéjiā jiù rènwéi, zhè fāngmiàn de yánjiū yǐjīng búshì zàiyǔ zhǔliú shèhuìxué*
This research has already become a milestone in the field of sociology	. A	well-known American sociologist holds that research in this area no longer lies within mainstream sociology […]

这篇文章讲的是一个动人的故事	。一位	名叫苏珊·斯蒂芬的母亲, 愿为她患肾炎的儿子捐出一个肾。
zhè piān wénzhāng jiǎng de shì yí ge dòngrén de gùshì	*. yí wèi*	*míng jiào Sūshān Sīdìfēn de mǔqīn, yuàn wéi tā huàn shènyán de érzi juānchū yí ge shèn*
This piece of writing tells a moving story	. A	mother named Susan Stephen is willing to donate a kidney to her son who suffers from nephritis.

Table 2 PKU corpus: first 5 occurrences of the string '。一个' (*. yí ge*)

[…] 社会正在进行一场新技术革命	。一个	国家生产力的发展, 国民经济的增长, 越来越依靠科学技术的进步 […]
shèhuì zhèngzài jìnxíng yì chǎng xīn jìshù gémìng	*. yí ge*	*guójiā shēngchǎnlì de fāzhǎn, guómín jīngjì de zēngzhǎng, yuè lái yuè yīkào kēxué jìshù de jìnbù*
[…] society is undergoing a new technological revolution	. A	country's productivity development and the growth of its national economy rely more and more on the progress of science and technology; […]
[…] 就是强调学校教育工作的时效性。(5)持久性	。一个	人所受的从幼儿园开始到大学的教育, 要经历17-18年的时间 […]
jiùshì qiángdiào xuéxiào jiàoyù gōngzuò de shíxiàoxìng (5) chíjiǔxìng	*. yí ge*	*rén suǒ shòu de cóng yòu'éryuán kāishǐ dào dàxué de jiàoyù, yào jīnglì 17-18 nián de shíjiān*
[…] it emphasises the timeliness of school education. (5) *Persistence.*	. A	person's education from kindergarten to university takes 17-18 years[…]
这是不少学者专家的共识	。一个	人, 作为生命个体, 从出生之日起, 就与周围环境 […]
zhè shì bù shǎo xuézhě zhuānjiā de gòngshì	*. yí ge*	*rén zuòwéi shēngmìng gètǐ, cóng chūshēng zhī rì qǐ, jiù yǔ zhōuwéi huánjìng*
[…] it is the internal driving force of individual development. This is the general consensus among several scholars and experts	. A	person, as an individual form of life, from the date of her birth, clashes with the surrounding environment […]
[…] 所谓自由、责任、义务, 都是幻想的名词	。一个	人对于社会的有用与否, 完全看遗传如何。
suǒwèi zìyóu, zérèn, yìwù, dōu shì huànxiǎng de míngcí	*. yí ge*	*rén duìyú shèhuì de yǒuyòng yǔ fǒu, wánquán kàn yíchuán rúhé*
[…] so-called freedom, responsibility, and obligation are all fantasy terms	. (Whether) a	person is useful to society depends entirely on her inheritance.
香港的幼儿教育（又称为学前教育）分为两个系统	。一个	是香港政府教育署管辖的幼稚园, 另一个系统是 […]
Xiānggǎng de yòu'ér jiàoyù (yòu chēng wéi xuéqián jiàoyù) fēn wéi liǎng gè xìtǒng	*. yí ge*	*shì Xiānggǎng zhèngfǔ jiàoyù shǔ guǎnxiá de yòuzhìyuán, lìng yí ge xìtǒng shì*
Early childhood education (also known as preschool education) in Hong Kong is divided into two systems	. One	consists of the kindergartens under the jurisdiction of the Education Department of the Hong Kong Government; the other system is […]

Such very preliminary data have little statistical relevance but open up interesting perspectives. First, SIIs do exist and are not statistically insignificant: results in all corpora are of the order of thousands; moreover, five out of five sentences in table 1 present sentence-initial NPs that receive a true indefinite reading. Second, corpora are tools that must be used *cum grano salis*: in table 2, the first four NPs are in fact generic, while only the fifth is a true indefinite. Hence, quantitative data will need to be filtered through a subsequent qualitative examination, to assess the extent to which sentence-initial NPs of the type of '一 *yī* CLF N' are true indefinites. Third, a striking difference is highlighted between a very common, generic classifier like 个 *ge* 'unit' and the highly specific classifier 位 *wèi*, i.e. the polite classifier for people: although 个 *ge* is much more frequent in absolute terms (its total occurrences as classifier in the ZHTenTen (ST) corpus is 9,265,680, as compared to 1,007,191 for 位 *wèi* – see table 3 below), the former occurs just little above twice as the latter in the '。一 *yī* CLF' pattern. This, together with the different ratio of true SIIs (100% *vs* 20%, respectively), suggests that the semantics of the classifier (e.g. the trait ±animate/±human) might also be relevant with respect to the acceptability degree/statistical relevance of SIIs. This hypothesis is supported by the cross-linguistic tendency of animate NPs to occur sentence-initially, regardless of their semantic role, syntactic function, and information status (non-agent, non-subject, and non-given animates still display this tendency). An experimental study carried out by Verhoeven on a sample of heterogeneous languages (German, Greek, Turkish, and Chinese) shows that "animate-first effects occur across languages" (2014, 129). This, according to Verhoeven, is an expected result under the view that "these effects come from asymmetries in the mental representation of the referents", which are independent from language-specific characteristics (2014, 129) – see also Van Bergen (2011) for a cross-linguistic overview of animacy and word order and Iemmolo and Arcodia (2014) for Chinese.

3.1 Research Questions and Scope

Against the background laid out so far, this study aims at answering the following research questions:

> RQ1 How significant is the phenomenon of SIIs from a quantitative/statistical perspective?
>
> RQ2: Does the trait of animacy play a role in the phenomenon?

The study focuses on indefinite NPs marked through the major indefiniteness encoding means in Chinese (Chen 2015, 409), i.e. a noun

phrase containing the string 一 *yī* 'one' + classifier (CLF),[7] that occurs sentence-initially. In fact, indefiniteness may be conveyed, more in general, by the string numeral + classifier (Li 1997, 18, among many others); however, indefinite NPs with numerals other than 一 *yī* 'one' (e.g. 三/几个学生 *sān/jǐ ge xuéshēng* 'three/some students') are excluded from the study, for two main reasons: the first is that the study itself would be more complex in terms of corpus queries; moreover, it would involve relying more on the accuracy of the tagging, which is not always high (see discussion in § 6) and is different in each corpus (e.g. the PKU is not POS-tagged), thus not allowing a comparison between the three corpora. Finally, numerals other than 'one' often emphasise the *quantity* or receive a *distributive* reading, as discussed by Li and Thompson with reference to (4a)-(4b) above, while the focus here is mainly on true indefinite readings. This implies that this study only accounts for singular indefinite NPs of the type of '一 *yī* CLF (N)' and that the number of SIIs identified in this study is smaller than those actually existing in the corpora.

Possible indefinite NPs may consist of simple patterns of the type of '一 *yī* CLF (N)', where the head noun may be overt or omitted. In some cases, the classifier may also be omitted; however, these cases are comparatively rarer and harder to detect, and thus will not be considered. This also implies that, again, the number of SIIs identified in this study is smaller than those existing in the corpora. Indefinite NPs may also include modifiers (nouns, adjectives, verbs, relative clauses etc.). These generally occur in two positions: between the classifier and the noun (14b) and to the left of the '一 *yī* CLF N' string (14c) – the former suggests a descriptive reading, the latter a restrictive one, see e.g. Chao (1968, 286-7):

14.	a.	[Numeral + CLF]		[Noun]
	b.	[Numeral + CLF]	[Modifier(s)]	[Noun]
	c. [Modifier(s)]	[Numeral + CLF]		[Noun]

Below are examples of SII types above. For pattern (14c), modifiers may include nouns/adjectives (15c), but also verbal elements occurring, for example, within a relative clause (15c'). Finally, other elements, such as time/location phrases, may occur to the left of the NP – see e.g. (12) above:

7 Indefinite NPs in Chinese may take two forms: nouns modified by a number + classifier structure and bare nouns, when postverbal (Li 1997, 18). Since the present article investigates the sentence-initial position, it focuses on the pattern '一 *yī* CLF N'.

15. a. 一位传记家赞叹道 […] (PKU)
yí *wèi* *zhuànjìjiā* *zàntàn-dào*
one CLF biographer admire-say
'A biographer said admiringly […]'

b. 一位著名的美国社会学家就认为 […] (PKU)
yí *wèi* *zhùmíng* *de* *Měiguó* *shèhuìxuéjiā*
one CLF famous SP American sociologist
jiù *rènwéi*
indeed think
'A famous American sociologist thinks that […]'

c. 加油站的一位工作人员说, 从下午三四点钟开始 […] (ZHTenTen (ST))
jiāyóuzhàn *de* *yí* *wèi* *gōngzuòrényuán* *shuō*
gas.station SP one CLF worker say
cóng *xiàwǔ* *sān-sì* *diǎnzhōng* *kāishǐ*
from PM 3-4 o'clock start
'A staff member of the gas station said that from 3-4 PM onwards […]'

c'. 刚来的一位天津大厨 […] (Wangyi News)[8]
gāng *lái* *de* *yí* *wèi* *Tiānjīn* *dàchú*
REL [just come SP] one CLF Tianjin chef
'A newly arrived chef from Tianjin […]'

3.2 Methodology and Data

Quantitative analysis. Identifying SIIs as described above involves examination of complex strings, including punctuation and sentence boundaries. Hence, for the quantitative analysis, three generalised, big-size corpora were chosen that allow such a query: the PKU corpus (470 million characters), the BCC corpus (15 billion characters), and the ZHTenTen simplified Chinese corpus mounted at Sketch Engine (Stanford Tagger subcorpus, 1,73 billion characters). Each corpus involves a different query system, and only the BCC and the ZHTenTen (Stanford Tagger, henceforth ST)[9] are POS-tagged; hence, the results are more or less fine-grained depending on the corpus. Specifically, while the BCC and the ZHTenTen (ST) corpora also allow queries through the POS tag for classifiers (*q* and *M*, respectively), in the

8 https://bit.ly/37wXhFe.

9 The ZHTenTen Stanford Tagger is POS tagged following the Part-Of-Speech Tagging Guidelines for the Penn Chinese Treebank. The corpus allows a rather detailed interrogation, lends itself to concordancing, collocation, and term extraction (Xu 2015).

PKU corpus the number of occurrences needs to be collected for each single classifier. To this end, Sketch Engine's wordlist tool was used to obtain a frequency list of the nominal classifiers listed in the 汉语量词词典 *Hanyu liangci cidian* (Chen et al. 1988): a total of 36 classifiers with more than 20 thousand occurrences as classifier in the ZHTenTen (ST) were identified. Units of measure, e.g. 元 *yuán* (RMB), 分 *fēn* (unit of length/area/money/time), 吨 *dūn* (ton), 亩 *mǔ* (unit of area), 公里 *gōnglǐ* (km) were excluded, in that they are mainly used to express specific quantities rather than indefiniteness. To tackle RQ2 (§ 3.1), particular attention was devoted to classifiers denoting animate nouns – marked as +A(nimate) – including 名 *míng*, 位 *wèi*, 只 *zhī* and 头 *tóu* (for animals), and 伙 *huǒ* (collective). Other classifiers used with people but also with inanimate nouns (±A) such as 个 ge, 行 *háng* (row), 家 *jiā* (for families and for shops), and 排 *pái* (line) were treated separately, as it is not possible to verify whether their frequency is connected with the occurrence of animate nouns. The classifier 对 *duì* 'couple', while compatible both with animates and inanimates, was marked as +A, in that a cursory examination of 150 random tokens of sentence-initial '一对 *yí duì*' NPs in all three corpora reveals that 90% of tokens introduce animate nouns. Table 3 shows the resulting list of examined classifiers, along with their frequency:

Table 3 List of classifiers

CLF	Animacy trait	Frequency as classifier in the ZHTenTen (ST) c.	CLF	Animacy trait	Frequency as classifier in the ZHTenTen (ST) c.
个	±A	9,265,680	座	–A	194,739
项	–A	1,458,480	本	–A	182,384
名	+A	1,156,327	系列	–A	174,548
条	±A	1,104,219	台	–A	174,530
位	+A	1,007,191	只	+A	164,721
级	–A	858,424	户	–A	160,875
家	±A	807,627	门	–A	114,744
批	±A	461,612	组	±A	105,680
件	–A	407,054	处	–A	104,857
份	–A	340,977	道	–A	85,349
期	–A	329,997	首	–A	81,823
所	–A	293,366	把	–A	79,768
篇	–A	278,140	对	+A	79,199
套	–A	260,345	班	±A	71,086
句	–A	234,465	间	–A	68,961
部	–A	216,625	头	+A	33,993
张	–A	214,591	排	±A	16,522
块	–A	208,768	伙	+A	6,596

For patterns (a) and (b) in (14), the string '一 *yī* CLF' is at the beginning of the sentence and can be easily detected with the appropriate syntax (i.e. (; |:|。|? |!)$一CLF in the PKU corpus; [。 ; ? !]一q/CLF in the BCC corpus; and <s> [word="一"][tag="M"] and <s> [word="一"][word="CLF"] in Sketch Engine). On the other hand, detection of pattern (c), where the modifier(s) occur(s) between the punctuation mark and the '一 *yī* CLF' string, is more complex and, in some cases, problematic. Specifically, modifiers such as relative clauses cannot be detected, as queries including verbs before the '一 *yī* CLF' string may both identify SIIs, as (15c'), but also postverbal indefinites, as in the following example:

16. 刚来了一位天津大厨

gāng	*lái-le*	*yí*	*wèi*	*Tiānjīn*	*dàchú*
just	arrive-PFV	one	CLF	Tianjin	cook

'A cook from Tianjin has just arrived'

To avoid that, the queries exclude verbal elements, but include adjectival and nominal modifiers (e.g. <s>[tag="JJ"][tag="N.*"]{1,7}[word="一"][word="CLF"&tag="M"], in the ZHTenTen (ST)). Finally, SIIs with leftmost time/location phrases separated by commas, as in (12), are hard to identify quantitatively and are not considered either. Again, this implies that the number of SIIs identified in the quantitative analysis does not include all possible patterns.

Qualitative analysis. As noted in § 2, while the string '一 *yī* CLF' is the most common formal marker for Chinese indefinite NPs, it does not always involve a true indefinite meaning, as the NP may display a quantitative (4a), distributive (4b), or generic (5) reading. The quantitative analysis as described above necessarily identifies all types, as they are formally identical. To determine the average ratio of true indefinites, as well as of NPs receiving a quantitative, distributive, or generic reading, a qualitative analysis was conducted on a random sample of sentences from the ZHTenTen (ST) corpus, collected[10] with the following query: <s>[tag="JJ|N.*"]{0,7}[word="一"][word="CLF1| CLF2|… "]. Each sample consists of 100 sentences for each subtype of classifiers (+A, ±A, –A), for a total of 300 sentences, a number that preserves the representativeness of the sample.

10 With the Sketch Engine function 'get a random sample', the same number of lines generated from a given concordance produces the same concordance lines: thus, the search can be easily repeated and reproduced.

4 Quantitative Results

The tables below show results for each corpus. In the paper, 'CLF' denotes each specific classifier, while 'CLF' indicates the word class. S.I. stands for 'sentence-initial', while *de* corresponds to the Chinese noun modifier marker 的 *de*, which may but need not be present. Orange, blue, and green mark +A, ±A, and –A classifiers, respectively (see § 3.2). Columns for pattern (c) as shown in (14) report figures of different modifiers patterns; the type and number of detectable patterns depend on the tools and CQL queries each corpus offers. The last column (ratio) shows the percentage of sentence-initial occurrences of each classifier in the pattern '一 *yī* CLF' over all occurrences of the pattern in any position in the sentence; in other words, it captures how often an indefinite noun phrase with a specific classifier occurs sentence-initially.

Table 4 ZHTenTen (ST) corpus

CLF	Any position	Patterns (a) – (b)	Pattern (c): S.I."一 *yī* CLF" occurrences with						All patterns			Ratio
	'*yī* CLF'	S.I. '*yī* CLF'	leftmost noun mod.	leftmost noun mod. +*de*	leftmost adj. mod.	leftmost adj. mod. +*de*	leftmost adj./ noun mod.	leftmost adj./noun mod.+*de*	Total detected without *de*	Total detected with *de*	Total detected	S.I. '*yī* CLF'/ '*yī* CLF'
名	207,535	6,035	878	190	58	2	52	13	8,005	619	8,624	3.48%
位	300,812	27,182	2,351	907	142	4	103	0	33,425	2,732	36,157	10.20%
只	64,460	1,887	205	22	35	3	15	1	2,228	56	2,284	3.36%
头	15,569	424	112	20	4	1	12	2	681	36	717	3.69%
伙	3,633	83	17	1	1	0	4	0	192	1	193	2.92%
对	40,725	1,065	151	26	32	2	13	1	1,427	57	1,484	3.17%
个	3,923,883	98,525	5,101	2,432	1,957	189	321	509	110,524	5,497	116,021	2.78%
条	204,214	2,575	437	99	119	35	24	6	3,397	209	3,606	1.61%
家	197,900	3,938	714	63	54	8	83	9	7,893	361	8,254	2.46%
批	253,206	2,841	342	24	96	6	48	6	3,578	90	3,668	1.33%
组	32,120	710	184	29	102	6	17	3	1,112	83	1,195	3.27%
班	11,040	150	152	0	51	0	6	0	419	3	422	3.25%
排	6,649	113	36	13	13	0	2	2	171	18	189	2.69%
项	236,816	3,059	431	313	208	17	17	46	4,401	1,272	5,673	1.73%
级	116,584	2,231	1,733	37	74	4	102	8	4,856	77	4,933	3.59%
件	115,770	1,360	142	37	98	10	7	4	1,668	61	1,729	1.43%
份	164,759	2,487	225	113	46	6	9	7	2,989	523	3,512	1.76%
期	52,922	2,259	1,053	14	274	0	159	6	5,324	47	5,371	7.11%
所	61,758	1,583	165	5	43	0	7	1	2,166	44	2,210	2.92%
篇	64,164	1,224	248	27	64	2	7	2	1,626	126	1,752	2.45%
套	105,632	956	0	18	41	7	20	7	1,221	50	1,271	0.99%
句	113,840	3,728	252	136	251	38	12	17	4,458	407	4,865	3.89%
部	73,383	2,252	195	11	37	2	18	2	2,651	60	2,711	3.43%
张	89,515	1,737	202	23	54	8	4	0	2,088	50	2,138	2.27%
块	74,084	816	141	26	38	7	6	3	1,060	58	1,118	1.40%
座	71,757	1,324	106	36	23	3	12	2	1,579	76	1,655	2.10%

本	68,435	1,536	158	10	65	4	9	1	1,890	43	1,933	2.61%
系列	163,248	1,750	50	104	61	7	3	15	1,940	250	2,190	1.22%
台	42,980	828	95	7	32	2	7	1	1,074	21	1,095	2.26%
户	13,017	135	46	1	1	0	4	1	215	7	222	1.44%
门	47,792	456	22	5	24	1	4	0	587	9	596	1.07%
处	37,630	190	199	8	12	0	9	2	518	30	548	1.12%
道	48,975	493	63	15	53	0	6	1	648	33	681	1.29%
首	36,766	1,256	122	25	42	3	15	5	1,525	139	1,664	3.99%
把	63,835	742	386	22	29	1	25	4	1,365	39	1,404	1.89%
间	21,696	385	100	24	7	2	10	2	547	50	597	2.44%

Thanks to the Corpus Query Language (CQL) option, the ZHTenTen (ST) is the corpus that allowed extraction of the most detailed data. Table 4 presents the number of occurrences for each classifier for patterns (14a)-(14b) (column 3) and some possible patterns for (14c), distinguishing different modifier types (adjective, noun, or both, and with or without 的 *de*); modifiers are up to 7 characters long. Columns 10 and 11 show the total amount of detected S.I. '一 *yī* CLF' patterns that occur without and with 的 *de*, respectively,[11] while column 12 (total detected) provides the sum of these two. The classifier with the highest total occurrences in the three patterns identified in (14) is 个 *ge* (116,021), followed by 位 *wèi* (36,157 – about one third). However, an inverse tendency is observable in the last column, which again captures how often an indefinite noun phrase with a specific classifier occurs sentence-initially: the classifier where this ratio is by far the highest is 位 *wèi* (more than 10%); other +A classifiers are all around 3%, followed by 个 *ge* that drops to 2.78%.

Table 5 BCC corpus

CLF	Any position	Patterns (a) – (b)	Pattern (c): S.I. '*yī* CLF' occurrences with						All patterns	Ratio
	'*yī* CLF'	S.I. '*yī* CLF'	leftmost noun mod.	leftmost noun mod. *+de*	leftmost adj. mod.	leftmost adj. mod. *+de*	leftmost adj./ noun mod.	leftmost adj./noun mod.*+de*	Total detected	S.I. '*yī* CLF'/ '*yī* CLF'
名	5,252	236	4	3	0	0	0	0	243	4.49%
位	29,484	1,673	26	12	4	2	0	0	1,717	5.67%
只	34,460	1,209	34	21	1	3	4	0	1,272	3.51%
头	8,676	161	20	26	2	1	2	0	212	1.86%
伙	1,540	83	2	0	0	0	0	0	85	5.39%
对	6,019	223	12	3	4	2	0	0	244	3.70%
个	351,862	14,327	275	53	37	27	13	2	14,734	4.07%

11 Used queries include: <s>[tag="JJ|N.*"]{0,7}[word="一"][word="CLF"] and <s>[tag="JJ|N.*"]{0,7}[word="的"] [word="一"][word="CLF"], respectively.

条	30,059	673	18	6	1	3	1	0	702	2.24%
家	14,639	329	44	4	1	1	3	0	382	2.25%
批	2,602	26	11	0	0	1	0	0	38	1.00%
组	690	15	2	0	0	0	0	0	17	2.17%
班	690	150	3	0	1	0	0	0	154	2.17%
排	3,131	91	5	2	0	0	0	0	98	2.91%
项	2,323	16	1	0	0	0	0	0	17	0.69%
级	782	6	2	0	1	0	0	0	9	0.77%
件	25,072	267	18	1	0	1	0	0	287	1.06%
份	6,353	42	2	0	0	0	0	0	44	0.66%
期	277	0	1	0	1	0	0	0	2	0.00%
所	2,613	16	2	0	0	0	0	0	18	0.61%
篇	3,916	41	0	1	0	0	0	0	42	1.05%
套	5,195	33	1	1	2	1	0	0	38	0.64%
句	24,806	376	25	1	4	2	0	0	408	1.52%
部	9,793	206	7	18	0	0	0	0	231	2.10%
张	23,339	519	16	5	6	0	1	0	547	2.22%
块	23,430	268	13	1	4	3	0	0	289	1.14%
座	10,229	222	2	1	4	1	1	0	231	2.17%
本	9,240	108	5	0	0	2	1	0	116	1.17%
系列	874	15	0	0	0	0	0	0	15	1.72%
台	988	31	1	0	0	0	0	0	32	3.14%
户	405	5	0	0	0	0	0	0	5	1.23%
门	1,195	9	2	0	0	0	0	0	11	0.75%
处	4,522	30	0	0	1	0	0	0	31	0.66%
道	10,229	275	10	5	0	0	0	0	290	2.69%
首	2,855	37	0	0	0	0	0	0	37	1.30%
把	13,777	106	1	0	0	0	1	0	108	0.77%
间	6,605	90	2	2	1	3	1	0	597	1.36%

In the BCC corpus **[tab. 5]**, it is more difficult to elaborate the query to include longer leftmost nominal or adjectival modifiers. Hence, detected modifiers are up to 2 characters long;[12] furthermore, composite queries to detect multiple patters (as in columns 9-10 of table 4) are not possible. This implies that the number of undetected tokens is higher than that in the ZHTenTen (ST) corpus. This is reflected in the figures, that are sensibly lower. The classifier with the highest ratio in the last column is still 位 *wèi*, although the ratio is lower (5.67%), about half the ratio in the ZHTenTen (ST) corpus.

12 Queries are of the type of [。; ? !](a/n/a n) (的)一CLF.

Table 6 PKU corpus

CLF	Any position	Patterns (a) – (b)	Pattern (c)	All patterns	Ratio	CLF	Any position	Patterns (a) – (b)	Pattern (c)	All patterns	Ratio
	'*yī* CLF'	S.I. '*yī* CLF'	Leftmost 2-character mod. +*de*	Total detected	S.I. '*yī* CLF'/ 'yī CLF'*yī* CLF'		'*yī* CLF'	S.I. '*yī* CLF'	Leftmost 2-character mod. +*de*	Total detected	S.I. '*yī* CLF / '*yī* CLF'
名	46,340	1,202	45	1,247	2.59%	所	8,724	194	7	201	2.22%
位	90,775	6,062	350	6,412	6.68%	篇	13,131	140	26	166	1.07%
只	26,904	597	51	648	2.22%	套	18,512	116	20	136	0.63%
头	11,513	154	25	179	1.34%	句	32,656	497	46	543	1.52%
伙	3,187	40	2	42	1.26%	部	54,013	781	53	834	1.45%
对	11,513	300	31	331	2.61%	张	27,310	401	27	428	1.47%
个	674,846	15,941	670	16,611	2.36%	块	27,310	286	23	309	1.05%
条	69,434	711	68	779	1.02%	座	26,148	272	13	285	1.04%
家	53,862	818	65	883	1.52%	本	18,022	337	20	357	1.87%
批	47,638	757	25	782	1.59%	系列	34,350	115	26	141	0.33%
组	6,981	62	3	65	0.89%	台	9,075	133	3	136	1.47%
班	3,709	29	3	32	0.78%	户	2,495	26	2	28	1.04%
排	4,069	92	3	95	2.26%	门	4,914	33	2	35	0.67%
项	53,679	354	68	422	0.66%	处	9,243	60	6	66	0.65%
级	11,528	79	6	85	0.69%	道	20,764	170	6	176	0.82%
件	34,619	283	11	294	0.82%	首	7,675	100	15	115	1.30%
份	27,932	194	25	219	0.69%	把	18,846	127	6	133	0.67%
期	9,047	194	4	198	2.14%	间	7938	94	11	105	1.18%

Since the PKU corpus is not tagged, complex queries involving nominal or adjectival modifiers highlighted in the previous corpora (pattern in (14c) are not possible **[tab. 4]**; however, the query (。| ? | ; | !) $2的一CLF was used to single out one/two-character modifiers (columns 4, 9). Such a query singles out, for example, modifiers such as the one in (17).

17. 我的一个好朋友他是浙江人 (PKU)
wǒ de yí ge hǎo péngyou tā shì Zhèjiāng-rén
1SG SP one CLF good friend 3SG be Zhejiang-man
'A good friend of mine (, he) comes from Zhejiang'.

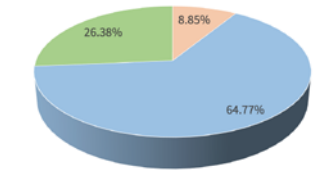
Such a limited interval minimises statistical possibilities of including verbal items and, hence, postverbal indefinites (see discussion in § 3.2). However, this involves that SIIs with longer modifiers – as in (15c') – are missing from the total count, hence the remarkably lower figures in table 4.

Discussion. Overall, results show that all examined classifiers occur with 一 *yī* in the sentence-initial position. Figures for pattern (14c) are higher in the ZHTenTen (ST) corpus, but this does not come as

a surprise, as leftmost modifiers detected in the ZhTenTen (ST) are up to 7 characters, while in the other two corpora they are up to two characters (see § 3.2). Let us focus on the two classifier 位 *wèi* and 个 *ge*: the former's total occurrences in the (14a-b-c) patterns are 36,157 in the ZHTenTen (ST), 1,717 in the BCC, and 6,412 in the PKU; the latter's are 116,021 in the ZHTenTen (ST), 14,734 in the BCC, and 16,611 in the PKU. Crucially, ratio-wise 位 *wèi* significantly outranks 个 *ge* (10.2% over 2.78% in the ZHTenTen (ST)): in other words, while the string '一位' *yí wèi* overall occurs far less than '一个' *yí ge*, in the sentence-initial position the former occurs much more frequently than the latter. Other classifiers with a relatively high ratio (last column), especially in the ZHTenTen (ST) corpus, include +A classifiers in general and ±A classifiers like 组 *zǔ* 'group' and 班 *bān* 'class' (highly compatible with +A nouns) – almost all show a ratio above 3% in the ZHTenTen (ST). Relatively high ratios are also displayed by some –A classifiers, such as 级 *jí* 'level' (3.59%), 期 *qī* 'period' (7.11%), 部 *bù* 'part' (3.43%), 句 *jù* 'line' (3.89%), and 首 *shǒu* 'piece (e.g. of poetry/lyric', 3.99%). Indefinite noun phrases with the first three classifiers (级 *jí* 'level', 期 *qī* 'period', 部 *bù* 'part') display an interesting common semantic trait related to partitivity: the referent may denote a part of a given whole, a level of a given multi-layered structure, a step of a given path, or else a phase of a given plan or project (see examples in sections below). The relatively high frequency of such NPs in the sentence-initial position might then be connected to the fact that the referent, although not identifiable, is at least *locatable* in a given set/whole/container that is comprehensible thanks to the semantics of each classifier (e.g. one level of a specific hierarchy, one step of a specific procedure etc.); it may also be specified in the previous context or, otherwise, be implicit (stage topics,[13] see discussion for sentence (4c)). This point will be examined in the qualitative analysis below. Conversely, 句 *jù* and 首 *shǒu* (classifiers for lines/quotes, and for songs/poems, respectively) come rather unexpected. We will look further into these classifiers through the qualitative analysis.

Let us now have a closer look at aggregated data with respect to the animacy trait (+A, ±A, and –A) in the ZHTenTen (ST) corpus **[tab. 7]**.

13 This is, in turn, related to the frame-containment property of topics (Chafe 1976; Her 1991; Morbiato 2020): topics express a frame of validity for the rest of the predication and are often a semantic container/whole/setting for what comes next.



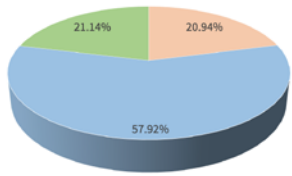


Chart 1 All '一 *yī* CLF' occurrences per animacy trait

Chart 2 Sentence-initial '一 *yī* CLF' occurrences per animacy trait

Table 7 Distribution of '一 *yī* CLF' patterns in the ZHTenTen (ST) corpus

	Sentence-initial position				**Any position**	**Ratio**
	(a) + (b)	(c) without *de*	(c) with *de*	all patterns	all patterns	
+A	35,666	10,292	3,501	49,459	632,734	7.82%
±A	108,852	18,242	6,261	133,355	4,629,012	2.88%
–A	32,787	13,609	3,472	49,868	1,885,358	2.65%
Total				**232,682**	**7,147,104**	**3.26%**

A total of 232,682 sentence-initial NPs introduced by '*yī* CLF' were detected in the corpus. As discussed, such a total includes neither NPs modified by relative clauses nor NPs preceded by modifiers longer than 7 characters and separated by commas (e.g. temporal/locative frame topics). Interestingly, almost 8% of animate NPs introduced by '一 *yī* CLF' are sentence-initial, while the ratio drops to 2.88% for ±A classifiers, and to 2.65% for –A classifiers. Charts below represent the percentage of '一 *yī* CLF' tokens over the total amount of tokens in all positions **[chart 1]** and in the sentence-initial position **[chart 2]**, divided per animacy trait: as can be seen, the percentage of +A tokens is significantly higher (more than double) in the sentence-initial position (8.8% *vs* 20.9%).

5 Qualitative Results

As discussed in § 3.2, a random sample of 300 '一 *yī* CLF' tokens was extracted from the ZHTenTen (ST) corpus, 100 for each type of classifiers: solely +A, (名 *míng*, 位 *wèi*, 只 *zhǐ*, 头 *tóu*, 伙 *huǒ*), ±A (个 *ge*, 条 *tiáo*, 家 *jiā*, 批 *pī*, 组 *zǔ*, 排 *pái*, 班 *bān*), and –A (项 *xiàng*, 级 *jí*, 件 *jiàn*, 份 *fèn*, 期 *qī*, 所 *suǒ*, 篇 *piān*, 套 *tào*, 句 *jù*, 部 *bù*, 张 *zhāng*, 块 *kuài*, 座 *zuò*, 本 *běn*, 系列 *xìliè*, 台 *tái*, 户 *hù*, 门 *mén*, 处 *chù*, 道 *dào*, 首 *shǒu*, 把

bǎ, 间 *jiān*). The referential properties of each NP introduced by '一 *yī* CLF' were analysed in all three subcorpora; results are in table 8.

Table 8 Referential properties of '一 *yī* CLF' tokens for each subcorpus of the ZHTenTen

	+A	±A	–A
SIIs	94	34	28
Generic	3	43	27
Referential	2	9	4
Referential SIIs	0	0	11
Numeral	0	9	25
Distributive	0	0	1
Wrong (postverbal)	1	5	4
Total	**100**	**100**	**100**

Let us first focus on SIIs: strikingly, 94% of +A tokens display an indefinite reading and hence are true SIIs. In other categories, conversely, the percentage of true SIIs drops to 34% for ±A and 28% for –A tokens. If we assume that the above figures are statistically relevant (although this would benefit from more tests conducted on different samples), we could consider these three percentages as coefficients that enable determining the true amount of SIIs from quantitative data presented in § 4. For data from the ZHTenTen (ST) corpus, results would be as follows:

Table 9 Percentage of true SIIs per +A, ±A, and–A animacy traits, ZHTenTen (ST)

	Total detected '*yī* CLF'	**Percentage of '*yī* CLF'**	**Samples' SII coefficient**	**Number of true SIIs**	**Percentage of true SIIs**
+A	49,459	21%	94%	46,491	44%
±A	133,355	58%	34%	45,341	43%
–A	49,868	21%	28%	13,963	13%

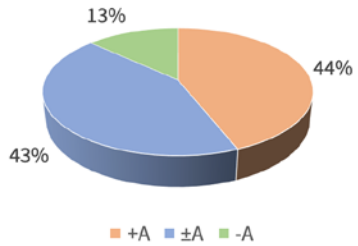


Chart 3 Percentage of true SIIs per animacy traits in the ZHTenTen (ST)

Figures in table 9 also show that animate SIIs in fact constitute a much higher percentage in the corpus, i.e. about 44% (see chart 3).

Let us now look more closely at the ±A subcorpus. First, the 100 tokens were analysed and differentiated according to the animacy trait of their head noun: 35 tokens consisted of +A NPs, 60 were –A NPs, while 5 were invalid tokens. Then, SIIs were identified in each group; figures are in table 10.

Table 10 Animate *vs* inanimate SIIs in the ±A subcorpus

±A	+A	–A
SII	12	22
Other	23	38
Total	35	60

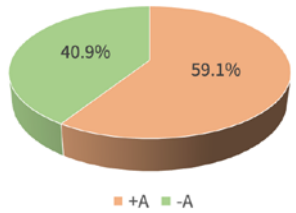


Chart 4 Percentage of true SIIs per +A and –A animacy traits, ZHTenTen (ST)

Interestingly, a reverse tendency can be observed with respect to +A tokens within the ±A subcorpus: only 12 (34%) are true SIIs (as compared to 94% in the +A subcorpus). Moreover, getting back to the comparison between 个 *ge* and 位 *wèi*, in the qualitative analysis, +animate (and +human) tokens introduced by 位 *wèi* tend to be referential/specific SIIs; conversely, for those introduced by 个 *ge*, generic NPs are twice as much as specific SIIs. This is very likely connected to their semantics: 位 *wèi* implies respect or courtesy and likely involves that the speaker knows the referent (specific indefi-

nite); 个 *ge*, on the other hand, means 'unit' and is more suitable to talk about a generic class, e.g. the NP 一个四川人 *yí ge Sìchuān-rén* 'A Sichuanese' in (18) from the ±A subcorpus:

18. 一个四川人可能很真诚的为"扬州十日"而 垂泪 […]

yí	***ge***	***Sìchuān-rén***	*kěnéng*	*hěn*	*zhēnchéng*	*de*
one	CLF	Sichuan-person	maybe	very	sincerely	SP
wèi	*Yángzhōu*	*Shí*	*Rì*	*ér*	*chuí-lèi*	
for	Yangzhou	10	days	SP	shed-tear	

'A Sichuanese may sincerely shed tears for the "Ten Days of Yangzhou" […]'

If we further split ±A SIIs into A+ and –A and add this data to percentages indicated in table 11, we obtain the following figures:

Table 11 Percentage of true SIIs per +A and –A animacy traits, ZHTenTen (ST)

	Number of true SIIs	Percentage of true SIIs
+A	62,494	59%
–A	43,301	41%

Such a projection suggests that, in the ZHTenTen (ST) corpus, a total of 105,795 SIIs can be detected. If compared to the total amount of '一 *yī* CLF' occurrences in the corpus, SIIs are 1.48%. Moreover, it suggests that, roughly, 6 SIIs out of 10 are animate. This proves that animacy is indeed a very significant trait for sentence-initial indefinite NPs. Again, this is in line with other cross-linguistic studies on the sentence-initial position and animacy.

Some examples. Let us now look at some of the most relevant examples of SIIs. As said, most are +animate (in fact, +human) and specific (known to the speaker but not to the hearer). A significant amount of examples involving +human SIIs introduce reported speech, either indirect (19) or direct (20). Verbs occurring in these sentences include: 提出 *tíchū* 'mention', 说 *shuō* 'say', 说明 *shuōmíng* 'explain', 坦言 *tǎnyán* 'say frankly', 告诉 *gàosù* 'tell', 表示 *biǎoshì* 'express'. Crucially, these verbs imply that the utterance is contextually situated in specific spatio-temporal coordinates, i.e. where and when the sentence is uttered (hence, it is locatable):

19. 一位人类学家曾经提出, 正常男女生交往的空间距离是 […]

yí	*wèi*	*rénlèixuéjiā*	*céngjīng*	*tíchū*	*zhèngcháng*
one	CLF	anthropologist	once	suggest	normal
nánnǚshēng	*jiāowǎng*	*de*	*kōngjiān*	*jùlí*	*shì*
male.female	interact	SP	spatial	distance	be

'An anthropologist once suggested that the normal spatial distance between boys and girls is […]'

20. 一名姓程的出租车司机说: "上下班时间是最多人打车的 […]
yì *míng* *xìng* *Chéng* *de* *chūzūchē* *sījī* *shuō*
one CLF surname Cheng SP taxi driver say
shàngxiàbān *shíjiān* *shì* *zuìduō* *rén* *dǎchē* *de*
commute time be most people take.taxi SP
'A taxi driver surnamed Cheng said: "Most people take taxis during commuting hours […]"'.

Reported speech SIIs are also found with inanimates, although such cases are much rarer:

21. 一项令人振奋的新研究表明 […]
yí *xiàng* *lìng* *rén* *zhènfèn* *de* *xīn*
one CLF cause people excite SP new
yánjiū *biǎomíng*
research show
'An exciting new study shows that […]'

Some +A SIIs are not specific; however, the context makes them at least *locatable* (see discussion in § 2). This is the case of (22): the referent of 一位父亲 *yí wèi fùqīn* 'a father' is not identifiable, but rather locatable within the temporal and spatial settings previously specified in the article, namely a dancing event at the Huazhong Agricultural University (cf. context). Similarly, in (23) the context makes it clear that the referent of 一位坐在最后一排的演 *yí wèi zuò zài zuìhòu yì pái de yǎnyuán* 'an actor sitting in the last row' cannot be identified, but rather located, within the given venue/group of 160 meeting participants:

22. [Context: article on a dancing event at the Huazhong Agricultural University; the previous two sentences contain no mentions of any event participant]
一位父亲领着自己刚及膝盖的女儿在场内跳着华尔兹 […]
yí *wèi* *fùqīn* *lǐng-zhe* *zìjǐ* *gāng* *jí* *xīgài* *de*
one CLF father lead-DUR REFL just reach knee SP
nǚ'ér *zài* *chǎng-nèi* *tiào-zhe* *huá'ěrzī*
daughter at field-in jump-DUR waltz
'A father with his daughter, who barely reaches his knees, dances waltz on the dancefloor […]'

23. [Context: meeting between a party committee and 160 employees in a huge venue]
一位坐在最后一排的演员站起来, 向市委宣传部副部长王立光提问 […]
yí *wèi* *zuò* *zài* *zuìhòu* *yì* *pái* *de* *yǎnyuán*
one CLF sit (be).at last one row SP actor
zhàn-qǐlái *xiàng* *Shìwěi*
stand-up towards Municipal.Party.Committee
Xuānchuán-bù *fùbùzhǎng* *Wáng* *Lìguāng* *tíwèn*
Propaganda-dept. vice.minister Wang Liguang ask
'An actor sitting in the last row stood up and asked Wang Liguang, Deputy Minister of the Municipal Party Committee Propaganda Department […]'

Other 'locatable' SIIs bear a partitive or whole-part relationship with previous sentences, as in (24). A partitive relationship is particularly frequent in occurrences of inanimate classifiers with an inherent partitive meaning (as hypothesised in § 4), e.g. 级 *jí* 'level' and 期 *qī* 'period, phase'.[14] In most cases, these receive a definite/numeral reading, e.g. 'the first phase' in (25).

24. [Context: story. The previous two sentences describe the protagonist looking at his own feet, and moving one to the wall's corner "一只移向墙角。" *yì zhǐ yí xiàng qiángjiǎo*]
一只移向门外 […]
yì *zhǐ* *yí* *xiàng* *mén-wài*
one CLF move towards door-out
'(I move) the other outside the door […]'

25. [Context: Text presenting an energy production plant]
一期装置拟建年产180万吨甲醇、68万吨烯烃。
yì *qī* *zhuāngzhì* *nǐ* *jiàn* *nián* *chǎn* *yībǎibāshí*
one CLF plant plan build year output 180
wàn *dūn* *jiǎchún* *liùshíbā* *wàn*
ten.thousand ton methanol 68 ten.thousand
dūn *xītīng*
ton olefin
'In the first phase, the plant is planned to produce 1.8 million tons of methanol and 680,000 tons of olefins per year'.

14 Qualitative data also reveal that the high frequency of patterns like '一级' *yì jí* is also connected to frequency in tables (tabs are also counted as sentence boundaries (<s>) in the ZHTenTen (ST) and are hard to rule out from the search).

A very interesting subtype found in –A tokens are referential SIIs, which come in three types: the first type (26) features a modifier that renders the referent uniquely identifiable, such as 最后 *zuìhòu* 'the last' or 最初 *zuìchū* 'the first'. The second type (27), also common in other languages (including English), is a sort of cross-clausal apposition linked to a referent mentioned in the previous context:

26. 最后一篇则包括了七个冥想练习 […]
zuìhòu *yì* *piān* *zé* *bāokuò-le* *qī* *ge*
last one CLF conversely include-PFV seven CLF
míngxiǎng *liànxí*
meditation exercise
'The last, on the other hand, includes seven meditation exercises […]'

27. [Context: the protagonist has just recalled a sentence pronounced by her grandmother]
一句看似无心的话, 却准确的[地]预测了我的未来
yí *jù* *kànsì* *wúxīn* *de* *huà* *què*
one CLF look.as unintentional SP word but
zhǔnquè *de* *yùcè-le* *wǒ* *de* *wèilái*
correctly SP predict-PFV 1SG SP future
'A seemingly unintentional sentence had in fact accurately predicted my future'.

The third type (28)-(29) interestingly features a proper name rather than a common name introduced by '一 *yī* CLF'. Classifiers occurring in this (not rare) pattern include 句 *jù* and 首 *shǒu*, thus explaining these classifiers' high sentence-initial ratios observed in table 4. This pattern had not been identified in our preliminary discussion, which confirms that corpora may help singling out new phenomena or patterns in a given language:

28. 一首《春天的故事》记录了1979年的那段往事 […]
yì *shǒu* *Chūntiān* *de* *Gùshì* *jìlù-le*
one CLF spring SP story record-PFV
yījiǔqījiǔ *nián* *de* *nà* *duàn* *wǎng-shì*
1979 year SP that CLF past-event
'A (the) (song) "The Story of Spring" recorded the events that happened in 1979 […]'

29. 一本《明朝那些事儿》可能就会让很多从来不看历史的人, 从此变成历史书的读者。
yì *běn* *Míng Cháo* *nà* *xiē* *shìr* *kěnéng*
one CLF Ming Dynasty that CLF(some) thing maybe
jiù *huì* *ràng* *hěn-duō* *cónglái* *bú* *kàn* *lìshǐ*
then will make very-many ever NEG read history

de rén cóngcǐ biànchéng lìshǐ shū de dúzhě
SP people from.now.on become history book SP reader
'A (the) book "Those Things Happened in the Ming Dynasty" may make many people who never read about history become readers of history books'.

We had found an example of such a pattern in table 1 above, reported in (30) below. In this case, the pattern occurs postverbally, but still features a proper noun (here, a title) introduced by the indefinite marker '一 *yī* CLF'.

30. 当时有两位大史学家 […]。一位是黄梨洲, 他著了一部《明夷待访录》 […]
dāngshí yǒu liǎng wèi dà shǐxuéjiā
that.time there.be two CLF great historian
yí wèi *shì Huáng Lízhōu tā zhù-le*
one CLF be Huang Lizhou 3SG.M write-PFV
yí bù Míngyí Dàifǎng Lù
one CLF Mingyi Daifang Lu
'At that time, there were two great historians […]. One is Huang Lizhou, who wrote a (the) *Mingyi Daifang Lu* […]'

If we look at this pattern from the perspective of its meaning, it seems to introduce unique referents, that are generally referred to with a proper name (such as book titles or pieces of poetry): in particular, while the speaker knows about that referent, (s)he might be not sure whether the interlocutor has some knowledge of it. Nonetheless, this would benefit from further research.

Generic readings are present in the +A subcorpus, as in (18), but are very rare (3%), while they are much more frequent with inanimates (43%), e.g. (31). Numeral (32) and distributive readings were found only in inanimate NPs:

31. 一篇短短的千字文, 往往凝结了作者十年的心血
yì piān duǎn-duǎn de qiān-zì wén
one CLF short-short SP thousand-character text
wǎngwǎng níngjié-le zuòzhě shí nián de xīnxuè
often condense-PFV author ten year SP blood
'A short thousand-word essay often condenses the author's ten years of hard work'.

32. 一套设备, 多种功能, 一本万利。
yí tào shèbèi duō zhǒng gōngnéng yì běn wànlì
one CLF device many CLF function one CLF profit
'One device, multiple functions, great profits'.

6 Conclusions and Limitations

The present study was designed to determine the statistical significance of SIIs in Chinese as well as the interconnections with features such as animacy and locatability. The quantitative and qualitative analyses discussed so far support our initial hypotheses.

Specifically, with reference to our initial research questions, this study shows that: (RQ 1) first, SIIs do exist in Chinese; statistically, their number is not unimportant. Statistical data and the analysis laid out so far suggest that, in the ZHTenTen (ST) corpus, a total of more than 100 thousands of true SIIs (i.e. sentence-initial '一 *yī* CLF' forms with a true indefinite reading) can be detected. If compared to the total amount of '一 *yī* CLF' occurrences in the ZHTenTen (ST) corpus, SIIs are 1.48%. Crucially, this analysis was not able to detect all SIIs (e.g. those introduced by numbers other than 一 *yī*, those with longer modifiers, or those modified by restrictive relative clauses as in (15c)): hence, the true amount of SIIs in the corpus is very likely to be higher. This has important implications: a theoretically sound account of the Chinese language and its word order should consider and discuss the existence and characteristics of this pattern. Similarly, SIIs should be introduced in Chinese grammars and teaching materials as well, explaining their peculiarities, tendencies, and restrictions. Of course, specific (cross-sectional or longitudinal) studies should be conducted to determine at what stage/proficiency level SIIs should be taught.

(RQ2) Animacy is indeed a factor that has significant impact on SIIs: the study shows that almost 8% of animate NPs introduced by '一 *yī* CLF' are sentence-initial, percentage that drops to 2.6 for non-animate NPs. Furthermore, roughly, 6 SIIs out of 10 are animate. Again, this is in line with other cross-linguistic studies on animacy and the sentence-initial position. Animacy was found to be a relevant factor in determining the order of event participants cross-linguistically. Studies conducted on different languages, including Spanish, Italian, Greek, Japanese, German, Dutch, Odawa (North America), and Yucatec, reveal that animate referents tend to occur before inanimate ones, regardless of their role in the event (see Van Bergen 2011 for an overview). When animate participants play the role of patients, speakers tend to produce passive sentences or to place the animate patient at the beginning of the sentence as a topic.

Finally, the above results confirm that corpora indeed contribute towards a better understanding of languages, even on topics with an established scholarship such as Chinese word order and referentiality, and allow finding new previously unobserved or underdescribed patterns in the language: the study has revealed a new reading for seemingly indefinite patterns of the type of '一 *yī* CLF N', i.e. those featuring a proper noun, as in (28) and (29).

On the other hand, the study has also highlighted some limitations of corpus tools. First, in this case a qualitative, sentence-by-sentence check was essential to refine, interpret, and validate quantitative results. Second, corpus design and POS tagging do not have a 100% reliability. For example the query "[。; ? !]n一对" in the BCC, corpus which should reveal only nominal modifiers, also identified the following (postverbal) token:

33. 若不是一对夫妇 […]
ruò bú shì ***yí duì fū-fù***
if NEG be one CLF husband-wife
'If they weren't a married couple […]'

All in all, the study clearly shows that SIIs are not only possible, but also do not constitute isolated exceptions, and that animacy and locatability indeed play a crucial role in increasing the acceptability of SIIs.

Bibliography

Bisang, W. (2016). "Chinese Syntax". Chan, S.-W.; Minett, J.; Li, W.Y.F. (eds), *The Routledge Encyclopedia of the Chinese Language*. Routledge Handbooks Online, 354-77. https://10.4324/9781315675541.ch20.

Chafe, W.L. (1976). "Givenness, Contrastiveness, Definiteness, Subjects, Topics, and Points of View". Li, C. (ed.), *Subject and Topic*. New York: Academic Press, 25-55.

Chao Y. (1968). *A Grammar of Spoken Chinese*. Berkeley: University of California Press.

Chen B. 陈保存 et al. (eds) (1988). *Hanyu liangci cidian* 汉语量词词典 (Chinese Classifiers Dictionary). Fuzhou: Fujian renmin chubanshe.

Chen P. (2015). "Referentiality and Definiteness in Chinese". Wang W.S.Y.; Sun C. (eds), *The Oxford Handbook of Chinese Linguistics*. New York: Oxford University Press, 404-13.

Chu C. 屈承熹 (2006). *Hanyu pianzhang yufa: Lilun yu fangfa* 汉语篇章语法: 理论与方法 (Mandarin Chinese Discourse Grammar. Theory and Practice). *Russian Language and Literature Studies*, 3(13), 1-15.

Erteschik-Shir, N. (2007). *Information structure. The Syntax-Discourse Interface*. Oxford: Oxford University Press.

Fan J. 范继淹. (1985). "Wuding NP zhuyu ju" 无定NP主语句 (Indefinite Subjects Sentences). *Zhongguo yuwen*, 5, 321-8.

Fang M. 方梅. (2019). "Cong huayu gongneng kan suowei 'wuding NP zhuyu ju'" 从话语功能看所谓"无定NP主语句" (So-Called "Indefinite-Subject Sentences" from a Discourse Perspective). *Shijie Hanyu jiaoxue*, 33(2), 189-200.

Fu Y. 付义琴 (2013). "Lun Hanyu 'wuding zhuyu ju' de jushiyi" 论汉语"无定主语句"的句式义 (A Syntactic Analysis of the Chinese Sentence with an Indefinite Subject). *Yunnan shifan daxue xuebao*, 11(5), 41-6. https://doi.org/10.16802/j.cnki.ynsddw.2013.05.008.

Her, O.-S. (1991). "Topic as a Grammatical Function in Chinese". *Lingua*, 84(1), 1-23. https://doi.org/10.1016/0024-3841(91)90011-S.

Ho, Y. (1993). *Aspects of Discourse Structure in Mandarin Chinese*. Lewiston; Queenston; Lampeter: Mellen University Press.

Hole, D.P. (2012). "The Information Structure of Chinese". Krifka, M.; Musan, R. (eds), *The Expression of Information Structure*. Berlin; Boston: De Gruyter Mouton, 45-70.

Huang S. 黄师哲 (2004). "Wuding mingci zhuyu tong shijian lunyuan de guanxi" 无定名词主语同事件论元的关系 (The Relationship Between Indefinite Subjects and Event Argument). Huang Z. 黄正德 (ed.), *Zhongguo yuyan xuelun cong* 中国语言学论丛 (Essays on Chinese Linguistics). Beijing: Shangwu yinshuguan, 93-100.

Iemmolo, G.; Arcodia, G.F. (2014). "Differential Object Marking and Identifiability of the Referent. A Study of Mandarin Chinese". *Linguistics*, 52(2), 315-34. https://doi.org/10.1515/ling-2013-0064.

Li, C.N.; Thompson, S.A. (1976). "Subject and Topic. A New Typology of Language". Li, C.N. (ed.), *Subject and Topic*. New York: Academic Press, 457-89.

Li, C.N.; Thompson, S.A. (1981). *Mandarin Chinese. A Functional Reference Grammar*. Berkeley; Los Angeles: University of California Press.

Li, W. (2005). *Topic Chains in Chinese. A Discourse Analysis and Application in Language Teaching*. München: Lincom Europa.

Li, Y.A. (1997). *Structures and Interpretations of Nominal Expressions*. Los Angeles: University of Southern California.

Liu A. 刘安春 (2003). "'Yi ge' de yongfa yanjiu" "一个"的用法研究 (Research on the Usage of 'yi ge'). *Zhongguo shehui kexue yanjiushengyuan boshi xuewei lunwen*. Dissertations published by the Graduate School, Chinese Academy of Social Sciences, Beijing.

Liu A. 刘安春; Zhang B. 张伯江 (2004). "Pianzhang zhong de wuding mingci zhuyu ju ji xiangguan jushi" 篇章中的无定名词主语句及相关句式 (The Discourse Function of the Sentence with an Indefinite NP Subject). *Journal of Chinese Language and Computing*, 14(2), 97-105.

Liu X. 刘晓亚 (2018). "Wuding NP zhuyu ju de shiyong tiaojian" 无定NP主语句的使用条件 (Conditions for the Use of Sentences with Indefinite Subject NPs). *Qingnian wenxuejia*, 20.

Lu, B.; Zhang, G.; Bisang, W. (2015). "Valency Classes in Mandarin". Malchukov, A.; Comrie, B. (eds), *Valency Classes in the World's Languages*. Berlin: Mouton De Gruyter, 709-64.

Lu S. 陆烁; Pan H. 潘海华 (2009). "Hanyu wuding zhuyu de yuyi yunzhun fenxi" 汉语无定主语的语义允准分析 (The Semantic Licensing Conditions of Indefinite Subjects in Mandarin Chinese). *Zhongguo yuwen*, 6, 528-37.

Morbiato, A. (2018). *Word Order and Sentence Structure in Mandarin Chinese. New Perspectives* [PhD dissertation]. Venice; Sydney: Ca' Foscari University of Venice; The University of Sydney.

Morbiato, A. (2020). "Cognitive and Functional Principles Shaping Chinese Linear Order. The Containment Schema". *Cognitive Linguistic Studies*, 7(2), 307-33.

Paul, W. (2015). *New Perspectives on Chinese Syntax*. Berlin, Boston: De Gruyter.

Shyu, S. (2016). "Information Structure". Huang, C.; Shi, D. (eds), *A Reference Grammar of Chinese*. Cambridge: Cambridge University Press, 518-76.

Tang C. 唐翠菊 (2005). "Cong jiwuxing yongdu kan Hanyu wuding zhuyu ju" 从及物性角度看汉语无定主语句 (Transitivity and Sentences with an Indefinite NP as Subject). *Yuyan jiaoxue yu yanjiu*, 3, 9-16.

Tang H. 唐或 (2011). "'Shu (liang) ming' wuding zhuyu ju" "数(量)名"无定主语句的使用特点分析 (Analysis of the Characteristics of the Use of "Num-

ber (Classifier) Noun" Indefinite Subject Sentences). *Xinan daxue xuebao*, 37(S1), 204-6.

Tsao, F.-F. (1977). *A Functional Study of Topic in Chinese. The First Step toward Discourse Analysis*. Los Angeles: University of Southern California.

Tsao, F.-F. (1989). "Comparison in Chinese. A Topic Comment Approach". *The Tsing Hua Journal of Chinese Studies*, New Series, 19(1), 151-89.

Van Bergen, G. (2011). *Who's First and What's Next. Animacy and Word Order Variation in Dutch Language Production* [PhD dissertation]. Nijmegen: Radboud University.

Verhoeven, E. (2014). "Thematic Prominence and Animacy Asymmetries. Evidence from a Cross-Linguistic Production Study". *Lingua*, 143, 129-61. https://doi.org/10.1016/j.lingua.2014.02.002.

Wang C. 王灿龙 (2003). "Zhiyue wuding zhuyu ju shiyong de ruogan yinsu" 制约无定主语句使用的若干因素 (Constraints of Indefinite-Subject Sentences). *Yufa yanjiu he tansuo* 语法研究和探索 (Research and Explorations into Grammar). Beijing: Shangwu yishuguan, 224-39.

Wei H. 魏红; Chu Z. 储泽祥 (2007). "'You dingju hou' yu xianshixing de wuding NP zhuyu ju" '有定居后'与现实性的无定NP主语句 (On the 'Postposition of Definite Subjects' and the Actual Sentences with Indefinite Subject NPs). *Shijie Hanyu jiaoxue*, 3, 38-51.

Wu, G. (1998). *Information Structure in Chinese*. Beijing: Peking University Press.

Xiong Z. 熊仲儒 (2008). "Hanyu zhong wuding zhuyu de yunzhun tiaojian" 汉语中无定主语的允准条件 (Licensing Conditions of Indefinite Subjects in Mandarin Chinese). *Anhui shifan daxue xuebao* (*Renwen shehui kexue ban*), 36(5), 541-8.

Xu, J. (2015). "Corpus-Based Chinese Studies". *Chinese Language and Discourse. An International and Interdisciplinary Journal*, 6(2), 218-44. https://doi.org/10.1075/cld.6.2.06xu.

Xu, L. (1995). "Definiteness Effects on Chinese Word Order". *Cahiers de Linguistique – Asie Orientale*, 24(1), 29-48. https://doi.org/10.3406/clao.1995.1465.

Xu, L. (1997). "Limitation on Subjecthood of Numerically Quantified Noun Phrases. A Pragmatic Approach". Xu, L. (ed.), *The Referential Properties of Chinese Noun Phrases*. Paris: Ecole des Hautes Etudes en Sciences Sociales, 25-44.

Xu L. 徐烈炯 (1999). "Mingcixing chengfen de zhicheng yongfa" 名词性成分的指称用法 (On the Semantic Content of Noun Phrases). Xu L. 徐烈炯 (ed.), *Gongxing yu gexing – Hanyu yuyanxue zhong de zhengyi* 共性与个性—汉语语言学中的争议 (Generality and Individuality. Controversies in Chinese Linguistics). Beijing: Beijing yuyan wenhua daxue chubanshe, 176-90.

Xu L. 徐烈炯; Liu D. 刘丹青 (2007). *Huati de jiegou yu gongneng* 话题的结构与功能 (Topic. Structural and Functional Analysis). Shanghai: Jiaoyu Chubanshe.

Zhang X. 张新华 (2007). "Yu wuding mingci zhuyu ju xiangguan de lilun wenti" 与无定名词主语句相关的理论问题 (On Indefinite-Subject Sentences). *Beijing daxue xuebao* (*Zhexue shehui kexue ban*), 44(6), 103-11.

Zhou S. 周思佳; Chen Z. 陈振宇 (2013). "'Yi liang ming' buding zhi mingci zhuyu ju yunzhun tiaojian jiliang yanjiu" "一量名" 不定指名词主语句允准条件计量研究 (A Quantitative Study of the Licensing Conditions of Indefinite Subjects Marked by "Yi(一) + Quantifier + Noun"). *Yuyan kexue*, 12(4), 371-82.

Zhu D. 朱德熙 (1982). *Yufa jiangyi* 语法讲义 (Lecture Notes on Grammar). Beijing: Shangwu yishuguan.

Corpus-Based Research on Chinese Language and Linguistics
a cura di Bianca Basciano, Franco Gatti, Anna Morbiato

Evidentiality 'In' and 'As' Context
Corpus-Based Insights About the Mandarin V-过 *guo* Construction

Vittorio Tantucci
Lancaster University, UK

Aiqing Wang
Lancaster University, UK

Abstract In this paper we argue that evidentiality can be a category of a linguistic system that emerges from the intersection between form, usage and 'contextual situatedness'. We provide a multivariate corpus-based case study about the usage of the V-过 *guo* construction in written Mandarin, and show how the text types in which the chunk appears significantly contribute to determine its pragmatic usage and its emergent meaning grounded in shared knowledge and collective recognition. This approach sheds new light on two critical issues. The first is that evidentiality is an important grammatical category of documentary, factual and academic prose in Mandarin Chinese. The second, much broader, claim of this paper is that generalisations about grammatical/semantic categories need to account for the usage of specific items in context. In this sense, 'physical and sociocultural situatedness' is as important a dimension as form and meaning in order to define categorial membership.

Keywords Chinese. Evidentiality. Context. Corpus-based. Multifactorial.

Edizioni Ca'Foscari

Sinica venetiana 6
e-ISSN 2610-9042 | ISSN 2610-9654
ISBN [ebook] 978-88-6969-406-6 | ISBN [print] 978-88-6969-407-3

Peer review | Open access
Submitted 2020-02-14 | Accepted 2020-10-14 | Published 2020-12-21

DOI 10.30687/978-88-6969-406-6/003

1 Introduction

It has been pointed out that the Mandarin[1] experiential marker 过 *guo*, originally expressing the past experience of a syntactic subject, has recently grammaticalised into an evidential construction (cf. Chappell 2001; Tantucci 2013, 2015a, 2015b, 2016c; Tantucci, Wang 2020b). In this study, we focus on the usage of V-过 *guo* in two comparable written corpora of Mandarin Chinese, namely the Lancaster Corpus of Mandarin Chinese (LCMC) (McEnery, Xiao 2004) and UCLA corpus of written Mandarin (Tao, Xiao 2012). The former includes texts from 1988 and 1992, whereas the latter includes texts from 2000 to 2005. Both corpora include one million words and are balanced with respect to the text types of which they are composed, so that they can be compared with one another. The aim of the present analysis is to shed light on the relationship between evidential reasoning and context and whether specific genres and textual environments favour the usage of evidential polysemies of V-过 *guo*. We are similarly interested in assessing whether the process of grammaticalisation of V-过 *guo* towards evidentiality is occurring at the expense of experiential usages of the same construct.

First of all, we can look at the formal and semantic differences between experiential and evidential usages of V-过 *guo*. Consider the two examples below:[2]

1. 她的鼻梁很细, 我从来没有看过人有这么细的鼻梁, 因而反把她年轻的、瘦削的脸衬得丰满起来。(LCMC / P: Romantic fiction)

tā	*de*	*bí-liáng*	*hěn*	*xì*	*wǒ*	*cónglái*	*méiyǒu*
she	SP	nasal-bridge	very	narrow	I	never	NEG
*kàn-**guo***	*rén*	*yǒu*	*zhème*	*xìde*	*bí-liáng*	*yīn'ér*	*fǎn*
see-EXP	person	have	such	narrow	nasal-bridge	thus	turn
bǎ	*tā*	*niánqīng*	*de*	*shòuxuēde*	*liǎn*	*chèn*	*de*
BA	she	young	SP	skinny	face	seem	DEG
fēngmǎn	*qǐlái*						
chubby	become						

'The bridge of her nose is very thin, I have never seen anyone with such a thin one, and it makes her young, skinny face look chubby'.

1 When Mandarin will be used in isolation, it will refer to present-day Mandarin (aka. currently spoken 普通话 *pǔtōnghuà* of Mainland China) throughout the present paper.

2 The glosses follow the general guidelines of the Leipzig Glossing Rules. Additional glosses include: BA = '*ba* construction particle'; DEG = 'complement of degree'; EMP = 'emphatic marker'; EVD = 'evidential'; EXP = 'experential'; ST = 'structural particle'.

2. 本世纪以来, 长江发生过三次严重的洪灾, 其中1931年和1935年两次大洪水, 分别淹地 5090万亩和2264万亩, 死亡 14.5万人和14.2万人。(LCMC / J: Academic prose)

běn	*shìjì*	*yǐlái*	*Chángjiāng*	*fāshēng-**guo***	*sān*	*cì*	
this	century	since	Yangtze.River	happen-EVD	three	time	
yánzhòng	*de*	*hóngzāi*	*qízhōng*	*1931*	*nián*	*hé*	*1935*
severe	SP	flood	among	1931	year	and	1935
nián	*liǎng*	*cì*	*dà*	*hóngshuǐ*	*fēnbié*	*yān*	
year	two	time	major	flood	respectively	inundate	
dì	*5090*	*wàn*	*mǔ*	*hé*	*2264*	*wàn*	*mǔ*
land	5,090	ten.thousand	mu	and	2,264	ten.thousand	mu
sǐwáng	*14.5*	*wàn*	*rén*	*hé*	*14.2*	*wàn*	*rén*
die	14.5	ten.thousand	person	and	14.2	ten.thousand	person

'Since the beginning of this century, there have been three severe floods in the Yangtze River, including two major floods in 1931 and 1935, which inundated 205,997 and 91,626 square meters of land and killed 145 thousand and 142 thousand people respectively'.

In (1), the speaker is genuinely expressing some subjective/personal impression that directly underpins his/her own personal experience, namely *that s/he normally has never seen a nose as fine* as the one of the character that is being narrated. S/he is therefore establishing reference to his/her own subjective experience and personal impressions about a specific event or state of affairs. In Pragmatics, the notion of perlocutionary effects regards *what a speaker intends an utterance to achieve in an addressee* (cf. Austin 1962; Searle 1976). The perlocutionary effects of (1) are clearly not the ones of informing the reader of a piece of documented information, but most likely to share his/her emotional/sensorial experience and/or personal affects. Simply put, the usage of 过 *guo* in (1) cannot express a piece of collective knowledge (it cannot be marked by evidential functions such as *it is known that*, or *as it seems*), but only personal experience and related emotions resulting of the speaking subject as an individual.

The usage of 过 *guo* in (2) is rather different. In this case the syntactic subject of the sentence is inanimate, and the event that is reported has not been necessarily experienced by the speaker. A completely different speech act is performed in this case. The speaker is no more referring to his/her personal affects, or the ones of a syntactic subject. Rather, s/he is reporting or presenting (cf. Faller 2002; Tantucci 2016a, 2016b, 2016c) a piece of information that s/he has somehow acquired and which s/he could potentially provide evidence for. Interestingly, the text types in which these two usages occur also differ substantially. In the former case the narration occurs in a fictional context, and it is therefore more likely to be aimed to entertain or empathise with the reader. In the latter usage, the V-过 *guo* construction is used in academic prose and is functional to mark a

piece of information as a fact that can be considered as reliable and documented/documentable. Intersubjectively, we could say that usages such as (1) tend to be aimed at establishing empathy among interlocutors, whereas utterances of the kind of (2) aim to be persuasive and reliable. Finally, it is important to note that both contexts of usage in (1) and (2) do indeed require the post-verbal marker 过 *guo* and could not be uttered with an evidentially/experientially neutral perfective marker such as 了 *le*[3] (Tantucci 2013, 225).

§ 2 provides an overview of the V-过 *guo* construction and its different usages. It also provides the operational criteria to disentangle experiential versus evidential senses. § 3 is based on a diachronic discussion about the grammaticalisation of the V-过 *guo* construction and the semasiological formation of different polysemies. The main case-study in § 4 is then centred on the relationship between evidential *vs* experiential usages of 过 *guo* and the text types in which they tend to occur. In particular, we will be focusing on the following research questions:

- What is the distribution in different text types of evidential versus experiential usages of V-过 *guo*?
- Have there been any significant changes in the partition of usages of V-过 *guo* in the last thirty years?
- What is the relationship between the different senses of V-过 *guo* and the textual environment in which they occur?

2 The Mandarin V-过 *guo* Construction

In the literature, V-过 *guo* is commonly considered as a polysemous construction. It can express directionality (e.g. Li, Thompson 1981; Chen 2008), therefore emphasising the actional (i.e. underpinning *Aktionsart*, see Vendler 1967) movement in space of dynamic verbs, as in 拿过 *náguò* 'to take/seize', 走过 *zǒuguò* 'to walk towards a certain direction', 递过 *dìguò* 'to hand over', and others (Tantucci 2015a, 69). It can express completivity (cf. Bybee, Perkins, Pagliuca 1994, 51; see also Dahl 1985, 95 on conclusives) or traversativity (Tantucci 2015a), thus describing the phasal meaning of "do[ing] something thoroughly and to completion", as conveyed by expressions such as *to shoot someone dead* or *to eat up*. The "lexical sources of completives [..] are all dynamic verbs or directionals, as they all suggest action or movement" (Bybee, Perkins, Pagliuca 1994, 59). They are actionally durative, as in 吃过 *chīguò* 'to finish eating' or 看过 *kànguò*

3 In the case of (1) this test would require a positive polarity.

'to end up watching'.[4] In example (3) below, V-过 *guo* expresses that the *action of eating the noodles has been completed* or '*traversed*' (Tantucci 2015a) so that a second action could be carried out or not.

3. 吃饭时没留意窗外, 吃过一碗刀削面走出小饭馆。(UCLA / G: Biography memoirs)
chī-fàn *shí* *méi* *liúyì* *chuāng* *wài* *chī-**guò*** *yì*
eat-meal while NEG pay.attention window outside eat-COMPL one
wǎn *dāoxiāomiàn* *zǒu-chū* *xiǎo* *fànguǎn*
bowl noodles walk-out small restaurant
'I did not pay attention to the outside of the window while eating; I finished a bowl of noodles and walked out of the small restaurant'.

These particular usages of V-过 *guo* do not contribute to the illocutionary force of the utterance, as they merely intervene lexically on the *Aktionsart* (Vendler 1957) – elsewhere alternatively called lexical aspect (Olsen 1997), transformativity (Johanson 2000) or situation aspect (Smith 1997) – of a verbal compound [VV]. Simply put, it only marks the temporal constituency or the internal phase structure IPS (Johanson 2000) of a predicate, i.e. whether an action has been brought to completion or to some resultant state.

A third function of V-过 *guo* is the "experiential perfect" usage (Comrie 1976, 58; Li, Thompson 1981; Dahl 1985, 141; Carey 1994; Yeh 1996; Dai 1997; Smith 1997; Dahl, Hedin 2000; Xiao, McEnery 2004; Lin 2006, 2007; Chen 2008; Wu 2008), whereby the construction indicates the past experience of the syntactic subject, as in example (1) (§ 1) or in expressions such as 我去过北京 *wǒ qù guo Běijīng* 'I have been to Beijing before', see also (4) below:

4. 或许林徽因的心情也是这般, 从来没有固执地想过要什么, 也没有刻意去拒绝什么。(UCLA / G: Biography memoirs)
huòxǔ *Lín Huīyīn* *de* *xīnqíng* *yě* *shì* *zhèbān* *cónglái*
perhaps Lin Huiyin SP mood also be like.this never
méiyǒu *gùzhíde* *xiǎng-**guò*** *yào* *shénme* *yě* *méiyǒu*
NEG stubbornly think-EXP want what also not
kèyì *qù* *jùjué* *shénme*
deliberately go refuse what
'Perhaps Lin Huiyin's mood is also like this; she never stubbornly thought about what she wanted, nor did she deliberately refuse anything'.

4 In Mandarin, both directional and completive usages of 过 *guò* retain the fourth tone, whereas more grammaticalised forms tend to be toneless. The *pinyin* notation of the rest of this paper will account for this distinction.

In (4) above, the function of V-过 *guo* is no more the one of expressing that a durative event has been completed, but rather to convey that the animate subject of the sentence, 林徽因 *Lín Huīyīn*, has never experienced a particular feeling, namely the one of *being obstinate in wanting something*. Table 2 below provides the diagnostics for identifying experiential usages of V-过 *guo*:

Table 1 Diagnostics for identifying 过 *guo* as an experiential (adapted from Tantucci 2015a, 87)

过 *guo* **as an experiential**
Profiles the syntactic subject's past experience.
Employed as a perfect in contexts where the syntactic subject has been through some experience before.
Frequently used with dynamic verbs.
Used generally in the first person, in negated statements or in second person questions (Dahl 1985; Dahl, Hedin 2000; Tantucci 2013).
It cannot collocate with the perfective post-verbal 了 *le*.*
It can collocate with the adverbials 曾经 *céngjīng* 'once' or 从来 *cónglái* 'never'.
It cannot collocate with inanimate subjects.
It can collocate with absolute-state predicates (rare).
Not felicitous when collocating with IE adverbials such as 据了解 *jù liǎojiě* 'it is understood that', 好像 *hǎoxiàng* 'apparently', 众所周知 *zhòngsuǒzhōuzhī* 'as everyone knows'.

* This is a diagnostic that helps distinguishing comparatively more grammaticalised usages of 过 *guo* (e.g. experiential and evidential) from cases where 过 *guò* is used as a completive or a directional complement, such as in 该联络的事宜都联络过了 *gāi liánluò de shìyí dōu liánluò guò le* 'all the arrangements that required contacts where dealt with' (LCMC / E14).

In Tantucci (2013; 2015a), it is also argued that 过 *guo* developed a more grammaticalised function underpinning knowledge ascription and evidentiality. At this stage of change of 过 *guo*, the notion current relevance for the here-and-now of the conversation underpins a presentative stance rather than an assertive one (Faller 2002). That is, while an assertive speech act has the sincerity condition that the speaker believes P and is unmarked with respect to its reliability, in the case of presentative utterances the speaker/writer merely 'introduces' a piece of knowledge s/he acquired somehow for the benefit of the addressee/reader. In this latter case, the speaker/writer marks the proposition as a piece of information that is somewhat 'reliable' and which can be potentially documented/confirmed. While experiential usages of 过 *guo* tend to occur in questions and in negative statements, evidential ones show a tendency to occur assertively, in the declarative mood (Tantucci 2013, 2015a; Tantucci, Wang 2020b). This functional and formal tendency is due to the presentative illocutionary force of evidential statements, and the fact that the perlo-

cutionary effects of P are distinctively the ones of informing a specific or generic addressee, rather than expressing subjective affective concern or empathy to the interlocutor. As a result, evidential usages of 过 *guo* tend to occur in the third person or in impersonal/subjectless constructions (Tantucci 2013, 2015a).

5. 在现实主义与古典主义之间, 出现过浪漫主义的 "叛乱"。(LCMC / J: Academic prose)
zài *xiànshízhǔyì* *yǔ* *gǔdiǎnzhǔyì* *zhījiān* *chūxiàn-**guo***
at realism and classicism between exist-EVD
làngmànzhǔyì *de* *pànluàn*
romanticism SP rebellion
'There used to exist a 'rebellion' of romanticism between realism and classicism'.

In the academic context of example (5) above, no experiential meaning is at issue. The author is not interested in sharing his/her own or someone else's past experience with the reader. Rather, s/he purposely marks the proposition as a piece of knowledge that bears some sort of social recognition and which can be potentially confirmed and verified. In other words, a different 'pragmeme' is at play, viz. a different "situational prototype capable of being executed in the situation" (Mey 2001, 221). In this paper, we will argue that "contextual situatedness" (cf. Mey 2010; Haugh 2012) is a fundamental dimension that inherently informs meaning, and in particular contributes to determine the polysemic status of the V-过 *guo* construction. In Pragmatics, it is stressed that the physical and cultural environment plays a fundamental role in the encoding of the illocutionary force of an utterance. In other words, speech acts "in order to have an effect, must be situated" (Mey 2010, 2883; Capone 2005; Tantucci 2016c). The different intersection between contextual situatedness and illocutionary force that we find in (5) above determines a distinctive evidential reading of the utterance. In fact, in the same context, the merely perfective marker 了 *le* would not be idiomatic (to some degree not grammatical), as it would lack added evidential meaning that marks the proposition as a piece of 'documented' evidence, which bears collective recognition (*出现了浪漫主义的 "叛乱" *chūxiàn le làngmànzhǔyì de pànluàn*) (Tantucci 2013, 255). In table 2 below, we report the formal and functional diagnostics for identifying evidential usages of V-过 *guo*:

Table 2 Diagnostics for identifying 过 *guo* as an interpersonal evidential (IE) (adapted from Tantucci 2015a, 88)

过 ***guo*** **as an evidential**
Profiles the speaking subject's (Benveniste [1958] 1971; Traugott 2003; Langacker 2008) acquired information.
Employed in contexts characterised by an epistemic or presentative stance (Mushin 2001; Faller 2002), that is, the speaker/writer markedly 'introduces' a particular piece of knowledge s/he has acquired somehow.
Frequently in third person declaratives.
It cannot collocate with the perfective post-verbal 了 *le*.
It can collocate with the adverbials 曾经 *céngjīng* 'once' or 从来 *cónglái* 'never'.*
It can collocate with inanimate subjects.**
It can collocate with absolute-state predicates (rare).
Felicitous when collocating with IE adverbials such as 据了解 *jù liǎojiě* 'it is understood that', 好像 *hǎoxiàng* 'apparently', 众所周知 *zhòngsuǒzhōuzhī* 'as everyone knows'.
* This indicates that 过 *guo* reached a grammaticalisation stage where it can express aspectual discontinuity or anti-resultativity (e.g. Plungian, van der Awera 2006; Tantucci 2015a), which in turn is not possible for completive and directional usages of the same form.
** This is an important diagnostic as what is at issue in evidential usages is a piece of documented and/or socially recognised information, rather than the subjective experience of an individual. Impersonal usages (absent at earlier stages of the grammaticalisation of 过 *guo*) are an important sign of this shift, as the absence of a syntactic subject is precisely due to the attempt to communicate *what has accordingly happened*, rather than *what has been once experienced by someone*, i.e. the syntactic subject of the sentence (Tantucci 2015a, 91).

Evidentiality has been defined as "the existence of a source of evidence for some information" (Aikhenvald 2004, 1), the "encoding of the speaker's (type of) grounds for making a speech act" (Faller 2002, 2), or the communication of a piece of "acquired knowledge" (Tantucci 2013, 214). Evidentials relativise or measure the information status of the sentence (Rooryck 2001a, 125; 2001b), yet in many languages, such as English, do not constitute a grammatical category and are generally communicated through adverbials or discourse markers such as *apparently* and *allegedly* (see Mushin 2001, 54; Narrog 2009, 10), predicates conveying an evidential meaning such as *it seems that*, *it appears that*, and *I saw that*, pragmatic strategies (see Aikhenvald 2004), or overtly expressed contextual elements providing some type of information. In our view, in languages where evidentiality does not correspond to a distinctive inflectional category, it is precisely the intersection between form, usage, and context that define an evidential reading. Similarly, it could also be argued that, even in languages where evidential systems are highly complex and grammaticalised (e.g. mostly spread through Northern, Central America, Eastern Europe, central and Southeast Asia; Aikhenvald 2004, 303), there is still a crucial intersection between contextually situatedness and usage of

those forms (see for instance the hybrid case of Gitksan evidentials, which are entirely optional and not paradigmatically organised; Peterson 2010). The inherent relationship between contextual situatedness and formal usage of some evidentials is an argument that has been put forward by Squartini (2012) in the discussion of the subcategory of circumstantial evidentiality, but also by Capone (2005; 2010) and Tantucci (2016c) concerning the crucial role of physical and sociocultural context for the encoding of so-called 'evidential pragmemes'.

A crucial dimension that is missing from the classification in table (2) above is therefore the one of 'contextual situatedness' of the V-过 *guo* construction. That is to say, the diagnostics that are reported in each table take into account formal and functional elements of usage, yet they overlook the textual and sociocultural environment of each polysemy. In this sense, a multivariate corpus-based analysis can shed important light on the holistic relationship between form, illocutionary force and context. Significant intersections of the variables subsumed by formal, pragmatic and contextual dimensions are referred to as **illocutional concurrences** (**IC**) (Tantucci, Wang 2018, 2020a, 2020b; Formato, Tantucci 2020). Namely, ICs encompass converging factors at different levels of verbal experience that contribute, both locally (i.e. at the morphosyntactic level) and peripherally (i.e. at the illocutionary level), to the encoding of contextually and culturally situated speech acts. The final discussion of this paper will be devoted to the inherent relationship between contextual situatedness and schematic categorisation of form and meaning. A specific focus will be placed on the interdependence of conventional association of linguistic functions and the situation type in which they are used as an important factor of semantic and grammatical change.

3 The Grammaticalisation of V-过 *guo*

In this brief section, we discuss the importance of context in the diachronic reanalysis of V-过 *guo* as an evidential construction. This claim will be further discussed in § 4, where we will provide a detailed multivariate analysis of the synchronic usage of V-过 *guo* in the LCMC and the UCLA corpora of Mandarin Chinese.

During the 唐 Tang dynasty (618-907 AD), 过 *guò* starts to occur in the second slot of [vv] constructions with a specific completive/traversative meaning (Cao 1995, 38), therefore expressing lexically the phase where an action has been completed/traversed. Different from early directional usages, this new function collocates with durative verbs that do not necessarily express physical movement:

6. 每至义理深微常不能解处, 闻醉僧诵过经, 心自开解。(纪闻 *Jìwén*, 太平广记 *Tàipíng guǎngjì*, 异人异僧释证卷 *Yìrén yìsēng shìzhèng juàn*, 第 *dì* 81-101 卷 *juǎn*, Cao 1995, 38)
   měi *zhì* *yìlǐ* *shēnwēi* *cháng* *bù* *néng*
   every arrive argumentation mysterious often NEG can
   jiě *chù* *wén* *zuì* *sēng* *song-**guò*** *jīng*
   comprehend place hear drunk monk recite-COMPL scripture
   xīn *zì* *kāi* *jiě*
   heart self open understand
   'Every time the argumentation would become too difficult and mysterious, all the parts that s/he could not comprehend would then become clear after s/he listened to that drunk monk reading through them'.

From (6) above, we can see that 过 *guò* now starts to convey completivity/traversativity, as it marks the phasal meaning of completing/traversing an action, rather than marking a syntactic subject's past experience. Cao notes that during the Tang dynasty the phasal meaning of 过 *guò* merely

> indicates the action itself, and never stresses the subsequent results of the event [...] this is evident from the missed co-occurrence with resultative verbs such as 关 *guān* 'to close', 锁 *suǒ* 'to lock', 盛 *chéng* 'to fill' or absolute states such as 老 *lǎo* 'be/grow old', 冷 *lěng* 'to be cold', 红 *hóng* 'to be red', 白 *bái* 'to be white' and others. (1995, 40)[5]

A possible operational model that can inform the stages of semantic and grammatical change of V-过 *guo* is the Invited inferencing theory of semantic change (IITSC) (Traugott 1999; Traugott, Dasher 2002, 5; see also Dahl 1985, 11). IITSC states that inferences pragmatically induced from the speaker/writer to the addressee/reader tend to become conventionalised and determine new semantic polysemies within a construction. In a subsequent stage of reanalysis, due to its semantic element of discontinuity to the present, the V-过 *guo* construction starts to be encoded as a perfect with a conventionalised meaning expressing past-experience of an animate subject. Earliest evidence of this is found between the Tang and the Song (960-1279 AD) dynasties whereby 过 *guo* starts to collocate with mental verbs or verbs referring to the syntactic subject's past experience, as in the case of 尝 *cháng* 'to taste', 验 *yàn* 'to experience', 问 *wèn* 'to ask' (Lin 2004, 45), albeit it is not frequently used before the Yuan dynasty (1271-1368 AC) (Cao 1995, 43; Lin 2004, 42):

5 Translated and readapted from Chinese. Unless otherwise indicated all translations are by the Authors.

7. 看文字须仔细, 虽是旧曾看过, 重温亦须仔细。(朱子语类 *Zhūzǐ yǔlèi*, 卷一〇 *jǔan yīlíng*, Cao 1995, 41)

kàn	*wénzì*	*xū*	*zǐxì*	*suíshì*	*jiù*	*céng*	*kàn-**guo***
see	character	must	careful	although	old	once	see-**EXP**
chóngwēn	*yì*	*xū*	*zǐxì*				
review	also	must	careful				

'When you look at a character you must be attentive, even if it is one that you saw before, you still have to be attentive'.

In the case of (7), 过 *guo* no longer simply intervenes on the *Aktionsart* of the predicate on a lexical level. It has now developed a new grammaticalised function of experiential perfect (e.g. Comrie 1976). It therefore expresses current relevance of a previous experience occurring in a vague, discontinuous past. The bulk of the literature focusing on the aspectual features of 过 *guo* is distinctively focused on this particular usage. The main aspectual features of the experiential V-过 *guo* that emerge from the literature are the following:

a. It marks an eventuality having at least one occurrence in the past.
b. It has a 'class' meaning, viz. refers to an event type, rather than a specific instantiation.
c. It expresses aspectual discontinuity to the present/or a reference time.
d. It encodes only repeatable eventualities.
e. It marks an event that is temporally independent from others in the discourse.
f. It expresses a past eventuality often in the form of someone's past experience (e.g. Chao 1968; Li, Thompson 1981; Yeh 1996; Dai 1997; Smith 1997; Lin 2006, 2007; Wu 2008; Li 2011).

In experiential usages of V-过 *guo*, the original actional meaning of 'having been through an action' that was originally encoded on a lexical level, has now turned into a more speaker-based meaning whereby some animate subject's past experience becomes at-issue for the here-and-now of the speech event.

While both completive or resultant states are attested to be common lexical sources of perfects (i.e. resultative, hot-news, existential, experiential meanings; see McCawley 1971; Portner 2003; Dahl, Hedin 2000), in the case of 过 *guo*, aspectual discontinuity and 'absence' of results are themselves the trigger of specifically experiential and subsequent evidential reanalyses of the chunk: i.e. 我年轻过 *wǒ niánqīng guo* 'I have been young (albeit I am not anymore)' (see Comrie 1976; Carey 1994; Dahl 1985; Dahl, Hedin 2000; Chappell 2001; Li 2011; Tantucci 2013 for specific discussions about the typological features of experiential perfects).

It is acknowledged that experiential and existential perfects express relevance to the present without expressing a resultative continuation of the past event up to the moment of speech. This is the case of a well-known example:

8. The Earth has been hit by giant asteroids before. (Portner 2003, 464)

Usages involving a discontinuous past such as (8) show that relevance needs to be intended as having a primarily discursive nature, rather than having to do with the actionality or some temporal/physical contiguity/continuity of the event to the utterance time. Most crucially, Portner notes that the experiential and existential perfects of the kind of (8) "provide evidence for something, not that it indicates any results" (2003, 464; cf. Rubovitz 1999 about the semantic-pragmatic correspondence between existential/experiential perfects and evidential reasoning).

The notion of discontinuity to the present becomes an important element of further semantic and grammatical reanalysis of V-过 *guo*. At this point in time, invited inferences being conveyed by the speaker/writer can be semantically and pragmatically associated with some reliability behind the proposition, whereby the truthfulness of P becomes markedly "at-issue" (Faller 2002; Tantucci 2016a, 2016b). In fact, due to the inherent anti-resultativity of the construction, an event marked with 过 *guo* is necessarily communicated either in the form of personal experience or as a piece of interpersonally shared knowledge (Tantucci 2015a). Crucially, earliest usages of V-过 *guo* as an experiential perfect seem to be limited to collocations with animate subjects, mental verbs or verbs profiling the syntactic subject's personal experience in the past (Cao 1995; Lin 2004; Liu 2009, 231). However, Tantucci (2013, 224-5; 2015a) notes that during the Qing dynasty (1644-1912 AD) V-过 *guo* undergoes a new stage of semantic and grammatical reanalysis. This is a stage where V-过 *guo* collocates with subjectless or impersonal constructions with a new interpersonal evidential (IE) meaning. At this stage, V-过 *guo* is no longer used to mark an event in the form of an animate subject's passed experience, but rather as a piece of knowledge shared by the speaker/writer together with a generic third party in society. Tantucci (2013, 2015a) notes that this trend is confirmed by the rise of the subjectless construction 发生过 *fāshēng-guo* 'it happened before that', as the valency of 发生 *fāshēng* in Mandarin normally does not include an experiencer. Earliest collocations of this verb with 过 *guo* are a clear sign of new evidential reanalysis of the chunk. Something similar is at stake for the verb 有 *yǒu* 'to exist, to be there', expressing an existential meaning rather than a possessive one. Early evidential usages of 有过 *yǒu-guo* 'there has been before' in the PKU-CCL-COR-

PUS[6] also date back to the Qing dynasty:

9. 这一天城里的街道,居然也打扫干净了,只怕从有上海城以来,也不曾有过这个干净的劲儿。(CCL / 清 *Qīng* / 二十年目睹之怪现状 *èrshínián mùdǔ zhī guài xiànzhuàng*)
zhè yī tiān chéng lǐ de jiēdào jūrán yě
this one day city in SP street unexpectedly also
dǎsǎo-gānjìng le zhǐ pà cóng yǒu Shànghǎi
clean-up.clean PFV only be.afraid since exist Shanghai
*chéng yǐlái yě bùcéng yǒu-**guo** zhè ge gānjìng de*
city since also never exist-EVD this CLF clean SP
jìn'er
degree/energy
'On this day, streets in the city had unexpectedly been cleaned thoroughly; I am afraid since the existence of Shanghai, the city has never been this clean'.

In example (9), there is not an animate syntactic subject to which some past experience is ascribed. The speaker/writer is similarly not referring to his personal life, as s/he cannot have experienced the full history of the city of Shanghai. S/he is rather referring to a piece of information that could be confirmed by other members of his/her own community of practice, thus expressing a proposition bearing collective recognition (cf. Searle 2010). Usages such as the one above are defined as interpersonal evidentials (IE) since,[7] as while no specific source of evidence is encoded by the construction, a piece of information is marked as shared knowledge within a community of practice, ideally paraphrasable as *it is known that*.

After the 民国 *Mínguó* period (1912-1949), the PKU-CCL-CORPUS includes a fairly balanced collection of texts, which is no longer limited to fictional registers, but also includes factual prose from press, academic journals and biographies. In Tantucci (2015a), it is shown that it is precisely in these textual environments that evidential usages of V-过 *guo* become increasingly frequent. From (9) above, we can observe that it is precisely the anti-resultativity of V-过 *guo* that

6 The PKU-CCL-CORPUS is one of the largest corpora of Mandarin Chinese available and includes both a balanced synchronic and a diachronic section of written language. The total size of corpus data is approximately 200 million Chinese characters. Texts written in traditional Chinese in PKU-CCL-CORPUS contain approximately 101 million Chinese characters (486 documents, 54 folders, 202,305,825 bytes), and the texts written in modern Chinese contain 115 million Chinese characters (157 documents, 23 folders, 229,700,435 bytes).

7 E.g. Tantucci 2013, 2015a, 2015b, 2016a, 2016c, 2017a, 2017b, 2020; Tantucci, Wang 2020a; Arslan et al. 2014; Jarque, Pascual 2015; Brugman, Macaulay 2015; Guardamagna 2017; Van Olmen 2019.

prompts further speculations concerning the evidence behind the proposition. In this sense, all the evidence that is provided subsequently is pragmatically aimed at filling a 'temporal gap' between the event and the reference time.

The diagram below summarises the present data about the grammaticalisation pathway of the V-过 *guo* construction:

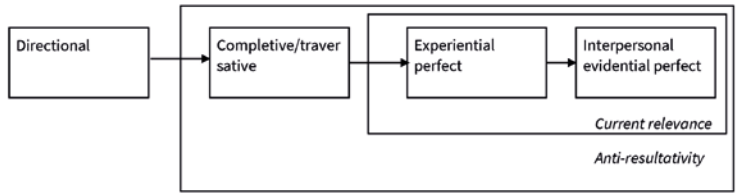


Figure 1 The pathway of change of the V-过 *guo* construction

As we can see from figure 1, a first step towards the grammaticalisation of V-过 *guo* is the transition from meaning expressing directionality of actions in space to a new aspectual meaning (completive/traversative) expressing that some action has been completed or 'traversed' by an animate subject **[fig. 1]**. This is an important stage of change of the construction, as the event is never conceptualised as entailing a resultative state. This element of anti-resultativity becomes crucial for further stages of change, as it persists (cf. Hopper 1982) in later usages conveying past experience of an animate subject. Anti-resultativity, in connection with discursive current relevance, contributes to express that an event has been experienced by the subject in a vague past, without specific reference to when this happened. Usages of the construction in the third person or in impersonal contexts contribute to a new evidential reading of the events that are referred to. In Fludernik, a distinction is made between "natural narrative proper" and "retelling of other people's stories" (2006, 14). The crucial grammatical distinction between the two consists in first-person versus third-person narration (Norrick 2013a). Norrick notes that differences in first-person versus third-person narration underpin idiosyncratic features of the two types of narratives in relation to their form and function. They reflect differences in terms of teller perspective, story introduction, epistemic authority, and function (Norrick 2013a, 2013b). Frequent third-person-shift and impersonal usages are here considered as a very important factor contributing to the rise of novel interpersonal evidential polysemies of V-过 *guo*. Formal features as such, intersecting with specific text types and 'contextual situatedness', holistically affected the last stage of grammaticalisation of the construction from experientiality to interpersonal evidentiality.

4 A Corpus-Based Account of V-过 *guo* in Context

In this section, we provide the results of a corpus-based study from two synchronic corpora of Mandarin Chinese:

- the synchronic Lancaster Corpus of Mandarin Chinese (LCMC) (McEnery, Xiao 2004), a one-million-word balanced corpus designed as a Chinese match of the Freiburg-LOB Corpus of British English (FLOB), including texts from the years 1988 and 1992.
- The UCLA corpus of written Mandarin (Tao, Xiao 2012), also a one-million-word balanced corpus, designed as a match of the LCMC, including texts from 2000 to 2005.

The partition of texts of the LCMC is reported in the table below:

Table 3 Text types of the LCMC

Code	Text category	Samples	Proportion
A	Press reportage	44	8.8%
B	Press editorials	27	5.4%
C	Press reviews	17	3.4%
D	Religion	17	3.4%
E	Skills/trades/hobbies	38	7.6%
F	Popular lore	44	8.8%
G	Biographies/essays	77	15.4%
H	Miscellaneous	30	6%
J	Science	80	16%
K	General fiction	29	5.8%
L	Mystery detective fiction	24	4.8%
M	Science fiction	6	1.2%
N	Martial arts fiction	29	5.8%
P	Romantic fiction	29	5.8%
R	Humour	9	1.8%
Total		**500**	**100%**

With this survey we aimed at answering three research questions:

- What is the distribution in the two corpora of evidential versus experiential usages of V-过 *guo*?
- Have there been any significant changes between the 1990s and the beginning of the 21st century in the partition of usages of V-过 *guo*?
- Is there a relationship between the formal and functional categories of V-过 *guo* and the textual environment in which it occurs?

4.1 Data Retrieval and Annotation

To answer each question it was necessary to design a solid annotation scheme that could grant a high inter-rater reliability (85%). We took into account a number of formal, functional and contextual dimensions, so that we could gather a holistic understanding of the behavioural profiles (cf. Gries 2010) of the construction. We therefore focused on: whether the polarity of the sentence was negative or positive; the corpus in which the chunk appeared; the verb (both as a token and as a type) collocating with 过 *guo*; the text–type where the V-过 *guo* was used; whether sentence final particles were present in the utterance; the type of the location of the force of each usage; the person of the verb (e.g. first singular, 3rd plural, and so on); and whether the function was evidential rather than experiential. The function of the construction was also the dependent variable of our analysis, and was based on the assessed set of criteria given in tables 1 and 2, in § 2. In table 4 below is given an example of one string of annotation:

Table 4 Example of an annotated string of the usage of V-过 *guo* in the LCMC and UCLA

Function	Person	Verb	Verb_type	Ill_force	Sent_final	Text_type	Corpus	Polarity
Evid	3s	说	say	asser	N	B	LCMC	P

The utterance in table 4 has been annotated as an evidential usage, in the third person singular, collocating with the verb 说 *shuō*, which is a verb of saying (annotated as 'say'). The illocutionary force of the utterance is assertive, it does not include sentence final particles, the text type corresponds to Press-editorials **[tab. 3]**, the corpus in which it occurs is the LCMC and the polarity of the sentence is positive.

We retrieved all the usages including verbs with the highest MI^3 score from both the LCMC and the UCLA. Mutual Information (MI) expresses the extent to which observed frequency of co-occurrence differs from expected frequencies. It measures the strength of association among specific words or word types (in our case the strength of association of 过 *guo* with a preceding verb). The MI^3 score is used to rebalance MI score so as to give more weight to frequent words and less to infrequent words, by 'cubing' observed frequencies (cf. Oakes 1998, 171-2).

4.2 Data Analysis

After the retrieval of the top 15 verbs with the highest MI^3 score from both corpora, we first seeked to answer our first two research ques-

tions, underpinning respectively the distribution in the two corpora of evidential versus experiential usages of V-过 *guo* and whether any changes between the 1990s and the beginning of the 21st century have occurred in the partition of usages of V-过 *guo*. We thus looked at the general distribution of experiential and evidential usages in the two corpora. We then performed a test of independence to assess whether there were significant mismatches based on chi-square and 'Pearson residuals'.

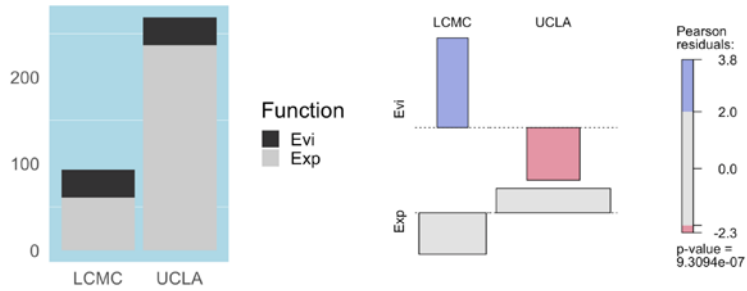


Figure 2 Distribution and test of independence of evidential *vs* experiential usages in LCMC and UCLA

The bar plot on the left hand side of figure 2 indicates a much more frequent usage of the V-过 *guo* construction **[fig. 2]**. It also shows a remarkably higher frequency of experiential usages (light grey) in contrast with the evidential ones (black) in the UCLA in comparison with the LCMC. This mismatch is statistically significant as indicated by the p-value (< 0.0005) from the chi-square test, given at the bottom right hand side of figure 2. To explain, the plot on the right-hand side above is called assocplot (R package: vcd, cf. Hornik, Zeileis, Meyer 2006) and allows the analyst to visualise significant mismatches between observed and predicted frequencies deriving from a chi-square test. These mismatches are commonly called 'Pearson residuals'. If the observed frequency is greater than expected, the residual is positive. If the observed frequency is smaller than expected, it is then negative (Levshina 2015, 218). A blue colour (if any) indicates a significantly positive mismatch, whilst a red colour (if any) indicates a negative one, while the width of the bars is based on frequency.

From figure 2 we can clearly conclude that the frequency of evidential usages of V-过 *guo* is significantly higher in the LCMC corpus in comparison with the UCLA. This first result is not an obvious one. If we consider the relatively recent development of evidential functions of V-过 *guo*, one may expect it to progressively increase throughout the decade in between the LCMC and the UCLA. Quite

the opposite emerges from figure 2, as it is the experiential function the one that increases dramatically. This tendency supports the idea that constructional change and grammaticalisation are not necessarily incremental (e.g. Tantucci, Culpeper, Di Cristofaro 2018; Tantucci, Di Crostofaro 2019). Once a division of labour among functions of one construction is established, the frequency of comparatively more recent usages (such as the case of V-过 *guo* used as an evidential) is not necessarily going to further increase at the expense of comparatively older ones (e.g. V-过 *guo* used as an experiential).

It is now time to bring to the fore the role of context and text types in the encoding of evidential rather than experiential functions of V-过 *guo*. To begin with, we plotted a multiple correspondence analysis (MCA) (e.g. Nenadic, Greenacre 2007) on a two-dimensional plane. In this model, associations among variables are measured by calculating the chi-square distance between different categories of the variables and between observations. These associations are then represented graphically as a map, which eases the interpretation of the structures in the data: the closer the distance between variables, the stronger the statistical correspondence (Levshina 2015).

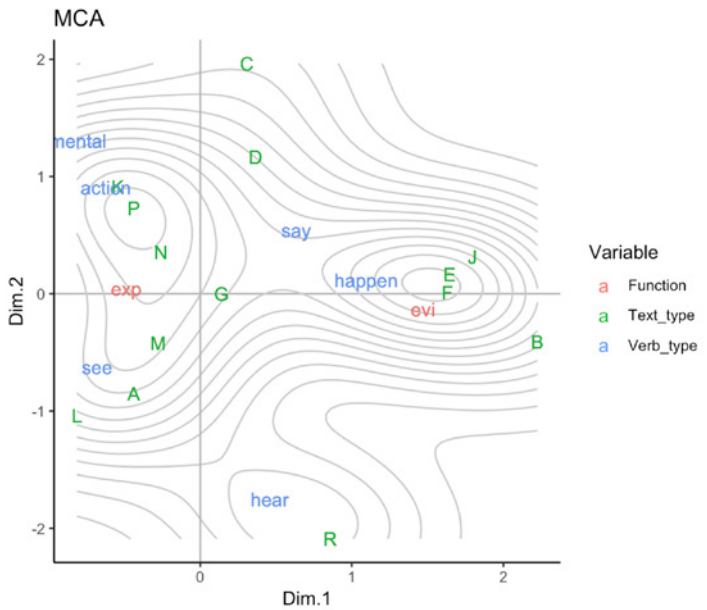


Figure 3 MCA of the relationship between function of 过 *guo*, text types and verb types

In the plot above the two dimensions represent 84.1% of variation among the three variables, which is a good approximation for MCA visualisation (Levshina 2015, 382). What counts for the interpretation of the data is the degree to which Function (i.e. Experiential *vs* Evidential), Text Type and Verb Type cluster together, therefore indicating a large-scale convergence in the way people use the V-过 *guo* construction, pragmatically, semantically and in contextually situated text types.

We can first note a clear division between the left and the right hand side of the plot, with two distinct clusters including text types (green) and verb types (blue) around respectively experiential and experiential usages (red). More specifically, at the left-hand side of the map there are experiential functions of V-过 *guo*, in turn attracting a different set of text types and verb types. More specifically, experientials are strongly attracted to verbs of action, or physical perception, such as 见 *jiàn* 'to see' or 看 *kàn* 'to watch', and mental verbs such as 想 *xiǎng* 'to think/plan'. These tend to form a cluster with text types K (General fiction), P (Romantic fiction), N (Martial arts fiction), M (Science fiction), A (Press reportage), L (Mystery detective fiction), and G (Biographies/essays). Most of these textual environments are fictional, whereby emotions and distinctive features of characters are often expressed through reference to their past experiences. The only exception regards A (Press reportage), which is undoubtedly a factual genre, yet also strongly based on a narrative stance of past events, which are very often experienced by the reporter or by other people who are being interviewed. Consider the extract below from the LCMC:

10. 这个问题是个很大的问题，因为我是从车间里滚出来的, 发电机我也开过, 上海二六轰炸的时候, 我开发电机, 我当厂长, 我都是干过活的。 (LCMC / A: Press reportage)

zhè	*ge*	*wèntí*	*shì*	*ge*	*hěn*	*dà*	*de*	*wèntí*
this	CLF	question	be	CLF	very	big	SP	question
yīnwèi	*wǒ*	*shì*	*cóng*	*chējiān*	*lǐ*	*gǔn-chūlái*	*de*	*fādiànjī*
because	I	be	from	workshop	in	work-out	SP	generator
wǒ	*yě*	*kāi-**guo***	*Shànghǎi*	*èrliùhōngzhà*			*de*	
I	also	operate-**EXP**	Shanghai	February.Sixth.Incident			SP	
shíhòu	*wǒ*	*kāi*	*fādiànjī*	*wǒ*	*dāng*	*chǎng*	*zhǎng*	*wǒ*
time	I	operate	generator	I	be	factory	director	I
dōu	*shì*	*gàn-**guo***	*huó*	*de*				
all	be	do-**EXP**	work	EMP				

'This is a very big question, because I used to work in a workshop and operated generators before; at the time of the February Sixth Incident in Shanghai, I operated generators and was a factory director—I was indeed engaged in my work'.

In the case above, the narrator is being interviewed about his previous experience working in a factory in Shanghai. This is a very interesting contextual environment. In fact, the usage of the construction is clearly experiential, yet, in this and similar contextual environments, someone's personal experience is not shared merely to establish empathy among interlocutors, but more specifically to count as evidence about some broader factual information that has been reported by the interviewer. Nonetheless, interpersonal evidential pragmatic markers, such as 据了解 *jù liǎojiě* 'it is understood that', 好像 *hǎoxiàng* 'apparently', 众所周知 *zhòngsuǒzhōuzhī* 'as everyone knows', would not be compatible with this usage, which indicates that V-过 *guo* in (10) can still be considered as prominently experiential.

Back to the map, we can see that evidential polysemies are rather attracted to verbs of saying (e.g. 说 *shuō* or 讲 *jiǎng*) or verbs inherently expressing the occurrence of some event, such as 出现 *chūxiàn* 'to appear', 发生 *fāshēng* 'to happen', 有 *yǒu* 'to exist, to occur', and so on. The convergence of these verb types and evidential usages of 过 *guo* is at stake in texts such as E (Skills/trades/hobbies), F (Popular lore), J (Science) and B (Press editorials). The latter all tend to be geared to registers whereby information needs to be reported as a piece of evidence, rather than some past event that contribute to shape the personality or the personal history of a specific persona/character. In this case, events are presented to the reader as facts that can be potentially verified. The per-locutionary effects of these usages are not the ones of getting to know someone better, but rather to inform the reader of a piece of socially shared knowledge.

11. 自上个世纪七十年代开始，有过四次较大的 METI 项目，但每一次都有人站出来反对。 (UCLA / J: Science)

zì	*shàng*	*ge*	*shìjì*	*qīshíniándài*	*kāishǐ*	*yǒu-**guo***	*sì*	
from	last	CLF	century	70s	begin	have-EVD	four	
cì	*jiào*	*dà*	*de*	*METI*	*xiàngmù*	*dàn*	*měi*	*yí*
time	relatively	big	SP	METI	project	but	every	one
cì	*dōu*	*yǒurén*		*zhàn-chūlái*	*fǎnduì*			
time	always	someone		stand-out	object			

'Starting from 1970s, there have been four relatively big METI projects, but every time there was always someone standing out to object'.

In (11) above, the stance of the speaker/writer is not centred on the identity of a specific persona, rather s/he uses 过 *guo* to report a piece of documented information that entails collective recognition, as adverbials of the kind of 据了解 *jù liǎojiě* 'it is understood that', 好像 *hǎoxiàng* 'apparently', 众所周知 *zhòngsuǒzhōuzhī* 'as everyone knows' would be perfectly idiomatic with this usage. This kind of usage is grounded in interpersonal evidentiality and is significantly associated with text types such as scientific essays or reports, as in the case above.

4.3 Evidential *vs* Experiential Categorisation in Context

Significant data-driven intersections of pragmatic, formal and contextual features are elsewhere defined as illocutional concurrences (IC) (cf. Tantucci, Wang 2018; 2020a; 2020b). IC are crucial to show that grammatical meaning is not independent from the pragmatic stance adopted by the interlocutors as well as the 'contextual situatedness' in which the speech event takes place.

This point is particularly evident in the last analysis of this paper below. In this case we plotted a conditional inference tree model (cf. Hothorn, Hornik, Zeileis 2006; Tagliamonte, Baayen 2012) gathering unbiased corpus-driven convergences of form, meaning, context and pragmatic effects, all contributing to the spontaneous encoding of either experiential or evidential usages of V-过 *guo*. We took in to account the function of the construction, the polarity (from table 1 in § 2 we can see how experiential usages of 过 *guo* are generally agreed to occur with negative polarity or in questions), the illocutionary force (whether the speech act occurs as a modalised evaluation – e.g. Tantucci, Wang 2018 – a question or a bare assertion), and the presence of sentence final particles, which could shed light on whether the construction is used in questions, or whether the utterance is characterised by modalised elements of intersubjectivity occurring at sentence periphery (cf. Traugott 2012, 2016; Tantucci 2017a, 2017b, 2020, forthcoming; Tantucci, Wang 2018, 2020a, 2020b).

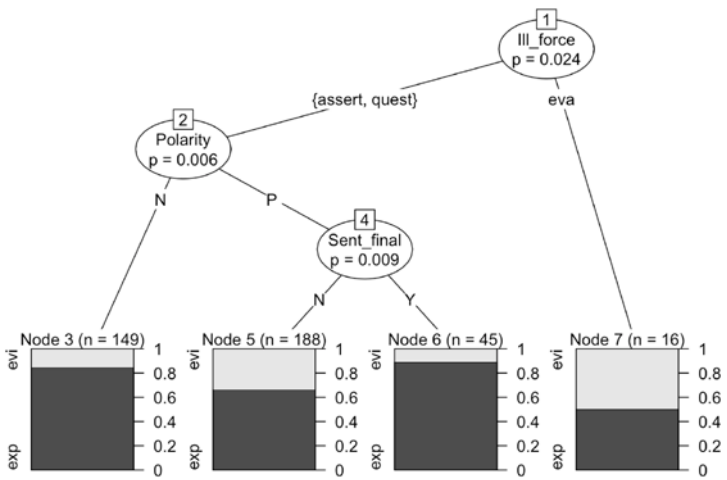


Figure 4 Conditional inference tree IC of evidential *vs* experiential usages of V-过 *guo*

The plot above is obtained with the 'ctree' function of the R package 'party' (Levshina 2015, 291). It is important to emphasise that the tree above has nothing to do with a generative one. Conditional dependencies among variables in figure 4 exclusively depend on statistical significance (the higher the node, the more significant the 'conditional decision') **[fig. 4]**. The descending order of each split computationally simulates a conditional 'decision' made by the speaker/writer based on degrees of significance of each covariant that comes into play when a speech act including experiential or evidential functions of 过 *guo* is realised. In other words, the plot above is completely usage-based and computes holistically probabilities among semantic, pragmatic together with formal variables. The p-value of each 'decision' is reported under each variable before every split (e.g. ill_force p = 0.024 at the top of the tree).

From the above we can see that one interesting IC has to do with illocutionary force being either assertive or interrogative, and the polarity being negative. Convergence of these two features is significantly (p = 0.006) connected to experiential usages of V-过 *guo*.

12. 婉姐, 我没见过什么世面, 啥也不懂。(UCLA / M: Science fiction)

wǎn	*jiě*	*wǒ*	*méi*	*jiàn-**guo***	*shénme*	*shìmiàn*	*shá*
Wan	sister	I	NEG	see-**EXP**	what	life	what
yě	*bù*	*dǒng*					
also	NEG	know					

'Wan, I haven't seen much of life, and I don't know anything'.

In (12) is given a negative assertion of the speaker referring to his/her past experience as a specific persona. This usage is distinctly narrative and occurs in a fictional text (M: Mistery fiction).

Another interesting IC has to do with presence of sentence final particles, which is not preponderant in neither of the two usages, yet still significantly more salient when experiences are narrated or enquired by the speaker/writer.

13. 你在海上清理垃圾的时候, 你想过有一天你会死在这件事上吗? (UCLA / L: Mystery detective fiction)

nǐ	*zài*	*hǎishàng*	*qīnglǐ*	*lājī*	*de*	*shíhòu*	*nǐ*
you	at	sea	clean.up	rubbish	SP	time	you
*xiǎng-**guo***	*yǒu*	*yī*	*tiān*	*nǐ*	*huì*	*sǐ*	*zài*	*zhè*	*jiàn*
think-EXP	have	one	day	you	would	die	at	this	CLF
shì-shàng	*ma*								
thing-on	Q								

'While cleaning up rubbish at sea, have you ever thought about one day you would die because of this?'

As (13) illustrates, experiential usages of the construction tend to occur in dialogic contexts and thus are more likely to be attracted to sentence final particles such as the interrogative 吗 *ma* above. This IC is significantly absent when evidential grounding is at play, as statements are given assertively as reported, potentially verifiable pieces of information. This underpins a clear division of labour between the two functions, one hinging on affective engagement with an animate subject's past experiences, the other being distinctively uttered to mark a proposition as an intersubjectively reliable piece of information. This case study has shed light on the holistic and multimodal factors that concur to the differentiation of the evidential *vs* experiential senses of the V-过 *guo* construction.

What emerged from this analysis is that speakers differentiate experiential and evidential meanings based on the context in which the construction is used, the illocutionary force of the linguistic act, the polarity and the presence of sentence final particles. This entails that meaning disambiguation occurs simultaneously at grammatical, semantic, pragmatic, and situational levels and results from the repeated ascription of a linguistic function to the situation type in which a lexeme is used. The present usage-based analysis of V-过 *guo* is relevant to a broader discussion about linguistic categorisation. In fact, from a usage-based perspective, categorisation is a process that arises as a result of single token instantiations of meaning. What this analysis suggests, is that speakers' ability to identify analogies and similarities among instantiations of meaning cannot be detached from the physical or sociocultural space in which each occurrence takes place. Put simply, context and conventions of usage inform grammatical categorisation. The role of context is thus a crucial one for conceptualisers' ability to establish categories at increasing levels of schematicity and grammatical specialisation. In this sense, the diachronic notion of upward strengthening regards the increased abstraction of a linguistic form leading to the progressive formation of grammatical categories (Hilpert 2015; Tantucci, Di Cristofaro 2019). When the latter reaches highly schematic nodes in a constructional network, it is then possible that context and 'situatedness' become progressively detached from schematic heuristics. This is the case of very abstract schemas such as transitivity, di-transitivity or resultativity, or even aspect or tense, in which conceptualisations of meaning are almost entirely schematic, and not metonymically attached to contextual state of affairs and sociocultural conventions. However, most linguistic functions are the result of a combination of single instantiations and schematic representation, and context does indeed play a crucial role in the speakers' ability to identify and express categorial membership. This is precisely the case of the evidential functions of V-过 *guo* in Mandarin Chinese, as speakers' ability to ascribe the relatively schematic notion of 'shared

knowledge' to the construction is inherently determined by the register and the sociocultural context in which those utterances occur (cf. text types of the kind of J, E, F, B in figure 3). This clearly entails that some degree of entrenchment (e.g. Langacker 1987; Schmid 2017; Tantucci, Culpeper, Di Cristofaro 2018; Tantucci, Di Cristofaro 2019) underpins the recurrent usage of 过 *guo* specifically in connection with text types that allow speakers to infer an evidential meaning rather than an experiential one. In turn, this means that entrenchment as such is also a process that is inherently context-driven and socioculturally situated, and not simply arising as the result of frequent co-occurrence of two or more items independently from contextual situatedness and pragmatic conventions (cf. Terkourafi 2015).

5 Conclusions

In this paper we argued that polysemy and categorial membership cannot be detached from 'contextual situatedness'. While we maintain that, at very high levels of abstraction, sociocultural context does not play a role for the identification of grammatical categories, we also suggest that the progressive formation of those categories is inherently determined by the sociocultural instantiations in which a particular form tends to occur. Entrenchment is therefore experienced as a socioculturally situated phenomenon, and the contextual and co-textual environment where a particular form occurs is a crucial factor for identifying a division of labour among its usages. In this paper we provided a detailed case-study centred on the V-过 *guo* construction in Mandarin Chinese. We showed that a clear division of labour is at stake among experiential and evidential usages of the construction. This categorial separation occurs as a result of features underpinning form, usage and 'contextual situatedness'. Evidentiality in Mandarin is therefore a category that emerges significantly from specific intersections among these three dimensions and from distinctive illocutional concurrences of conventionalised behaviour.

Bibliography

Aikhenvald, A.Y. (2004). *Evidentiality*. Oxford: Oxford University Press.

Arslan, S. et al. (2014). "Finite Verb Inflections for Evidential Categories and Source Identification in Turkish Agrammatic Broca's Aphasia". *Journal of Pragmatics*, 70, 165-81. https://doi.org/10.1016/j.pragma.2014.07.002.

Austin, J.L. (1962). *How to Do Things with Words*. Oxford: Oxford University Press.

Benveniste, E. [1958] (1971). "Subjectivity in Language". Transl. by M.E. Meek. *Problems in General Linguistics*. Coral Gables: University of Miami Press, 223-30.

Brugman, C.M.; Macaulay, M. (2015). "Characterizing Evidentiality". *Linguistic Typology*, 19(2), 201-37. https://doi.org/10.1515/lingty-2015-0007.

Bybee, J.; Perkins, R.; Pagliuca, W. (1994). *The Evolution of Grammar. Tense, Aspect, and Modality in the Languages of the World*. Chicago: University of Chicago Press.

Cao G. 曹广顺 (1995). *Jindai Hanyu Zhuci* 近代汉语助词 (The Auxiliary Particles of Modern Chinese). Beijing: Yuwen chubanshe.

Capone, A. (2005). "Pragmemes (a Study with Reference to English and Italian)". *Journal of Pragmatics*, 37(9), 1355-71. https://doi.org/10.1016/j.pragma.2005.01.013.

Capone, A. (2010). "On the Social Practice of Indirect Reports (Further Advances in the Theory of Pragmemes)". *Journal of Pragmatics*, 42(2), 377-91. https://doi.org/10.1016/j.pragma.2009.06.013.

Carey, K. (1994). *Pragmatics, Subjectivity and the Grammaticalization of the English Perfect* [PhD Dissertation]. San Diego: University of California.

Chao, Y.R. (1968). *A Grammar of Spoken Chinese*. Oakland: University of California Press.

Chappell, H. (2001). "A Typology of Evidential Markers in Sinitic Languages". Chappell, H. (ed.), *Chinese Grammar. Synchronic and Diachronic Perspectives*. New York: Oxford University Press, 56-85.

Chen Q. 陈前瑞 (2008). *Hanyu timao yanjiu de leixingxue shiye* 汉语体貌研究的类型学事业 (A Study on the Aspectual System of Mandarin from a Typological Perspective). Beijing: The Commercial Press.

Comrie, B. (1976). *Aspect*. Cambridge: Cambridge University Press.

Dahl, Ö. (ed.) (2000). *Tense and Aspect in the Languages of Europe*. Berlin: De Gruyter Mouton. https://doi.org/10.1515/9783110197099.

Dahl, Ö.S. (1985). *Tense and Aspect Systems*. Blackwell: Oxford.

Dahl, Ö.S.; Hedin, E. (2000). "Current Relevance and Event Reference". Dahl 2000, 385-402. https://doi.org/10.1515/9783110197099.3.385.

Dai Y. 戴耀晶 (1997). *Xiandai Hanyu shiti xitong yanjiu* 汉语时体系统研究 (A Study of Aspect in Modern Chinese). Hangzou: Zhejiang Educational Press.

Faller, M. (2002). *Semantics and Pragmatics of Evidentials in Cuzco Quechua* [PhD Dissertation]. Stanford (CA): Standford University.

Fludernik, M. (2006). *Towards a 'Natural' Narratology*. New York: Routledge.

Formato, F.; Tantucci, V. (2020). "Uno. A Corpus Linguistic Investigation of Intersubjectivity and Gender". *Journal of Language and Discrimination*, 4(1), 51-73. https://doi.org/10.1558/jld.40129.

Guardamagna, C. (2017). "Reportative Evidentiality, Attribution and Epistemic Modality. A Corpus-Based Diachronic Study of Latin Secundum NP ('According to NP')". *Language Science*s, 59, 159-79. https://doi.org/10.1016/j.langsci.2016.09.001.

Gries, S.T. (2010). "Behavioral Profiles. A Fine-Grained and Quantitative Approach in Corpus-Based Lexical Semantics". *The Mental Lexicon*, 5(3), 323-46. https://doi.org/10.1075/ml.5.3.04gri.

Haugh, M. (2012). "Conversational Interaction". Allan, K.; Jaszczolt, K.M. (eds), *The Cambridge Handbook of Pragmatics*. Cambridge: Cambridge University Press, 251-74. https://doi.org/10.1017/CBO9781139022453.014.

Hilpert, M. (2015). "From *Hand-Carved* to *Computer-Based*. Noun-Participle Compounding and the Upward Strengthening Hypothesis". *Cognitive Linguistics*, 26(1), 113-47. https://doi.org/10.1515/cog-2014-0001.

Hopper, P.J. (1982). "Aspect Before Discourse and Grammar". Hopper, P.J. (ed.), *Tense-Aspect. Between Semantics and Pragmatics*. Amsterdam: John Benjamins, 3-18.

Hornik, K.; Zeileis, A.; Meyer, D. (2006). "The Strucplot Framework. Visualizing Multi-Way Contingency Tables with VCD". *Journal of Statistical Software*, 17(3), 1-48. https://doi.org/10.18637/jss.v017.i03.

Hothorn, T.; Hornik, K.; Zeileis, A. (2006). "Unbiased Recursive Partitioning. A Conditional Inference Framework". *Journal of Computational and Graphical Statistics*, 15(3), 651-74. https://doi.org/10.1198/106186006x133933.

Jarque, M.J.; Pascual, E. (2015). "Direct Discourse Expressing Evidential Values in Catalan Sign Language". *eHumanista. Journal of Iberian Studies*, 8, 421-45. https://www.ehumanista.ucsb.edu/sites/secure.lsit.ucsb.edu.span.d7_eh/files/sitefiles/ivitra/volume8/4.monograficIV/5_JarquePascual.pdf.

Johanson, L. (2000). "Viewpoint Operators in European Languages". Dahl 2000, 27-188. https://doi.org/10.1515/9783110197099.1.27.

Langacker, R.W. (1987). *Foundations of Cognitive Grammar. Theoretical Prerequisites*, vol. 1. Stanford (CA): Stanford University Press.

Langacker, R.W. (2008). *Cognitive Grammar. A Basic Introduction*. Oxford: Oxford University Press.

Levshina, N. (2015). *How to Do Linguistics with R. Data Exploration and Statistical Analysis*. Amsterdam: John Benjamins.

Li, C.; Thompson, S.A. (1981). *Mandarin Chinese. A Functional Reference Grammar*. Berkeley: University of California Press.

Li, D.C.S. (2011). "'Perfective Paradox': A Cross-linguistic Study of the Aspectual Functions of *-guo* in Mandarin Chinese". *Chinese Language and Discourse*, 2(1), 23-57. https://doi.org/10.1075/cld.2.1.02li.

Lin, J.-W. (2006). "Time in a Language without Tense. The Case of Chinese". *Journal of Semantics*, 23, 1-53. https://doi.org/10.1093/jos/ffh033.

Lin, J.-W. (2007). "Predicate Restriction, Discontinuity Property and the Meaning of the Perfective Marker Guo in Mandarin Chinese". *Journal of East Asian Linguistics*, 16(3), 237-57. https://doi.org/10.1007/s10831-007-9013-5.

Lin X. 林新年 (2004). "Shixi Tang Song shiqi de 'guo' yufahua jincheng chihuan de yuanyin" 试析唐宋时期的"过"语法化进程迟缓的原因 (An Analysis of the Slowdown in the Grammaticalisation Process of *guo* During the Tang and Song Periods). *Yuyan Kexue*, 6, 42-52.

Liu J. 刘坚 (2009). "Shitai zhuci de yanjiu yu VO guo" 时态助词的研究与 VO 过 (A Study on Tense Particles and the VO *guo* Construction). Feng, L 冯力; Yang, Y. 杨永龙; Zhao, C. 赵长才 (eds), *Hanyu shiti de lishi yanjiu* 汉语时体的历时研究 (Diachronic Study on the Tense and Aspect System of Chinese). Beijing: Yuwen Chubanshe, 229-34.

McCawley, J.D. (1971). "Tense and Time Reference in English". Fillmore, C.; Langendoen, T. (eds), *Studies in Linguistic Semantics*. New York: Holt, Rinehart and Winston, 96-113.

McEnery, A.; Xiao, Z. (2004). "The Lancaster Corpus of Mandarin Chinese. A Corpus for Monolingual and Contrastive Language Study". *Religion*, 17, 3-4.

Mey, J.L. (2001). *Pragmatics. An Introduction*. 2nd ed. Oxford: Blackwell.

Mey, J.L. (2010). "Reference and the Pragmeme". *Journal of Pragmatics*, 42(11), 2882-8. https://doi.org/10.1016/j.pragma.2010.06.009.

Mushin, I. (2001). *Evidentiality and Epistemological Stance*. Amsterdam: John Benjamins.

Narrog, H. (2009). *Modality in Japanese. The Layered Structure of the Clause and Hierarchies of Functional Categories*, vol. 109. Amsterdam: John Benjamins.

Nenadic, O.; Greenacre, M. (2007). "Correspondence Analysis in R, with Two- and Three-Dimensional Graphics. The ca Package". *Journal of Statistical Software*, 20(3). https://doi.org/10.18637/jss.v020.i03.

Norrick, N.R. (2013a). "Narratives of Vicarious Experience in Conversation". *Language in Society*, 42(4), 385-406. https://doi.org/10.1017/s0047404513000444.

Norrick, N.R. (2013b). "Stories of Vicarious Experience in Speeches by Barack Obama". *Narrative Inquiry*, 23(2), 283-301. https://doi.org/10.1075/ni.23.2.04nor.

Oakes, M.P. (1998). *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University.

Olsen, M.B. (1997). *A Semantic and Pragmatic Model of Lexical and Grammatical Aspect*. New York: Garland Press.

Peterson, T. (2010). *Epistemic Modality and Evidentiality in Gitksan at the Semantics-Pragmatics Interface* [Phd Dissertation]. Vancouver (BC): University of British Columbia.

Plungian, V.A.; van der Auwera, J. (2006). "Towards a Typology of Discontinuous Past Marking". *Language Typology and Universals*, 59(4), 317-49. https://doi.org/10.1524/stuf.2006.59.4.317.

Portner, P. (2003). "The (Temporal) Semantics and (Modal) Pragmatics of the Perfect". *Linguistics and Philosophy*, 26(4), 459-510. https://doi.org/10.1023/A:1024697112760.

Rubovitz, T. (1999). "Evidential-Existentials. The Interaction Between Discourse and Sentence Structure". *Journal of Pragmatics*, 31(8), 1025-40. https://doi.org/10.1016/s0378-2166(99)00038-7.

Rooryck, J. (2001a). "Evidentiality, Part I". *GLOT International*, 5(4), 125-33. https://bit.ly/3qHhaTZ.

Rooryck, J. (2001b). "Evidentiality, Part II". *GLOT international*, 5(4), 161-8. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.461.2545&rep=rep1&type=pdf.

Schmid, H.-J. (2017). "A Framework for Understanding Linguistic Entrenchment and Its Psychological Foundations". Schmid, H.J. (ed.), *Entrenchment and the Psychology of Language Learning. How We Reorganize and Adapt Linguistic Knowledge*. Washington, DC: American Psychological Association; Berlin: Walter de Gruyter, 9-38. https://doi.org/10.1515/9783110341423-002.

Searle, J.R. (1976). "A Classification of Illocutionary Acts". *Language in Society*, 5(1), 1-23. https://sites.duke.edu/conversions/files/2014/09/Searle_Illocutionary-Acts.pdf.

Searle, J. (2010). *Making the Social World: The Structure of Human Civilization*. Oxford: Oxford University Press.

Smith, C. (1997). *The Parameter of Aspect*. 2nd ed. Dordrecht: Kluwer Academic Publishers.

Squartini, M. (2012). "Evidentiality in Interaction. The Concessive Use of the Italian Future between Grammar and Discourse". *Journal of Pragmatics*, 44(15), 2116-28. https://doi.org/10.1016/j.pragma.2012.09.008.

Tagliamonte, S.A.; Baayen, R.H. (2012). "Models, Forests, and Trees of York English. Was/Were Variation as a Case Study for Statistical Practice". *Language Variation and Change*, 24(2), 135-78. https://doi.org/10.1017/s0954394512000129.

Tantucci, V. (2013). "Interpersonal Evidentiality. The Mandarin V-过 *guo* Construction and Other Evidential Systems beyond the 'Source of Information'". *Journal of Pragmatics*, 57, 210-30. https://doi.org/10.1016/j.pragma.2013.08.013.

Tantucci, V. (2015a). "Traversativity and Grammaticalization. The Aktionsart of 过 *guo* as a Lexical Source of Evidentiality". *Chinese Language and Discourse*, 6(1), 57-100. https://doi.org/10.1075/cld.6.1.03tan.

Tantucci, V. (2015b). "Epistemic Inclination and Factualization. A Synchronic and Diachronic Study on the Semantic Gradience of Factuality". *Language and Cognition*, 7(4), 590-90. https://doi.org/10.1017/langcog.2014.34.

Tantucci, V. (2016a). "Textual Factualization. The Phenomenology of Assertive Reformulation and Presupposition during a Speech Event". *Journal of Pragmatics*, 101, 155-71. https://doi.org/10.1016/j.pragma.2016.06.007.

Tantucci, V. (2016b). "Towards a Typology of Constative Speech Acts. Actions beyond Evidentiality, Epistemic Modality and Factuality". *Intercultural Pragmatics*, 13(2), 181-209. https://doi.org/10.1515/ip-2016-0008.

Tantucci, V. (2016c). "The Multimodal Marking of Evidentiality. Pragmemes of Circumstantial Inference and Mandarin Written News Report". *Pragmemes and Theories of Language*, (6), 477-91. https://doi.org/10.1007/978-3-319-43491-9_24.

Tantucci, V. (2017a). "From Immediate to Extended Intersubjectification. A Gradient Approach to Intersubjective Awareness and Semasiological Change". *Language and Cognition*, 9(1), 88-120. https://doi.org/10.1017/langcog.2015.26.

Tantucci, V. (2017b). "An Evolutionary Approach to Semasiological Change. Overt Influence Attempts through the Development of the Mandarin 吧 *-ba* Particle". *Journal of Pragmatics*, 120, 35-53. https://doi.org/10.1016/j.pragma.2017.08.006.

Tantucci, V. (2020). "From Co-actionality to Extended Intersubjectivity. Drawing on Language Change and Ontogenetic Development". *Applied Linguistics*, 41(2), 185-214. https://doi.org/10.1093/applin/amy050.

Tantucci, V. (forthcoming). *Language and Social Minds. The Semantics and Pragmatics of Intersubjectivity*. Cambridge: Cambridge University Press.

Tantucci, V.; Culpeper, J.; Di Cristofaro, M. (2018). "Dynamic Resonance and Social Reciprocity in Language Change. The Case of Good Morrow". *Language Sciences*, 68, 6-21. https://doi.org/10.1016/j.langsci.2017.09.004.

Tantucci, V.; Di Cristofaro, M. (2019). "Entrenchment Inhibition. Constructional Change and Repetitive Behaviour Can Be in Competition with Large-Scale 'Recompositional' Creativity". *Corpus Linguistics and Linguistic Theory*, 16(3), 547-79. https://doi.org/10.1515/cllt-2019-0017.

Tantucci, V.; Wang, A. (2018). "Illocutional Concurrences. The Case of Evaluative Speech Acts and Face-Work in Spoken Mandarin and American English". *Journal of Pragmatics*, 138, 60-76. https://doi.org/10.1016/j.pragma.2018.09.014.

Tantucci, V.; Wang, A. (2020a). "Diachronic Change of Rapport Orientation and Sentence-Periphery in Mandarin". *Discourse Studies*, 22(2), 146-73. https://doi.org/10.1177/1461445619893777.

Tantucci, V.; Wang, A. (2020b). "From Co-actions to Intersubjectivity Throughout Chinese Ontogeny. A Usage-Based Analysis of Knowledge Ascription and Expected Agreement". *Journal of Pragmatics*, 167, 98-115. https://doi.org/10.1016/j.pragma.2020.05.011.

Tao, H.; Xiao, R. (2012). *The UCLA Chinese Corpus*. 2nd ed. Lancaster: University Centre for Computer Corpus Research on Language.

Terkourafi, M. (2015). "Conventionalization. A New Agenda for Im/politeness Research". *Journal of Pragmatics*, 86, 11-18. https://doi.org/10.1016/j.pragma.2015.06.004.

Traugott, E.C. (1999). "From Subjectification to Intersubjectification". Paper presented at the *Fourteenth International Conference on Historical Linguistics* (Vancouver, 9-13th August 1999).

Traugott, E.C. (2003). "From Subjectification to Intersubjectification". Hickey, R. (ed.), *Motives for Language Change*. Cambridge: Cambridge University Press, 124-39. https://doi.org/10.1017/cbo9780511486937.009.

Traugott, E.C. (2012). "Intersubjectification and Clause Periphery". *English Text Construction*, 5(1), 7-28. https://doi.org/10.1075/bct.65.02trau.

Traugott, E.C. (2016). "On the Rise of Types of Clause-Final Pragmatic Markers in English". *Journal of Historical Pragmatics*, 17(1), 26-54. https://doi.org/10.1075/jhp.17.1.02tra.

Traugott, E.C.; Dasher, R.B. (2002). *Regularity in Semantic Change*. Cambridge: Cambridge University Press.

Van Olmen, D. (2019). "A Diachronic Corpus Study of Prenominal zo'n 'so a' in Dutch. Pathways and (Inter) Subjectification". *Functions of Language*, 26(2), 216-47. https://doi.org/10.1075/fol.16017.van.

Vendler, Z. (1967). *Linguistics in Philosophy*. Ithaca (NY): Cornell University Press.

Yeh, M. (1996). "An Analysis of the Experiential *guo*$_{exp}$ in Mandarin: A Temporal Quantifier". *Journal of East Asian Linguistics*, 5, 183-215. https://doi.org/10.1007/BF00215072.

Xiao, R.; McEnery, T. (2004). *Aspect in Mandarin Chinese. A Corpus-Based Study*. Amsterdam: John Benjamins.

Wu, J.-S. (2008). "Terminability, Wholeness and Semantics of Experiential *Guo*". *Journal of East Asian Linguistics*, 17(1), 1-32. https://doi.org/10.1007/s10831-007-9018-0.

Semantics

Corpus-Based Research on Chinese Language and Linguistics
edited by Bianca Basciano, Franco Gatti, Anna Morbiato

Manual Action Metaphors in Chinese
A Usage-Based Constructionist Study

Heidi Hui Shi
University of Oregon, USA

Sophia Xiaoyu Liu
University of Oregon, USA

Zhuo Jing-Schmidt
University of Oregon, USA

Abstract This article examines Chinese manual motor metaphors involving manual object manipulation as the source domain. Specifically, we use corpus data to investigate two transitive constructions, [抓紧 *zhuājǐn* 'grab tightly, clutch' NP] and [把住 *bǎzhù* 'grasp firmly' NP], and a causative construction, [把 *bǎ* NP 捧 *pěng* COMPL] 'lift NP with deliberation', where the referent of the NP does not lend itself to manual manipulation in the literal sense and must be interpreted as metaphoric in the unity of semantic domains. Results from both quantitative and qualitative analyses show that the two transitive grasping actions are systematically used to abstract actions requiring a keen sense of urgency and/or importance, and that the causative action of lifting systematically conceptualises over-promotion of an undeserving entity. The findings point to the bodily origin of social cognition and the embodiment of conceptualisation.

Keywords Manual Motor Metaphor. Object Manipulation. Embodiment. Chinese.

Edizioni Ca'Foscari

Sinica venetiana 6
e-ISSN 2610-9042 | ISSN 2610-9654
ISBN [ebook] 978-88-6969-406-6 | ISBN [print] 978-88-6969-407-3

Peer review | Open access
Submitted 2020-02-17 | Accepted 2020-04-09 | Published 2020-12-21

DOI 10.30687/978-88-6969-406-6/004

1 Introduction[1]

Metaphor is not just a phenomenon of language. It is a way of knowledge representation. This idea was articulated by Jakobson as early as 1956 (Jakobson [1956] 2003) and was subsequently elaborated by Lakoff and Johnson (1980) in a systematic and theoretically significant way that gave rise to the Conceptual Metaphor Theory (CMT). The essence of CMT in terms of experientialism or the bodily basis of abstract thought is now the consensus on metaphor as a cognitive phenomenon, supported by research over the last three decades in cognitive linguistics and cognitive science. More recent work on the relationship between conceptualisation and sensory perception has further consolidated the notion of embodiment understood as the grounding of conceptualisation in physical and perceptual experiences (Johnson 2017; Barsalou 1999, 2008; Gibbs 2006; Gallese, Lakoff 2005).

Manual object manipulation requires the coordinated use of the hands and the arms as the effectors of action. As a tool-using species, humans have evolved extraordinary manual dexterity and sophisticated skills of manual praxis (Darwin 1871). There is accumulating evidence that human manual praxis is closely related to the evolution of the human brain and the development of vocal language (Bradshaw 1991; Gibson, Ingold 1993; Steele, Ferrari, Fogassi 2012). Iriki and Taoka (2012) attribute the development of abstract cognitive functions in humans to cortical plasticity that enabled the recruitment of cortical areas originally involved in computing sensorimotor transformations for reaching and grasping actions to serve higher cognitive functions, including language.

The evolutionary significance of manual object manipulation leaves stamps on languages. To get a sense of the conceptual reach of manual actions in language, one need to look no further than the vocabulary of English. The verb *hold* is one of the most polysemous verbs in English, with over two dozen essentially metaphoric senses ranging from 'control' to 'sustain' to 'continue', all derived from the basic manual meaning of "grasp, carry, or support with one's arms or hands" (`www.dictionary.com`) and used in a rich array of phraseological configurations. Similarly, we use *grasp* metaphorically when talking about *grasping* an idea or concept. These examples have counterparts in other languages. Germans speak of *eine Idee begreifen* 'to comprehend an idea' whereby *begreifen* is a complex verb derived from the manual action verb stem *greifen* 'grasp'. In fact, the abstract noun *Begriff* 'concept' itself is derived from the same verb denoting grasping.

1 The glosses follow the general guidelines of the Leipzig Glossing Rules. Additional glosses include: ASSOC = 'associative'; OM = 'object marker'.
Further in-text abbreviations include: COMPL = 'complement'; NP = 'noun phrase'.

Another German compound verb, *ergreifen*, which also features the manual action verb stem *greifen* 'grab', frequently collocates with *eine Chance* 'a chance, an opportunity'. Similarly, in Korean, the manual action verb 잡다 *jabda* 'hold, grasp, catch' can be used metaphorically in collocation with the abstract noun 기회를 *gihoeleul* 'opportunity'.

Neuroimaging studies in cognitive neuroscience provide evidence that brain regions of sensory and motor perception are activated when participants read metaphors with sensory motor actions as source domain. Desai et al. (2011) compared neural responses to descriptions of literal action (e.g. *grasped the flowers*), metaphoric action (e.g. *grasped the concept*), and abstract mental action (e.g. *understood the concept*). They found that sentences describing literal and metaphoric actions but not abstract actions activated motor regions involved in action planning. In particular, metaphoric action sentences recruited secondary sensory-motor regions and less familiar action metaphors engaged primary motor regions, suggesting a role of metaphor conventionality in motor activation. Boulenger, Shtyrov and Pulvermüller (2012) conducted a MEG study on the time-course of cortical motor activation during the comprehension of literal and figurative sentences involving arm and leg action verbs. They reported early motor activations to both figurative and literal action sentences whereby arm action verbs (*scrape*, *pick*, and *catch*) more reliably recruited the corresponding motor region than leg action verbs (*kick*, *walk*, and *jump*). In a subsequent fMRI study that aimed to clarify how the extent to which the figurative stimuli are conventionalised influences sensory-motor activation, Desai et al. (2013) also included idiomatic action sentences with conventionalised action metaphors, comparing four experimental conditions involving the verbs *grasp* and *lift*: (1) literal (e.g. *grasping the steering wheel very tightly/lifted the pebble from the ground*), (2) metaphorical (e.g. *grasping the state of the affairs/lifted this nation out of poverty*), (3) idiomatic (e.g. *grasping at straws in the crisis/lifted the veil on its nuclear program*), and (4) abstract as control (e.g. *causing a big trade deficit/wanted the plan for a nuclear program*). Their results showed a trend of decreasing sensory-motor activation from literal to metaphoric to idiomatic to abstract action sentences. Similarly, Romero Lauro et al. (2013) conducted an fMRI study of literal, metaphoric, and idiomatic action sentences in Italian, with abstract mental action sentences as a control condition. They found that the degree of cortical motor activation was a function of the degree of perceived concreteness of the motor action, a result consistent with Desai et al. (2013). Interestingly, their results also indicated a stronger motor activation effect for arm actions than leg actions, converging with Boulenger, Shtyrov and Pulvermüller (2012). The authors interpreted this effect as consistent with the perception that arm motions are more concrete and specific than leg motions.

These neurolinguistic studies show that the motor system facilitates the processing of linguistic representations of motor actions, including metaphorical motor actions, albeit with reduced effect of activation correlating with a higher degree of conventionality. What stands out from these studies is the prominence of motor actions involving the hand/arm in the way their linguistic representations trigger activations of cortical motor regions. This comes as no surprise given the fundamental role of primate tool use in the co-evolution of the human brain and language (Steele, Ferrari, Fogassi 2012).

The Chinese lexicon has been shown to lexicalise abstract experiences based on manual action effectors including the hand, the palm, and the finger as metaphoric and metonymic sources. For example, Yu (2003) discussed the extensive presence of 手 *shǒu* 'hand' not only in compound nouns that refer to aptitude, means, manners, and people, but also in compound verbs that describe operations, transactions etc. by way of metaphor and metonymy. Yu (2000) showed how Chinese compounds and idioms involving the morphemes 指 *zhǐ* 'finger' and 掌 *zhǎng* 'palm' that conceptualise abstract experiences are grounded in the acts of pointing and holding. Specifically, 'finger' is involved in verbs of abstract actions such as demonstrating and designating, while 'palm' is found in compound verbs denoting control. Gao (2001) offers a broader coverage of the bodily foundation of physical action verbs in Chinese. While not directly focusing on the metaphoric uses of action verbs, Gao argues that the semantic patterning of action verbs mirrors the anatomical limitations of the body parts employed in executing the actions, which has implications for the embodiment of conceptualisation. These studies shed light on the role of body parts in the metaphorical and metonymical conceptualisation of abstract experiences in Chinese. What remains largely unexplored, but equally intriguing, is how manual actions as a basic experiential domain contribute to the conceptualisation of abstract actions and behaviours.

The present study goes above and beyond lexical semantics and takes a usage-based constructionist approach to metaphor analysis. This approach is grounded in the theoretical and methodological integration of Construction Grammar and usage-based linguistics. Construction Grammar treats language as a structured inventory of constructions, which are form-meaning pairings that occupy a continuum from morphemes and lexical units, over phrasal constructions, partially schematic constructions, to fully abstract argument structure constructions and discourse units (Fillmore 1988; Fillmore, Kay, O'Connor 1988; Goldberg 1995, 2006, 2019; Croft 2001). This view effectively blurs the boundary between lexicon and syntax and allows for the accounting of linguistic knowledge in its entirety (Goldberg 2013; Hilpert 2014). Usage-based linguistics views language as emergent from experiences with language use and generalisations over re-

current usage events (Barlow, Kemmer 2000; Tomasello 2003; Bybee 2013). On this approach, linguistic knowledge comprises a vast storage of both specific exemplars and abstract patterns in a linked network whereby frequency of use plays a central role in the representation of linguistic knowledge (Bybee 2006; Ellis 2002, 2013; Gries 2012; Goldberg 2019). The usage-based constructionist approach is optimally suited for the analysis of metaphors if our goal is to explore patterns of conceptual mapping and the prototypes and productivity of those patterns in a systematic way. In particular, Croft pointed out that the syntactic construction is the structural site of metaphorical meaning, which can be identified only by way of the "conceptual unity of domains", in the sense that "all of the elements in a syntactic unit must be interpreted in a single domain" (Croft 2003, 162). Recent research shows systematic lexical grammatical alignments in metaphorical expressions, systematic correspondences between grammatical dependency within a metaphorical construction, and source-target dependency in metaphorical mapping (Lederer 2019; Sullivan 2013, 2016).

In this study, we examine verbal constructions that encode metaphorical manual object manipulation. We aim to understand the semantic categories of the metaphorical objects collocating with the metaphorical hand actions described by these constructions, as well as the productivity of their uses as manual object manipulation metaphors. One of the constructions in question is [把 *bǎ* NP 捧 *pěng* COMPL] 'lift NP with deliberation', such that NP undergoes change of location or state, which is a type of the 把 *bǎ*-construction that dramatises how a definite object undergoes change as a result of the action described by the verb (Jing-Schmidt 2005). The lifting action is described by 捧 *pěng* 'lift with deliberation on the joint surfaces of both palms'. This verb encodes the deliberate manner of lifting, the spatial configuration of the manual effectors, and implies an undeserved assignment of value to the object being lifted (Jing-Schmidt 2010). Consider (1) as an example:

1. 绝不要一高兴起来就把孩子捧上了天
   juébúyào yī gāoxìng-qǐlái jiù bǎ háizi pěng-shàng le tiān
   never once happy-up then OM child lift-up PFV sky
   'Don't worship the child just because all of a sudden you are in a good mood'

In this example, the description of lifting the child to the sky is not meant to be literal. We can tell this from the conceptual contradiction between the physical domain of lifting a child and the domain of location change described by the postverbal complement 上了天

shàng le tiān 'up to the sky'. Following Croft (2003), the lifting action involving a child as object and the location change as a result of the action must be interpreted in a unity of the two domains where lifting someone up to the sky hyperbolically conceptualises the act of worshiping or overpraising.

The other constructions included in this analysis are two transitive constructions that involve object grasping/grabbing as the experiential basis on which to conceptualise abstract experiences with intangible objects. They are [抓紧 *zhuājǐn* 'grab tightly' NP] and [把住 *bǎzhù* 'grasp firmly' NP], each with a compound verb describing a grasping motion and a resultative morpheme describing the tightness of the grip. Because of their similarity in surface lexical semantics, the two manual action verbs may come across as synonyms. However, as our usage-based constructionist analysis will reveal, the semantic categories of the metaphorical objects in the respective constructions are very different.

2 Data and Methods

The corpus data were retrieved from the online BCC corpus (Xun et al. 2016). We used the search syntax 把* 捧* in the balanced subcorpus (多领域 *duō lǐngyù*) to maximally extract all uses of construction [把 *bǎ* NP 捧 *pěng* COMPL]. The asterisk designates any structure of unspecified size that occurs in the respective slots of NP and COMPL in [把 *bǎ* NP 捧 *pěng* COMPL]. A total of 1,667 concordances were obtained from the initial search. Two coders conducted independent annotations of this sample to identify metaphorical uses by eliminating (1) syntactic false positives and (2) semantic false positives. Syntactic false positives contained the target lexemes 把 *bǎ* and 捧 *pěng*, but did not match the structural requirement of the 把 *bǎ*-construction, such as 一把一把地捧了出去 *yībǎyībǎ-de pěng-le chūqù* 'lift and put outside by the handful', where 把 *bǎ* is used as a measure word (handful). Semantic false positives are those sentences that meet the structural requirement but describe physical, and therefore not metaphorical, lifting such as 把餐具捧上来 *bǎ cānjù pěng-shànglái* 'hold the utensils in both hands and bring them up here'. A total of 736 false positives were removed and a total of 931 tokens of the metaphorical uses were obtained. To retrieve tokens of the transitive construction [抓紧 *zhuājǐn* 'grab tightly, clutch' NP], we searched for "抓紧n" to extract concordances with the object noun immediately following the verb, and the research returned 8,022 tokens. Two of the authors conducted independent manual annotations to identify metaphorical uses by removing (a) items that describe physical grasping of objects by hand such as 缰绳 *jiāngshéng* 'bridle' and (b) syntactically labile words that are tagged in the wrong parts of speech

in the corpus, such as 移民 *yímín* 'emigrate'. After removal of a total of 693 false positives, 7,335 tokens remained, out of which 1,000 tokens were selected as a sample for the analysis. The same search process was conducted for the construction [把住 *bǎzhù* 'grasp firmly' NP] and a total of 655 concordances were retrieved. Independent manual annotations by two of the authors removed 143 false positives that describe physical grasping of objects by hand, such as 舵 *duò* 'rudder' and 方向盘 *fāngxiàngpán* 'steering wheel'. A total of 512 metaphorical uses were retained for the analysis.

Both quantitative and qualitative analyses were adopted in this study. The quantitative analyses focused on measuring the productivity of the three constructions. One way to measure productivity is to count the type frequency of the open slot(s) in a construction. Type frequency is the "number of distinct lexical items that can be substituted in a given slot in a construction" (Ellis 2002, 166). It has been argued that high type frequency in the input facilitates the formation of a schematic pattern and productive expansion of the pattern to novel uses (Goldberg 1995; Bybee 2006; Ellis 2011). In fact, Goldberg's (2006, 5) definition of 'construction' has evolved to include "sufficient frequency" of use as an independent criterion of constructionhood. Gries (2012, 505) considers the skewness of the type-token distributions with a Zipfian power tendency as a way to 'operationalise' Goldberg's notion of sufficient frequency. Following this proposal, we analysed rank-frequency distributions of the open slot(s) in each construction to identify skewness as a measure of productivity. Quantitative data processing, analysis, and graphing was conducted in R (3.6.2) and R-studio (1.2.5033) with the additional software packages *stringr*, *qdapRegex*, *dplyr*, and *fs*.

The qualitative analysis aimed to investigate the mutual selection of the verb and the open object and/or complement slot(s) in each of the constructions with a focus on identifying the semantic subclasses of these open slot(s) based on the patterns identified in the quantitative analysis. This focus was informed by the theoretical insight that semantic coverage plays a role in providing confidence in generating new instances in language use (Osherson et al. 1990; Goldberg 2006). From a usage-based perspective, semantic subclasses are generalisations over usage events at the level of knowledge representation. Similar items used in an open slot of the same construction "are classified together by general categorization processes" and novel items are used based on perceived similarity to members of existing clusters (Goldberg 1995, 133).

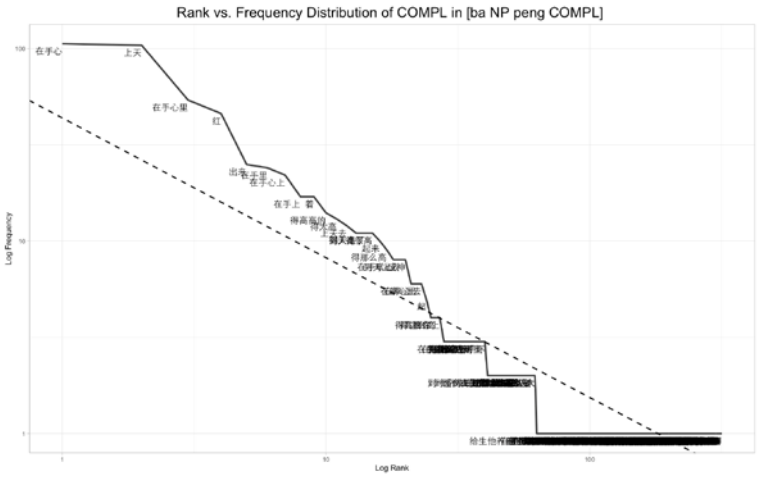


Figure 1 Power Distribution of COMPL in [把 *bǎ* NP 捧 *pěng* COMPL]

3 Results

3.1 The Construction [把 *bǎ* NP 捧 *pěng* COMPL]

The identification of 931 metaphor uses from the retrieved 1,667 concordances of [把 *bǎ* NP 捧 *pěng* COMPL] yielded a better than chance probability (55%) for this construction to be used metaphorically. This tendency finds further confirmation in the productivity of the metaphor uses measured by the type frequencies of the NP and the COMPL, as well as their frequency distribution patterns. The 931 tokens fall into 349 types of NP and 317 types of COMPL. Apart from the high type frequencies of the NP and the COMPL slots, the distributions in both slots show high skewedness. The rank-frequency distributions of the nouns in the NP slot, as shown in figure 1, and the complements in the COMPL slot, as shown in figure 2, display Zipfian skewedness characterised by an entropy-reducing spike with a long tail of low-frequency types. Specifically, the top five most frequent types of NP, which is slightly over 1% of all the 349 types, make up 48% of the entire dataset of 931 tokens. By contrast, 312 (89%) of the 349 types are *hapax legomena*, i.e. items that occur only once in the data. These *hapax legomena* cluster into a dark long tail at the bottom of the frequency rank in figure 1 **[fig. 1]**. Similarly, the five top-ranked types of COMPL, which is 1.5% of all 317 types, make up 37% of the data whereas 255 (80%) of the 317 types are *hapax legomena* forming the dark long tail at the bottom of the frequency rank in fig-

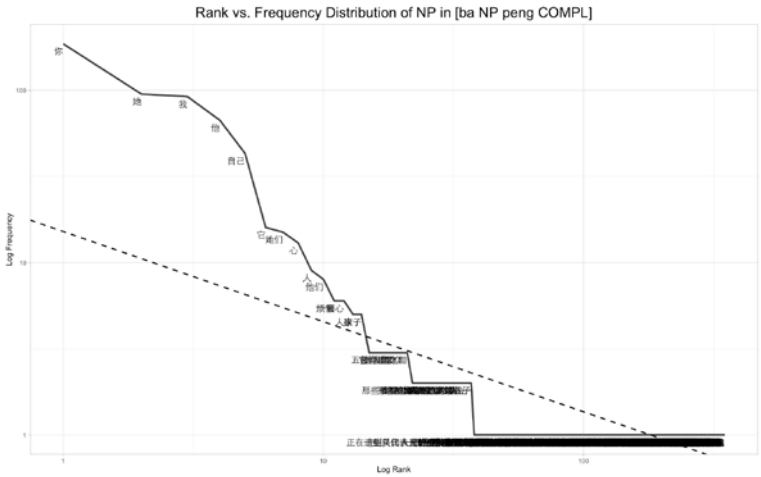


Figure 2 Power Distribution of NP in [把 *bǎ* NP 捧 *pěng* COMPL]

ure 2 **[fig. 2]**. The Zipfian certainty and reduced entropy as seen in the rank-frequency distributions suggest that the NP and the COMPL slots are productive and can readily admit new items. Together, the high type frequencies and the Zipfian power distributions of the open NP and COMPL slots in [把 *bǎ* NP 捧 *pěng* COMPL] demonstrate the productivity of the metaphorical uses of this construction.

Turning now to the semantic subclasses in the NP slot, we found that 267 (77%) of the NP types are human referents, which make a total of 819 (88%) of the 931 tokens. The top five most frequent items are all personal pronouns: 你 *nǐ* 'you', 她 *tā* 'she/her', 我 *wǒ* 'I/me', 他 *tā* 'he/him', and 自己 *zìjǐ* 'self'. The non-human nouns that make up 12% of the dataset refer to human-made cultural products such as literary works, movies, music etc., and abstract concepts such as human behaviours, experiences, accomplishments, performances, ideas, technology etc., all of which are human-generated. As such the objects of the metaphorical action described by the construction [把 *bǎ* NP 捧 *pěng* COMPL] cannot be literally lifted by hand. By the "unity of domains" in Croft's (2003) terms, the manual action of 捧 *pěng* together with its complement (COMPL) that describes change must be interpreted metaphorically.

Our analysis of the semantic subclasses in the COMPL slot employed the major categories identified for the 把 *bǎ*-construction in Jing-Schmidt, Peng and Chen (2015, 120). These are (i) locative encoding change of absolute location, (ii) directional encoding change of spatial orientation, (iii) resultative encoding change of state and

(iv) metamorphic describing change of identity or appearance. Among these, the locative is the most productive subclass with a type frequency of 106, or 33% of all the distinct types of complement in the data. The most frequently used tokens in the locative type are 在手心 *zài shǒuxīn* 'in the centre of the palm' and 上天 *shàngtiān* 'up to the sky'. The former accentuates the perceived value of a cherished object, as in (2). The latter emphasises the degree of admiration afforded an object of perceived value by way of the hyperbolic use of a spatial metaphor, UP IS GOOD, an example of which is (1) discussed in the previous section. The resultative is the second most productive subclass with 90 different types, or 28% of the total COMPL types. For example, the resultative 红 *hóng* 'red, hot, popular' in (3) features a colour metaphor of popularity. The metamorphic complement in the form of 成 *chéng*/为 *wéi* NP 'become/turn into NP' is the third most productive subclass with 73 different types, or 23% of the total COMPL types. As illustrated in (4), the complement 成一个神 *chéng yí-gè shén* 'become a deity' describes the perceived excess with which honour and praise are afforded the person in question. A close English translation would be 'put someone up on a pedestal', which itself is a metaphor of uncritical worship.

2. 把烦恼当宝一样捧在手心
   bǎ *fánnǎo* *dāng* *bǎo* *yíyàng* *pěng* *zài* *shǒu-xīn*
   OM distress as treasure same lift in hand-centre
   'Hold on to one's distress like treasure'

3. 我们一定会尽人事, 把你捧红
   wǒmen *yídìng* *huì* *jìn* *rénshì* *bǎ* *nǐ*
   1PL certainly will exhaust human.affair OM 2SG
   pěng-hóng
   lift-red
   'We will certainly do everything we can to make you popular'

4. 把任长霞捧成一个神
   bǎ *Rén* *Chángxiá* *pěng-chéng* *yí-gè* *shén*
   OM Ren Changxia lift-become one-CLF deity
   'Put Ren Changxia up on a pedestal'

In general, the construction [把 *bǎ* NP 捧 *pěng* COMPL] represents a systematic and productive conceptual mapping from LIFTING NP WITH DELIBERATION to WORSHIPPING OR CHERISHING NP whereby NP refers to a person or an abstract entity associated with a person.

3.2 The Construction [抓紧 *zhuājǐn* 'grab tightly, clutch' NP]

The metaphorical uses of this construction make up 91% of the entire sample of 1,000 tokens. This is strong evidence that [抓紧 *zhuājǐn* 'grab tightly, clutch' NP] is much more productive in its metaphorical sense than in its literal sense. Its productivity as metaphor can also be seen in the type frequency of NP and its distributions. Specifically, a total of 196 types of NP were identified in the 1,000 tokens. Notably, as shown in figure 3, the top three items make up nearly 80% of the dataset whereby the top-ranked item 时间 *shíjiān* 'time' takes the lion's share, forming an entropy-reducing spike with 70% of the entire dataset **[fig. 3]**. On the other hand, 84% of all the 196 types form a long tail of *hapax legomena*. This is a highly skewed distribution pattern that fits a Zipfian power law, suggesting that the construction [抓紧 *zhuājǐn* 'grab tightly, clutch' NP] is highly productive in its metaphorical use.

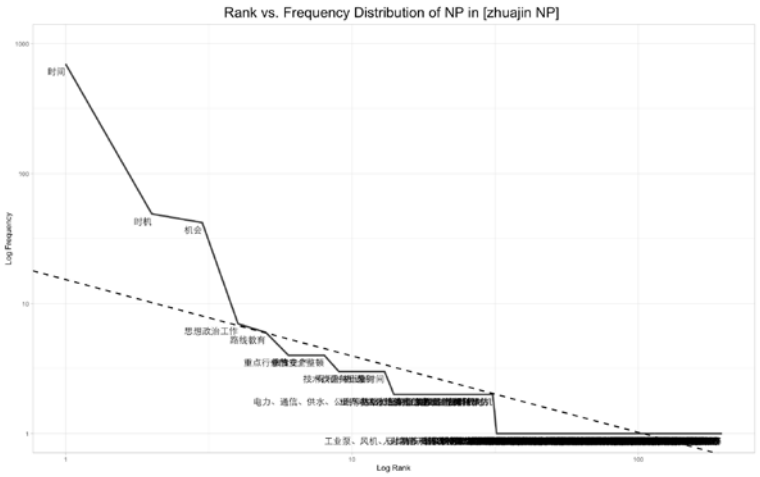


Figure 3 Power Distributions of NP in [抓紧 *zhuājǐn* NP]

In terms of the semantic subclasses of the NP, two observations can be made. First, the concept of time or timing stands out as the dominant semantic subclass. In addition to the top-ranked type 时间 *shíjiān* 'time', there are 17 time-related types referring to units of time, such as 分分秒秒 *fēnfēn miǎomiǎo* 'minutes and seconds' and 每一天 *měi yī tiān* 'every day'. There are 11 types referring to opportunity, which is defined in terms of timing and the perceived possibility it holds. Both the second and third ranked nouns, 时机 *shíjī* 'opportunity' and 机会 *jīhuì* 'opportunity, chance', belong to this subclass. Second, all the other abstract nouns form a semantic cluster

that can be characterised as referring to tasks or activities of perceived importance and urgency, such as 建设 *jiànshè* 'construction', 改造 *gǎizào* 'reform', 生产 *shēngchǎn* 'production', 训练 *xùnliàn* 'training', 工作 *gōngzuò* 'work', and 教育 *jiàoyù* 'education', most of which are deverbal nominals. Examples of these usages are in (5)-(7):

5. 抓紧时间, 学习, 学习, 再学习
   zhuājǐn *shíjiān* *xuéxí* *xuéxí* *zài* *xuéxí*
   grab.tightly time study study again study
   'Hurry up, study, study, and study some more'

6. 要提高, 就要抓紧一切机会学习
   yào *tígāo* *jiù* *yào* *zhuājǐn* *yíqiè*
   want improve then must grab.tightly all
   jīhuì *xuéxí*
   opportunity study
   'If we want to improve, we must grab every opportunity to study'.

7. 抓紧党内的思想教育
   zhuājǐn *dǎng-nèi-de* *sīxiǎng* *jiàoyù*
   grab.tightly party-inside-ASSOC thought education
   'Act urgently on thought education within the Party'

From the analysis of the semantic subclasses, it is obvious that the manual object manipulation metaphor [抓紧 *zhuājǐn* 'grab tightly, clutch' NP] profiles conceptually intangible entities such as time, opportunities, and priorities as moving physical objects that may escape our grip unless grabbed tightly. On the other hand, it makes sense to grab something that is precious but does not often come along. Therefore, it is reasonable to suggest that the ACTING WITH URGENCY AS GRABBING metaphor, especially the subclass that profiles time and opportunity as objects, invokes two ontological metaphors: TIME AS A MOVING OBJECT and TIME AS A COMMODITY, as discussed in Lakoff and Johnson (1980).

3.3 The Construction [把住 *bǎzhù* 'grasp firmly' NP]

The fact that 512 (78%) out of a total of 655 tokens of [抓紧 *zhuājǐn* 'grab tightly, clutch' NP] retrieved from the corpus are metaphorical suggests the productivity of the construction as a conventional metaphor. Again, this productivity is further confirmed by the type frequency of NP and its type-token frequency distributions. The 512 concordances fall into 272 types. As can be seen in figure 4, the ranked frequencies of the NP fit a Zipfian distribution **[fig. 4]**. The top ranked four types make up 34% of the entire dataset, whereby the item in the highest rank, 质量关 *zhìliàngguān* 'quality control checkpoint', is

more than twice as frequent as the second ranked type, 关口 *guānkǒu* 'checkpoint, control', whereas the overwhelming majority (86%) of all the types are *hapax legomena* that cluster into a dark long tail at the bottom of the frequency rank. It is obvious that the construction is productively used in its metaphorical sense.

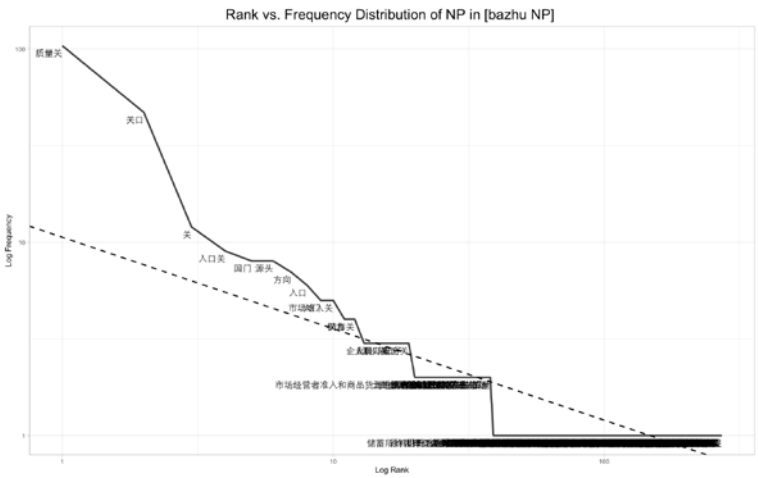


Figure 4 Power Distributions of NP in [把住 *bǎzhù* NP]

Semantically, the NP slot displays a strong preference for nouns that essentially signal control. The primary subclass is metaphorically represented by lexemes such as 关 *guān* 'checkpoint, control', 关口 *guānkǒu* 'checkpoint', and 入口 *rùkǒu* 'entrance' that refer to checkpoint and entrance where tight control is exercised. A related subclass consists of abstract nouns the referents of which are deemed central to organisational policy and are therefore necessary to be kept under control, such as 权力 *quánlì* 'power', 大局 *dàjú* 'overall situation', 方向 *fāngxiàng* 'direction' etc. Underlying all these uses is the GRASPING AS CONTROLLING metaphor, examples of which are shown in (8)-(9):

8. 帮助饲料企业把住质量关
bāngzhù sìliào qǐyè bǎzhù zhìliàng-guān
help fertiliser company grasp.firmly quality-control
'Help the fertiliser companies to perform firm quality control'.

9. 一些部门把住权力不放
yìxiē bùmén bǎzhù quánlì bú fàng
some sector grasp.firmly power not release
'Some sectors hold on to power and won't let go'.

This GRASPING AS CONTROLLING metaphor is similar to the English idiomatic expression 'to get a (firm) grip on something' that conveys the abstract idea of taking control of something, as in *get a grip on your finances*. The concept of 'taking control' is motivated by and embodied in our physical experience with the functions of the hand as a neuromuscular system of controlling manual motions and forces for automatic object manipulation.

4 Discussion

Taylor and Schwarz noted that "the human hand represents a mechanism of the most intricate fashioning and one of great complexity and utility" (1955, 22). It goes without saying that the hand as an automatic system that governs the motions and forces of manual actions is instrumental to human evolution and individual development (Steele, Ferrari, Fogassi 2012). While the role of Chinese manual body part concepts (e.g. 手 *shǒu* 'hand', 掌 *zhǎng* 'palm', and 指 *zhǐ* 'finger') in lexical semantic representations of abstract human experiences is well documented, manual object manipulation actions have been largely off the radar of Chinese metaphor research. This corpus-based study filled the gap. We demonstrated that the three manual actions LIFTING WITH DELIBERATION, TIGHTLY GRABBING/CLUTCHING, and GRASPING FIRMLY specialise in systematic metaphorical representations of the respective abstract domains of human experience: OVERPRAISING OR WORSHIPPING, ACTING WITH URGENCY, and CONTROLLING. In other words, these metaphors draw on manual motor actions as the sensory motor basis of abstract cognition. The three manual action constructions are not only conventionalised, they are productive in their metaphorical usages and can readily admit new items into their open slots. These results add to the existing and accumulating evidence of embodied conceptualisation, namely that language concepts are rooted in sensory perceptions and motor actions (Barsalou 1999, 2008; Gallese, Lakoff 2005; Glenberg, Kaschak 2002; Grush 2004; Pecher, Zwaan 2005; Simmons et al. 2007; van Dantzig et al. 2008; Kiefer et al. 2008).

Our results are also significant from a crosslinguistic perspective. On the one hand, the findings revealed mapping patterns that have been observed across languages. For example, 'opportunity' as a metaphorical object of grabbing is common across languages, as noted in the Introduction. On the other hand, convergence in conceptual mapping is often partial if not superficial. As we have pointed out previously, [把住 *bǎzhù* NP] 'hold fast, grasp firmly' is reminiscent of *get a grip on something* in English. Yet the Chinese metaphor clearly attracts nouns referring to matters related to organisational policy rather than personal affairs, which cannot be said of its putative Eng-

lish counterpart. Similarly, although both Chinese and English utilise lifting metaphors to conceptualise uncritical praising and admiring, they draw on different conceptual resources. The Chinese metaphor [把 *bǎ* NP 捧 *pěng* COMPL] employs what Rüschemeyer, Pfeiffer and Bekkering (2010) call a "body schema", with specifications of hand posture and spatial configuration, whereas English *put someone on the pedestal* relies on our encyclopedic knowledge of 'pedestal' as the central element of a culturally motivated imagery as the metaphor source domain. Following from this discussion, the notion of embodiment as a universal cognitive mechanism shall be understood as going hand in hand with, and as being under the influence of, experiences specific to social groups and communities that bear the stamp of culture (Gibbs 1999; Kövecses 2005).

Finally, previous research indicates the flexibility and contextual dependency of embodied representations in the sense that neural activations are relative and non-automatic (e.g. Rüschemeyer, Brass, Friederici 2007; Boulenger, Shtyrov, Pulvermüller 2012; Van Dam et al. 2012). Our results on the proportion of the metaphorical uses in the sample of data on each construction indicate a gradation of conventionality: 55% of [把 *bǎ* NP 捧 *pěng* COMPL], 78% of [把住 *bǎzhù* NP], and 91% of [抓紧 *zhuājǐn* NP] are metaphorical. Will these metaphors vary in their ability to trigger sensorimotor brain areas as a result of their differing degrees of conventionality? By establishing the relative conventionality of these Chinese motor action metaphors, this study lays the groundwork for in-depth experimental research on the involvement of the motor system in the comprehension of Chinese object manipulation metaphors in relation to conventionality and contextuality.

5 Conclusion

This study provides a usage-based constructionist perspective on manual motor metaphors in Chinese. An immediate implication to be drawn from this study is the methodological importance of quantitative usage data in establishing the conventionality, productivity, and semantic subclassification of metaphors encoded in syntactic patterns. The present cognitive semantic analysis of the three constructions lays an empirical foundation for future behavioural and neuroimaging research on the extent to which Chinese verbal metaphors of manual object manipulation engage cortical sensorimotor regions in the brain. Finally, this study holds an implication for language learning and teaching. As Jing-Schmidt (2015) suggested, the usage-based constructionist approach to language provides a toolbox for teachers as well as learners. This is particularly true of the teaching and learning of figurative language the conventionality of which defies

compositionist bottom-up comprehension and acquisition. Exposing learners to the high-frequency tokens, together with the dominant semantic subclasses of a metaphorical construction, can contribute to acquisition by facilitating prototype-based learning.

Bibliography

Barlow, M.; Kemmer, S. (eds) (2000). *Usage-Based Models of Grammar*. Stanford (CA): Center for the Study of Language and Information.

Barsalou, L.W. (1999). "Perceptual Symbol Systems". *Behavioral and Brain Sciences*, 22(4), 577-660. https://doi.org/10.1017/s0140525x99002149.

Barsalou, L.W. (2008). "Grounded Cognition". *Annual Review of Psychology*, 59, 617-45. https://doi.org/10.1146/annurev.psych.59.103006.093639.

Boulenger, V.; Shtyrov, Y.; Pulvermüller, F. (2012). "When Do You Grasp the Idea? MEG Evidence of Simultanous Idiom Understanding". *Neuroimage,* 59, 3502-13. https://doi.org/10.1016/j.neuroimage.2011.11.011.

Bradshaw, J.L. (1991). "Animal Asymmetry and Human Heredity: Dextrality, Tool Use and Language in Evolution–10 Years after Walker (1980)". *The British Journal of Psychology*, 82(1), 39-59.

Bybee, J. (2006). "From Usage to Grammar. The Mind's Response to Repetition". *Language*, 82(4), 711-33. https://doi.org/10.1353/lan.2006.0186.

Bybee, J. (2013). "Usage-Based Theory and Exemplar Representations of Constructions". Hoffmann, Trousdale 2013, 49-69. https://doi.org/10.1093/oxfordhb/9780195396683.013.0004.

Croft, W. (2001). *Radical Construction Grammar. Syntactic Theory in Typological Perspective*. Oxford: Oxford University Press.

Croft, W. (2003). "The Role of Domains in the Interpretation of Metaphors and Metonymies". Dirven, R.; Pörings, R. (eds), *Metaphor and Metonymy in Comparison and Contrast*. Berlin; New York: Mouton De Gruyter, 161-206. https://doi.org/10.1515/9783110219197.161.

Darwin, C. (1871). *The Descent of Man, and Selection in Relation to Sex*. London: John Murray.

Desai, R.H.; Binder, J.R. ; Conant, L.L. ; Mano, Q.R.; Seidenberg, M.S. (2011). "The Neural Career of Sensory-Motor Metaphors". *Journal of Cognitive Neuroscience*, 23(9), 2376-86. https://doi.org/10.1162/jocn.2010.21596.

Desai, R.H.; Conant, L.L.; Binder, J.R.; Park, H.; Seidenberg, M.S. (2013). "A Piece of the Action: Modulation of Sensory-motor Regions by Action Idioms and Metaphors". *NeuroImage*, 83, 862-69.

Ellis, N.C. (2002). "Frequency Effect in Language Processing. A Review with Implications for Theories of Implicit and Explicit Language Acquisition". *Studies in Second Language Acquisition*, 24(2), 143-88. https://doi.org/10.1017/s0272263102002024.

Ellis, N.C. (2011). "Frequency-Based Accounts of SLA". Gass, S.; Mackey, A. (eds), *Handbook of Second Language Acquisition*. London: Routledge, 193-210. https://doi.org/10.1002/9780470756492.ch7.

Ellis, N.C. (2013). "Construction Grammar and Second Language Acquisition". Hoffmann, Trousdale 2013, 365-78. https://doi.org/10.1093/oxfordhb/9780195396683.013.0020.

Fischer, M.H.; Zwaan, R.A. (2008). "Embodied Language. A Review of the Role of the Motor System in Language Comprehension". *Quarterly Journal of Experimental Psychology*, 61(6), 825-50. https://doi.org/10.1080/17470210701623605.

Fillmore, C.J. (1988). "The Mechanisms of Construction Grammar". *Berkeley Linguistics Society*, 14, 35-55. https://doi.org/10.3765/bls.v14i0.1794.

Fillmore, C.J.; Kay, P.; O'Connor, M.C. (1988). "Regularity and Idiomaticity in Grammatical Constructions. The Case of 'Let Alone'". *Language*, 64(3), 501-38. https://doi.org/10.2307/414531.

Gallese, V.; Lakoff, G. (2005). "The Brain's Concepts. The Role of the Sensory-Motor System in Conceptual Knowledge". *Cognitive Neuropsychology*, 22(3-4), 455-79. https://doi.org/10.1080/02643290442000310.

Gao, H. (2001). "The Physical Foundation of the Patterning of Physical Action Verbs. A Study of Chinese Verbs". *Travaux de l'institut de linguistique de Lund XLI* [PhD dissertation]. Lund: Lund University.

Gibson, K.R.; Ingold, T. (eds) (1993). *Tools, Language and Cognition in Human Evolution*. New York: Cambridge University Press.

Gibbs, R.J. (1999). "Taking Metaphor out of Our Heads and Putting It into the Cultural World". Gibbs, R.; Stehen, G. (eds), *Metaphor in Cognitive Linguistics*. Amsterdam: John Benjamins, 145-66. https://doi.org/10.1075/cilt.175.09gib.

Gibbs, R.J. (2006). *Embodiment and Cognitive Science*. Cambridge: Cambridge University Press.

Glenberg, A.L.; Kaschak, M.P. (2002). "Grounding Language in Action". *Psychonomic Bulletin & Review*, 9, 558-65. https://doi.org/10.3758/BF03196313.

Goldberg, A. (1995). *Constructions. A Construction Grammar Approach to Argument Structure.* Chicago: University of Chicago Press.

Goldberg, A. (2006). *Constructions at Work. The Nature of Generalization in Language*. Oxford: Oxford University Press.

Goldberg, A. (2013). "Constructionist Approaches". Hoffmann, Trousdale 2013, 15-31. https://doi.org/10.1093/oxfordhb/9780195396683.013.0002.

Goldberg, A. (2019). *Explain Me This. Creativity, Competition, and the Partial Productivity of Constructions*. Princeton (NJ): Princeton University Press.

Gries, S.T. (2012). "Frequencies, Probabilities and Association Measures in Usage-/Exemplar-Based Linguistics. Some Necessary Clarifications". *Studies in Language*, 11(3), 477-510. https://doi.org/10.1075/sl.36.3.02gri.

Grush, R. (2004). "The Emulation Theory of Representation. Motor Control, Imagery, and Perception". *Behavioral and Brain Sciences*, 27(3), 377-442. https://doi.org/10.1017/s0140525x04000093.

Hilpert, M. (2014). *Construction Grammar and Its Application to English*. Edinburgh: Edinburgh University Press.

Hoffmann, T.; Trousdale, G. (eds) (2013). *The Oxford Handbook of Construction Grammar*. Oxford: Oxford University Press.

Iriki, A.; Taoka, M. (2012). "Triadic (Ecological, Neural, Cognitive) Niche Construction. A Scenario of Human Brain Evolution Extrapolating Tool Use and Language from the Control of Reaching Actions". *Philosophical Transactions Of The Royal Society B-Biological Sciences*, 367(1585), 10-23. https://doi.org/10.1098/rstb.2011.0190.

Jakobson, R. [1956] (2003). "The Metaphoric and Metonymic Poles". Dirven, R.; Pörings, R. (eds), *Metaphor and Metonymy in Compari-*

son and Contrast. Berlin: Mouton de Gruyter, 41-7. https://doi.org/10.1515/9783110219197.1.41.

Jing-Schmidt, Z. (2005). *Dramatized Discourse. The Mandarin Chinese Ba-Construction*. Amsterdam; Philadelphia: John Benjamins.

Jing-Schmidt, Z. (2010). "Seven Events in Three Languages. Culture-Specific Conceptualizations and Their Pedagogical Implications". De Knop, S.; Boers, F.; De Rycker, A. (eds), *Fostering Language Teaching Efficiency through Cognitive Linguistics*. Berlin: Mouton de Gruyter, 137-66. https://doi.org/10.1515/9783110245837.137.

Jing-Schmidt, Z. (2015). "The Place of Linguistics in CSL Teaching and Teacher Education. Toward a Usage-Based Constructionist Theoretical Orientation". *Journal of Chinese Language Teachers Association*, 50(3), 1-22.

Jing-Schmidt, Z.; Peng, X.; Chen, J.Y. (2015). "From Corpus Analysis to Grammar Instruction. Toward a Usage-Based Constructionist Approach to Constructional Stratification". *Journal of Chinese Language Teachers Association*, 50(2), 109-38.

Johnson, M. (2017). *Embodied Mind, Meaning, and Reason. How Our Bodies Give Rise to Understanding*. Chicago: University of Chicago.

Kiefer, M.; Sim, E.; Herrnberger, B.; Grothe, J.; Hoenig, K. (2008). "The Sound of Concepts. Four Markers for a Link between Auditory and Conceptual Brain Systems". *The Journal of Neuroscience*, 28(47), 12224-30. https://doi.org/10.1523/jneurosci.3579–08.2008.

Kövecses, Z. (2005). *Metaphor in Culture*. Cambridge: Cambridge University Press.

Lakoff, G.; Johnson, M. (1980). *Metaphors We Live By*. Chicago: The University of Chicago Press.

Lederer, J. (2019). "Lexico-Grammatical Alignment in Metaphor Construal". *Cognitive Linguistics*, 30(1), 165-203. https://doi.org/10.1515/cog–2017–0135.

Osherson, D.N.; Smith, E.E.; Wilkie, O.; López, A.; Shafir, E. (1990). "Category-Based Induction". *Psychological Review*, 97(2), 185-200.

Pecher, D.; Zwaan, R.A. (2005). *Grounding Cognition. The Role of Perception and Action in Memory, Language, and Thinking*. Cambridge: Cambridge University Press.

Romero Lauro, L.J.; Mattavelli, G.; Papagno, C.; Tettamanti, M. (2013). "She Runs, the Road Runs, My Mind Runs, Bad Blood Runs Between Us. Literal and Figurative Motion Verbs. An FMRI Study". *NeuroImage*, 83, 361-71. https://doi.org/10.1016/j.neuroimage.2013.06.050.

Rüschemeyer, S.-A.; Brass, M.; Friederici, A.D. (2007). "Comprehending Prehending. Neural Correlates of Processing Verbs with Motor Stems". *Journal of Cognitive Neuroscience*, 19(5), 855-65. https://doi.org/10.1162/jocn.2007.19.5.855.

Rüschemeyer, S.-A.; Pfeiffer, C.; Bekkering, H. (2010). "Body Schematics. On the Role of the Body Schema in Embodied Lexical-Semantic Representations". *Neuropsychologia*, 48(3), 774-81. https://doi.org/10.1016/j.neuropsychologia.2009.09.019.

Simmons, W.K.; Ramjee, V.; Beauchamp, M.S. ; McRae, K.; Martin, A.; Barsalou, L.W. (2007). "A Common Neural Substrate for Perceiving and Knowing about Color". *Neuropsychologia*, 45(12), 2802-10. https://doi.org/10.1016/j.neuropsychologia.2007.05.002.

Steele, J.; Ferrari, P.F.; Fogassi. L. (2012). "From Action to Language. Comparative Perspectives on Primate Tool Use, Gesture and the Evolution of Human Language". *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 367(1585), 4-9. https://doi.org/10.1098/rstb.2011.0295.

Sullivan, K. (2013). *Frames and Constructions in Metaphoric Language*. Amsterdam; Philadelphia: John Benjamins.

Sullivan, K. (2016). "Integrating Constructional Semantics and Conceptual Metaphor". *Constructions and Frames*, 8(2), 141-65. https://doi.org/10.1075/cf.8.2.02sul.

Taylor, C.L.; Schwarz, R.J. (1955). "The Anatomy and Mechanics of the Human Hand". *Artificial Limbs*, 2(2), 22-35. http://www.oandplibrary.org/al/pdf/1955_02_022.pdf.

Tomasello, M. (2003). *Constructing a Language. A Usage-Based Theory of Language Acquisition*. Cambridge (MA): Harvard University Press.

van Dam, W.O. ; van Dijk, M.; Bekkering, H.; Rueschemeyer, S.-A. (2012). "Flexibility in Embodied Lexical-Semantic Representations". *Human Brain Mapping*, 33(10), 2322-33. https://doi.org/10.1002/hbm.21365.

Van Dantzig, S.; Pecher, D.; Zeelenberg, R.; Barsalou, L.W. (2008). "Perceptual Processing Affects Conceptual Processing". *Cognitive Science*, 32(3), 579-90. https://doi.org/10.1080/03640210802035365.

Xun, E.; Rao, G.; Xiao, X.; Zang, J. (2016). "Research and Development of the BCC Corpus in the Background of Big Data". *Corpus Linguistics*, 3(1), 93-109, 118.

Yu, N. (2000). "Figurative Uses of Finger and Palm in Chinese and English". *Metaphor and Symbol*, 15(3), 159-75. https://doi.org/10.1207/s15327868ms1503_3.

Yu, N. (2003). "The Bodily Dimension of Meaning in Chinese. What Do We Do and Mean with 'Hands'?". Casad, E.H.; Palmer, G.B. (eds), *Cognitive Linguistics and Non-Indo-European Languages*. Berlin; New York: Mouton De Gruyter, 337-62.

Corpus-Based Research on Chinese Language and Linguistics
edited by Bianca Basciano, Franco Gatti, Anna Morbiato

The Factuality Status of Chinese Necessity Modals
Exploring the Distribution Via Corpus-Based Approach

Carlotta Sparvoli
Alma Mater, Università di Bologna, Italia

Abstract This paper is intended to test the deontic *vs* anankastic hypothesis outlined by Sparvoli 2012. The stipulation is that, in past contexts, deontic modals trigger a counterfactual inference, while anankastic modals (here called 'goal-oriented modals') either trigger an actuality entailment effects ('only possibility' modals) or a generic non-factual reading ('mere necessity' modals). The result of this corpus-based study conducted in a Chinese-English parallel corpus confirm the crucial role played by the deontic *vs* goal-oriented contrast in the marking of factuality in Chinese and shows that the factuality value decreases across a cline from goal-oriented to deontic modals.

Keywords Actuality entailment. Counterfactuality. Deontic modality. Goal-oriented modality.

Edizioni Ca'Foscari

Sinica venetiana 6
e-ISSN 2610-9042 | ISSN 2610-9654
ISBN [ebook] 978-88-6969-406-6 | ISBN [print] 978-88-6969-407-3

Peer review | Open access
Submitted 2020-03-04 | Accepted 2020-10-14 | Published 2020-12-21

DOI 10.30687/978-88-6969-406-6/005

1 Introduction

The framework here adopted relies on the differentiation between deontic and anankastic modalities. Based on von Wright (1963), this theory postulates that modals pertaining to duty and necessity are distributed within a semantic domain having two poles (Sparvoli 2012): namely, the *deontic*, which expresses an obligation (ancient Greek *déon*) and is related to a moral duty, grounded on a principle, as in (1a); and the *anankastic* (from *anánkē*, literally 'rope, wire'), which indicates a practical necessity, linked to a specific purpose, as in (1b).

1. a. 'We should be modest and prudent'.
   (translated from Alleton 1984, 200)
   [deontic]
   b. *To get to the station you have to take bus 66*.
   (Van der Auwera, Plungian 1998, 80)
   [anankastic]

Anchored in the notion of 'inevitability', the anankastic expresses what 'cannot be done otherwise' and makes it possible to establish a unique and consistent class for expressions which are commonly related to different modalities, such as the necessity depending on natural law, circumstances or a given goal (or wish). Rough equivalents of the anankastic modality are found in the "participant-external non-deontic" (Van der Auwera, Plungian 1998), the "goal-oriented or teleological" (von Fintel, Iatridou 2007) and in the "neutral" or "circumstantial dynamic" modality (Palmer 1990). Importantly, the anankastic domain includes markers of different binding force, ranging from weak to strong anankastic modals (as 'must' and 'cannot but', respectively) (Sparvoli 2012).

Along these lines, this paper focuses on the factuality[1] reading triggered by Chinese modals in past contexts. The working hypothesis is that (i) deontic modals such as 应该 *yīnggāi* 'should' yields counterfactuality, that is, they trigger the inference that "the speaker believes a certain proposition not to hold" (Iatridou 2000, 231) and such meaning is understood via an inference; (ii) the strongest anankastic modals, such as 不得不 *bùdébù* 'cannot but' or 只好 *zhǐhǎo* 'can only', trigger an uncancellable inference that the event took place in the actual world, therefore they are implicative, yield actuality entailments (Bhatt 1999; Hacquard 2006) and have a factual reading; (iii) 必须 *bìxū* 'have to' preferably gets a factual interpretation; (iv) weaker anankastic modals, such as 得 *děi* and 要 *yào* 'must', have a distribution similar to imper-

1 For an account of equivalent labels of 'factuality', such as 'actuality', see Giannakidou, Mari 2016, 82.

fective modals in French or Italian, thus they are not implicative and are compatible with both counterfactual and factual interpretations.

This hypothesis, already outlined in Sparvoli (2012), will be explored through a corpus-based study. To facilitate the identification of Chinese modals in past contexts, we selected the most prominent English (counter)factual necessity markers, respectively, *should have* and *had to*, to then identify the Chinese equivalent in the bilingual token thus retrieved. We browsed two subsets of the *E-C English-Chinese Parallel Concordancer*, published by the Hong Kong Institute of Education,[2] namely, the *E-C English Novels* (0.807 million words) and the *E-C Chinese Novels* (0.181 million words). In total, we processed 795 tokens and manually tagged the valid ones (527) against five types of eventualities (counterfactual, factual, habitual, non-factual in matrix position, non-factual embedded). Finally, we filtered the tokens including modal markers (387) for analysing their distribution across those types of eventualities.

§§ 2, 3, 4 and 5 illustrate, respectively, the theoretical framework, the prediction, the method and the study. The results show that the factuality reading of Chinese modals of duty and necessity is gradient: it extends from a unique factual reading for strong anankastic modals to a unique counterfactual reading for the deontic. Between these two poles are located the weaker anankastic modals, which can also have habitual reading and thus have a similar distribution of the imperfective form of the Italian *dovere*.

2 Background

2.1 The Deontic *vs* Anankastic Contrast

Though interchangeable in a positive context, the classification into deontic or anankastic modality is based on the different interaction with negation (Sparvoli 2012). Namely, the negation of a prominent[3] deontic marker produces a Prohibition, like 'should not', while the negation of the anankastic produces an Exemption, like 'don't have to', 'need not'. In other words, deontic modals scope over negation, while anankastic modals scope under negation (Lü [1942] 1944). In Chinese, the categorisation into either one of these two modalities, though expressed in different terminology, is already found in the modality in-

2 Further details on the corpus are provided in the Bibliography.

3 The underlying principle of the concept of "modal prominence" (Li 2004, 176) is that the different modal meanings of polysemous markers can be ranked into four categories: namely, prominent markers (that is, prototypical, as for 应该 *yīnggāi* in the deontic and epistemic modalities); frequent but non-prominent; non-frequent; not used.

vestigation prior to 1949 (Sparvoli 2012). In this literature, the prominent markers of these two modalities are the deontic (应)该/当 (*yīng*) *gāi*/*dāng* 'should' (2a), and the anankastic 必须 *bìxū* 'must' and 得 *děi* 'have to' (2b); the latter two are positive polarity items, negated via suppletive forms expressing Exemption, like 不必 *búbì*, 无需 *wúxū* 'don't have to' or 不用 *bùyòng*, 甭 *béng* 'need not'.[4] The classification of 要 *yào* is more difficult, since it can have the meaning of 必要 *bìyào* 'must', 需要 *xūyào* 'need', 想要 *xiǎngyào* 'would like to', 快要 *kuàiyào* 'is going to', or 将要 *jiāngyào* 'will' (Li 2004, 162). In a normative context, following von Wright, who "classified 'must' as anankastic but 'must not' as deontic" (1963, VIII-2, 157), we labelled 要 *yào* as a weak anankastic and 不要 *búyào* as a deontic. It must be noted that, in this corpus-based study (see Chart 1),[5] 要 *yào* also occurs as a dynamic marker, indicating some "necessity internal to a participant engaged in the state of affairs" (Van der Auwera, Plungian 1998, 80), as in (2c).[6]

2. a. 我们应该/应当谦虚谨慎。 (Alleton 1984, 200) [deontic]
wǒmen ***yīnggāi/yīngdāng*** *qiānxū* *jǐnshèn*
we should be.modest be.prudent
'We should be modest and prudent'.

b. 去火车站得坐第六六路公共汽车。 (Li 2004, 107) [anankastic]
qù *huǒchē-zhàn* ***děi*** *zuò* *dìliùliù* *lù* *gònggōngqìchē*
go train-station have.to sit 66 CLF bus
'To get to the station you have to take bus 66'. (Van der Auwera, Plungian 1998, 80)

c. 鲍里斯每晚要睡十个小时才能正常活动。 (Li 2004, 107) [dynamic]
Bàolǐsī *měi* *wǎn* ***yào*** *shuì* *shí* *ge* *xiǎoshí*
Boris every night need sleep ten CLF hour
cái *néng* *zhèngcháng* *huódòng*
then can normally function
'Boris needs to sleep ten hours every night for him to function properly'. (Van der Auwera, Plungian 1998, 80)

4 Concerning the status of 得 *děi*, Lü Shuxiang clarified that the negative form of 得 *děi* is 不用 *bùyòng*, 甭 *béng* 'need not': "[*děi* 得] 表示否定用'不用、甭', 不能用'不得'" (1984, 143). In other words, in Chinese linguistics prior to 1949, the homograph 得 is considered to have three distinct forms, *dé*, *de* and *děi*, wherein the latter surfaced only in Modern Chinese; such later usage can be also considered as a "second split" in the grammaticalisation process of the lexical verb *dé* 'to obtain' (Ziegeler 2003, 251).

5 In this study, 要 *yào* also occurs with volitional or futurity readings, especially when retrieved with the token *should have* **[tab. 5]**.

6 The glosses follow the general guidelines of the Leipzig Glossing Rules. Additional glosses include: BA = 'preposition introducing the object in the *ba*-construction'; DE = 'structural particle *de*'; INC = 'inchoative'; SFP = 'sentence-final particle'.

In a cartographic perspective, adjusting our terminology and taxonomy into Tsai's (2015) proposal, the anankastic 必须/要 *bìxū*/*yào* are hosted in the inflectional layer, between the outer and the inner subject, while the deontic 应该 *yīnggāi* is hosted in the complementiser layer, as its epistemic counterpart. Finally, Sparvoli (2012) identified a set of symmetrical traits of the deontic/anankastic contrast. In this context, the more relevant is related to the different behaviour in perfective contexts, where anankastic modals trigger actuality entailment while the deontic get a counterfactual reading. This corpus-based study is therefore aimed at testing this stipulation, but before presenting the method and the results, we need to present the issue related to the factual reading of modalised expression and introduce the notion of 'actuality entailment'.

2.2 Modals and Factuality

Since Kiefer (1987) and Chung and Timberlake (1985) and, even before, with Lü Shuxiang ([1942] 1944, 187), modality has been related to the notion of 'non-factuality', implying that when an eventuality is *possible* or *necessary*, it is by default *non-factual*. However, the implicative feature of the semi-modal *get* and the lexical verb *manage to* has been identified already by Karttunen (1971), who observed that, in a past environment, sentences like (3a) imply (3b) and express that a given event was actualised; therefore, they are not compatible with a continuation which negates the actualisation of the state of affairs.

3. a. John **managed/got/happened** to solve the problem,
   #but he didn't solve it.
   b. = John solved the problem.
   (Karttunen 1971, 342, 346 slightly modified)

From a typological approach to modality, Van der Auwera and Plungian (1998, 103-4) underscored that most markers, such as *manage*, in the perfective form mark the completion of the process.[7] From the possible world semantics, Bhatt (1999) describes this phenomenon as "actuality entailment" (hereafter AE), referred to a modalised proposition whose event holds in the actual world. Hacquard (2006) provided a unified account where AE is inferred contextually through the combination of two ingredients: the scopal properties

7 Van der Auwera and Plungian (1998, 103-4) classified *manage* as a demodalised marker expressing participant-internal "actuality" and underscored that most markers of participant-internal actuality, in the perfective form, when paralleled to their imperfective counterparts, mark the completion of the process

of the modal and the identity of the event. If the modal scopes below aspect and the event is anchored in a bound interval, then we have AE. In this way, the actuality implication is analysed not only with reference to Ability – that is, considering perfective ability modals as underlyingly implicative, *à la* Bhatt – but it is also accounted for other root modalities:

> modal interpretations that did yield actuality entailments were those with a *circumstantial* modal base (abilities, goal-oriented and pure circumstantials); the ones that didn't were those with an *epistemic* or a (truly) *deontic* interpretation.[8] (Hacquard 2006, 113)

The circumstantial feature seems to play a crucial role in the actuality reading of modalised expressions in past environment.[9] Moreover, in languages with perfective-imperfective morphology, a deontic modal occurring with an anankastic interpretation, as *devoir* in (4a), in the perfective form yields AE. In the imperfective form instead (4b), depending on the context and the continuation, it can have a counterfactual, progressive/habitual or generic interpretation (Hacquard 2006, 103).

4. a. *Pour aller au zoo, Jane* ***a dû*** *prendre le*
to go to.the zoo Jane ***must*.PST.PFV** take the
train... **[#but did not]**
train
b. *Pour aller au zoo, Jane* ***devait*** *prendre le*
to go to.the zoo Jane **must.PST.PFV** take the
train... **[but did not]**
train (Hacquard 2006, 14)

In Hacquard's framework, the implicative reading arises from the perfective aspect outscoping the modal. More specifically, aspect starts as an argument of the verb and moves out yielding two nodes of type *t*: TP and VP. This allows a root modal to appear either right above TP or right above VP, with aspect moving right above the modal (Hacquard 2017, 52).[10] When low, the modal is bound by the aspect of the VP event; when high, it is bound by the speech event or, in embedded contexts, by attitude events. This, in turn, implies that

8 Hacquard (2006, 41) uses the label 'real' deontic with reference to someone granting permission or imposing an obligation on someone else.

9 The circumstantial reading is also underscored by Van der Auwera and Plungian (1998, 103-4) with reference to participant-internal actuality.

10 For a cartographic account on the scopal property with respect to aspect of Chinese modals, see Tsai 2015.

in each configuration, the modal has different relational time: it is anchored, respectively, to the event time, the utterance time and the attitude time. As a result, AE effect is not expected when the modal occurs in embedded sentences. In § 5.2, we will take into account this feature while discussing our results concerning the tokens in embedded position (shown in chart 2).

2.3 Counterfactuality and Temporal Orientation

For both Bhatt (1999) and Hacquard (2017), the lack of AEs in imperfective modals, as in (4b), is due to an additional layer of modality associated with the latter. In Reischenbachian terms, the difference between perfective and imperfective aspect is accounted for with reference to the specular relation between reference and event time whereby the perfective locates the event within the *reference time*, whereas the imperfective locates the reference time within the *time of the event*, hence its typical features of ongoingness, repetition, and regularity. We do not need to discuss here in more detail the perfective/imperfective contrast, but we should recall that the imperfective morphology can give rise to a number of different readings, such as the *progressive* and *non-progressive continuous* interpretations, the *habitual* (including generic/dispositional meanings), and also the *circumstantial habitual*. The latter encompasses "a type of discourse in which a type of setting is first introduced, and then sequences of events that typically occur within that setting are enumerated" (Carlson 2012, 838).

Moreover, in modalised expressions, the imperfective can trigger a past counterfactual interpretation. Generated by the opposite inference of AE, the counterfactual reading conveys that "the speaker believes a certain proposition not to hold" (Iatridou 2000, 231); a counterfactual interpretation implies that the situation at stake has already been 'settled', and that such an (unactualised) state of affairs cannot be reversed. In other words, past counterfactual modals tell us how the world *should* or *could have* turned out to be, if a state of affairs had obtained (Condoravdi 2002), as in (5):

5. At that point he **should/might** (still) have won the game but he didn't in the end. (Condoravdi 2002, 62 slightly modified)

As emphasised by Condoravdi, (5) conveys that "we are now located in a world whose past included the (unactualised) possibility of his winning the game" (2002, 60); in general terms, *should have* expresses that it is "necessary at the present moment that a certain state of affairs obtained in the past" (60) and is thus compatible with both the epistemic and counterfactual interpretation. The latter reading

stems from a future temporal orientation of the modal combined with a past perspective, that is, its reference time is an interval "starting at some past time and extending to the end of time" (75). These elements point to a future-in-the-past orientation of the counterfactual construal. Now that we have set the main coordinates of the theoretical framework, we can turn our attention to the language-specific issues related to counterfactual and AE in Chinese, which will be addressed, respectively, in §§ 2.4 and 3.

2.4 Counterfactuality in Chinese

Since Bloom (1981), the investigation on the encoding of counterfactuality in Chinese (Nevins 2002; Jiang 2000, 2019a; Yong 2016; Jing-Schmidt 2017; Liu 2019, among others) has been primarily focused on counterfactual conditionals, as in (6). Using the terminology adopted in § 2.2, we could say that in these constructions, the antecedent conveys an hypothesis (as 'it had rained yesterday') which is opposite to what happens (or happened) in reality; the consequent instead states what would or would have turned out to be, if that state of affairs had obtained (that is, 'I would have gone' in (6)).

6. 要是昨天下雨了, 我(就)回去。 (Liu 2019, 41)

yàoshi	*zuótiān*	*xià*	*yǔ*	***le***	*wǒ*	*jiù*	*huí*	*qù*
if	yesterday	fall	rain	SFP	I	(then)	return	go

'If it had rained yesterday, I would have gone'.
NOT: *'If it rained yesterday, I will go'.

While in Indo-European languages the reality status of each proposition is typically signalled through tense morphology, the Chinese encoding of counterfactuality can hardly be captured by a clear-cut syntactic account. The relevant literature has in fact shed light on the role of the combination of hypothetical conjunctions like 要不是 *yàobushì* 'were it not for' with other markers, such as the aspectual and the sentence final particle 了 *le*, the temporal marker 早 *zǎo* 'early', negative operators or discourse markers such as 真的 *zhēnde* 'really'.[11] Due to the diverse elements at stake, the investigations on Chinese counterfactual conditionals are characterised by a constructionist approach and typically aim at producing a pragmatic or semantic account, without relying on a specific syntactic derivation. This composite scenario is described as a "cluster of unnoticeable weak features or lexical items that contribute, sometimes jointly, to reaching of counterfactual meaning" (Jiang 2019b, 283). For instance, in (7),

11 For a detailed account of this topic, see Jiang 2019b, 284 ff.

we have the combination of a conditional conjunction 要是 *yàoshi*, a past time-reference and the distal 那个 *nàge* 'that' which contributes to locating the event in a hypothetical past event. As observed by Jiang, by replacing it with the proximal 这个 *zhège*, the sentence could be interpreted as "if this free-kick is in, the match will go into overtime" (285). The subtle, though essential, contribution of the distal 那个 *nàge* is thus a good example of what is meant by 'weak feature', that is, a feature which is neither sufficient nor essential but yet contributes to the 'construction' of the counterfactual interpretation.

7. 要是那个任意球罚进了, 就会踢加时赛了。 (Jiang 2019b, 285)

yàoshi	***nà-ge***	*rènyì-qiú*	*fá-jìn*	***le***	*jiù*	*huì*	*tī*
if	that-CLF	free-kick	shoot-in	SFP	hence	will	kick
jiā-shí-sài		***le***					
extra-time-match		SFP					

'If that free-kick had been in, the match would have gone into overtime'.
NOT: *'if this free-kick is in, the match will go into overtime'.

Despite this 'weak feature', unified accounts are being formulated, especially with reference to past counterfactuals, which, starting from Ziegeler, are considered as the only environment in which the "counterfactual construal can be obtained reliably" (2000, 104), as in (6). Similarly, Liu (2019) stressed the role of the combination of the past time reference and the conditional setting, while Jiang (2019a) highlighted the "tense mismatch" which locates the event in a hypothetical past, obtained either by pointing to a relative tense (as in 7) or by the use of time adverbs as 早 *zǎo* 'early'.[12] It must be emphasised that the proposals above are consistent with Condoravdi's emphasis on the combination between a past perspective and a future temporal orientation of the modal, as the aspectual 了 *le* in the antecedent, and 会 *huì* in the consequent, in (6) and (7).

In a corpus-based approach, Yong (2016) shed light on the correlation with past-oriented temporality, negation, emphatic modal adverbs, optative mood, first person pronouns, and demonstratives. Focusing on the pragmatic dimension, Jing-Schmidt (2017) paired a set of discourse functions with five bi-clausal hypothetical constructions and provided an analysis of the co-occurring modality markers, including modal verbs, adverbs, and modal particles. Based on 3,698 tokens of 要不是 *yàobushì*, she singled out 35 modal items (Jing-Schmidt (2017, 37) wherein the two highest ranked expressions are the futuri-

12 Jiang (2019a) also mentioned a second type of encoding of counterfactual conditional, having impossible or absurd antecedents, where the counterfactual meaning is only triggered by a 'pure inference', but those instances are not relevant in the context of this paper.

ty markers 不会 *búhùi* 'won't' and 会 *hùi* 'will'.[13] Further discussion is in order on the contribution of 会 *huì*, which can be classified either as a futurity marker or, following Jing-Schmidt, as a speaker stance marker signalling 'epistemic certainty'. In the discussion of current data, we will address this topic in § 5.1. Here we need to recall that Jing-Schmidt observed that those 35 modal combinations uniformly signal speaker stance; thus, she emphasised the evaluative nature of this construal, describing it as the result of the idiosyncratic combination of different counterfactual ingredients.

To conclude, in the study on Chinese counterfactual, the issue of the contribution offered by necessity modals is addressed only peripherally. Importantly, Feng and Yi (2006), following Wu (1994), included 原来应该 *yuánlái yīnggāi*, glossed as 'should have been', among the markers used to elicit a counterfactual reading by the participants in their study; for two out of three respondents, the deontic modal preceded by 原来 *yuánlái* proved to be the most productive marker, triggering counterfactual reading in 92% of the 200 statements. This result directly leads us to the working hypothesis of present studies.

3 Hypothesis and Prediction

3.1 Anankastic Strength and Actuality Entailment

We propose that in Chinese, in past contexts, deontic and anankastic modals can be a likely index of the (counter)factual reading (Sparvoli 2012).[14] For outlining our proposal, we will start by focusing on the factuality reading of necessity modals in past contexts.

In a formal semantic perspective, Chen (2012) observed a lack of AE of 应该 *yīnggāi* and 必须 *bìxū* due to a covert prospective aspect of Mandarin deontic and anankastic (in her terminology, "goal-oriented") modals. From a typological framework and based on the semantic contents of the notional ideas underlying modalities, our working hypothesis is that AE effects are correlated to the modal prominence of the necessity marker: it is high with anankastic markers and it is

13 Jing-Schmidt labels them as "modals that express high epistemic certainty" (2017, 36). In the framework entertained here, futurity is a post-modal marker (Van der Awera, Plungian, 1998, 194 ff.), developed from epistemic necessity (Li 2004, 256).

14 As an anticipation of this claim, cf. Alleton 1984 and Myhill, Smith 1995, 266, who underscored the counterfactual value played by 该 *gāi*. For a diachronic account, cf. Meisterernst 2017. Liu (2019) also suggested the need for more investigation on the role of modality in the making of counterfactual reading.

null with pure deontic ones (Sparvoli 2012, 2015).[15] Our framework suggests that full-fledged AE is typically found with negative forms or forms combined with the exclusive focus marker 只 *zhǐ* 'only, just' (Sparvoli 2019). Regarding the latter, it must be stressed that:

> 表示可能的词, 加一"只"字, 如 "只能"、"只好"、"只得"、"只会", 把他的可能性缩小, 就成为表示必要或必然。
>
> By adding the character 只 *zhǐ* before words expressing possibility, as in 只能 *zhǐnéng*, 只好 *zhǐhǎo*, 只得 *zhǐdé*, 只会 *zhǐ huì*, their possibility feature is reduced, and they are turned into expressions of necessity or certainty. (Lü Shuxiang [1942] 1944, 256)[16]

As emphasised by Li Renzhi, in these cases we do not have a real semantic shift into the necessity domain, but rather the extension of a possibility expression "to its extreme" (2004, 190). The underlying principle is that there is a continuum from possibility to necessity. Along the same lines, we propose a cline from deontic to strong anankastic modals, based on their anankastic strength.

Table 1 Anankastic strength of necessity modals (Sparvoli 2012, 217; 293)

Anankastic strength	**Necessity modal**	**Modality**	**Logic implication in a conditional period**
++++	不得不 *bùdébù*, 非得 *fēiděi*, 只好 *zhǐhǎo*	Anankastic necessity	Only possibility
+++-	必须 *bìxū*		Mere necessity (necessary condition)
++--	得 *děi*		
+---	要 *yào*		Sufficiency condition*
----	应该 *yīnggāi*	Deontic necessity	Simple implication, alternatives are available

* Typically, bouletic meaning in the antecedent of a conditional period. In the consequent it typically occurs combined with the focus marker 只 *zhǐ* expressing sufficiency condition. For a more detailed account of the different modal distribution in conditional construction, in combination with 才 *cái* and 就 *jiù*, see Sparvoli 2012, 273 ff.

15 Sparvoli (2019) suggests that the occurrence of AE in the negative form points to an aspectual coercion, arguably the neutralisation of the modal prospectivity feature, triggered by the negation.

16 Unless otherwise indicated all translations are by the Author.

3.2 The Working Hypothesis

We have seen that, with a circumstantial reading, the perfective forces the complement to hold in the actual word (Hacquard 2006, 14), and that an imperfective modalised form is typically compatible with a counterfactual, habitual/circumstantial, progressive, and generic reading. In Chinese, morphological tense marking is not available, while anankastic and deontic modalities are lexicalised in two sets of items displaying opposite scopal properties with reference to negation (Lü [1942] 1944; Sparvoli 2012) and aspect (Tsai 2015). The working hypothesis of this paper is that, in such heavily isolating language, the strategy for denoting (counter)factuality could be offered by the shift to a different necessity modal. Practically speaking, a contrast like (4a) and (4b) above would be expressed shifting from a deontic marker, as 应该 *yīnggāi*, 该 *gāi*, 应当 *yīngdāng*, to an anakastic marker, as 不得不 *bùdébù*, 只好 *zhǐhǎo*, 必须 *bìxū*, 得 *děi*. This paper attempts to verify such an hypothesis through a corpus-based study. If confirmed, this proposal would make it possible to outline a tripartite typological classification of (counter)factual marking:

a. in languages perfective/imperfective morphology (French, Italian, Catalan, Bulgarian, Greek, Hindi): *mood* and *tense* shift (Hacquard 2006);
b. in languages lacking perfective/imperfective morphology but having morphological tense-marking (English): both *mood*, *tense* and *modal shift*;
c. in heavily isolating languages like Chinese: *modal shift* combined with *temporal markers*.

Now we can turn again to the prototypical examples by Hacquard (2006), mentioned in (3-4) and propose their Chinese equivalents as visible in (8), (9) and (10) below.

8. Factual [AE effect, 'Jane did take the train']
   a. *Pour aller au zoo, Jane a **dû** prendre le train.*
      [Indicative, past perfective, deontic *devoir*]

   b. To go to the zoo, Jane **had to** take the train.
      [Indicative, past, anankastic *have to*]

   c. (那时候)去动物园珍妮不得不坐火车。
      (***nà shíhou***) *qù dòngwùyuán Zhēnnī **bùdébù** zuò huǒchē*
      that time go zoo Jane cannot.but sit train
      [Temporal marker + strongest anankastic marker 不得不 *budébu* 'cannot but']

9. Non-factual [maybe Jane took the train, or maybe not]
   a. *Pour aller au zoo, Jane **devait** prendre le train.*
   [Indicative, past imperfective, deontic, *devoir*]

   b. To go to the zoo, Jane **would have had** to take the train.
   [conditional, past, anankastic *have to*]

   c. (那时候)去动物园珍妮得坐火车。
   (***nà shíhou***) *qù dòngwùyuán Zhēnnī **děi** zuò huǒchē*
   that time go zoo Jane need.to sit train
   [Temporal marker + anankastic 得 *děi* 'need to']

10. Counterfactual [Jane did not take the train]
   a. *Pour aller au zoo, Jane **aurait dû** prendre le train.*
   [Conditional, past, deontic, *devoir*]

   b. To go to the zoo, Jane **should have taken** the train.
   [Conditional, past, deontic *should*]

   c. (那时候)去动物园珍妮[本来]应该坐火车。
   (***nà shíhou***) *qù dòngwùyuán Zhēnnī* [***běnlái***] *yīnggāi zuò huǒchē*
   that time go zoo Jane originally should sit train
   [Temp. marker + (counterfactual adverbial) + deontic 应该 *yīnggāi* 'should']

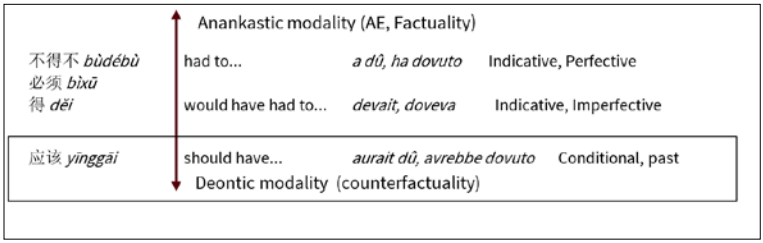


Figure 1 From Counterfactuality to Factuality (Sparvoli 2015)

3.3 The Prediction

Along these lines, the predictions are that: (i) the Chinese equivalents of the counterfactual occurrences of *should have* are marked by pure deontic markers such as (应)当/该 (*yīng*)*dāng*/*gāi* 'should', alone or in combination with the counterfactual marker 本(来) *běn*(*lái*); (ii) stronger anankastic markers, such as 不得不 *bùdébù* 'cannot but' or

只好 *zhǐhǎo* 'can only', are banned in counterfactual environments; (iii) 必须 *bìxū* 'have to' preferably gets a factual interpretation; (iv) weaker anankastic modals, such as 得/要 *děi*/*yào* 'must', have a distribution similar to imperfective modals in French or Italian, thus they are compatible with both counterfactual and factual environments, without yielding AE.

Table 2 Prediction: the distribution of Chinese necessity modal in (counter)factual statements

Modal	Counterfactual	Habitual	Non-factual	Factual*
不得不 *bùdébù*, 非得 *fēiděi*, 只好 *zhǐhǎo*, 只得 *zhǐdé*, 只能 *zhǐ néng*	x	x	x	√
必须 *bìxū*	x	√	√	√
得 *děi*	√	√	√	√
要 *yào*	√	√	√	√
应该 *yīnggāi*	√	x	x	x

* By factual we intend a proposition that can only be understood as actualised, which would typically happen when we have a modal yielding AE effect.

4 The Method

To test our predictions, we browsed two subsets of the *E-C English-Chinese Parallel Concordancer*. More specifically, we consulted the datasets named *E-C English Novels* (0.807 million words) and the *E-C Chinese Novels* (0.181 million words), wherein each pair of source and target text is aligned at the sentence level. To facilitate the identification of Chinese modals in past contexts, we selected the most prominent English (counter)factual necessity markers (*should have* and *had to*), to then identify their Chinese equivalents in the bilingual tokens thus retrieved. In total, we processed 795 bilingual tokens; after filtering the invalid tokens, the remaining 527 valid ones were tagged against five types of eventualities. Table 3 shows the token distributions and the list of Chinese equivalents encountered for each type of eventuality.[17]

17 The specific distribution of Chinese markers per each eventuality is visible in Chart 2, which provides a comprehensive overview of the results. The distribution obtained for each keyword, separately, is shown in table 5 (*should have*) and table 7 (*had to*).

Table 3 Tokens and types of eventualities

Keywords	E-C subsets	Filtered tokens	Types					
			Counterfactual	Factual	Habitual	Non-factual (matrix)	Non-factual (embedded)	Total
should have (385)	English novels	151	173 应该/当 *yīnggāi/dāng*; 最好 *zuìhǎo*; 要 *yào*; 一定 *yídìng*; 准 *zhǔn*; 不见得 *bújiàndé*; 可以 *kěyǐ*; 能 *néng*; 会 *huì*; 本来 *běnlái*; 早就 *zǎo jiù*; other	1 竟然 *jìngrán*				325
	Chinese novels	34	26 应该/当 *yīnggāi/dāng*; 本来 *běnlái*; 可以 *kěyǐ*; 想要 *xiǎng yào*; 须 *xū*; other					60
had to (410)	English novels	42		112 不得/能不 *bùdé/néng bù*; 只好 *zhǐhǎo*; 非得 *fēiděi*; 必须 *bìxū*; 得 *děi*; 需要 *xūyào*; 要 *yào*; other	18 要 *yào*; 得 *děi*; 必须 *bìxū*; 非得 *fēiděi*; other	15 必须 *bìxū*; 得 *děi*; 要 *yào*; 早就会 *zǎo jiù huì*; 非 *fēiděi*; other	48 必须 *bìxū*; 不得不 *bùdé bù*; 应该 *yīnggāi*; 需要 *xūyào*; 能 *néng*; 须 *xū*; 得 *děi*; other	235
	Chinese novels	41	3 应该 *yīnggāi*; 不必 *búbì*; other	97 不得/能不 *bùdé/néng bù*; 只好/能 *zhǐhǎo/néng*; 非不可 *fēi bùkě*; 必须 *bìxū*; 得 *děi*; 要 *yào*; 需要 *xūyào*; other	7 要 *yào*; 会 *huì*	20 要 *yào*; 必须 *bìxū*; 得 *děi*; 该 *gāi*; other	7 应该 *yīnggāi*; other	175
		268	202	207	25	35	58	795

The high rate of invalid tokens (34%, no. 268) is due to the characteristics of the major datasets used in this study. The *E-C English Novels Large Corpus* includes 13 classics from 19th-century English literature and their Chinese translation (typically conducted just before the turn of this century, see Appendix). In that variety of English, the usage of our first token, *should have*, encompassed a heterogeneous range of meanings, thus requiring an attentive process of selection for isolating the relevant tokens (as we will clarify below). Moreover, in that repertoire, even when occurring with a counterfactual meaning, *should have* is often used as an equivalent of *would have*, as in (11), thus providing data related to conditional counterfactuals rather than modalised counterfactual. However, since conditional counterfactuals attract a conspicuous number of deontic modals (Jing-Schmidt 2017), we also included this type of token in the scope of our analysis.

11. "and the effort which the formation and the perusal of this letter must occasion, *should have* been spared, *had not* my character required it to be written and read". (Jane Austen, *Pride and Prejudice*)
[counterfactual conditional, should have=would have]

On the other hand, while the sampling size is limited, this repertoire offers the advantage of being easily accessible in full narrative context and in a variety of languages. Focusing on widely translated, easily accessible and relatively familiar classics facilitated the process of disambiguation of the factuality reading. In fact, when necessary, we also double-checked the results of our disambiguation analysing the perfective-imperfective morphology found in the Italian translation of the relevant passage. In this way, we could disambiguate each token in the light of the context of narration, independently from the morphology and the modal classes of the keyword. For instance, (12) was retrieved from the *E-C Chinese Novels* by selecting *had to*; in light of the continuation in full narrative context, the token including 该 *gāi* 'should' was tagged in the counterfactual type.

12. The Kianghsi bus did not cross over, so they *had to* transfer to the Hunan bus, which departed at noon.
江西公路车不开过去了, 他们该换坐中午开的湖南公路车。
Jiāngxī gōnglùchē bù kāi guo qu le tāmen gāi
Jiangxi bus NEG drive cross go SFP they should
huàn zuò zhōngwǔ kāi de Húnán gōnglùchē
transfer sit noon depart DE Hunan bus
Continuation: The next morning they arrived at Chiehhualung, on the border between the provinces of Kinaghsi and Hunan. The Kianghsi bus did not cross over, so they **had to transfer** to the Hunan bus, which departed at noon. Of all the buses they had taken on the way, none had arrived at a station so promptly as this one; so rather than quarrel about

the short distance they felt that they'd come out a good half-day ahead and **decided to take a night's rest instead of catching the bus that day**. (Qian Zhongshu, *Wei cheng*. Engl. transl. *Fortress Besieged*, 2017, 255)

The token visible in (13), instead, has been retrieved with the keyword *should have* but tagged as factual, given the reading of *should have*, rendered in Chinese with the evaluative modal 竟然 *jìngrán*.

13. "It is astonishing […] that my heart *should have* been so insensible!" (Jane Austen, *Sense and Sensibility*)
简直令人吃惊, 我的心竟然那么麻木不仁!

jiǎnzhí	*lìngrénchījīng*	*wǒde*	*xīn*	*jìngrán*
simply	shocking	my	heart	unexpectedly
name	*mámùbùrén*			
like.that	insensitive			

= I was insensitive

The first step in the disambiguation process was filtering all the invalid segments wherein the Chinese target does not correspond to the English source text or vice versa. When possible, we tried to retrieve the correct target segments. A case in point is (11), repeated in (14), which was already mentioned in the previous section. Such a segment has been classified as counterfactual and tagged as a conditional, namely, a case where *should have* is rendered in Chinese with the possibility modal 可以 *kěyǐ* 'can, may' preceded by a hypothetical conjunction.

14. "and the effort which the formation and the perusal of this letter must occasion, *should have* been spared, had not my character required it to be written and read". (Jane Austen, *Pride and Prejudice*)

"我曾经衷心地希望我们双方会幸福, 可是我不想在这封信里再提到这些, 免得使你痛苦, 使我自己受委屈。" Correct match: 我所以要写这封信, 写了又要劳你的神去读, 这无非是拗不过自己的性格, 否则便可以双方省事, 免得我写你读。

Entries wherein *should have* occurs as the conditional of the lexical verb 'to have', as (15), have also been filtered:

15. "As to the future," said the Doctor, recovering firmness, "I *should have* great hope". (Charles Dickens, *A Tale of Two Cities*)

The second step in the disambiguation process was filtering the segments whose reading is not counterfactual. As a point of fact, *should have* does not necessarily force the counterfactual meaning. It can also have an epistemic reading, as in (16a), and, in embedded claus-

es, a deontic meaning (16b). Considering the variety of English offered by the corpus, it also occurs in future-in-the-past interpretations, as in (16c).

16. a. "This wine-shop keeper was a bull-necked, martial-looking man of thirty, and he *should have* been of a hot temperament, for, although it was a bitter day, he wore no coat, but carried one slung over his shoulder". (Charles Dickens, *A Tale of Two Cities*)

    b. "My mother", said Monks, in a louder tone, "did what a woman *should have* done". (Charles Dickens, *Oliver Twist*)

    c. "She had asked him not to leave London on any account, until he *should have* seen her again". (Charles Dickens, *David Copperfield*)

Moreover, in a substantial group of filtered segments, *should have* has a purely illocutionary function. In these cases, the Chinese rendering relies on discourse markers, such as 我相信 *wǒ xiāngxìn* 'I think', as in (17).

17. ["Oh me, oh me!" exclaimed the wretched Emily,][18] in a tone that might have touched the hardest heart, I *should have* thought. (Dickens, *David Copperfield*)
    [...] 那声音我相信就连最铁石的硬心肠人听了也会被感动的

nà	*shēngyīn*	***wǒ***	***xiāngxìn***	*jiù*	*lián*	*zuì*	*tiě-shí*
that	sound	I	believe	then	even	most	iron-stone
de	*yìng-xīncháng*	*rén*	*tīng-le*	*yě*	*huì*	*bèi*	
DE	hard-heart	person	hear-PFV	also	FUT	PASS	
gǎndòng	*de*						
move	DE						

Table 4 Filtered tags (*should have*: All English novels)

Misalignments	Lexical verb 'to have'	Illocutionary	Future-in-the-past	Epistemic	Deontic	Total
28	43	17	45	16	2	151
19%	28%	11%	30%	11%	1%	100%

The segment with future-in-the-past reading covers 30% of the filtered items **[tab. 4]**, and 14% of the entire 325 tokens retrieved from the *E-C English Novels* via *should have*.

18 In order to provide the contextual information needed for the factuality judgement, we included the relevant source text between square brackets.

5 The Study

5.1 Keyword 1. *Should Have*

In this section, we will first present the data retrieved from the *E-C English Novels*, that is, the English Chinese language combination. The first observation is that the tokens with counterfactual interpretation are embedded in the same environment described in the literature on Chinese counterfactual conditionals (see § 2.4), as 应该 *yīnggāi* in the consequent of a conditional construction, in (18).

18. "Well, sir, I think I *should have* known you, if I had taken the liberty of looking more closely at you". (Charles Dickens, *David Copperfield*)
"哦, 先生我相信, 如果我刚才能看你更仔细些, 我应该认出你。"

ó	*xiānshēng*	*wǒ*	*xiāngxìn*	***rúguǒ***	*wǒ*	*gāngcái*	***néng***
oh	sir	I	believe	if	I	just	could
kàn	*nǐ*	*gèng*	*zǐxì*	*xiē*	*wǒ*	***yīnggāi***	*rènchū*	*nǐ*
look	you	more	closely	a.bit	I	should	recognise	you

= I did NOT recognise you.

The results of the interrogation show that among the counter-factual tokens retrieved through the keyword *should have*, the most frequent non-epistemic necessity modal is the deontic (应)该/当 (*yīng*) *gāi*/*dāng*, followed by 要 *yào* and 最好 *zuìhǎo*. In the taxonomy, 最好 *zuìhǎo* is classified as deontic (Sparvoli 2012, 263), and it can safely be said that among the equivalents of *should have* with counterfactual meaning, anankastic modals are not found.

It also appears that the counterfactual reading is contributed by a number of other markers (see table 5, 'Non-modals') that typically occur in counterfactual conditionals, such as conditional conjunctions, focus markers, and temporal deictics that locate the sentence in a past context (Jiang 2000; Jing-Schmidt 2017; Liu 2019, among others).

Table 5 Modal distribution, counterfactual tokens (*should have*, *E-C English Novels*)[19]

Domain	Marker	No.		%
Deontic necessity (non-anankastic)	(应)该/当 (*yīng*)*gāi*/*dāng* 'should'	29	17%	18%
	最好 *zuìhǎo* 'had better'	1	1%	
Volition + Futurity*	要 *yào* 'want', 'is going to'	6		3%

19 Each modal can occur in combination with other counterfactual ingredients, such as a conditional constructions or other markers typically found in Chinese counterfactuals.

Epistemic necessity	一定 *yídìng* 'certainly'	6	3%	6%
	准 *zhǔn* 'certainly'	3	2%	
	不见得 *bújiànde* 'not necessarily'	1	1%	
Non-epistemic possibility	可以 *kěyǐ* 'may'	6	3%	11%
	能 *néng* 'can'	14	8%	
Futurity	会 *huì* 'futurity'	68		39%
Non-modals	本(来) *běn*(*lái*) 'originally'	5	3%	18%
	早就 *zǎo jiù* 'earlier, before'	8	5%	
	Conditionals	11	6%	
	Others	7	4%	
Underspecified	ND	8		5%
		173		**100%**

* This dataset consists of bilingual segments translated from English into Chinese, obtained with the keyword *should have*; in this type of repertoire, 要 *yào* occurs in sentences with a first-person subject, as a 'subjective necessity marker', with volitional or futurity meaning, thus having the meaning of 想要 *xiǎngyào* 'would like to', 快要 *kuàiyào* 'to be going to', or 将要 *jiāngyào* 'will'. For a comprehensive account of all 要 *yào* tokens, see chart 3.

The study also confirmed the crucial role of counterfactual chunks (Jiang 2019) like 早就 *zǎo jiù* in (19).

19. "I should have cried out, if I could". (Charles Dickens, *Great Expectations*)
如果我能够叫出声, 我早就大叫了起来。
rúguǒ wǒ nénggòu jiào-chu shēng wǒ ***zǎo jiù***
if I be.capable yell-exit voice I earlier then
dà jiào le qǐlai
greatly yell PFV start
= I did NOT yell

The constructionist feature of Chinese counterfactual is well represented by (20), which, paraphrasing Wang and Jiang (2011), displays virtually all the "ingredients of counterfactuality", in addition to the deontic 该 *gāi*:

20. "I *should have* said this sooner, but for my long mistake".
(Charles Dickens, *Great Expectations*)
"要不是我一向对这事情的误解, 我本该早就说了。"
yàobúshì *wǒ yíxiàng duì zhè shìqíng de wùjiě wǒ*
if.not.be I always towards this matter DE misread I
běn gāi zǎo jiù *shuō* ***le***
originally should early then tell SFP

There are also entries in which the counterfactual meaning is underspecified in Chinese (here signalled with 'nd'), thus confirming a phenomenon already observed by Yong (2016).[20] An example from the present study is (21).

21. "[mimicking his poverty, his boots, his coat, his mother,] everything belonging to him that they **should have had** consideration for". (Charles Dickens, *David Copperfield*, 242)
[…] 一切他们注意到的属于他的, 都被他们取笑。

yīqiè	*tāmen*	*zhùyì-dào*	*de*	*shǔyú*	*tā*	*de*
all	they	notice-RES	DE	belong.to	he	DE

Importantly, as highlighted by Jing-Schmidt (2017), the futurity marker 会 *huì* is the most common equivalent (39%) of the counterfactual *should have* **[tab. 5]**. The typical scenario of the occurrence of 会 *huì* is in the consequent of a conditional period. In such an environment, the counterfactual reading is derived by implicature and signalled by a number of *weak features* described in § 2.3, such as a past temporal orientation combining with a negative or adversative presupposition, typically provided contextually or in the continuation of the narration (as in (22)) and, thus, difficult to capture syntactically.

22. "If I could have seen my mother alone, I should have gone down on my knees to her and besought her forgiveness". (Charles Dickens, *David Copperfield*)
如果我可以单独看到母亲, 我会向她跪下, 请求她原谅

rúguǒ	*wǒ*	*kěyǐ*	*dāndú*	*kàn-dào*	*mǔqīn*	*wǒ*	*huì*	*xiàng*
if	I	can	alone	see-RES	mother	I	FUT	towards

tā	*guìxia*	*qǐngqiú*	*tā*	*yuánliàng*
she	kneel.down	plea	she	forgive

Further contextual information: "but I saw no one […] during the whole time" / "可是在那段日子里 […]我看不到任何人" *kěshì zài nà duàn rìzi li* […] *wǒ kànbudào rènhé rén*.

Jing-Schmidt relates Chinese counterfactuals to the prominence of the epistemic stance of the viewer. While agreeing in the epistemic nuance of futurity as conveyed by 会 *huì*, and in the modal component of the semantic of future in general (Giannakidou, Mari 2016), we prefer to single out the futurity reading from the epistemic certainty. This choice is based on two main reasons. Firstly, 10% of 会 *huì* occurrenc-

20 In a corpus-based study, Yong (2016) used 13 different hypothetical conjunctions as keywords and, after collecting 3,000 conditionals, disambiguated 245 counterfactuals. Yong's investigation also includes data from a parallel corpus, observing a tendency towards "counterfactual cancellation" occurring after being translated into Mandarin (Yong 2016, 909, 912).

es are in combination with necessity epistemic markers such as 一定 *yídìng* and 准 *zhǔn*, which would confirm classic modal stacking *epistemic necessity* > *futurity* (23). Secondly, even though there are contexts in which 会 *huì* could be interpreted epistemically or even dynamically, as in (23), it could also be argued that without 会 *huì* the event would be anchored to the time of utterance ("I now know what you meant") rather than to the event time ("at that time, I would have known what you meant"). Paraphrasing Condoravdi (2002), it could be said that 会 *huì* sets the reference time in an interval "starting at some past time and extending to the end of time". Therefore, in the composite mechanism of Chinese counterfactuality, 会 *huì* expresses how the world *would have* turned out to be if a state of affairs had obtained.

23. a. "If I had never seen Charles, my father, I should have been quite happy with you". (Charles Dickens, *A Tale of Two Cities*)
"若是我没遇到查尔斯, 爸爸, 我跟你也一定会很幸福的。"
ruòshì *wǒ* *méi* *yùdào* *Chá'ěrsī* *bàba* *wǒ* *gēn* *nǐ*
if I not meet Charles dad I with you
yě ***yídìng*** ***huì*** *hěn* *xìngfú* *de*
also certainly FUT quite happy DE

b. "If you [had sent the message, 'Recalled to Life', again," muttered Jerry, as he turned,] "I *should have* known what you meant, this time". (Charles Dickens, *A Tale of Two Cities*)
"即使你 […] 我也会懂得你的意思的。"
jíshǐ *nǐ* […] *wǒ* ***yě*** ***huì*** *dǒngdé* *nǐde*
even.though you I also FUT understand your
yìsi *de*
meaning DE

Moreover, the data also include examples wherein 会 *huì* cannot be spelled out with any other meaning than futurity. A case in point is (24), which refers to the topic of love commitment. The addressee is telling a third person that, even though Estella's personality had been ruined, had she married him, he would have loved Estella anyway. Our understanding of the sentence in its narrative context is that the speaker's heart here is crying out "I will always love her", without the slightest *epistemic weakening* (Giannakidou, Mari 2017).

24. "I *should have* loved her under any circumstances—Is she married?" (Charles Dickens, *Great Expectations*)
我在任何情况下都会爱她。[她现在结婚了吗?]
wǒ *zài* *rènhé* *qíngkuàng* *xià* *dōu* *huì* *ài* *tā*
I in whatever situation under even FUT love she
tā *xiànzài* *jiéhūn* *le* *ma?*
she now marry PFV Q

In summary, the results suggest that, in past conditionals, 会 *huì* can be considered as the equivalent of *would* in future-in-the-past expressions and that the combination with weak features as the past temporal orientation, the negative presupposition and the first person subject (Ziegeler 2000; Yong 2016) trigger a counterfactual inference.

5.1.1 Past Counterfactual of Wish

The data collected selecting the keyword *should have* in the English-Chinese combination seem to confirm Ziegeler's (2000, 104) claim that: "it is only in past temporal conditionals that a counterfactual construal may be reliably obtained in Chinese". But we also encountered examples where (应)该/当 (*yīng*)*gāi*/*dāng* does not occur in conditional contexts, as in (25). Such examples are labelled as counterfactual wishes, "whereby the subject expresses a desire for things to be different from what they are or were" (Iatridou 2000, 231).

25. "I might have been too reserved, and *should have* patronised her more". (Charles Dickens, *Great Expectations*)
我是太谨小慎微了。我应该多关怀她, 更加地真诚友好
wǒ shì tài jǐnxiǎoshènwēi le wǒ ***yīnggāi*** *duō*
I be too cautios PFV I should more
guānhuái tā gèngjiā-de zhēnchéng yǒuhǎo
take.care she even.more-ly be.sincere be.friendly

Even though it is clear that no linguistic category is independently responsible for the counterfactual interpretation (just as for any other construction, it could be said), the data also show that by adding an appropriate temporal marker such as 那时候 *nàshíhòu* 'at that time', the shift from counterfactual to factual reading can be obtained by replacing 应该 *yīnggāi* with 只好 *zhǐhǎo*; with the latter an AE effect is triggered and the sentence gets a factual reading (26).

26. a. 我是太谨小慎微了。[那时候]我应该多关怀她
wǒ shì tài jǐnxiǎoshènwēi le nàshíhòu wǒ ***yīnggāi***
I be too cautious SFP that.time I should
duō guānhuái tā
more take.care she
'I had been too reserved. At that time, I **should have** taken more care of her'.

b. 我是太谨小慎微了。[那时候]我只好多关怀她，更加地真诚友好
*wǒ shì tài jǐnxiǎoshènwēi le nàshíhòu wǒ **zhǐhǎo***
I be too cautious SFP that.time I can.only
duō guānhuái tā
more take.care she
'I had been too reserved. At that time, I **had to** take more care of her'.

The data from the *E-C English Novels* thus suggest that (i) unlike in anankastic modals, the cluster (应)该/当 (*yīng*)*gāi*/*dāng* is attracted by conditional counterfactual **[tab. 5]** and that (ii) (应)该/当 (*yīng*)*gāi*/*dāng* plays a crucial role in conveying a counterfactual meaning of the 'past wishes' type, as in (26a) and (26b).

5.1.2 Past Counterfactual of Reprimand

More evidence about the contribution of deontic modal in counterfactual environment is found by selecting the keyword *should have* in the *E-C Chinese Novels* (0.181 million words). In this way, we collected 60 tokens from texts originally written in Chinese, and then rendered in English via *should have*. Of the total 60, only 26 have counterfactual interpretation; moreover, in addition to these 26, we also found 5 tokens in which the counterfactual interpretation is present only in the English rendering. Importantly, while processing texts originally written in Chinese and subsequently rendered with the English *should have*, we found that out of 19 tokens including (应)该/当 (*yīng*)*gāi*/*dāng* only 2 are in conditional constructions. Moreover, in this repertoire, the prevailing *nuance* of the deontic tokens is the expression of reproach or reprimand (16 out 20 tokens) that performs the discourse function described by Myhill and Smith, in which "the speaker expresses dissatisfaction with the listener's failure to do something" (1995, 266). In a past context, this discourse function obtained a counterfactual reading, as in (27a). Though mostly addressed to second-person subjects, the reprimand can also be referred to a third party, as in (27b).

27. a. 方先生，你应该知道出典，你不比我们呀! (Qian Zhongshu, *Wei cheng*)
*Fāng xiānshēng nǐ **yīnggāi** zhīdào-chu diǎn nǐ*
Fang mr. you **should** know-RES classics you
bùbǐ women ya
be.unlike us SFP
'Mr Fang, you *should have* recognised the allusion. You're not like us!'
Continuation: 为什么也一窍不通?你罚两杯，来! *Wèishéme yě yīqiàobùtōng? Nǐ fá liǎng bēi, lái!* 'How come you didn't have the faintest idea about it either? You're fined two glasses. Come on'.

b. […] 说鸿渐父亲当初该要求至少两间里有一间大房。
(Qian Zhongshu, *Wei cheng*)
shuō Hóngjiàn fùqīn dāngchū ***gāi*** *yāoqiú zhìshǎo*
tell Hongjian father originally **should** request at.least
liǎng jiān li yǒu yī jiàn dà fang
two CLF in have one CLF big room
'[…] commenting that Hung-chien's father *should have* insisted that at least one of the two rooms be a large one'.

Table 6 Modality distribution, counterfactual tokens (should have, *E-C Chinese Novels*)

Deontic necessity		**Volition + FUT**	**Possibility**	**Anankastic**	**No modal devise**		
	(应)该/当 ***(yīng)gāi/dāng***	想要 ***xiǎng yào***	可以 ***kěyǐ***	须 ***xū***	本(来) ***běn(lái)***	**other**	**Total**
16 *reprimand* counterfactuals 4 *past wish* counterfactuals	20	1	1	1	1	2	26
	73%	**4%**	**4%**	**4%**	**4%**	**12%**	**100%**

The distribution of modal markers in the tokens from the *E-C Chinese Novels* attests to the prominence of (应)该/当 (*yīng)gāi/dāng*, present in 20 out of 26 counterfactual tokens (73%). However, contrary to expectations, there is also one anankastic modal, 须 *xū* in (28), occurring in first-person direct speech, in a prose poem by Lu Xun (死火 *Sǐ huǒ*, Dead Fire, 1925).

28. 倘使你不给我温热, 使我重行烧起, 我不久就须灭亡。 (Lu Xun, *Sǐ huǒ*)
tǎng shǐ nǐ bù gěi wǒ wēnrè shǐ wǒ chóng
if cause you NEG to me warm cause me again
xíng shāo qǐ wǒ bùjiǔ jiù ***xū*** *mièwáng*
do burn INC I not.long then **must** perish
'If you had not warmed me and made me burn again, before long I *should have* perished'.

Other unexpected results found in first-person direct speech will be discussed in § 5.2.2.

5.2 Keyword 2. *Had to*

Selecting *had to*, 410 tokens were retrieved from the two datasets. Once filtered the invalid and irrelevant entries (83 in total), we obtained 327 segments in which *had to* occurs with a modal meaning.

The perfective morphology of *had to* does not necessarily force perfective aspect, being also compatible with habitual, generic, and progressive readings. Moreover, as emphasised by Hacquard (2017), AE is typically neutralised when the modalised proposition is an embedded clause (§ 2.2). Along these lines, each entry was manually tagged as *factual*, *habitual/generic/circumstantial*, *non-factual*, or *non-factual* (*embedded*), as in table 7.

Table 7 Token distribution for the keyword *had to*

	E-C English Novels													
	不得不 *bùdébù*	不能不 *bùnéngbù*	只好 *zhǐhǎo* 只得 *zhǐdé*	非可 *fēikě*	必须 *bìxū*	须 *xū*	得 *děi*	需要 *xūyào*	要 *yào*	(应)该 (*yīng*) *gāi*	不必 *búbì*	可以 *kěyǐ*	**Other**	**Tot**
Factual	21	3	16	3	21	0	20	1	2	0	0	0	25	112
Habitual	0	0	0	1	2	0	4	0	7	0	0	0	4	18
Non-factual	0	0	0	1	6	0	3	0	1	0	0	0	3	15
Non-factual embedded	6	0	0	0	7	1	1	2	13	3	1	1	14	48
Subtotal	27	3	16	5	36	1	28	3	23	3	1	1	46	193

	E-C Chinese Novels													
	不得不 *bùdébù*	不能不 *bùnéngbù*	只好 *zhǐhǎo* 只得 *zhǐdé*	非可 *fēikě*	必须 *bìxū*	须 *xū*	得 *děi*	需要 *xūyào*	要 *yào*	(应)该 (*yīng*)*gāi*	不必 *búbì*	可以 *kěyǐ*	**Other**	**Tot**
Factual	3	3	25	2	2	0	6	1	11	0	0	0	44	97
Habitual	0	0	0	0	0	0	0	0	6	0	0	0	1	18
Non-factual	0	0	0	0	2	0	2	0	10	1	0	1	4	15
Non-factual embedded	0	0	0	0	0	0	0	0	3	2	0	0	2	48
Counterfactual	0	0	0	0	0	0	0	0	0	2	1	0	0	3
Subtotal	3	3	25	2	4	0	8	1	30	5	1	1	51	134
Total	**30**	**6**	**41**	**7**	**40**	**1**	**36**	**4**	**53**	**8**	**2**	**2**	**97**	**327**

We identified 112 tokens having *factual reading*. Excluding 3 tokens with dynamic prominent modals (需要 *xūyào* 'need', and 要 *yào* 'must'), all the other modalised tokens (84 in total) include strong anankasticmodals, such as 只好 *zhǐhǎo* in (29).

29. "he *had to* keep swallowing, he was so like to choke". (Mark Twain, *Tom Sawyer*)
为了嗓子不哽塞住,只好把泪水往肚子里咽。

wèile	*sǎngzi*	*bù*	*gěng*	*sè-zhù*	***zhǐhǎo***
for	throat	NEG	choke	stop-RES	can.only
bǎ	*lèishuǐ*	*wǎng*	*dùzi*	*lǐ*	*yàn*
BA	tears	to	stomach	in	swallow

Habitual entries also encompass *circumstantial habituals* (see § 2.3), that is, a sequence of events is enumerated within a setting previously created, as 'cleaning and scraping', introduced by 要 *yào* with a dynamic necessity meaning, as in (30):

30. "[…] The spoons *had to* be *cleaned* and the frying-pan *scraped*, and the mugs and pudding-basin **swilled** in the lake". (Arthur Ransome, *Swallows and Amazons*)
[…] 有汤匙要清洗, 煎锅要刮洗, 还有杯子及布丁盘浸泡在湖里。
yǒu tāngchí ***yào qīngxǐ*** *jiān-guō* ***yào guā-xǐ***
exist spoon need clean frying-pan need scrape-clean
hái yǒu bēizi jí bùdīng pán jìnpào zài hú lǐ
also exist mugs and pudding basin soak to.be.at lake in

We have included in the habitual class also entries like (31), where an episode is depicted as something happening with a certain regularity (有时 *yǒushí* 'now and then') in a given setting. In languages with rich tense morphology, habitual eventualities are typically rendered with the imperfective; therefore, for double checking the reading, when available, we consulted their Italian translation, and found the indicative imperfective of *dovere* 'must', which is typically used for expressing a habitual ongoing event in the past, such as *doveva* in (31).

31. "You'd see [a muddy sow and a litter of pigs come lazying along the street and whollop herself right down in the way,] where folks *had to* walk around her". (Mark Twain, *The Adventures of Huckleberry Finn*)
有时你会看见 […] 人们走过时必须绕过它走。
yǒushí *nǐ huì kànjiàn* […] *rénmen zǒu guo shí*
sometime you might see people walk pass time
bìxū rào guo tā zǒu
must go.round PASS it walk
Ecco una scrofa coperta di fango che se ne andava a spasso per la via trotterellando con tutta la figliata dei maialini appresso, e la gente ci ***doveva*** **must.IND.IPFV** ***girare*** *attorno*. (It. transl., 221)

The following is an example of *generic habitual*, expressing a generalisation which obtained some time in the past, as that for the duty of "a common servant" in (32).

32. "[But next minute I whirled in on a kind of an explanation how a valley was different from a common servant and] *had to* go to church […] *on account of its being the law*". (Mark Twain, *The Adventures of Huckleberry Finn*)
[…] 他非得上教堂去 […] 因为这是法律上有了规定的。
tā fēiděi shàng jiàotáng qù […] *yīnwèi zhè shì*
he must go church go because this be

fălǜ *shàng* *yǒu-le* *guīdìng* *de*
law on exist-PFV rule DE
Ma un attimo dopo mi sono lanciato in una spiegazione di come un valletto è diverso da un servo qualsiasi, ed ***era costretto***[be.forced to.IND.IPFV] *ad andare in chiesa volente o nolente, e a sedersi con la sua famiglia, perché* ***così voleva***[want.IND.IPFV] ***la legge***. (It. transl., 266)

(33) is an example of *non-factual reading*. Notwithstanding the perfective morphology in English, the full context reveals that the subject hasn't left the island yet (Ransome [1930] 2012, 486); therefore, the entry is tagged as *non-factual*.

33. "Besides, she *had to* say good-bye to the island". (Arthur Ransome, *Swallows and Amazons*)
而且, 她也必须和小岛说再见。
érqiě *tā* *yě* ***bìxū*** *hé* *xiǎo* *dǎo* *shuō* *zàijiàn*
beside she also must with small island say goodbye

A considerable number of entries (tagged as 'others') are not modalised and convey factuality through other means, such as resultative constructions, perfective 了 *le* and the focus marker 才 *cái* 'only then, not until', as in (34).

34. 这是远绕了三十里路才找到的。 (Lu Xun, *Bēn yuè*)
zhè *shì* *yuǎn* *rào* *le* *sānshí* *lǐ* *lù* ***cái***
this be far go.round PFV thirty *li* road only.then
zhǎodào *de*
find DE
'I *had to* go an extra thirty *li* to find it'.

5.2.1 Temporal Feature Bleach in Embedded Position

The eventuality types observed for deontic and anankastic modals in embedded position are in line with predictions (i) and (ii): as an equivalent of *had to*, 应该/当 *yīnggāi*/*dāng* is found only in this environment in which the AE effect is not triggered (cf. Hacquard 2017, 52; see § 2.2). In these cases, modals retain their non-factual orientation and their specific flavour, as for (35), where 该 *gāi* has a fully-fledged deontic reading without shifting to counterfactual reading.

35. "[Nor, did I look towards Wemmick] until I had finished all I *had to* tell". (Charles Dickens, *Great Expectations*)
一直等我说完了我该说的话
yìzhí *děng* *wǒ* *shuō-wán* *le* *wǒ* ***gāi***
straight.to wait I say-RES PFV I should

shuō *de* *huà*
say DE word

Similarly, in the same environment, the strongest anankastic modals occur without triggering AE, as in (36), having a futurity temporal orientation, as confirmed by the past conditional in the Italian translation.

36. I walked the last mile, **thinking** as I went along **of** what I *had to* do. (Charles Dickens, *David Copperfield*)
我边走边考虑我不得不去做的事
wǒ *biān* *zǒu* *biān* ***kǎolǜ*** *wǒ* ***bùdébù*** *qù*
I while walk while think I cannot.but go
zuò *de* *shì*
handle DE matter
Percorsi a piedi l'ultimo miglio pensando, lungo il cammino, a quello che ***avrei fatto***[do.PST.COND] (It. transl., 749)

Another interesting phenomenon is related to the counterfactual reading of 不必 *búbì* in past contexts, as an equivalent of 'would not have had to'. Just as all the modals triggering AE are possibility markers combined with the negation or with the focus marker 只 *zhǐ*, in a similar and symmetric way, the anankastic negation 不必 *búbì* 'there is no need to' seems to yield a counterfactual reading. This is another element pointing to the role of focus-sensitive operators in the expression of factuality and counterfactuality (Sparvoli 2019), a topic that will need to be discussed separately.

5.2.2 Unexpected Data. Backshift in First-Person Narrative

Although the modal distribution in the factual domain meets the prediction, we did find one token in which 要 *yào* marks the anankastic modality and obtains a factual reading – recall that in our prediction the weak anankastic 要 *yào* should convey a non-factual meaning, open to both a factual and counterfactual reading, or a habitual reading. The case in point is (37), in which the event, described in a direct speech first-person narrative context, is only compatible with factual interpretation, as it can be inferred by the continuation ('it produced various effects') and confirmed by the perfective indicative (*passato remoto*) of the Italian *dovere* 'must' (*dovemmo*). Similarly, to the unexpected counterfactual reading of 须 *xū*, (28), it appears that, in first-person direct speech, the reading of necessity modals is elusive.

37. "said Traddles: '[…], [after Sarah was restored], we still *had to* break it to the other eight; [and it produced various effects upon them of a most pathetic nature]'". (Charles Dickens, *David Copperfield*)
特拉德尔说道, "[…] 我们还要告诉其余那八个"
Tèlādé'ěr *shuōdào* […] *wǒmen* *hái* ***yào*** *gàosù* *qíyú*
Traddler say we still must inform the.others
nà *bā* *ge*
that eight CLF
Protestò Traddles: "[…] *Quando Sarah si fu ripresa, dovemmo*[must.IND.IPFV] *affrontare le altre otto".*
(Charles Dickens, *David Copperfield*, It transl., 563)

Another unexpected behaviour, again found in a first-person narrative context, is shown in (38) where 非得 *fēiděi* 'must' gets a non-factual interpretation.

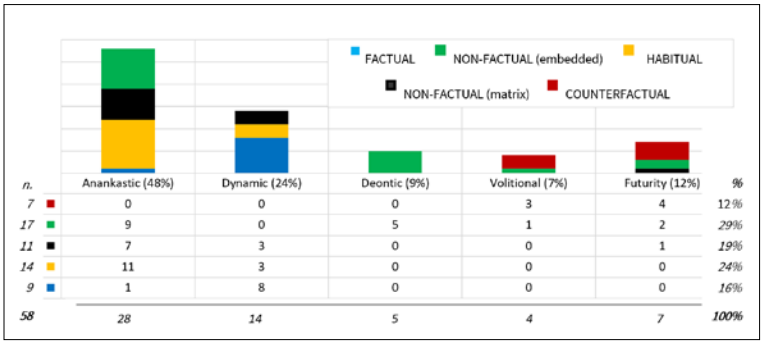
38. "We'd GOT to find that boat now – *had to* have it for ourselves". (Mark Twain, *The Adventures of Huckleburry Finn*)
我们得把那条小船找到, 马上找到——非得找来给我们自己用。
wǒmen *děi* *bǎ* *nà* *tiáo* *xiǎochuán* *zhǎodào*
we have.to BA that CLF boat find
mǎshàng *zhǎodào* ***fēiděi*** *zhǎo* *lái* *gěi*
immediately find must find come to
wǒmen *zìjǐ* *yòng*
we REFL use
Ora davvero ***dovevamo***[must.IND.IPFV] *trovare quella barca – per noi stessi.*
(It. transl., 114)

These phenomena, observed in first-person narrative contexts, could be interpreted as a temporal backshift of the speaker viewpoint. More precisely, in a modalised context, the evaluation of necessity is set back at a past time, that is, in (38), before finding the boat. Along these lines, the AE effect stemming from the strong anankastic is neutralised and the event is described as an ongoing state – as also suggested by the imperfective (*imperfetto*) of the Italian *dovere* 'must' (*dovevamo*).[21]

21 Two types of backshifts, in the scenarios of *justification for a past action* and in the *narration context*, have been described by Hacquard (2017, 59) with reference to the epistemic modals.

5.3 Distribution of the 要 *yào* Tokens

Before presenting our concluding data, we need to focus on the modal distribution of the 要 *yào* tokens, which surface with five different meanings (see § 2.1). As shown in chart 1, the 要 *yaò* tokens display a set of related behaviours which are consistent with our predictions for the anankastic and with the account by Bhatt (1999), Hacquard (2006) and Tsai (2015) for the dynamic domain. Firstly, the reality status of the segments including 要 *yào* is evenly distributed in all the types of eventualities, with the most frequent occurrences in habitual reading (34% in matrix position and 24% including embedded tokens). Secondly, the factual reading is mainly visible in the dynamic domain (8 out of 9, 89%); in the anankastic contexts, we only have one token, shown in (37). Thirdly, given the past contexts of all the tokens, 要 *yaò* is compatible with the deontic meaning only in embedded position (see § 2.2); finally, 要 *yaò* gets counterfactual reading only when occurring with a volitional or futurity reading, thus confirming the non-factual feature of this weak anankastic modal.

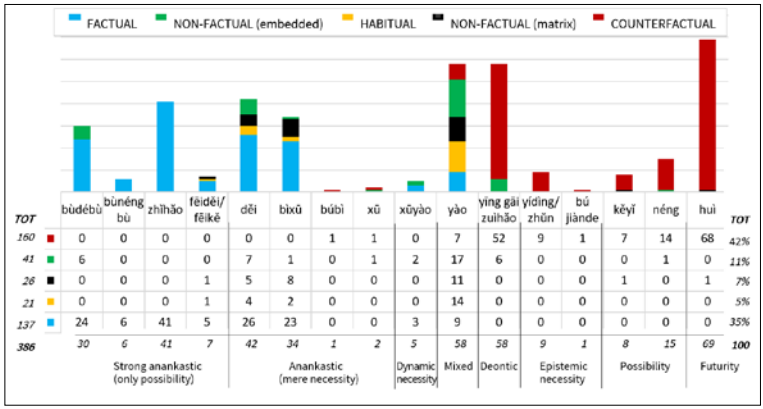


n.	Anankastic (48%)	Dynamic (24%)	Deontic (9%)	Volitional (7%)	Futurity (12%)	%
7	0	0	0	3	4	*12%*
17	9	0	5	1	2	*29%*
11	7	3	0	0	1	*19%*
14	11	3	0	0	0	*24%*
9	1	8	0	0	0	*16%*
58	*28*	*14*	*5*	*4*	*7*	*100%*

Chart 1 Distribution of 要 *yào*: Eventuality types per modal reading (58 tokens)

Finally, by aggregating all the data retrieved with the two keywords *should have* and *had to*, we obtained a tentative picture of the factuality reading of 386 tokens including Chinese modals, shown in chart 2.[22] By including also modals in embedded position, we could observe that, consistent with what was anticipated in § 2.2, in such an environment strong anankastic modals do not have implicative reading, as in (36), while deontic modals retain their meaning without shift-

22 It should be noted that the data displayed in Chart 2 are the result of a filtering process: from the total of 795 tokens, we excluded 268 non-relevant tokens and, from the remaining 567, we also filtered 141 tokens whose Chinese segment does not include a modal marker, thus obtaining 386 tokens including modals in matrix and embedded position.



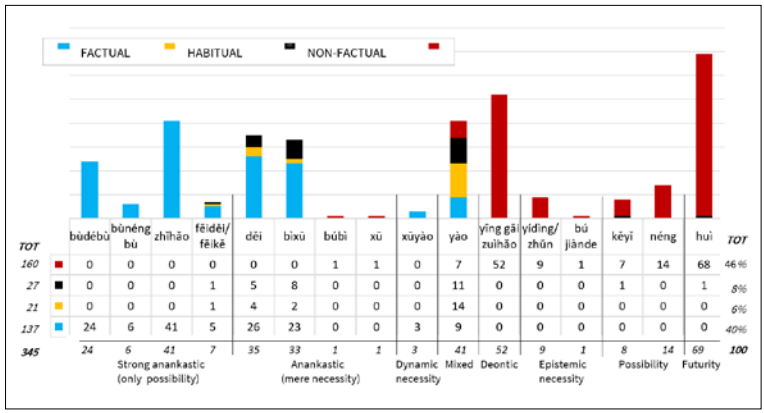


Chart 2 Eventuality types per Chinese modal (386 tokens of *had to* and *should have*)

Chart 3 Eventuality types of Chinese modals in matrix position (345 tokens)

ing to counterfactual reading, as in (35). Finally, to get a clearer picture of the modal distribution per eventuality type, we excluded the tokens in embedded position (41,11%), as seen in chart 3.

6 Conclusion

The results of the aggregated data for modals in matrix position **[chart 3]** show a gradient cline in which the two extreme poles obtain a unique reading: past counterfactual for pure deontic and factual for strong anankastic modals. In terms of factuality, the modal categories here observed are not discrete. Each class presents one mark-

er that partially overlaps with the adjacent modality. For instance, the distribution of the habitual reading ranges from the dynamic 要 *yào* (3.14%) to the anankastic 要 *yào* (11.52%), and can also be seen, albeit less frequently, with other anankastic markers such as 得 *děi* (4.19%) and 必须 *bìxū* (2.10%), and even the strong anankastic 非得 *fēiděi* (1.5%), as seen in (32). Since each modality contains a marker that shares (to a lesser extent) one reading with the adjacent class, the factuality value decreases across a cline from anankastic to deontic modals.

The results confirm our prediction (i): namely, pure deontic markers such as (应)该/当 (*yīng*)*dāng*/*gāi*, alone or in combination with the counterfactual marker 本(来) *běn*(*lái*) are the equivalents of counterfactual *should have*. As shown in chart 2, we can see that, out of all 160 tokens with counterfactual meaning, the deontic is the most prominent full-fledged modality, and it allows for counterfactual reading also when occurring without 本(来) *běn*(*lái*). However, the counterfactual distribution is twofold. On the one hand, deontic markers prevail in the Wish and Reprimand Counterfactuals retrieved by browsing the texts originally written in Chinese **[tab. 6]**. On the other hand, the data retrieved from material originally written in English and then translated into Chinese mainly returned counterfactual conditionals wherein the prominent role is played by the futurity marker 会 *huì* **[tab. 5]**. This latter result supports the constructionist view of Chinese counterfactual conditionals and points to the prominent role of futurity markers (Ziegeler 2000; Jiang 2000; Jing-Schimidt 2017; Liu 2019, among others). It also attests to a future-in-the-past orientation of the counterfactual construal, thus confirming Condoravdi's (2002) account. In this sense, we could say that, in the typical makeup of Chinese counterfactual conditional, the choice between a possibility modal (能 *néng*, 可以 *kěyǐ*), a deontic necessity modal (应)该/当 (*yīng*)*gāi*/*dāng* or a futurity marker (会 *huì*) tells us, respectively, how the world *could*, *should* or *would* have turned out to be if only the given state of affairs had obtained.

Prediction (ii) stipulated that stronger anankastic markers, such as 不得不 *bùdébù* 'cannot but' or 只好 *zhǐhǎo* 'can only', are banned from counterfactual environments. The data confirm this hypothesis, but we must also mention the occurrence of 非得 *fēiděi* with a non-factual reading. The relevant entry occurs in a first-person narrative context, thus it could be interpreted as a backshift, but we also found one token with generic habitual reading; therefore, it appears that, contrary to the predictions, 非得 *fēiděi* patterns more with the mere necessity markers than with the only-possibility ones.

We obtained a problematic result for prediction (iii), positing that 必须 *bìxū* 'have to' preferably gets a factual interpretation. We found one token with a counterfactual 须 *xū* (first-person direct speech), and the data point to a weaker anankastic strength of 必须 *bìxū* com-

pared with 得 *děi*. Prediction (iv), on the other hand, is confirmed. In general, mere necessity modals have a distribution similar to imperfective markers in Italian since they are compatible and commonly found in habitual and non-factual sentences. In sum, the data show a slightly different order in anankastic strength, namely, 只好 *zhǐhǎo* > 不得不/不能不 *bùdébù*/*bùnéngbù* > 非得 *fēiděi* > 得 *děi* > 必须 *bìxū* > 要 *yào*, whereas more data need to be collected for analysing the factuality of 须 *xū*.

Notwithstanding some minor discrepancies with the prediction, the data confirm the crucial role played by the deontic *vs* anankastic contrast in the marking of factuality in Chinese. Lastly, some pedagogical implications may be emphasised with reference to the equivalents of the tensed forms of the Italian *dovere* 'must'. Namely, the two poles getting unique factual (只好 *zhǐhǎo*, 不得不 *bùdébù*) and counterfactual ((应)该 (*yīng*)*gāi* cluster) readings can be mapped onto, respectively, the past indicative and the past conditional of *dovere;* a good candidate as an equivalent of the imperfective of *dovere* can be found in 要 *yào* (especially for direct speech) or 得 *děi*. Finally, the data point to the equivalence between the role of the English *would* and 会 *huì* in past contexts.

Bibliography

E-C Concord (2008). *English Chinese Parallel Concordancer.* `https://corpus.eduhk.hk/paraconc/search`. The Hong Kong Institute of Education. Project leader: Dr. Wang Lixun. Program designers: Chris Greaves, Wang Lixun.

Primary sources

Dickens, C. (1859). *A Tale of Two Cities*. London: Chapman & Hall. Transl. by S. Spaventa Filippi as *Le due città*. Roma: Newton Compton editori, 2012.

Dickens, C. (1872). *The Personal History and Experience of David Copperfield the Younger*. London: Chapman & Hall. Transl. by U. Déttore as *David Copperfield*. Milano: Garzanti, 2012.

Dickens, C. (1881). *Great Expectations*. Boston: Estes & Laurait. Transl. by M.F. Melchiorri as *Grandi speranze*. Roma: Newton & Compton editori, 2010.

Lu, X. 鲁迅 (2005). *Lu Xun quan ji. Di er juan* 鲁迅全集　第二卷 (The Complete Works of Lu Xun Vol. Two). Beijing: Renmin wenxue chubanshe. Transl. by X. Yang and G. Yang as *Selected Works. Vol. One*. Beijing: Foreign Language Press, 1957. `https://archive.org/details/in.ernet.dli.2015.184581/page/n1/mode/2up`.

Lu, X. 鲁迅 [1925] (2005). "Ben yue" 奔月 (The Flight to the Moon). Lu 2005, 370-80. Transl. by Yang, X., Yang, G, 1957, 283-95.

Lu, X. 鲁迅 [1926] (2005). "Si huo" 死火 (Dead Fire). Lu 2005, 200-1. Transl. by Yang, X., Yang, G, 1957, 342-4.

Qian Z. 钱钟书 [1947] (2017). *Wei cheng* 围城. Beijing: Renmin wenxue chubanshe. Transl. by J. Kelly and N.K. Mao as *Fortress Besieged*. Bloomington; London: Indiana University Press, 1979.

Ransome, A. [1930] (2012). *Swallows and Amazons*. London: Random House.
Twain, M. (1876). *The Adventures of Tom Sawyer.* Toronto: Belford Brothers. Transl. by L. Bigiaretti as *Le avventure di Tom Sawyer.* Milano: Giunti Editore, 2007.
Twain, M. (1885). *Adventures of Huckleberry Finn* (*Tom Sawyer's Comrade*). New York: C.L. Webster. Transl. by R. Reim as *Le avventure di Huckleberry Finn*. Roma: Newton Compton editori, 2007.

Secondary sources

Alleton, V. (1984). *Les auxiliaires de mode en chinois contemporain*. Paris: Editions de la Maison des sciences de l'homme.
Bhatt, R. (1999). *Covert Modality in Non-Finite Contexts* [PhD Dissertation]. Philadelphia: University of Pennsylvania.
Blaszack, J. et al. (eds) (2016). *Tense, Mood, and Modality. New Perspectives on Old Questions*. Chicago: University of Chicago Press.
Bloom, A.H. (1981). *The Linguistic Shaping of Thought. A Study in the Impact of Language on Thinking in China and the West*. Hillsdale (NJ): Lawrence Erlbaum Associates.
Carlson, G. (2012). "Habitual and Generic Aspect". Binnick, R. (ed.), *The Oxford Handbook of Tense and Aspect*. Oxford: Oxford University Press, 828-52.
Chen, S. (2012). "The Interaction of Modals and Temporal Marking in Mandarin Chinese". Goto, N.; Otaki, K.; Sato, A.; Takita K. (eds), *Proceedings of GLOW-in-Asia IX 2012*. Mie (Japan): Mie University, article 4.
Chung, S.; Timberlake, A. (1985). "Tense, Aspect and Mood". Shopen, T. (ed.), *Grammatical Categories and the Lexicon*. Cambridge: Cambridge University Press, 241-58.
Condoravdi, C. (2002). "Temporal Interpretations of Modals. Modals for the Present and for the Past". Beaver, D. et al. (eds), *Stanford Papers on Semantics*. Stanford: CSLI Publications, 59-87.
Feng, G.; Yi, L. (2006). "What if Chinese Had Linguistic Markers for Counterfactual Conditionals? Language and Thought Revisited". Sun, R. (ed.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (CogSci 2006). Mahwah, NJ: Lawrence Erlbaum Associates, 1281-6.
Giannakidou, A.; Mari, A. (2016). "Epistemic Future and Epistemic MUST. Nonveridicality, Evidence and Partial Knowledge". Blaszack, J. et al. 2016, 75-117.
Hacquard, V. (2006). *Aspects of Modality* [PhD dissertation]. Boston: Massachusetts Institute of Technology.
Hacquard, V. (2017). "Modals. Meaning Categories?". Blaszack, J. et al. 2016, 45-74.
Iatridou, S. (2000). "The Grammatical Ingredients of Counterfactuality". *Linguistic Inquiry*, 31(2), 231-70. https://doi.org/10.1162/002438900554352.
Jiang Y. 蒋严 (2000). "Hanyu tiaojian ju de weishi jieshi" 汉语条件句的违实解释 (On the Counterfactual Reading of Chinese Conditionals). Zhang B. (ed.), *Yufa yanjiu he tansuo, 10* 语法研究和探索(十) (Studies and Investigations on Chinese Grammar). Beijing: The Commercial Press, 257-79.
Jiang, Y. (2019a). "Ways of Expressing Counterfactual Conditionals in Mandarin Chinese". *Linguistics Vanguard*, 5(3). https://doi.org/10.1515/lingvan-2019-0009.

Jiang, Y. (2019b). "Chinese and Counterfactual Reasoning". Huang C.-R.; Jing-Schmidt, Z.; Meisterernst, B. (eds), *The Routledge Handbook of Chinese Applied Linguistics*. Abingdon; Oxford; New York: Routledge, 276-93.

Jing-Schmidt, Z. (2017). "What Are They Good for? A Constructionist Account of Counterfactuals in Ordinary Chinese". *Journal of Pragmatics*, 113, 30-52. https://doi.org/10.1016/j.pragma.2017.03.004.

Karttunen, L. (1971). "Implicative Verbs". *Language*, 47(2), 340-58. https://doi.org/10.2307/412084.

Kiefer, F. (1987). "On Defining Modality". *Folia Linguistica*, 21(1), 67-94. https://doi.org/10.1515/flin.1987.21.1.67.

Li, R. (2004). *Modality in English and Chinese. A Typological Perspective*. Boca Raton: Dissertation.com.

Liu, H. (2019). "Encoding Counterfactuality in Chinese, Syntactically". *International Journal of Chinese Linguistics*, 6(1), 27-45. https://doi.org/10.1075/ijchl.18002.liu.

Lü S. 吕叔湘 [1942] (1944). *Zhongguo wenfa yaolüe* 中国文法要略 (Outline of Chinese Grammar). Shanghai: Shangwu yinshuguan.

Lü, S. 吕叔湘 (1984). *Xiandai Hanyǔ babai ci* 现代汉语八百词 (Modern Chinese 800 Words). Hong Kong: Shangwu yinshuguan.

Meisterernst, B. (2017). "Modality and Aspect and the Thematic Role of the Subject in Late Archaic and Han Period Chinese. Obligation and Necessity". *Lingua Sinica*, 3, art. 10. https://doi.org/10.1186/s40655-017-0027-2.

Myhill, J.; Smith, L.A. (1995). "The Discourse and Interactive Functions of Obligation Expressions". Bybee, J.; Fleischman, S. (eds), *Modality in Grammar and Discourse*. Amsterdam: John Benjamins, 239-91.

Nevins, A. (2002). "Counterfactuality without Past Tense". Hirotani, M. (ed.), *The Proceedings of NELS* 32. Amherst (MA): GLSA University of Massachusetts, 441-51.

Palmer, F.R. (1990). *Modality and the English Modals*. 2nd ed. London; New York: Longman.

Sparvoli, C. (2012). *Deontico e anankastico. Proposta di ampliamento della tassonomia modale basata sull'analisi dei tratti distintivi dei modali cinesi inerenti dovere e necessità* [PhD Dissertation]. Venice: Ca' Foscari University of Venice. http://hdl.handle.net/10579/1228.

Sparvoli, C. (2015). "Actuality Entailment in the Necessity Domain, a Case for Chinese". *9th Conference of the European Association of Chinese Linguistics* (Stuttgart, 25 September 2015). Stuttgart University.

Sparvoli, C. (2019). "The Temporal Orientation of Chinese Necessity Modals. Some Observations on the Negated Form". *XXXIIèmes Journées de Linguistique d'Asie Orientale* (Paris 29 June 2019). Institut National des Langues et Civilisations Orientales (INALCO).

Tsai, W.-T.D. (2015). "On the Topography of Chinese Modals". Shlonsky, U. (ed.), *Beyond Functional Sequence. The Cartography of Syntactic Structures*, vol. 10. Oxford: Oxford University Press, 275-94.

Wang Y. 王宇婴; Jiang Y. 蒋严 (2011). "Hanyu weishi yi de goucheng yinsu" 汉语违实语义的构成因素 (The Ingredients of Counterfactuality in Chinese). Jiang, Y. 蒋严 (ed.), *Zoujin xingshi yuyongxue* 走近形式语用学 (Approaching Formal Pragmatics). Shanghai: Shanghai Educational Publishing House, 366-412.

Wang, Y. (2013). *The Ingredients of Counterfactuality in Mandarin Chinese*. Beijing: China Social Science Press.

Wu, C.H.-F. (1994). *"If Triangle Were Circles,…" – A Study of Counterfactuals in Chinese and English*. Taipei: The Crane Publishing.

Yong, Q. (2016). "A Corpus-Based Study of Counterfactuals in Mandarin". *Language and Linguistics*, 17(6), 891-915. https://doi.org/10.1177/1606822x16660505.

van der Auwera, J.; Plungian, A.V. (1998). "Modality's Semantic Map". *Linguistic Typology*, 2, 79-124.

von Fintel, K., Iatridou, S. (2007). "Anatomy of a Modal Construction". *Linguistic Inquiry*, 38(3), 445-83. https://doi.org/10.1515/lity.1998.2.1.79.

von Wright, G.H. (1963). *The Varieties of Goodness 1958-1960*. London: Routledge & Kegan Paul. https://www.jstor.org/stable/40071397.

Ziegeler, D.P. (2000). *Hypothetical Modality. Grammaticalisation in an L2 Dialect*. Studies in Language Companion Series 51. Amsterdam; Philadelphia: John Benjamins.

Ziegeler, D.P. (2003). "Redefining Unidirectionality. Insights from Demodalisation". *Folia Linguistica*, 37(Issue Historica-vol-24-1-2), 225-66. https://doi.org/10.1515/flih.2003.24.1–2.225.

Appendices

E-C English Novels

Files included in the consulted corpus.
Wordcount, title, author and translator retrieved from https://corpus.eduhk.hk/paraconc/info.

Title	Author and translator	English words	Chinese characters
Alice in Wonderland 爱丽丝梦游仙境	Lewis Carroll. Translator: not mentioned.	7,393	12,924
A Tale of Two Cities 双城记	Charles Dickens. Translator: not mentioned.	136,382	231,837
David Copperfield 大卫•科波菲尔	Charles Dickens. Translators: Shi Dingle 石定乐, Shi Dingrou 石定柔 (1999). 湖南文艺出版社. Changsha: Hunan wenshu chubanshe.	357,668	579,714
Oliver Twist 雾都孤儿	Charles Dickens. Translator: He Wenan 何文安 (1999). 译林出版社. Nanjing: Yilin chubanshe.	159,045	289,200
Great Expectations 远大前程	Charles Dickens. Translator: Luo Zhiye 罗志野 (2001). 译林出版社. Nanjing: Yilin chubanshe.	186,424	364,840
Pride and Prejudice 傲慢与偏见	Jane Austen. Translator: not mentioned.	126,950	225,307
Sense and Sensitivity 理智与情感	Jane Austen. Translator: not mentioned.	120,333	207,710

Swallows and Amazon 小水手探险记	Arthur Ransome. Translators: Lee Hsingchin 李幸瑾; Wang Lixun 王立勋 (2004). 台湾商务. Taibei: Taiwan shangwu.	99,291	177,225
The Adventures of Tom Sawyer 汤姆. 索亚历险记	Mark Twain. Translators: Cao Xiaohong 曹晓红; Yu Xiaoguang 于晓光 (1999). 大众文艺出版社. Beijing: Dazhong wenyi chubanshe.	70,622	126,651
The Adventures of Huckleberry Finn 赫克尔贝里• 芬历险记	Mark Twain. Translator: Xu Ruzhi 许汝禳 (1998). 译林出版社. Nanjing: Yilin chubanshe.	111,209	199,525
Dog Tale 一只小狗的故事	Mark Twain. Translator: Zhang Yousong 张友松 (2010). 光明日报出版社. Beijing: Guangming ribao chushe.	4,355	8,161
The $30,000 Bequest 三万元遗产	Mark Twain . Translator: not mentioned.	10,977	19,818
	Total	**1,390,649**	**2,442,912**

E-C Chinese Novels

Title	**Author and translator**	**English words**	**Chinese characters**
绿化树 *Mimosa*	Zhang Xianliang 张贤亮. Translator: not mentioned.	46,625	102,820
芙蓉镇 *A Small Town Called Hibiscus*	Gu Hua 古华. Translator: Gladys Yang (2015). Beijing: Foreign Languages Press.	66,346	136,974
鲁迅小说 *Lu Xun's Novels*	Lu Xun (鲁迅) Translator: not mentioned.	66,027	94,464
围城 *Fortress Besieged*	Qian Zhongshu 钱锺书. Translators: Jeanne Kelly, Nathan K. Mao (1979). Bloomington; London: Indiana University Press.	220,000	219,996
	Total	**398,998**	**554,254**

Corpus-Based Research on Chinese Language and Linguistics
edited by Bianca Basciano, Franco Gatti, Anna Morbiato

Pope Francis' *Laudato Si'*: A Corpus-Based Study of Modality in the English and Chinese Versions

Adriano Boaretto
Università Ca' Foscari Venezia, Italia

Erik Castello
Università degli Studi di Padova, Italia

Abstract This paper compares the use of modal expressions in the English and Chinese versions of Pope Francis' Encyclical Letter *Laudato Si'* (2015). It explores the Encyclical Letter as a corpus through the study of word lists and parallel concordance lines. The research also benefits from the close parallel reading of extracts from the two versions. It focuses on the semantic areas of prediction/volition/intention, lack of possibility/ability/permission and obligation. The results confirm predictable parallel expressions (e.g. *will* and 会 *huì*, *cannot* and 不能 *bùnéng*, *be called to* and 召 *zhào*) and bring to light less predictable renderings – e.g. *zero* (in English) and 会 *huì*, *cannot* and 无法 *wúfǎ*, the noun *vocation* and 召 *zhào*. They also suggest that some translation choices are due to the translator's attempt to make the text explicit and to adapt it to the target culture.

Keywords Chinese-English modality. Corpus-based study. Explicitation. Laudato Si'.

Edizioni Ca'Foscari

Sinica venetiana 6
e-ISSN 2610-9042 | ISSN 2610-9654
ISBN [ebook] 978-88-6969-406-6 | ISBN [print] 978-88-6969-407-3

Peer review | Open access
Submitted 2020-03-27 | Accepted 2020-10-14 | Published 2020-12-21

DOI 10.30687/978-88-6969-406-6/006

1 Introduction[1]

This paper explores *Laudato Si'*, Pope's Francis' second Encyclical Letter, issued in 2015. Novelist and essayist Amitav Ghosh (2016) compares it to the *Paris Agreement on Climate Change*, which was also released in 2015 by diplomats and delegates from the United Nations. He claims that both texts "occupy a realm that few texts can aspire to: one in which words effect changes in the real world" (Ghosh 2016, 150). They are both founded on the results of research produced by climate science, yet they diverge sharply in linguistic terms. The Encyclical is "remarkable for the lucidity of its language and the simplicity of its construction", while the *Paris Agreement* is "highly stylised in its wording and complex in structure" (Ghosh 2016, 151). Ghosh goes on to say that "mass organisations will have to be in the forefront of the struggle. And of such organisations, those with religious affiliations possess the ability to mobilise people in far greater numbers than any others" (Gosh 2016, 160). The Papal document thus appears to be particularly meaningful and worth investigating from a linguistic perspective: it lucidly discusses climate change issues and has the potential to effectively put forward insightful religious, cultural, social and economic lines of action against it.

The recent branch of linguistics called "ecolinguistics" attempts to raise awareness on "discourses that have (or potentially have) a significant impact not only on how people treat other people, but also on how they treat the larger ecological systems that life depends on" (Stibbe 2014, 118). In line with this approach, Castello and Gesuato (2019) explore the language of the English version of *Laudato Si'* using corpus-based methods. Among their findings is the frequent use of modality in the text, with the modal verbs *must*, *cannot*, *need*, *needs*, *should*, *can* figuring among the keywords they obtained. They also identified a number of other expressions of modality, including *fail to* and *be called to*. They claim that

> the modal items identified and their patterns of occurrence suggest that *Laudato Si'* is mainly oriented towards the expression of deontic (participant external) modality, qualifying the degree of human involvement in and responsibility for the well-being of the planet. Additionally, […] the text draws attention to the possibility for humankind to perceive and become aware of the planet's present condition and future prospects. (Castello, Gesuato 2019, 139-40)

1 For academic purposes, Adriano Boaretto is responsible for §§ 1, 2, 3, 6.2 and 6.3; Erik Castello is responsible for §§ 4, 5, 6.1, 6.4 and 7.

The notion of modality has been dealt with from various theoretical perspectives, including the functional, the formal syntactic and the semantic ones (see Nuyts, van der Auwera 2016 for an overview). This paper adopts a semantic approach to this phenomenon, and refers to the domains of 'epistemic' modality and 'non-epistemic' modality, which can in turn be subdivided into "participant-external modality" and "participant-internal modality" (Chappell, Peyraube 2016, 300). It also takes into account the closely related notion of negation (Nuyts 2016, 3-4). As is well known, it is often difficult to decide which sense should be attributed to a given English modal item in a sentence (Huddleston 2002, 177). For example, the modal verb *can* (and its negative counterpart *cannot*) can be used epistemically to make suppositions, participant-externally to express (lack of) permissions, or participant-internally to indicate (lack of) ability. Analogously, in Chinese most modal verbs display a high degree of polysemy, e.g. the modal verb 能 *néng* can indicate, among others, the ability of the subject (non-epistemic participant-internal modality) or the permission given to somebody due to circumstances (non-epistemic participant-external modality) (Chappell, Peyraube 2016, 299-300). During the translation process, translators have to make out the correct interpretation of the meaning of a given modal marker and then choose the most suitable item or a construction from those available in the target language that conveys it.

Like all encyclical letters, *Laudato Si'* is available in different languages. Teubert, who studies a corpus of papal documents, suggests that a linguistic comparison of the various versions of an encyclical letter "can be a fruitful exercise in itself" (2007, 95), which is exactly what the present paper sets out to do with reference to the English and the Chinese versions of *Laudato Si'*. A parallel close reading of them suggests that the Chinese version was translated from the English one,[2] and, consequently, that the former is highly likely to present features of translated language, such as explicitation and simplification (e.g. Laviosa 2002). From a methodological perspective, this paper adopts a corpus-based translation approach (e.g. Xiao, Wei 2014) for the investigation of a selection of modal expressions in the English version vis-à-vis the Chinese one, including the 'quasi-modal' verb *be called to*. It attempts to identify and categorise the "meaningful correspondences" (Tognini-Bonelli 1996, 199) between the instances of the selected English and Chinese modal items, and to explore the semantic space that they cover. Finally, it investigates the hypothesis that at least some of these translation choices might represent cases of explicitation of the modal meanings expressed in the source text.

2 The Authors have read the English, Italian and Chinese versions of the Letter, and noticed that many parts of the Chinese version are more adherent to the English one.

§ 2 provides a brief introduction to *Laudato Si'*, while § 3 presents the concept of modality and its realisation in English and Chinese. § 4 introduces corpus-based translation studies of English and Chinese, and § 5 describes the features of the two texts and how they are investigated as corpus data. Finally, § 6 discusses the results, starting from general observations and then focusing on three areas of modality and a selection of modal items.

2 The Encyclical Letter *Laudato Si'*. Religious Writing about Ecological Issues

Jorge Mario Bergoglio, Pope Francis, was elected Pope of the Catholic Church on 13 March 2013. He published his first Encyclical Letter, *Lumen Fidei*, on 29 June 2013 and issued his second and latest one, *Laudato Si'*, on 24 May 2015. *Laudato Si'* is a complex document, probably resulting from the writing of several authors (Tilche, Nociti 2015, 5) writing in different languages, which is the case for most papal texts. Encyclicals are normally released in one modern language, mainly French, German or Italian, while their Latin version, the authoritative one, is usually produced at a later stage (Teubert 2007, 95). *Laudato Si'* is currently available in fourteen languages, including Italian, Latin, English, and Chinese.[3] The Chinese translation is released both in simplified characters, Chinese (China), and in traditional characters, Chinese (Taiwan).

Laudato Si' consists of a Preamble, six chapters and two final prayers, "A Prayer for Our Earth" and "A Christian Prayer in Union with Creation". Chapters one, three, four and five appear to have a stronger economic and ecological slant, while chapters two and six share a more religious and pastoral thrust (Castello, Gesuato 2019, 134). The Preamble provides an overview of the Pope's thought, of Saint Francis' view of beauty and fraternity, and of the ethical and spiritual roots of environmental problems. It calls for a spiritual change of humankind and expresses the Pope's openness to a dialogue with science (Tilche, Nociti 2015, 2). The first chapter draws a picture of the problems *our common home* (Chinese: 我们的共同家园 *wǒmen de gòngtóng jiāyuán*)[4] is now facing, including the changes affecting humanity and our planet, the *throwaway culture* (Chinese: 丢弃文化 *diūqì wénhuà*), and *climate as a common good* (气候乃是大众福

3 The versions are available on the Vatican website in the following languages: Arabic, Belarusian, Chinese (China), Chinese (Taiwan), English, French, German, Italian, Latin, Polish, Portuguese, Russian, Spanish, Ukrainian: `http://www.vatican.va/content/francesco/en/encyclicals.html`.

4 Simplified Chinese characters and the *Pinyin* romanisation system have been used throughout the article.

社 *qìhòu nǎi shì dàzhòng fúzhǐ*). Subsequently, it describes some features of climate change using "correct but non-scientific language" (Tilche, Nociti 2015, 3), such as the pressure on water resources and the loss of biodiversity, and finally it addresses the human and social dimension of the ecological crisis. The second chapter re-reads biblical texts concerning the relationship between God, humankind and nature. It focuses on the mystery of the universe and on the conception of creation as a gift from God. It ends up claiming that creation is bound up with the mystery of Christ. The third chapter explores the ultimate causes of the ecological crisis with reference to philosophy and science and to the global phenomena known as technocratic paradigm and power. It then looks at the consequences of modern anthropocentrism, that is practical relativism, at the need to protect employment, and finally considers new biological technologies. The fourth chapter gets to the core of Pope Francis's message and proposes *integral ecology* (整体生态学 *zhěngtǐ shēngtàixué*) as the fruitful combination of scientific, environmental, economic and social perspectives on ecology. The Pope also puts forward the concepts of *cultural ecology* (文化生态学 *wénhuà shēngtàixué*) and the *ecology of daily life* (日常生活的生态学 *rìcháng shēnghuó de shēngtàixué*), in view of the *principle of the common good* (公益原则 *gōngyì yuánzé*) and of the need of justice between the generations (Spadaro 2015). The fifth chapter claims that a series of patterns of dialogue should be pursued with a view to escaping the current spiral of self-destruction: dialogue in the international community, dialogue for new national and local policies, dialogue and transparency in decision-making, dialogue between politics and economy for human fulfilment, dialogue between religions and science. The sixth chapter posits that an *ecological conversion* (生态皈依 *shēngtài guīyī*) is needed. People should change their lifestyle and overcome selfishness. They should be educated for the covenant between humanity and the environment, which should bring them joy and peace, reflected in a balanced lifestyle and a deeper understanding of life. The Eucharist and the day of rest should motivate people's concerns for the environment.

3 Modality in English and Chinese

Modality is a semantic category which is "centrally concerned with the speaker's attitude towards the factuality or actualisation of the situation expressed by the rest of the clause" (Huddleston 2002, 172-3). By contrast, mood is a

> formally grammaticalized category of the verb which has a modal function. [Mood is] expressed inflectionally, generally in distinct sets of verbal paradigms, e.g. indicative, subjunctive, optative, im-

> perative, conditional etc., which vary from one language to another. (Bybee, Fleischmann 1995, 2)

English modality has been studied extensively from various perspectives, including the semantic (e.g. Lyons 1977; Bybee, Fleischman 1995; Palmer 2001; Portner 2009), the descriptive (e.g. Quirk et al. 1985; Huddleston 2002) and the functional one (e.g. Halliday 1976, 2004). This phenomenon has also been addressed in the field of Chinese linguistics, and various proposals have been put forward to categorise Chinese modality (e.g. Tsang 1981; Peng 2007; Tang 2000; Chappell, Peyraube 2016). Scholars have also explored Chinese modality in relation to English modality from the contrastive and typological perspective (e.g. Li 2004; Hsieh 2005) and the functional perspective (e.g. Chen 2017). A large number of studies have also availed themselves of corpus-based methods (Coates 1983; Biber et al. 1999; Carter, McCarthy 2006) for the study of modality.

From the semantic perspective, von Wright (1951) breaks down modality into "epistemic", "deontic", and "dynamic" modality. Epistemic modality is concerned with "the speaker's attitude to the truth-value or factual status of the proposition", deontic modality "relates to obligation or permission emanating from an external source", while dynamic modality "relates to the ability or willingness which comes from the individual concerned" (Palmer 2001, 9-10). This terminology has been frequently elaborated and revised. For example, Chappell and Peyraube (2016, 299-300) follow van der Auwera and Plungian's (1998) framework and distinguish between epistemic and "situational" (non-epistemic) modality. More specifically, they divide situational modality into "participant-internal" and "participant-external". Furthermore, they associate epistemic modality with the semantic fields of possibility, probability, certainty. and necessity, participant-external modality with possibility, permission, obligation, requirement, and necessity, and, finally, participant-internal modality with ability, willingness, volition, and intention. The subdivision between participant-internal and participant-external modality partly overlaps with that between dynamic and deontic modality (e.g. Palmer 2001), yet in Chappell and Peyraube's (2016) framework the main discriminating factor lies in whether the modal meaning is related to the subject of the sentences or to an external participant. Chappell and Peyraube's (2016) semantic categorisation is reproduced in table 1:

Table 1 Categories for modality markers (slightly adapted from Chappell and Peyraube 2016, 300)

Epistemic	**Situational (non-epistemic)**	
	Participant-external	**Participant-internal**
possibility	possibility	
	permission	ability
probability	obligation	willingness
certainty	requirement	volition
necessity	necessity	intention

In English, modality is primarily expressed by core modal auxiliaries (e.g. *must*, *will*, *should*) and marginal auxiliaries or quasi-modals (e.g. *have to*, *need to*, *be bound to*) (Quirk et al. 1985, 237). English modal auxiliaries display special features, including the fact that they have no *-s* form for the third person singular (e.g. **cans*, **musts*), take negation directly (e.g. *can't*/*cannot*, *mustn't*), do not admit co-occurrence (e.g. **may will*), and take inversion without *do* (e.g. *can I?*, *must I*) (Coates 1983, 4). Quasi-modals do not share these features with modal auxiliaries and are much closer to lexical verbs. Modality is also conveyed by "lexical modals", a broad category comprising items that do not belong to the class of auxiliary verbs. It includes adjectives (e.g. *possible*, *necessary*), adverbs (e.g. *perhaps*, *possibly*), lexical verbs (e.g. *hope*, *want*), and nouns (e.g. *possibility*, *necessity*) (Huddleston 2002, 173).

Chinese expresses modality by means of grammatical, lexical and syntactic devices. It shares with English the use of modal auxiliary verbs (variously named, e.g. 情态助动词 *qíngtài zhùdòngcí* or 能愿动词 *néngyuàn dòngcí*) and lexical modals, such as modal adverbs (态度副词 *tàidù fùcí*). It also employs the so-called modal particles (语气助词 *yǔqì zhùcí*) and the potential construction, also known as potential verb compound (Hsieh 2005, 38; Chappell, Peyraube 2016, 297, 312-14).

The category of modal auxiliary verbs[5] include: 能 *néng*, 能够 *nénggòu*, 可以 *kěyǐ*, 得 *dé*, 会 *huì*, and 可能 *kěnéng*,[6] used to express possibility, permission and ability; 要 *yào*, 应 *yīng*, 应该 *yīnggāi*, 应当 *yīngdāng*, 该 *gāi*, 当 *dāng*, 得 *děi*, 需要 *xūyào*, 必须 *bìxū*, and 须要 *xūyào* to express obligation and necessity; and 要 *yào*, 想 *xiǎng*, 想

5 The status of Chinese modal auxiliary verbs is debated in the literature. Tang (2000), for example, does not even ascribe them to the category of auxiliary verbs and calls them 情态动词 *qíngtài dòngcí* 'modal verbs'.

6 The status of 可能 *kěnéng* is controversial. Some authors consider it an adverb (Li, Thompson 1983, 168), yet some others consider it a modal verb (Li 2004, 138).

要 *xiǎngyào*, 愿 *yuàn*, 愿意 *yuànyì*, 肯 *kěn* to express volition (intention) (e.g. Chao 1968, 731-48; Chapell, Peyraube 2016, 301-2; Abbiati 2014, 213-21).

Adverbs such as 竟 *jìng*, 居然 *jūrán*, 究竟 *jiūjìng*, 或许 *huòxǔ*, and 显然 *xiǎnrán* belong to the category of modal adverbs (e.g. Chao 1968, 780-90; Li, Thompson 1983, 267-8). Modal or sentence particles (e.g. 吗 *ma*, 呢 *ne*, 啊 *a*, 吧 *ba*, 了 *le* and 嘛 *ma*) are morphemes uttered in the neutral tone occurring at the end of an utterance with the aim of adding modal and attitudinal meanings to it (Chao 1968, 796; Abbiati 2014, 58). Finally, potential constructions (verb compounds) derive from both resultative and directional verb compounds and can indicate either ability or possibility, as can be seen from example (1):[7]

1. 听得懂
   tīng *de* *dǒng*
   hear POT understand
   'can understand'

Li and Thompson (1981, 182-3) suggest a series of functional correspondences between Chinese and English modal auxiliaries. Sparvoli (2012, 209) elaborates on their proposal, and puts forward a possible mapping of modal Chinese/English pairs of auxiliaries onto van der Auwera and Plungian's (1998) semantic categories. Table 2 is an adaptation of Sparvoli's list of correspondences, and will be the starting point for the study presented in this paper. Differently from Sparvoli (2012), the categories "participant-internal volition, intention", "epistemic possibility" and "epistemic necessity, certainty" have been included. Also, a wider repertoire of Chinese and English modal auxiliaries is presented, as they are relevant to this study.[8]

Table 2 Hypothesised correspondences between a selection of English and Chinese modal auxiliaries

English	Chinese	Categories
will, shall, be going to	会 *huì*	Epistemic possibility, probability Participant-internal willingness, intention

7 The glosses used in this paper follow the general guidelines of the Leipzig Glossing Rules. Additional glosses include: DIR = 'directional complement or verb'; DISP = 'disposal construction marker'; LIG = 'ligature' (genitive, relative clause or attributive marker); P = 'particle'; POT = 'potential marker'.

8 The Chinese modal 要 *yào* has been added, although Li and Thompson (1981), for example, do not include it into their list of modal auxiliaries. The English modal verb *can*, the quasi-modal *be called to*, and its hypothesised Chinese equivalent 召 zhào have also been included.

can, be able to	会 *huì*, 能 *néng*, 能够 *nénggòu*, 可以 *kěyǐ*	Epistemic possibility Participant-internal ability
can, may	能 *néng*, 可以 *kěyǐ*	Participant-external possibility
must, should, have to, need to, ought to, be called to	要 *yào*, 应 *yīng*, 应该 *yīnggāi*, 应当 *yīngdāng*, 该 *gāi*, 需要 *xūyào*, 召 *zhào*	Epistemic necessity, certainty Participant-external necessity, obligation, requirement
want to, will/shall	要 *yào*, 想 *xiǎng*, 想要 *xiǎngyào*, 愿 *yuàn*	Participant-internal volition, intention

From table 2, the polysemous nature of some auxiliary verbs is apparent, as they straddle one or more semantic categories. This is the case of *will* and 会 *huì*, *can* and 能 *néng*, 可以 *kěyǐ* and 要 *yào*.

The English modal auxiliary *will* can alternatively indicate epistemic possibility/probability or participant-internal willingness and intention (Coates 1983, 170-1; Huddleston 2002, 188-91). *Shall* can be used with first person subjects either singular or plural, as an alternative of *will* to ask for the intention or volition of the addressee. Also, in more formal and prescriptive contexts, *will* and *shall* can convey obligation (participant-internal modality) (Coates 1983, 185-6). In this last sense, *will*/*shall* correspond to the Chinese auxiliary 要 *yào* and to other verbs indicating participant-internal volition/intention.

The Chinese modal 会 *huì* can take on three main meanings: 1) 'know how to, have the ability to'; 2) 'be good at'; 3) 'there is the possibility (that...)' (our translation) (Lǚ 2004, 278-9). In the first two senses it overlaps semantically with the English auxiliary core modal *can* and the quasi-modal *be able to*, and indicates participant-internal ability, while in the third sense it covers part of the semantic area of *will* and *shall*.

The modal auxiliary *can* has the potential to express epistemic possibility, participant-internal ability or participant-external possibility and permission, and thus it overlaps semantically with the Chinese auxiliaries 能 *néng* and 可以 *kěyǐ*. Interpreting whether the use of *can* is epistemic, participant-internal or participant-external can be hard in some contexts, as suggested, for example, by Biber et al. (1999, 491-3) with regard to academic prose.

Finally, as seen above, not only can 要 *yào* be employed to convey participant-internal volition or intention, but also participant-external necessity, obligation, and requirement, and thus corresponds to, for instance, English *must*, *should*, and *need to*.

As noticed by Coates (1983, 20), the negative forms of some English modal auxiliaries are unavailable in the language, and alternative ones have to be used to make up for them. For example, in British English the negative form of epistemic *must* is *cannot* and not **mustn't*. This phenomenon, also known as 'suppletion', can be found in Chinese as well, in that some modal auxiliaries have a negative

counterpart which differs from the positive one for all or some of their meanings (Sparvoli 2012, 171). For example, 可以 *kěyǐ* takes on the negative forms 不能 *bù néng*, 不行 *bù xíng*, 不成 *bù chéng* or 不值得 *bù zhídé* when it indicates negative participant-external possibility. The auxiliaries 要 *yào*, 必须 *bìxū* and 得 *děi* are negated by 不用 *búyòng* or 不必 *búbì* in contexts in which they express participant-external necessity. Furthermore, the verb 要 *yào*, indicating participant-internal volition and intention, is negated with 不想 *bù xiǎng*, 不会 *bú huì*, or 不可能 *bù kěnéng* (Abbiati 2014, 213-20).

In spite of these shared functional and semantic aspects, many authors have pointed out typological differences between modality in English and Chinese, especially from the morphosyntactic perspective (e.g. Li, Thompson 1981; Tang 2000; Li 2004). In this respect, Li claims that:

> modal verbs in English and Chinese are very different things [...] They constitute a grammatical category belonging to "auxiliary verbs". However, apart from the component of the modals, the auxiliary verbs of the two languages share little resemblance. The "helping" functions of English auxiliaries in aspect, phase, and voice do not exist with Chinese auxiliaries. "Auxiliary verb" is a suitable term for the intermediate category between verbs and modal verbs in English, but not for that in Chinese. Chinese has no auxiliary verbs in the English sense. (2004, 316)

4 Corpus Linguistics for the Study of English and Translated Chinese

Language corpora are naturally occurring language data, stored as computer files. An important distinction can be drawn between general corpora, representing a language as a whole, and specialised corpora, focusing on a specific language variety. Depending on the type of language under examination and the research questions the corpus is designed to address, one might need to restrict the number of texts that make up a corpus (Baker 2010, 12-14). Pierini (2015), for example, carries out a study of the translation of English compound adjectives from English into Italian and chooses to study only one text, Stephen King's novel *Under the Dome* and its Italian translation. She claims that while it is true that "a small corpus provides a partial insight into a phenomenon" it "can be scanned manually so that the collection of data does not leave out any [...] pattern" (Pierini 2015, 22). Corpus linguistics can be defined as a series of methods, techniques, and processes for the investigation of language corpora, including the analysis of word frequencies, concordances, collocations, keywords and the dispersion of words and keywords (Baker 2010, 5, 19-30).

Some studies have applied corpus-based methods to the investigation of translated language. These are known as Corpus-Based Translation Studies and are based on bilingual parallel corpora and comparable corpora of native and translated texts. This research attempts "to uncover evidence to support or reject the so-called translation universal hypotheses" (Xiao, Wei 2014, 3), including the existence of translation phenomena such as explicitation and simplification (e.g. Laviosa 2002). Explicitation, in particular, is "an overall tendency to spell things out rather than leave them implicit in translation" (Baker 1996, 180).

Xiao (2010) examines features of translated Chinese emerging from the study of a corpus of translated texts compared to original Chinese texts. His analysis reveals the presence of "properties which are specific to English-to-Chinese translation due to translation shifts", including significantly lower lexical density and a lower proportion of lexical words over function words than in native Chinese (Xiao 2010, 29). Xiao and Dai reevaluate the "English-based" translation universal hypotheses and suggest that:

> some [hypotheses] (e.g. explicitation) are supported in Chinese while others are not fully supported (e.g. simplification) […]. More specifically, translational language is more explicit semantically, lexically, grammatically and logically. But simplification is not a pure, simple phenomenon in that translated texts may be simpler in some aspects but more complicated in others vis-à-vis comparable native texts. (2014, 50)

Xiao and Wei call for further corpus-based translation and cross-linguistic studies of "genetically distant languages such as English and Chinese" (2014, 5), as they can have important implications for linguistic theorisation.

Corpus-based translation studies can also have practical aims and implications. Lian and Jiang (2014), for example, examine the use of modality in a parallel corpus of Chinese laws and regulations of international exchanges and their translations into English. Such legal texts have become increasingly important in our globalised world, and more attention should be paid to their translation, as translators tend to use the "modal operator" *shall* excessively and to misuse other English modal operators. Furthermore, they tend to overuse synonymous words to avoid repetitions, but in this way they violate the principles of consistency, accuracy, and authority of the law (Lian, Jiang 2014, 502).

Finally, corpus linguistics methodologies have also informed the study of the writings of the Catholic Church. Teubert (2007), for instance, examines concordances extracted from a corpus of encyclical letters and other texts about the social doctrine of the Church and explores the evolution of the meaning of concepts such as 'natural law', 'human rights', and 'property' over time. The author claims

that not only can corpus linguistics help to identify the regularities of language use, but also to observe the construction of social reality in a given discourse at a given time (Teubert 2007, 89).

5 The Data and the Analysis

The English and the Chinese versions of the Encyclical Letter were downloaded from the Vatican website as PDF files and converted into .txt files. We tokenised the Chinese text with the aid of the software *SegmentAnt* (Anthony 2018), as Chinese is written as running strings of characters without spaces delimiting words (Xiao 2010, 14). We checked the output of the software manually and made some changes to it. For example, Some sets of characters had been treated by the software as single units, while for semantic and syntactic reasons we decided to separate them and put a space between them, e.g. 一些 *yī xiē*, 就是 *jiù shì*, 不可 *bù kě*, 不能 *bù néng*.The first string is composed of a numeral followed by a classifier and the remaining ones of an adverb followed by a verb. By contrast, we decided to write idiomatic expressions with no space between their characters, e.g. 若无其事 *ruòwúqíshì* 'as if it did/does not concern him'. In dubious cases, we consulted the 现代汉语词典 *Xiandai Hanyu Cidian - The Contemporary Chinese Dictionary* (2014). Once the two versions were ready for analysis, we processed them by means of the software *AntConc* (Anthony 2019), and obtained word lists and concordances for a selection of both English and Chinese modal expressions. The word lists provided information about the frequency of all the words in each corpus, while concordances presented all the occurrences of a given modal item within their linguistic contexts.

We first identified parallel expressions that encode modal meanings in the two languages (cf. Tognini-Bonelli 1996, 198). Subsequently, we attempted to "locate meaningful correspondences and build up a network of semantic relations across the two languages"; however, as is often the case, some "mismatches [came] to light […]: these are just as important as the similarities between the two languages" (Tognini-Bonelli 1996, 199). Using an Excel spreadsheet, we matched each line in a concordance with the corresponding "co-text" in the other version of the Letter and inserted the parallel expressions into to two adjacent columns for further analysis. This procedure provided us with a framework for the study of translation equivalence in the English and in the Chinese version with regard to modality.

As can be seen from table 3, the number of word types (i.e. unique words) and word tokens (i.e. running words) in the two versions is similar, and so is the type/token ratio, that is the ratio between the number of types and the number of tokens (Xiao 2010, 17).

Table 3 Quantitative data about the English and the Chinese version of *Laudato Si'*

Laudato Si'	**English**	**Chinese**
Word types	4,846	4,861
Word tokens	36,911	35,547
Type/token ratio	7.6	7.3

The two research questions explored in this study are:

1. Which are the most important 'meaningful correspondences' of a selection of the most frequent English modal expressions in the two versions, and how can they help understand the semantic space covered by each expression in *Laudato Si'*?
2. Can any differences in the use of modal items be detected which might not only be due to typological contrasts between the two languages but also, or exclusively, to attempts to make the target text more explicit?

6 An Analysis of Modality in *Laudato Si'*

This section first looks at the overall use of modality in the English and Chinese versions of *Laudato Si'* (§ 6.1). It then zooms in on the use of a selection of frequently occurring modal expressions indicating epistemic possibility and probability and participant-internal willingness, intention (§ 6.2), lack of participant-internal ability or participant-external possibility (§ 6.3), and participant-external obligation and requirement (§ 6.4).

6.1 Modality in the English and Chinese Versions. General Observations

Table 4 lists the most frequent modal expressions found on the English and Chinese word lists, respectively. On the one hand, the modal expressions occurring at least 30 times in the English version are *can*, *will*, *would*, *must*, *cannot*, *should* and *may*, the lemmas NEED (verb) and CALL (verb).[9] On the other hand, the ones that stand out quantitatively in the Chinese version are the modal verbs 能 *néng*, 会 *huì*, 可 *kě*, 要 *yào*, 应 *yīng*, 必须 *bìxū*, and 可以 *kěyǐ*, the modal verb/noun 需要 *xūyào*, the adverb 将 *jiāng* and the compound verb 无法 *wúfǎ*. We

9 Capital letters indicate lemmas, that is, groups of all inflectional forms related to one stem that belong to the same word class (Kučera, Francis 1967, 19). NEED (verb) stands for *need*, *needs*, *needed*, *needing*, and CALL (verb) stands for *call*, *calls*, *called*, *calling*.

decided to also include the occurrences of NEED (noun), which are very frequent in the Letter, and also those of HOPE (noun) and CALL (noun),[10] because their equivalent Chinese translations 需要 *xūyào*, 希望 *xīwàng*, 召 *zhào* and its compound forms (indicated as 召* *zhào**) are used as both verbs and nouns. The raw frequencies are provided along with the normalised frequencies per number of word tokens.

Table 4 The most frequent modal expressions in the English and Chinese versions of *Laudato Si'*

English	Freq.	%	Chinese (1)	Freq.	%	Chinese (2)	Freq.	%
can	179	0.48	能 *néng*	198	0.56	召* *zhào** *	17	0.05
will	94	0.25	会 *huì*	140	0.39	需 *xū*	14	0.04
NEED (verb)	75	0.20	可 *kě*	132	0.37	应该 *yīnggāi*	11	0.03
would	64	0.17	需要 *xūyào*	102	0.29	想 *xiǎng*	11	0.03
must	58	0.16	要 *yào*	97	0.27	愿意 *yuànyì*	9	0.03
cannot	54	0.15	应 *yīng*	74	0.21	毋须 (无须) *wúxū*	7	0.02
NEED (noun)	47	0.13	无法 *wúfǎ*	58	0.16	不得不 *bùdé bù*	7	0.02
should	41	0.11	必须 *bìxū*	57	0.16	难以 *nányǐ*	6	0.02
CALL (verb)	34	0.09	将 *jiāng*	42	0.12	须 *xū*	6	0.02
may	32	0.09	可以 *kěyǐ*	30	0.08	想要 *xiǎngyào*	6	0.02
could	19	0.05	可能 *kěnéng*	21	0.06	懂得 *dǒngdé*	5	0.01
shall	11	0.03	必要 *bìyào*	21	0.06	易 *yì*	5	0.01
might	7	0.02	要求 *yāoqiú*	18	0.05	难 *nán*	5	0.01
HOPE (noun)	10	0.03	希望 *xīwàng*	18	0.05	愿 *yuàn*	5	0.01
HOPE (verb)	4	0.01	能够 *nénggòu*	17	0.05	宜 *yí*	2	0.01
CALL (noun)	2	0.01				**Total**	**1,141**	**3.2**
Total	**731**	**1.98**						

* 召* *zhào** stands for: 召唤 *zhàohuàn*, 召叫 *zhàojiào*, 号召 *hàozhào*.

For space constraints, we decided to focus on the following selection of English modal expressions: *will*/*shall* (*not*), *cannot* and *may*/*might not* and CALL (verb and noun, expressing a modal meaning). The auxiliaries *will*/*shall* and *cannot* (*may not*) were chosen because of their polysemous nature, that is, because of their potential to cover more than one of the meanings identified in table 2 above. The quasi-modal CALL, on the other hand, was chosen because previous research had identified it as a marker of modality in *Laudato Si'*.

Starting from these English modals, we first investigated how their instances are rendered into Chinese, and came up with lists of

10 NEED (noun) stands for the forms *need* and *needs*, HOPE (noun) stands for *hope* and *hopes*, and CALL (noun) for *call* and *calls*.

Chinese equivalents for each one of them. As predictable, in almost all cases each identified Chinese modal translates various source expressions and not just the ones from which we started. Therefore, we also created and analysed lists of source items corresponding to the most frequent Chinese equivalents. §§ 6.3 to 6.5 illustrate in detail the results of this 'bi-directional' analysis, which aims at shedding light on the semantic space covered by each of these English modal verbs with respect to their Chinese translation equivalents and at exploring possible instances of explicitation.

As can be noticed from table 4, the number of modal verbs identified in the Chinese version of *Laudato Si'* is higher than those in the English one. This may be due to two main reasons. The first one is that some modal expressions used in the Chinese version do not correspond to any explicit modal expression in English, as illustrated by example (2):

2. Some forms of pollution **Ø** are part of people's daily experience.
   每人在日常生活中均会接触到不同形式的污染。
   měi rén zài rìcháng shēnghuó zhōng jūn ***huì***
   every person at daily life inside all can
   jiēchù-dào bù tóng xíngshì de wūrǎn
   come.into.contact-RES NEG similar form LIG pollution

The second one is that in our corpus a large number of English adjectives (e.g. *possible*, *probable*, *able*) used in impersonal constructions, such as the one in example (3), are translated into Chinese with a modal verb:

3. It is **possible** that we do not grasp the gravity of the challenges now before us.
   我们很可能仍未理解到目前的挑战有多么严峻。
   wǒmen hěn ***kěnéng*** *réng wèi lǐjiě-dào*
   1PL very can still NEG comprehend-RES
   mùqián de tiǎozhàn yǒu duōme yánjùn
   at.present LIG challenge have how.much severe

It stands to reason that a complete correspondence between the English and the Chinese modal expressions in the two versions cannot be expected, as a given modal meaning in one language can be phrased in the other language in various ways, according to the specific contextual (and typological needs) and the translator's preferences. Furthermore, the original English (co-)texts often differ from the translated ones in various other respects, including syntactic aspects. For example, in the parallel sentences in excerpt (4), the English modal verb *can* in the main clause is rendered in Chinese with the verb 会 *huì*. Also, the Chinese version adds the modal verb 能 *néng* in the subordinate clause, which has no explicit equivalent in the English

version. Finally, the main clause and the subordinate if-clause are inverted in the Chinese version with respect to the English one:

4. Local legislation can be more effective, too, if agreements exist between neighbouring communities to support the same environmental policies.
若能与邻近地区达成协议, 支持相同的环境政策, 本地立法则会更有效力。

ruò	*néng*	*yǔ*	*línjìn*	*dìqū*	*dáchéng*	*xiéyì*
if	can	with	close	area	reach	agreement
zhīchí	*xiāngtóng*	*de*	*huánjìng*	*zhèngcè*	*běndì*	
support	similar	LIG	environment	policy	this.place	
lìfǎ	*zé*	*huì*	*gèng*	*yǒu*	*xiàolì*	
legislation	then	can	still.more	have	effect	

6.2 Will/Shall. Epistemic Possibility and Probability; Participant-Internal Willingness and Intention

Table 5 lays out the translations of the instances of *will* and *shall* in the Letter.

Table 5 The use of *will* and *shall* in the English version and their corresponding translations into Chinese

English >	Freq.		Chinese	Freq.
will (not)	96	→	Ø	37
			会 *huì*	26
			将 *jiāng*	9
			无法 *wúfǎ*	4
			能 *néng*	4
			将(不)会 *jiāng (bù) huì*	3
			others	12
			Sub-total	**95**
shall (not)	11	→	不可 *bù kě*	7
			不应 *bù yīng*	2
			应 *yīng*	1
			会 *huì*	1
			Sub-total	**11**
			Grand total	**105**

As can be noticed, 37 occurrences of *will* are not translated into Chinese altogether, 26 are translated with the verb 会 *huì*, 9 with the adverb 将 *jiāng*, 4 with 能 *néng*, 3 with the adverb/verb combination 将会 *jiāng huì* or its negative counterpart 将 (不) 会 *jiāng* (*bu*) *huì*. Finally, 无法 *wúfǎ* translates negative uses of *will* in four cases. As for *shall* (*not*), all the instances but one are part of citations from the Bible or from other documents. Only one case of *shall* conveys epistemic modality and is translated as 会 *huì*, while the others express participant-external modality. We will deal with some instances of them in § 6.3 below.

会 *huì* is the second most used modal verb in the Chinese version after 能 *néng* **[tab. 2]**. As seen in § 3, 会 *huì* can indicate epistemic possibility and probability as well as participant-internal ability, while 能 *néng* expresses both participant-internal ability and external possibility (Abbiati 2014, 213).

An interesting modal item is the adverb 将 *jiāng*,[11] which is used in formal written Chinese to indicate imminent future reference or certainty about a future situation (Lǚ 2004, 300). Generally speaking, future tense and modality are strongly linked. With regard to *will* and *shall*, for instance, Coates points out that "it would be meaningless to be willing or to intend to do something which has already been done" (1983, 233-4). Furthermore, Lehmann notices that from a diachronic perspective "often the future may arise through the grammaticalisation of a desiderative modal", of which "*will* is a known example" (2002, 26). That is, although modal expressions signal epistemic possibility and probability or participant-internal ability rather than future time *per se*, they are used with reference to future events or states.[12]

The translation of *will*/*shall* (*not*) with 会 *huì* and 将 *jiāng* was expected, while the correspondence with 无法 *wúfǎ* was not, both because of its meaning (see the description in § 6.3) and because, like 将 *jiāng*, it is not often mentioned in studies on modality. The frequent use of 会 *huì* and 将 *jiāng* suggests that epistemic possibility and probability and participant-internal willingness and intention are the main semantic areas covered by *will* in the Encyclical Letter. Examples (5) and (6) show the use of 会 *huì* as a translation of *will*, while example (7) illustrates how 将 *jiāng* is used to this end:

11 Some authors, including Smith and Erbaugh (2005, 731), consider 将 *jiāng* as a modal verb.

12 For a more in-depth treatment of modality in relation to tense, see Portner 2009, 236-41.

5. I **will** briefly turn to what is happening to our common home.
[…] 我会略述我们共同家园的现况。

wǒ	***huì***	*lüè*	*shù*	*wǒmen*	*gòngtóng*	*jiāyuán*
1SG	can	sketchy	narrate	1PL	common	home
de	*xiànkuàng*					
LIG	present.situation					

6. Greater scarcity of water **will** lead to an increase in the cost of food and the various products which depend on its use.
水资源不足若进一步恶化, 会使食物, 以及各种制造过程中需要用水的产品的成本增加。

shuǐ	*zīyuán*	*bù*	*zú*	*ruò*	*jìnyībù*	*èhuà*
water	resource	NEG	sufficient	if	further	deteriorate
huì	*shǐ*	*shíwù*	*yǐjí*	*gè*	*zhǒng*	*zhìzào*	*guòchéng*
can	let	food	as.well	each	CLF	produce	process
zhōng	*xūyào*	*yòng*	*shuǐ*	*de*	*chǎnpǐn*	*de*
inside	need	use	water	LIG	product	LIG
chéngběn	*zēngjiā*					
cost	increase					

7. […] politicians **will** inevitably clash with the mindset of short-term gain and results which dominates present-day economics and politics.
从政者[…], 将无可避免地与现今经济和政治以短期利益和成效为目标的心态相冲突。

cóngzhèngzhě	***jiāng***	*wú-kě-bìmiǎn*	*de*	*yǔ*	*xiànjīn*
politician	will	NEG-can-avoid	LIG	and	present
jīngjì	*hé*	*zhèngzhì*	*yǐ*	*duǎnqī*	*lìyì*	*hé*
economy	and	politics	with	short.term	profit	and
chéngxiào	*wéi*	*mùbiāo*	*de*	*xīntài*	*xiāng*	*chōngtū*
effect	be	target	LIG	mindset	mutually	clash

Example (5) is an extract from the "Preamble" and expresses the Pope's intention to address a given topic later on in the Letter, while example (6) predicts that a given event will happen in the future. 将 *jiāng* in example (7) also conveys the meaning of epistemic possibility and probability rather than imminent future reference or certainty about a future situation, which suggests that the semantic spaces covered by 将 *jiāng* and 会 *huì* are very close. However, the two of them are also used together in the combination 将会 *jiāng huì* to translate some other instances of *will*, which suggests that their meanings do not fully overlap and that, if used together, they complement each other, such as in extract (8):[13]

13 We are undecided about whether in this particular case the hierarchical structure is [[将会]是] or [将[会是]], and leave the question to future investigation.

8. Eternal life *will* be a shared experience of awe […]
永生将会是共享的美事。

yǒngshēng	***jiāng***	*huì*	*shì*	*gòngxiǎng*	*de*	*měi-shì*
eternal.life	will	can	be	share	LIG	beautiful-thing

A large number of instances of *will* (37) are not translated into Chinese with an explicit modal expression. The reason for this choice is not easy to explain, yet three observations can be made. Firstly, on some occasions the original English text had to be rephrased to meet the needs of Chinese syntax and discourse, which also involved omitting the translation of the modality. This is especially the case of many English restrictive relative clauses which were translated into Chinese as pre-modifying structures, as example (9) shows (the relative clauses are underlined):

9. Those who **will** have to suffer the consequences of what we are trying to hide will not forget this failure of conscience and responsibility.
那些因我们的隐瞒实情而Ø受害的人, 将不会忘记我们的埋没良知和欠缺承担。

nà	*xiē*	*yīn*	*women*	*de*	*yǐnmán*	*shíqíng*	*ér*	***Ø***
those	CLF	because	1PL	LIG	conceal	truth	and	
shòu	*hài*	*de*	*rén*	*jiāng*	*bú*	*huì*	*wàngjì*	*wǒmen*
suffer	harm	LIG	person	will	NEG	can	forget	1PL
de	*máimò*	*liángzhī*			*hé*	*qiànquē*	*chéngdān*	
LIG	cover.up	intuitive.knowledge			and	lack	assume	

As can be noticed, the relative construction pre-modifying the noun 人 *rén* 'person' does not explicitly render *will*. This can be related to a general tendency in Chinese to avoid the use of grammatical markers in such constructions, including the perfective aspectual marker 了 *le* and modal particles.

Secondly, some other instances of *will* are not explicitly translated when the verb *hope* (Chinese 希望 *xīwàng* and 盼望 *pànwàng*) is used in the main clause to introduce another clause expressing futurity with *will*, such as in example (10):

10. Can we **hope**, then, that in such cases, legislation and regulations dealing with the environment **will** really prove effective?
在这种情况之下, 我们仍能希望有关环境的立法和规定Ø真正有效用吗?

zài	*zhè*	*zhǒng*	*qíngkuàng*	*zhīxià*	*women*	*réng*	*néng*
at	this	CLF	situation	under	1PL	still	can
xīwàng	*yǒuguān*	*huánjìng*		*de*	*lìfǎ*		*hé*
hope	regard	environment		LIG	legislation		and
guīdìng	***Ø***	*zhēnzhèng*	*yǒu*	*xiàoyòng*	*ma*		
regulation		really	have	effect	Q		

Hope implies the speaker's attitude towards the future (cf. Portner 2009, 6), which is arguably the reason why the translator did not feel the need to translate *will* explicitly.

Thirdly, when a quasi-modal (e.g. *be able to*) is used in combination with *will*, only the meaning of the quasi-modal is translated.[14] Example (11) illustrates that 能 *néng* translates the meaning of *be able to* but not that of *will*:

11. Only by cultivating sound virtues **will** people **be able** to make a selfless ecological commitment.
唯有藉培养良好的品德, 人才Ø能作出无私的生态承诺。
wéiyǒu *jiè* *péiyǎng* *liánghǎo* *de* *pǐndé*
only make.use.of cultivate good LIG moral.character
rén *cái* ***Ø*** ***néng*** *zuò-chū* *wúsī* *de* *shēngtài* *chéngnuò*
person only can make-DIR unselfish LIG ecology promise

Four cases of *will* were rendered with the verb 能 *néng* expressing participant-internal ability or epistemic possibility (see example (12)), while four cases of *will* plus a negative element were translated with 无法 *wúfǎ*, functioning as a marker of negative participant-internal ability (see example (13)). Obviously, as is always the case, it is the overall meaning emerging from the unfolding discourse rather than that of a single word (e.g. the modal verb *will*) that leads a translator to make a given translation choice.

12. […] ecological problems **will** solve themselves […]
[…] 则生态问题自然能迎刃而解。
zé *shēngtài* *wèntí* *zìrán* ***néng*** *yíng-rèn-ér-jiě*
then ecology problem naturally can meet-blade-and-solve

13. Unless we do this, other creatures **will not** be recognised for their true worth […]
除非我们这样做, 否则无法认识其它受造物的真正价值 […]
chúfēi *women* *zhèyàng* *zuò* *fǒuzé* ***wúfǎ*** *rènshi*
unless 1PL this.way do otherwise not.have.way know
qítā *shòuzàowù* *de* *zhēnzhèng* *jiàzhí*
other creature LIG true worth

Some more instances of *will* are translated with a Chinese modal verb preceded by a time adverbial, thus adding to the epistemic probability meaning of the sentence and making the reference to the fu-

14 According to Chao (1968, 732), two or more auxiliary verbs, including 会能 *huì* and 能 *néng*, can occur in succession. The translator clearly did not opt for this use in this case.

ture even more explicit. For example, in excerpt (14) the adverb 永远 *yǒngyuǎn*, which, unlike the English adverb *never*, can only refer to the future, occurs before 无法 *wúfǎ*:

14. […] so too living species are part of a network which we **will never** fully explore and understand.
[…] 生物物种之间也是如此, 它们属于一个我们永远无法完全探索和明白的网络的一部分。

shēngwù	*wùzhǒng*	*zhījiān*	*yě*	*shì*	*rúcǐ*		*tāmen*
living.being	species	between	also	be	this.way		3PL
shǔyú	*yī*	*ge*	*wǒmen*	***yǒngyuǎn***	***wúfǎ***	*wánquán*	
belong	one	CLF	1PL	forever	not.have.way	fully	
tànsuǒ	*hé*	*míngbai*	*de*	*wǎngluo*	*de*	*yī*	*bùfen*
explore	and	understand	LIG	net	LIG	one	part

The compound 无法 *wúfǎ* will be dealt with in more detail in § 6.3 below as a translation equivalent of *cannot*. The other translations of *will* are not discussed here, as they occur only once each. They include the modal auxiliaries 应 *yīng*, 不可能 *bù kěnéng*, 可 *kě*, 可能 *kěnéng*, 必要 *bìyào*, 要 *yào*, 足以 *zúyǐ* and the adverbs 未必 *wèibì* and 决 *jué*.

The right-hand side of table 6 below summarises the English modal expressions that were translated into Chinese with 会 *huì*, 将 *jiāng* and 将会 *jiāng huì* and their frequencies. The analysis of these translation equivalents aims to illuminate the semantic space covered by these three Chinese modal expressions further, with reference to the original modal expressions and their co-texts.

Table 6 The use of 会 *huì*, 将 *jiāng* and 将(不)会 *jiāng (bù) huì* in Chinese and the corresponding source expressions

English	**Freq.**		**Chinese**	**Freq.**
Ø	58			
will	26			
can	21			
would	10	→	会 *huì*	132
end up	6			
may	4			
others	7			
Sub-total	**132**			

Ø	14		
will	9		
would	4	→ 将 *jiāng*	29
could	1		
may	1		
Sub-total	**29**		
Ø	3		
will	3		
would	2	→ 将(不)会 *jiāng (bù) huì*	9
can	1		
Sub-total	**9**		
		Grand total	**170**

The data shows that 58 cases of 会 *huì*, 14 of 将 *jiāng*, and 3 of 将会 *jiāng huì* do not correspond to any explicit modal element in the original version, while 26 of 会 *huì*, 9 of 将 *jiāng*, and 3 of 将会 *jiāng huì* translate the verb *will*. The other source modal verb that these three forms have in common is *would*. What is also noticeable is that 21 instances of *can*, 6 of the verb *end up* and 4 of *may* are associated with 会 *huì*.

The 58 instances of 会 *huì* that do not translate any overt English modal marker (Ø) need a tentative explanation, as they might represent attempts of explicitation of the source meaning. An analysis of the concordance lines for 会 *huì* reveals that in many such cases this modal translates statements which in English are couched in the simple present and indicate a general truth, which is either habitual or bound to happen, such as in examples (15) and (16):

15. Valuable works of art and music now **make use** of new technologies.
现时具价值的艺术品和音乐也会运用新科技。

xiànshí	*jù*	*jiàzhí*	*de*	*yìshùpǐn*	*hé*	
current.time	possess	value	LIG	work.of.art	and	
yīnyuè	*yě*	***huì***	*yùnyòng*	*xīn*	*kējì*	
music	also	can	utilise	new	technology	

16. Yet God's infinite power **does not** lead us to flee his fatherly tenderness [...]
天主无限的威能总不会令我们逃离祂父爱的温柔 [...]

Tiānzhǔ	*wúxiàn*	*de*	*wēinéng*	*zǒng*	*bú*	***huì***	*lìng*
God	infinite	LIG	power	after.all	NEG	can	let

women	*táolí*	*tā*	*fù*	*ài*	*de*	*wēnróu*
1PL	flee	3SG	father	love	LIG	tenderness

The addition of the modal disambiguates the original meaning and appears to make the Chinese version more transparent and therefore explicit. The analysis also suggests that in other cases the explicit translation of modality with 会 *huì* is triggered by the conditional meaning of the sentence it occurs in, such as in example (17):[15]

17. If we do not, we **burden** our consciences with the weight of having denied the existence of others.
如果我们不这样做, 会因否定他人的存在而受良知的譴责。

rúguǒ	*women*	*bù*	*zhèyàng*	*zuò*	***huì***	*yīn*	*fǒudìng*
if	1PL	NEG	this.way	do	can	because	negate
tā-rén	*de*	*cúnzài*	*ér*	*shòu*	*liángzhī*		
other-person	LIG	existence	and	suffer	intuitive.knowledge		
de	*qiǎnzé*						
LIG	condemn						

Finally, instances of 会 *huì* corresponding to no modal marker in the original text are found in clauses complementing the meaning of verbs such as 相信 *xiāngxìn* (see example 18). This verb translates the source text *believe*, which, like the verb *hope* discussed above, implies the speaker's attitude towards the future.

18. There is also the fact that people no longer seem to **believe** in a happy future.
此外人类似乎不再相信会有快乐的未来。

cǐwài	*rénlèi*	*sìhū*	*bù*	*zài*	***xiāngxìn***	***huì***
moreover	humanity	seemingly	NEG	again	believe	can
yǒu	*kuàilè*	*de*	*wèilái*			
there.be	joyful	LIG	future			

The occurrences of 会 *huì* that translate English *can* and *may* are less unexpected and confirm that 会 *huì* shares with these English modals the semantic areas of participant-internal ability and epistemic possibility and probability, as illustrated by example (19):

15 This is in line with Chappell and Peyraube (2016, 306), who found that also the cognate Cantonese modal verb 會 *wúih* is highly compatible with conditional and counterfactual clauses. For more information about the relation between conditionals and modality, see Portner 2009, 247-57.

19. [...] for we know that things **can** change.
[...] 因为我们知道事情是会改变的。

yīnwèi	*women*	*zhīdào*	*shìqing*	*shì*	***huì***	*gǎibiàn*	*de*
because	1PL	know	thing	be	can	change	P

Another parallel expression of 会 *huì* emerging from table 6 that deserves some attention is the lexical verb *end up*. This verb is used epistemically in the English version to make a prediction through a general statement, and is translated into Chinese with 会 *huì* in six cases. It must be said that the adverb 最终 *zuìzhōng* is used in four such instances out of six to reinforce the telicity of *end up*, as in example (20):

20. The alliance between the economy and technology **ends up** sidelining anything unrelated to its immediate interests.
经济和科技结盟, 最终会将与其当时利益无关的一切弃之不顾。

jīngjì	*hé*	*kējì*	*jié*	*méng*	***zuìzhōng***	***huì***	
economy	and	technology	unite	alliance	finally	can	
jiāng	*yǔ*	*qí*	*dāngshí*	*lìyì*	*wú*	*guān*	de
DISP	with	3SG/PL	then	profit	NEG.have	relation	LIG
yīqiè	*qì-zhī-bú-gù*						
all	abandon-3SG/PL-NEG-care						

To sum up, with regard to the Encyclical Letter the semantic space of 会 *huì*, 将 *jiāng*, and 将会 *jiāng huì* covers the areas of epistemic possibility and probability and participant-internal willingness and intention. However, the hypothesised correspondence between *will* (*shall*) and these Chinese expressions is only partial, as the data reveals that they also cover the meanings conveyed by the English verbs *can*, *end up*, *may*, *would* and *could*. Finally, the large number of cases in which the three Chinese modal markers do not translate any overt English modals may be due to typological differences between the two languages, to the translator's attempt to make such modal meanings more explicit, or to both.

6.3 Cannot and May not. Participant-Internal Ability and Participant-External Possibility

Table 7 below shows how the 55 instances of *cannot*[16] and the 2 instances of *may not* are translated into Chinese.

Table 7 The use of *cannot* and *may not* in the English version and their corresponding translations into Chinese

English	Freq.		Chinese	Freq.
cannot	55	→	不能 *bù néng*	19
			无法 *wúfǎ*	12
			Ø	5
			不可 *bù kě*	4
			不应 *bù yīng*	4
			不可能 *bù kěnéng*	2
			必须 *bìxū*	2
			不得不 *bù dé bù*	1
			不容 *bù róng*	1
			others	5
			Sub-total	**55**
may not	2	→	未必会 *wèibì huì*	1
			未必能 *wèibì néng*	1
			Grand total	**57**

If used epistemically, *cannot* can be paraphrased as 'it is not possible that [...]'. Not only is it used to negate epistemic *can*, but also epistemic *must* and *may* (see § 3). By contrast, epistemic *may not* can be paraphrased as 'it is possible that [...] not', that is, it negates the truth of the proposition (Coates 1983, 100-2). When *cannot* expresses participant-internal ability, it can be paraphrased as 'inherent properties [do not] allow me to do it', while it takes on the meaning 'external circumstances [do not] allow me to do it', if it expresses participant-external possibility (Coates 1983, 93).

16 The informal contracted form *can't* is not used in the Encyclical Letter.

The translation choices 不能 *bù néng* (19 occurrences), 不可 *bù kě* (4 occurrences), 不可能 *bù kěnéng* (2 occurrences) were expected, as they are among the direct Chinese equivalents of *cannot*, covering its main semantic areas (e.g. Abbiati 2014, 213-14). By contrast, the negated form of 应 *yīng* (不应 *bù yīng*) (3 occurrences), the modal verb 必须 *bìxū* (2 occurrences), the cases of zero translation (5 occurrences), and especially 无法 *wúfǎ* (12 occurrences) were less predictable and deserve some attention. In particular, 无法 *wúfǎ* is a verb composed of two morphemes: the classic Chinese negative form of the modern Chinese verb 有 *yǒu* 'have', that is 无 *wú*, followed by its object 法 *fǎ*. Literally, it means 'to have no means of (doing something)', and therefore it mainly indicates lack of participant-internal ability and participant-external possibility.

The four instances of 不应 *bù yīng* represent a translation choice whereby the ambiguous use of English *cannot* is interpreted as explicit participant-external necessity[17] (see example 21).

21. If an artist **cannot** be stopped from using his or her creativity […]
正如艺术家不应被禁止发挥他或她的创意 […]

zhèng	*rú*	*yìshùjiā*	***bù***	***yīng***	*bèi*	*jìnzhǐ*	*fāhuī*
just	as	artist	NEG	should	PASS	forbid	bring.into.play
tā	*huò*	*tā*	*de*	*chuàngyì*			
3SG.M	or	3SG.F	LIG	creativity			

The marker 必须 *bìxū* makes the meaning of two other uses of *cannot* more explicit. For instance, in example (22) it spells out the meaning of *cannot* (*fail*) (with *fail* also having a negative meaning) as participant-external necessity:

22. We **cannot fail** to praise the commitment of international agencies and civil society organisations […]
我们必须赞扬一些国际机构和公民社会组织的努力 […]

wǒmen	***bìxū***	*zànyáng*	*yī*	*xiē*	*guójì*	*jīgòu*
1PL	must	praise	one	CLF	international	organisation
hé	*gōngmín*	*shèhuì*	*zǔzhī*		*de*	*nǔlì*
and	citizen	society	organisation		LIG	make.effort

The analysis of the concordance lines for 无法 *wúfǎ* suggests that in this case this compound verb unambiguously signals the sense of negative participant-internal ability of *cannot*, such as in example (23):

17 Participant-external necessity and obligation can be difficult to tell apart. If negated, necessity or obligation express a prohibition, like in this case (cf. Sparvoli 2012, 263 ff.).

23. […] we **cannot** adequately combat environmental degradation unless […]
[…] 除非我们 […], 否则无法抵抗环境的恶化。
chúfēi *women* *fǒuzé* ***wúfǎ*** *dǐkàng* *huánjìng*
unless 1PL otherwise cannot resist environment
de *èhuà*
LIG deteriorate

Table 8 below presents the original sources of four of the most frequent translation equivalents of *cannot*: 无法 *wúfǎ*, 不能 *bù néng*, 不可 *bù kě* and 不可能 *bù kěnéng*. Not only does 无法 *wúfǎ* translate 12 instances of *cannot*, but it also renders several other expressions of negated participant-internal ability, such as the adjectives *incapable*, *irretrievable* and *unsustainable*, the verbs *fail* and *not succeed*, and the noun *inability*. These equivalent expressions confirm that the semantic space covered by 无法 *wúfǎ* is mainly lack of participant-internal ability.

Table 8 The use of 无法 *wúfǎ*, 不能 *bù néng*, 不可 *bù kě* and 不可能 *bù kěnéng* in the Chinese version and the corresponding source expressions in English

English	**Freq.**		**Chinese**	**Freq.**
neg. adjective	16			
cannot	12			
fail	6			
neg. will	6			
can + negative element	5	→	无法 *wúfǎ*	58
could not	2			
inability	2			
lack	2			
others	7			
Sub-total	**58**			
cannot	19			
neg. adjective	6			
can + negative element	5	→	不能 *bù néng*	40
Ø	5			
others	5			
Sub-total	**40**			

shall not	7			
cannot	4			
should not	3			
can + negative element	2	→	不可 *bù kě*	24
demand	2			
neg. adjectives	2			
others	4			
Sub-total	**24**			
cannot	2	→	不可能 *bù kěnéng*	8
will not	2			
others	4			
Sub-total	**8**			
			Grand total	**130**

Example (24) illustrates how the meanings of the morphemes in the de-verbal adjective *incalculable* are rendered into Chinese. As can be noted, the negative meaning of the prefix *in-* and that of the suffix *-able* are conveyed by the Chinese morphemes 无 *wú* and 法 *fǎ*, while the stem *calcula*(*te*) is rendered by the verb 计算 *jìsuàn* 'calculate'. These words are inserted in the '是 ... 的 *shì ... de*' construction, which literally means 'belonging to the class of things for which there is no way to calculate':

24. […] the values involved are **incalculable**.
所涉及的价值是无法计算的。

suǒ	*shèjí*	*de*	*jiàzhí*	*shì*	***wúfǎ***	***jìsuàn***	*de*
NMLZ	involve	LIG	value	be	cannot	calculate	NMLZ

The item 无法 *wúfǎ* also renders some instances of *can* used in combination with negative elements (e.g. the negative quantifier *no* and the adverb *never*), such as in example (25):

25. There **can** be **no** renewal of our relationship with nature without a renewal of humanity itself.
人类若不自我更新, 人类与大自然的关系则无法更新。

rénlèi	*ruò*	*bú*	*zìwǒ*	*gēngxīn*	*rénlèi*	*yǔ*	*dàzìrán*
humanity	if	NEG	self	renew	humanity	and	nature
de	*guānxi*	*zé*	***wúfǎ***	*gēngxīn*			
LIG	relation	then	cannot	renew			

Table 8 shows that 不能 *bù néng* is the most frequent translation equivalent of *cannot*. Like 无法 *wúfǎ*, it often translates negative deverbal adjectives and instances in which *can* collocates with a negative element, and, differently from it, it has the potential to express all of the meanings covered by *cannot*. It also shows that five occurrences of 不能 *bù néng* translate source co-texts with zero modality, thus making the target meaning more precise and explicit (see example (26)):

26. Man **does not** create himself.
人不能自我创造。
rén ***bù*** ***néng*** *zìwǒ* *chuàngzào*
man NEG can self create

不可 *bù kě* covers the field of participant-external necessity or obligation (prohibition). Its source expressions range from *cannot* and *can* plus a negated element, through *should not*, to *shall not*. Most of the instances of *shall not*, in particular, are quotations from the Bible, like the one in example (27):

27. "When you reap the harvest of your land, you **shall not** reap your field to its very border […]
"当你们收割田地的庄稼时, 你不可割到地边 [...]
dāng *nǐmen* *shōugē* *tiándì* *de* *zhuāngjia* *shí* *nǐ* ***bù***
When 2PL reap land LIG crops time 2SG NEG
kě *gē-dào* *dì* *biān*
can reap-RES land edge

Finally, 不可能 *bù kěnéng* represents a choice whereby the translator conveys an epistemic reading of the source modals *cannot*, *will not* and of other forms such as *impossible* and *not possible*. Extract (28) exemplifies how *impossible* is translated into Chinese:

28. It becomes almost **impossible** to accept the limits imposed by reality.
要接受现实的掣肘几乎是不可能的。
yào *jiēshòu* *xiànshí* *de* *chèzhǒu* *jīhū* *shì*
want accept reality LIG hold.back.by.the.elbow almost be
bù ***kěnéng*** *de*
NEG possible NMLZ

To conclude, in *Laudato Si'*, 不能 *bù néng* straddles the areas of negative participant-external possibility and negative participant-internal ability expressed by *cannot*. By contrast, 无法 *wúfǎ* appears to be an indicator of negative participant-internal ability, 不可能 *bù kěnéng* of epistemic modality, and 不可 *bù kě*, 不应 *bù yīng*, 必须 *bìxū* of participant-external obligation, necessity or requirement (prohibition).

The selective uses of these last modal expressions can be viewed as attempts to explicate the source meanings of *cannot*.

6.4 CALL. Participant-External Necessity, Obligation, and Requirement

Castello and Gesuato define the specific pattern 'someone is called to do something', used in the English version of *Laudato Si'*, as "a near-modal expression of obligation, which represents yet another linguistic realisation of the Pope's call for commitment to ecology and ecological spirituality" (2019, 138-9). An examination of the concordance lines for the instances of the lemma CALL (verb) revealed the presence of other patterns in which CALL (verb) is used, the most important of which are 'someone/something call(s) for something' and 'someone/something call(s) someone to'. These uses of *call* are reminiscent of citations from the Letters of Paul, such as "Christians are called to be saints" (Romans 1: 7) and "[…] yourself who are called to belong to Jesus Christ" (Romans 1: 6). They also recall phrases from the Gospel, such as "the call to repentance" (Luke 10: 13) and "the call to be a disciple" (Luke 14: 25).[18]

Table 9 presents the renderings of the forms of CALL (verb) and CALL (noun) into Chinese. In the English version CALL (verb) totals 34 occurrences and CALL (noun) two. They are translated into Chinese as 召 *zhào* or its compound forms 召唤 *zhàohuàn*, 召叫 *zhàojiào* and 号召 *hàozhào* in twelve cases. Quantitatively speaking, therefore, in the Encyclical Letter 召* *zhào** [19] represents the nearest semantic equivalent of CALL, and its use adds to the biblical and pastoral register of the text. According to the 现代汉语词典 *Xiandai Hanyu Cidian* (2014, 545-6, 1645), 召 *zhào* and its variant forms mean "call together, convene, summon someone" (our translation). Also the core meaning of 呼吁 *hūyù* and 呼唤 *hūhuàn* is similar to that of 召 *zhào* and indicate "appeal, call on somebody" and "call or shout to someone" (our translation). The twenty-four other renderings of CALL (verb and noun) in the text clearly represent less direct ways of rephrasing its core meaning. As can be seen, they are all modal verbs or no modal expression at all.

18 The quotations from the Gospel and the New Testament Letters were found at `http://www.vatican.va/archive/ENG0839/_INDEX.HTM`.

19 The asterisk after 召* *zhào* is used to indicate the base form 召 *zhào* and the three compounds 召唤 *zhàohuàn*, 召叫 *zhàojiào* and 号召 *hàozhào*.

Table 9 The use of CALL as a semi-modal in the English version and the corresponding translations into Chinese

English CALL	**Freq.**		**Chinese**	**Freq.**
called	14		召 *zhào*	6
calls	12		召唤 *zhàohuàn*	3
call (verb)	6		召叫 *zhàojiào*	2
call (noun)	2		号召 *hàozhào*	1
calling	2		**Total** 召* ***zhào****	**12**
Total CALL	**36**		需要 *xūyào*	6
			要求 *yāoqiú*	4
		→	必须 *bìxū*	3
			要 *yào*	2
			应 *yīng*	2
			呼吁 *hūyù*	2
			Ø	2
			应该 *yīnggāi*	1
			呼唤 *hūhuàn*	1
			会 *huì*	1
			Grand total	**36**

The 14 instances of the verb form *called* are used as part of the passive construction 'someone is called to do something'. Only four of these are rendered in the passive voice in Chinese. It is interesting to note that in passive clauses only the monosyllabic form 召 *zhào* is employed after a passive marker, such as 被 *bèi* in example (29):

29. As Christians, we **are** also **called** "to accept the world as a sacrament […]
身为基督徒, 我们被召「视世界为共融的圣事[…]
*shēn wéi jīdūtú women bèi **zhào** shì shìjiè wéi*
self be Christian 1PL PASS summon watch world be
gòngróng de shèngshì
common.harmony LIG sacrament

By contrast, the other occurrences of *called* as well as the other forms of CALL (verb) are translated by using the active voice and either a compound form of 召 *zhào* or a modal verb indicating participating-external modality, as examples (30) and (31) show:

30. God, who **calls** us to generous commitment and to give him our all […]
天主, 祂召唤我们慷慨大方献上自己和给予一切 […]。
Tiānzhǔ *tā* ***zhàohuàn*** *women* *kāngkǎi* *dàfāng* *xiàn-shàng*
God 3SG summon 1PL generous liberal offer-DIR
zìjǐ *hé* *jǐyǔ* *yīqiè*
self and give all

31. Together with our obligation to use the earth's goods responsibly, we **are called** to recognize that […]
除了要有责任地善用大地的产物外, 我们也必须明白 […]
chúle *yào* *yǒu* *zérèn* *de* *shànyòng* *dàdì*
besides must have responsibility LIG properly.use earth
de *chǎnwù* *wài* *women* *yě* ***bìxū*** *míngbai*
LIG product besides 1PL also must understand

The choice of the Chinese modal auxiliary verbs 需要 *xūyào*, 要求 *yāoqiú*, 必须 *bìxū*, 要 *yào*, 应 *yīng*, 应该 *yīnggāi* as translations of the other instances of CALL (verb and noun) stresses the participant-external nature of these 'religious' near-modal expressions.

Looking at how the lemmas CALL (verb) and CALL (noun) are translated as 召 *zhào* and its compound forms **[tab. 9]** does not provide a full picture of the meanings and functions they convey, as there could be other uses of them in the Chinese version which do not translate CALL (verb and noun) but other words. Table 10 explores this possibility:

Table 10 The use of 召* *zhào** in the Chinese version and the corresponding source expressions in English

English CALL	**Freq.**		**Chinese**	**Freq.**
call (noun)	2		召 *zhào*	7
called	7		召唤 *zhàohuàn*	6
calling	1		召叫 *zhàojiào*	2
calls (verb)	2		号召 *hàozhào*	2
Total CALL	**12**	→	**Total** 召* ***zhào****	17
a summons (号召 *hàozhào*)	1			
vocation (召唤 *zhàohuàn*)	2			
beckons (召唤 *zhàohuàn*)	1			
carried up (召 *zhào*)	1			
Grand total	**17**			

The table shows that 召 *zhào* and its compound forms translate the source expressions *a summons*, *vocation*, *to beckon*, *carried up* as well, which arguably also encode a near-modal obligation meaning. Excerpt (32) illustrates the context of use of *a vocation* and is followed by its translation:

32. We were created with a **vocation** to work.
人从受造开始以来, 就有工作的召唤。

rén	*cóng*	*shòu*	*zào*	*kāishǐ*	*yǐlái*	*jiù*
person	from	PASS	creation	start	from	right.away
yǒu	*gōngzuò*	*de*	***zhàohuàn***			
have	work	LIG	summon			

In short, in *Laudato Si'*, the 'religious' quasi-modal CALL (verb and noun) is either turned into 召* *zhào** or into an auxiliary verb conveying participant-external modality. Furthermore, four source 'religious' terms (e.g. *vocation*) are expressed with 召* *zhào*. Both the use of Chinese modal auxiliaries to render some instances of quasi-modal CALL and that of 召* *zhào* to translate specific Catholic religious terms can be viewed as instances of explicitation. That is, they can be interpreted as a way of spelling things out for the sake of clarity and for the benefit of the target Chinese readership, who might not be familiar with such concepts of the Catholic doctrine.

7 Conclusions

This paper has investigated the use of some of the most frequent modal expressions in the English and Chinese versions of the Encyclical Letter *Laudato Si'*, a document in which the Pope presents possible scenarios due to climate change and directs his readership to action. Using corpus-based methods, word lists for both versions were obtained and checked for the most frequent English and Chinese modal expressions. A general quantitative analysis brought to light that the Chinese version contains a larger variety of modal auxiliaries than the English one, and a selection was made of frequent items covering different areas of modality. Subsequently, meaningful translation correspondences were investigated with the aim of defining their semantic space (research question one) and of detecting possible cases of explicitation (research question two). The first areas that were explored are epistemic probability and possibility and participant-internal willingness and intention, as prototypically expressed by *will*/*shall* in English and by their hypothesised main equivalent 会 *huì*. The analysis revealed further translation correspondences: i.e. that between *will* and 将 *jiāng* and 将会 *jiāng huì* to signal epistemic possibility and probability, and the one between *will not* and 无法 *wúfǎ* to express lack of participant-

internal ability; finally, that between *end up* and 会 *huì* to indicate the end state of a situation. Furthermore, the frequent cases of 会 *huì*, 将 *jiāng* and 将会 *jiāng huì* that do not pair up with any overt modal expression in the original version lend support to the explicitation hypothesis. The second group of semantic areas investigated are lack of epistemic possibility or probability, lack of participant-internal ability, participant-external possibility and obligation conveyed by *cannot* and its predictable equivalents 不能 *bù néng*, 不可 *bù kě*, 不可能 *bù kěnéng*. The main finding in this respect is the extensive use of 无法 *wúfǎ* to render instances of *cannot* mainly indicating lack of participant-internal ability. On the one hand, 不可 *bù kě* translates English modals expressing participant-external obligation and necessity, including *shall not* from biblical quotations. The third area under scrutiny was participant-external necessity, obligation and requirement, as conveyed by the near-modal CALL (verb and noun). The verb 召 *zhào* has proved to be its main translation equivalent in passive constructions, while its compound forms occur only in the active voice. The translation of the other instances of CALL (verb and noun) by means of Chinese modal auxiliaries of participant-external obligation/necessity stresses the deontic nature of these religious near-modal items. Finally, the rendering of religious terms such as *summons* and *vocation* with 召 *zhào* can be considered as attempts to explicate their meaning.

Table 11 summarises the main results of the study and maps the most frequent English and Chinese modal expressions identified in *Laudato Si'* onto the semantic categories they belong to:

Table 11 The English and Chinese modal expressions discussed in this study mapped onto the semantic categories

English	Chinese	Categories
will, can, would, end up, may, Ø	会 *huì* 能 *néng* 将会 *jiāng huì* 将 *jiāng*	Epistemic possibility, probability or Participant-internal willingness, intention
cannot	不可能 *bù kěnéng*	Epistemic lack of possibility
cannot fail can + negative element neg. adjective (e.g. not possible) could not inability lack	不能 *bù néng* 不能 *bù néng* 无法 *wúfǎ* Ø 不可 *bù kě* 不应 *bù yīng* 必须 *bìxū*	Lack of participant-internal ability
shall not cannot should not can + negative element demand	不可 *bù kě* 不能 *bù néng* 不应 *bù yīng*	Participant-external necessity, obligation, requirement

CALL (verb and noun) a summons vocation beckons carried up	召 *zhào* 召唤 *zhàohuàn* 召叫 *zhàojiào* 号召 *hàozhào* 需要 *xūyào* 要求 *yāoqiú* 必须 *bìxū* 要 *yào* 应 *yīng* 呼吁 *hūyù*	Participant-external necessity obligation, requirement

This study has shown that even the translation of highly grammaticalised items like modal expressions need to undergo processes of interpretation and adaptation, which involve choosing a suitable expression or a combination of various linguistic resources to render a given meaning in the target text. This is especially true of the text type analysed in this study, i.e. a piece of writing about Catholic doctrine, with which the Chinese and the Taiwanese readerships might not be familiar. This study has also discussed cases of modal expressions in the target text that seem to explicate the modal meanings implicit in the source text. However, the extent to which this is not only due to typological differences between the two languages but also to specific translation choices is a matter of debate, and could be investigated further by other corpus-based studies.

The corpus-based analyses carried out in this study have revealed a network of semantically connected modal expressions which a close reading of the two versions of *Laudato Si'* would have hardly managed to bring to light. This method has helped us identify the linguistic choices made by the writer and the translator to convey the intended semantic meanings. Parallel concordancing software, such as the online corpus-analysis tool *Sketchengine*,[20] could help speed up this type of analysis, yet human scrutiny and judgement would still be needed. Future corpus-based research endeavours could explore modal expressions and other lexical, grammatical or semantic phenomena in larger corpora. Specifically, research on the translation/adaption of Catholic/religious writing into Chinese would benefit from the analysis of bigger parallel corpora of texts concerning the Catholic doctrine and the Holy Scriptures.

20 https://www.sketchengine.eu/quick-start-guide/parallel-concordance-lesson.

Bibliography

Abbiati, M. (2014). *Grammatica di cinese moderno*. Venezia: Cafoscarina.

Anthony, L. (2018). *SegmentAnt* (Version 1.1.3) [Computer Software]. Tokyo: Waseda University. http://www.laurenceanthony.net.

Anthony, L. (2019). *AntConc* (Version 3.5.8) [Computer Software]. Tokyo: Waseda University. https://www.laurenceanthony.net/software.

Baker, M. (1996). "Corpus-Based Translation Studies. The Challenges that Lie ahead". Somers, H. (ed.), *Terminology, LSP and Translation. Studies in Language Engineering in Honour of Juan C. Sager.* Amsterdam; Philadelphia: John Benjamins, 175-86.

Baker, P. (2010). *Sociolinguistics and Corpus Linguistics*. Edinburgh: Edinburgh Sociolinguistics.

Biber, D. et al. (1999). *Longman Grammar of Spoken and Written English*. London: Longman.

Bybee, J.; Fleischman, S. (eds) (1995). *Modality in Grammar and Discourse*. Amsterdam: John Benjamins.

Carter, R.; McCarthy, M. (2006). *Cambridge Grammar of English*. Cambridge: Cambridge University Press.

Castello, E.; Gesuato, S. (2019). "Pope Francis's *Laudato Si'*. A Corpus Study of Environmental and Religious Discourse". Manca, E.; Bianchi, F.; Milizia, D. (eds), *Representing and Redefining Specialised Knowledge. Corpora and LSP, Lingue e Linguaggi*, 29, Special Issue, 121-45. https://doi.org/10.1285/i22390359v29p121.

Chappell, H.; Peyraube, A. (2016). "Modality and Mood in Sinitic". Nuyts, van der Auwera 2016, 296-329.

Chao, Y.R. (1968). *A Grammar of Spoken Chinese*. Berkeley; Los Angeles; London: University of California Press.

Chen, S.-K. (2017). "From Explicit to Implicit Orientation. Mapping Rank Scale to Modality in English and Chinese". *Functional Linguistics*, 4(15), 1-20. https://doi.org/10.1186/s40554-017-0049-1.

Coates, J. (1983). *The Semantics of the Modal Auxiliaries*. London and Canberra: Croom Helm.

Ghosh, A. (2016). *The Great Derangement*. Chicago: The University of Chicago Press.

Halliday, M.A.K. (1976). "Modality and Modulation in English". Kress, G. (ed.), *Halliday. System and Function in Language*. London: Oxford University Press, 189-213.

Halliday, M.A.K. (2004). *An Introduction to Functional Grammar*. 3rd ed. Revised by C. Matthiessen. London: Arnold.

Huddleston, R. (2002). "The Verb". Huddleston, R.; Pullum, G. (eds), *The Cambridge Grammar of the English Language: Selected Papers*. Cambridge: Cambridge University Press, 71-212.

Hsieh, C.-L. (2005). "Modal Verbs and Modal Adverbs in Chinese. An Investigation into the Semantic Source". *University System of Taiwan Working Papers in Linguistics*, 1, 31-58. http://web.ntnu.edu.tw/~clhsieh/2_Research/2.1_Publication/A02_2005.07_UST.pdf.

Kučera, H.; Francis, N. (1967). *Computational Analysis of Present-Day American English*. Providence: Brown University Press.

Laviosa, S. (2002). *Corpus-Based Translation Studies. Theory, Findings, Applications*. Amsterdam: Rodopi.

Lehmann, C. (2002). "Thoughts on Grammaticalization". 2nd revised ed. Arbeitspapiere des Seminars für Sprachwissenschaft der Universität Erfurt 9. Erfurt: Philosophische Fakultät Universität. https://doi.org/10.17169/langsci.b88.98.

Li, C.N.; Thompson, S.A. (1981). *Mandarin Chinese. A Functional Reference Grammar*. Berkeley: University of California Press.

Li, C.N.; Thompson, S.A. (1983). *Hanyu yufa* 漢語語法 (Chinese Grammar). Transl. by Huang X. 黃宣範. Taibei: Wenhe chuban youxian gongsi. Chinese transl. of: *Mandarin Chinese. A Functional Reference Grammar*. Berkeley: University of California Press, 1981.

Li, R. (2004). *Modality in English and Chinese. A Typological Perspective* [PhD Dissertation]. Antwerp: University of Antwerp. https://search.proquest.com/docview/305290918?accountid=17274.

Lian, Z.; Jiang, T. (2014). "A Study of Modality System in Chinese-English Legal Translation from the Perspective of SFG". *Theory and Practice in Language Studies*, 4(3), 497-503. https://doi.org/10.4304/tpls.4.3.497-503.

Lǚ S. 呂叔湘 et al. (2004). *Xiandai Hanyu babai ci* 現代汉语八百词 (800 Words of Modern Chinese). Beijing: Shangwu yinshuguan.

Lyons, J. (1977). *Semantics*, vol. 2. Cambridge: Cambridge University Press.

Nuyts, J. (2016). "Surveying Modality and Mood. An Introduction". Nuyts, van der Auwera 2016, 1-8.

Nuyts, J.; van der Auwera, J. (eds) (2016). *The Oxford Handbook of Modality and Mood*. Oxford: Oxford University Press.

Palmer, F. (2001). *Mood and Modality*. 2nd ed. Cambridge: Cambridge University Press.

Peng L. 彭利贞 (2005). *Xiandai Hanyu qingtai yanjiu* 现代汉语情态研究 (Research on Modern Chinese Modality). Shanghai: Fudan University.

Pierini, P. (2015). "Translating English Compound Adjectives into Italian. Problems and Strategies". *The International Journal for Translation & Interpreting Research*, 7(2), 17-29. https://doi.org/10.12807/ti.107202.2015.a02.

Pope Francis (2015). *Encyclical Letter Laudato Si' of the Holy Father Francis on Care for Our Common Home*. http://www.vatican.va/content/francesco/en/encyclicals/documents/papa-francesco_20150524_enciclica-laudato-si.html.

Portner, P. (2009). *Modality*. Oxford: Oxford University Press.

Quirk, R. et al. (1985). *A Comprehensive Descriptive Grammar of the English Language*. London; New York: Longman.

Smith, C.S.; Erbaugh, M.S. (2005). "Temporal Interpretation in Mandarin Chinese". *Linguistics*, 43(4), 713-56. https://doi.org/10.1515/ling.2005.43.4.713.

Spadaro, A. (2015). "Laudato Si'. Guida alla lettura dell'enciclica di papa Francesco". *La Civiltà Cattolica*, 3961(III), 3-22. https://www.laciviltacattolica.it/articolo/laudato-si-guida-alla-lettura-dellenciclica-di-papa-francesco.

Sparvoli, C. (2012). *Deontico e anankastico. Proposta di ampliamento della tassonomia modale basata sull'analisi dei tratti distintivi dei modali cinesi inerenti dovere e necessità* [PhD Dissertation]. Venice: Ca' Foscari University of Venice. http://hdl.handle.net/10579/1228.

Stibbe, A. (2014). "An Ecolinguistic Approach to Critical Discourse Studies". *Critical Discourse Studies*, 11(1), 117-28. https://doi.org/10.1080/17405904.2013.845789.

Tang T. 湯廷池 (2000). "Hanyu qingtai fuci. Yuyi neihan yu jufa gongneng" 漢語的情態副詞——語意內涵與句法功能 (Chinese Modal Adverbs. Semantic Connotation and Syntactic Function). *Zhongyang Yanjiuyuan Lishi Yuyan Yanjiusuo Jikan* 中央研究院歷史語言研究所集刊 (Collective volume of the 'Institute of History and Philology - Academia Sinica'), 71(1), 199-219.

Teubert, W. (2007). "Natural and Human Rights, Work and Property in the Discourse of Catholic Social Doctrine". Hoey, M.; Mahlberg, M.; Stubbs, M.; Teubert, W. (eds), *Text, Discourse and Corpora*. London; New York: Continuum, 89-126.

Tilche, A.; Nociti, A. (2015). "Laudato Si'. The Beauty of Pope Francis' Vision". *S.A.P.I.EN.S.*, 8(1), 1-5. http://journals.openedition.org/sapiens/1704.

Tognini-Bonelli, E. (1996). "Towards Translation Equivalence from a Corpus Linguistics Perspective". *International Journal of Lexicography*, 9(3), 197-217. https://doi.org/10.1093/ijl/9.3.197.

Tsang, C.-L. (1981). *A Semantic Study of Modal Auxiliary Verbs in Chinese* [PhD Dissertation]. Stanford, CA: Stanford University.

van der Auwera, J.; Plungian, V. (1998). "Modality's Semantic Map". *Linguistic Typology*, 2(1), 79-124. https://doi.org/10.1515/lity.1998.2.1.79.

von Wright, E.H. (1951). *An Essay in Modal Logic*. Amsterdam: North Holland.

Xiandai Hanyu Cidian 现代汉语词典 (The Contemporary Chinese Dictionary). 6th ed. (2014). Beijing: Shangwu Yinshuguan.

Xiao, R. (2010). "How Different Is Translated Chinese from Native Chinese? A Corpus-Based Study of Translation Universals". *International Journal of Corpus Linguistics*, 15(1), 5-35. https://doi.org/10.1075/ijcl.15.1.01xia.

Xiao, R.; Dai, G. (2014). "Lexical and Grammatical Properties of Translational Chinese. Translation Universal Hypotheses Reevaluated from the Chinese Perspective". *Corpus Linguistics and Linguistic Theory*, 10(1), 11-55. https://doi.org/10.1515/cllt-2013-0016.

Xiao, R.; Wei, N. (2014). "Translation and Contrastive Linguistic Studies at the Interface of English and Chinese. Significance and Implications". *Corpus Linguistics and Linguistic Theory*, 10(1), 1-10. https://doi.org/10.1515/cllt-2013-0015.

Morphology and the Lexicon

Corpus-Based Research on Chinese Language and Linguistics
edited by Bianca Basciano, Franco Gatti, Anna Morbiato

Co-Varying Collexeme Analysis of Chinese Classifiers 棵 *kē* and 株 *zhū*

Aneta Dosedlová
Masaryk University, Brno, Czechia

Wei-lun Lu
Masaryk University, Brno, Czechia

Abstract The numeral classifier is a grammatical category in plenty of East Asian languages, with Chinese being one of the most widely reported. In Chinese, there are many classifiers that are near-synonymous, meaning that certain classifiers may be interchangeable in certain contexts. However, these classifiers are used with semantically similar nouns and, as a result, the distinction between the various usages is not always clear. In view of this issue, we propose to study near-synonymous classifiers using the co-varying collexeme method and the Euclidean distance, by exploring the case of the classifiers 棵 *kē* and 株 *zhū*. We report results that not only partially confirm but also complement what has been found in previous raw-frequency-based research.

Keywords Categorization. Collostructional analysis. Co-varying collexeme analysis. Eluclidean distance. Near-synonymy. Prototype.

Edizioni Ca'Foscari

Sinica venetiana 6
e-ISSN 2610-9042 | ISSN 2610-9654
ISBN [ebook] 978-88-6969-406-6 | ISBN [print] 978-88-6969-407-3

Peer review | Open access
Submitted 2020-04-17 | Accepted 2020-10-14 | Published 2020-12-21

DOI 10.30687/978-88-6969-406-6/007

1 Near-Synonymy. What It Is and the State of the Art[1]

The linguistic issue of near-synonymy is never an easy one. For decades, there have been different approaches trying to discuss and settle how different words have similar meanings and in what situations they do, based on conceptual semantic discussions, usage dictionaries, or a scrutiny of a body of linguistic samples. Among the numerous types of efforts, recent decades have witnessed the rise of corpus linguistics, which offers a methodological opportunity to approach linguistic phenomena in a way that can be faithful to how a word is actually used in real-world context. Based on the principle that one should "know a word by the company it keeps" (Firth 1957, 11), there have been numerous studies applying such rubric in the study of lexical semantics, generalising the contextual information over a number of usages of a particular word, in order to understand the lexical and grammatical company kept by the word at issue.

In corpus linguistics, there are several methods used to study similar and potentially confusing words, with the one most relevant to the present study being *collostructional analysis* (Stefanowitsch, Gries 2003; Schmid 2010; Schmid, Küchenhoff 2013), which is a family of corpus-based quantitative methods that helps measure mutual attraction between lexemes and constructions. Collostructional methods do not simply rely on numbers of lexical frequencies, but also measure the degree of probability that the patterns of analysed frequencies are due to chance. Such analyses work under the rubrics of *construction grammar* (Goldberg 1995), which claims that lexical and grammatical constructions are symbolic form-meaning pairings.[2] Collostructional analyses compare the strength of association between the analysed constructions and the chosen lexical elements in the actual use found in linguistic corpora.

In the present study, we employ the collostructional method called *co-varying collexeme analysis* (Stefanowitsch, Gries 2005;

1 The completion of this paper was partially supported by the grant "The influence of socio-cultural factors and writing system on perception and cognition of complex visual stimuli" (GC19-09265J), of which the second Author is a member. The analysis of this paper is based on the raw data obtained from the first Author's master's thesis research. We especially thank Dr. Alvin Cheng-hsien Chen for his kind advice on the statistical methods used in this paper. Thanks also go to the editors of the volume and the anonymous reviewers. All correspondences and requests for reprints should be addressed to the second Author at wllu@med.muni.cz.
Author contributions: both Authors conceptualised the study (main responsibility being with the first Author). The data collection and annotation were done by the first Author. All the sections were jointly written by both Authors.

2 Interested readers are referred to an overview of the position of synonymy research within Cognitive Linguistics in Glynn 2014.

Tang 2016), due to the nature of the linguistic phenomenon that we investigate. We will return to this point in § 3.

2 Classifier Constructions in Chinese and Their Near-Synonymy

Classifiers are linguistic devices that help humans categorise objects in the world. In language, classifiers are words that encode "salient perceived or imputed characteristic of the entity to which the associated noun refers" (Allan 1977, 285). Tai (1994) takes a similar stance and argues that Chinese classifiers are used to denote a group of perceptually- or functionally- based attributes associated with a given noun. Among all the systems of classifiers, the numeral classifier system is one of the most commonly recognised type (Aikvenhald 2003; Saalbach, Imai 2012). The usage of numeral classifiers is mostly compulsory with counting objects in a classifier language, which is also the case for Chinese. In a classifier language, a typical classifier construction consists of a numeral, a classifier, and a noun (Allan 1977, 288). In Chinese, the grammatical schema of such construction is [QUAN]-[CLF]-[N], exemplified by (1) below.[3]

1. 一只狗
   yī *zhī* *gǒu*
   one CLF dog
   'one dog'

The choice of a numeral classifier is never random but is based on the perceived properties of the head noun (Tai 1994; Jiang 2017). For the choice of a classifier in a usage like (1), when a speaker of Chinese (or a learner of Chinese as a second language) expresses the quantity of a noun such as 狗 *gǒu*, the noun needs to take a suitable classifier from the conceptual category of ANIMACY[4] that captures the imputed characteristics associated with DOG. As there are multiple classifiers in each linguistic category and as some of them overlap in meaning, by using a classifier, the speaker *profiles* (Langacker 2008, 66) a perceptual or a functional aspect of the noun. For instance, the classifiers for PLANT 棵 *kē* and 株 *zhū* are near-synonymous and interchangeable in certain contexts, as exemplified by (2a) and (2b) (cited from Dosedlová, Lu 2019, 115).

3 The glosses in this paper follow the general guidelines of the Leipzig Glossing Rules, with the addition of LK = 'linker'. Further in-text abbreviations include: N = 'noun'; QUAN = 'quantifier'.

4 We follow the typographic convention in Cognitive Linguistics, which uses lower caps to represent a concept.

2. a. 爸爸买了两棵巨大的圣诞树
bàba *mǎi-le* *liǎng-kē* *jùdà-de* *shèngdàn-shù*
father buy-PFV two-CLF big-LK Christmas-tree
'Father bought two huge Christmas trees'.

b. 爸爸买了两株巨大的圣诞树
bàba *mǎi-le* *liang-zhū* *jùdà-de* *shèngdàn-shù*
father buy-PFV two-CLF big-LK Christmas-tree
'Father bought two huge Christmas trees'. (constructed from (2a))

In their study, Dosedlová and Lu argue that 棵 *kē* and 株 *zhū* conceptually profile slightly different aspects of PLANT – by observing the span of nouns the classifiers co-occur with, the authors report that 株 *zhū* occasionally co-occurs with nouns of PLANT that invoke SMALL and VULNERABLE, such as 苗 *miáo* 'seedling' and 花 *huā* 'flower', and nouns of MICRO-ORGANISM, such as 霉 *méi* 'mold', 细菌 *xìjùn* 'bacterium', 病毒 *bìngdú* 'virus', and so on, but that pattern is not seen among the nouns that co-occur with 棵 *kē* as a classifier. However, a methodological insufficiency of that paper is that the observations are based merely on separate raw frequency counts of *each* of the slots in the classifier construction, while no attention is paid to how the multiple slots in the construction interact.[5] Therefore, to investigate the interaction between different slots within a construction, an alternative must be sought.

From an onomasiological point of view, it will be useful to find out the interaction and the detailed relationship between the classifier and the noun within [QUAN]-[CLF]-[N]. Therefore, we would like to focus on how the two slots in that particular construction (and *only* in that particular construction, *not elsewhere* in the language/corpus) co-vary. After all, a word with classifier as part of its syntactic function may occur in various grammatical constructions in Chinese, which is the case for 只 (also as an adverb when pronounced as *zhǐ* or as a noun when pronounced as *zhī*), 棵 *kē* (also as a noun), and 株 *zhū* (also as a noun or a verb), among numerous others, but that is something we would certainly like to exclude in order to achieve a more statistically-precise result. For this purpose, we consider it suitable to conduct the so-called co-varying collexeme analysis. Such an analysis always *begins with a construction* and studies which lexemes tend to be attracted to that particular construction and which do not. A typical collostructional analysis relies on frequency measures of tokens of different types of lexemes extracted from a corpus. Once obtained from the language sample, the frequencies are

5 A similar general observation from studies done in cognitive semantics is made in Stefanowitsch, Gries 2005, 1.

used for calculating the *p*-values of the list of collexemes (lexemes that may be attracted to a particular construction), which show the degree of association between the collexemes and the construction. Each lexeme analysed has its own *p*-value, which indicates its collocational strength with the construction. The calculation is done via the Fisher-Yates Exact test.

3 Co-Varying Collexeme Analysis and Euclidean Distance

In a co-varying collexeme analysis, it is important to identify the association strength between pairs of lexical items appearing in two different slots of the same construction. In our study, the lexical slots to examine are the CLF and the N within the [QUAN]-[CLF]-[N] construction. To conduct such an analysis, we first need to find out the span of lexemes that may occur in each of the slots investigated. We also need the frequency of the construction (C) investigated (which is the total number of concordance lines included in the sample), the frequency of the first target word (L1) in a particular slot (S1) in C in the sample, and the frequency of the second target word (M1) in the other slot (S2) in C in the sample. A template is shown in table 1 below.

Table 1 A schematic distribution table for a co-varying collexeme analysis (adapted from Stefanowitsch, Gries 2005, 9)

	M1 in S2 of C	**Other words (M2, M3…) in S2 in C**	**Total**
L1 in S1 in C	frequency of S2(M1) and S1(L1) in C	frequency of S2(⌐M1) and S1(L1) in C	total frequency of S1(L1) in C
other words (L2, L3…) in S1 in C	frequency of S2(M1) and S1(⌐L1) in C	frequency of S2(⌐M1) and S1(⌐L1) in C	total frequency of S1(⌐L1) in C
total	total frequency of S2(M1) in C	total frequency of S2(⌐M1) in C	total frequency of C

We illustrate such a template with the case study of the distribution of the causing event and the resulting event in the English *into* causative (Stefanowitsch, Gries 2005), as in *we must not fool ourselves into thinking there is no longer any problem*. To determine the extent of the correlation between *fool* (as the causing event) and *think* (as the resulting event) in *fool into thinking*, a distribution table for this pair of lexemes is given in table 2.

Table 2 Information needed for studying the correlation between *fool* and *think* in *fool into thinking* (Stefanowitsch, Gries 2005, 10)

	think	**Other verbs**	**Total**
fool	46 (7)	31 (70)	77
Other verbs	101 (140)	1,408 (1,369)	1,509
Total	147	1,439	1,586

Such a table is submitted to a contingency test and the whole procedure is done for *each* word pair appearing in the construction in question. The data of the tables is submitted to Fisher-Yates Exact test. The result of this test is a *p*-value that indicates the association strength between the lexeme and the construction. The strongest mutual association between a lexeme and a construction is the one with the smallest *p*-value (Desagulier 2014, 157). Co-varying collexemes are those pairs of words that co-occur more frequently than by pure chance (Stefanowitsch, Gries 2003, 2005). The final result can be submitted to further analysis, such as *cluster analysis* (Divjak 2010; Divjak, Fieller 2014), for a more detailed understanding of the results. Table 3 shows the information needed for studying the correlation between a classifier and the noun in [QUAN]-[CLF]-[N].

Table 3 Information needed for studying the correlation between CLF and N in [QUAN]-[CLF]-[N]

	CLF1 in S1 in [QUAN]-[CLF]-[N]	**Other words (CLF2, CLF3…) in S1 in [QUAN]-[CLF]-[N]**	**Total**
N1 in S2 in [QUAN]-[CLF]-[N]	frequency of S1(CLF1) and S2(N1) in [QUAN]-[CLF]-[N]	frequency of S1(¬CLF1) and S2(N1) in [QUAN]-[CLF]-[N]	total frequency of S2(N1) in [QUAN]-[CLF]-[N]
other words (N2, N3…) in S2 of [QUAN]-[CLF]-[N]	frequency of S1(CLF1) and S2(¬N1) in [QUAN]-[CLF]-[N]	frequency of S1(¬CLF1) and S2(¬N1) in [QUAN]-[CLF]-[N]	total frequency of S2(¬N1) in [QUAN]-[CLF]-[N]
total	total frequency of S1(CLF1) in [QUAN]-[CLF]-[N]	total frequency of S1(¬CLF1) in [QUAN]-[CLF]-[N]	total frequency of [QUAN]-[CLF]-[N]

Cluster analysis is a family of statistical methods used for deciding the distance and similarities between entities, which may be applied to the study of language to measure the internal structure of a set of synonymous lexical constructions. Divjak and Gries (2006), for in-

stance, study nine Russian verbs that all share the tentative meaning of TRY. The paper examines 1,585 concordance lines by tagging the individual usages using morphosyntatic cues that may influence the behavioural profile of the nine verbs. The authors find that the nine verbs form three groups and that each group exhibits similar internal behaviours, which means that the members in a group have smaller conceptual semantic distances with each other than with members outside the group.

The first step in conducting a cluster analysis is to choose the variables. There are several kinds of variables to choose from, which can be numerical, categorical, or ordinal.[6] We illustrate this with a simplified example below. Let us suppose we have four constructions (C1, C2, C3 and C4) to analyse. We also assume there are four possible variables that may factor in learning about the conceptual semantic distance between the four words, including: frequency in the corpus, co-occurrence with Word *x*, co-occurrence with Word *y*, and co-occurrence with an adjective. The hypothetical situation is put forth in table 4.

Table 4 A possible scenario with four constructions and four variables for a cluster analysis

	C1	**C2**	**C3**	**C4**
frequency in corpus	379	254	468	342
co-occurrence with *x*	257	159	374	285
co-occurrence with *y*	53	49	85	62
co-occurrence with adjective	81	37	103	64

The next step is to decide on a method for calculating the similarities among the words involved. In a cluster analysis, one of the most common methods for calculating distances (similarities) is *Euclidean distance*. The result of such method is a dissimilarity matrix table, which shows the distances among all the entities within a dataset.

The Euclidean distance between two objects is gained by summing the squared differences between the pairs of corresponding values for the two individuals and taking the square root of the sum (Divjak, Fieller 2014, 417). The formula for the calculation of Euclidean distance is as follows:

$$d_y = \sqrt{\sum_{k=1}^{n} \left(x_{ik} - x_{jk}\right)^2}$$

6 Interested readers are referred to Divjak, Fieller 2014 for a detailed discussion on how to choose the variables.

Following the hypothetical situation outlined in table 4, a Euclidean distance analysis can be conducted using the above formula for the set of the target words. For instance, the similarity distance between C1 and C2 can be figured out as follows:

$$d_{c1c2}=\sqrt{(379-254)^2+(257-159)^2+¿(53-49)^2+(81-37)^2}=164.9¿$$

The same can be done between each two of the four: the results are summarised in table 5. The lowest number in each column in bold indicates the smallest distance (or the highest degree of similarity) between words. As the table shows, the closest items are C1 and C4, with a distance of 50.23 (underlined, in bold), and the most dissimilar items are C2 and C3, with a distance of 312.5 (underlined only).

Table 5 Summarised result of the Euclidean distances based on table 4

	C1	**C2**	**C3**	**C4**
C1	0	164.9	152.0	**50.23**
C2	164.9	0	312.5	156.6
C3	152.0	312.5	0	160.8
C4	**50.23**	156.6	160.8	0

Having introduced the related statistical algorithms, now we move on to a detailed description of the research issues and the research steps.

4 Research Issue, Scope, and Steps

In this paper, we address the following issues: first of all, what can we learn about the relationships between a pair of synonymous classifiers using a co-varying collexeme analysis? In what way does the Euclidean distance help? We believe that the relationships between the synonymous classifiers can be made available based on the nouns that collocate with each of these classifiers and that a co-varying collexeme analysis will provide useful data related to the behaviour of the classifiers involved, including the collocational strength and certain association measures. Such results are what we may further submit for a cluster analysis in order to explicate the internal structure of the synonymous set. Secondly, does the co-varying collexeme analysis and an analysis based on the Euclidean distance tell us anything beyond an analysis informed only by a raw frequency count of the lexical items in question?

To answer the questions above, we chose to investigate the classifiers 棵 *kē* and 株 *zhū*, which had already been examined based on a raw frequency approach in Dosedlová and Lu (2019). In that paper, the authors used data extracted from Sketch Engine[7] and observed the types of nouns that occurred in their language sample, and the token frequencies of each of the nouns, which allowed the authors to come up with the conceptual similarities and differences between the two classifiers. In order to see how a different methodological approach may shed alternative light on the same linguistic phenomenon, we extracted the collocating nouns and analysed the data to calculate their T-score, MI score and logDice. After that, we calculated the Euclidean distance between the nouns in the dataset. The steps are outlined below.

In order to properly sample the usages of each of the classifiers investigated, we built a corpus for each of the classifiers by extracting random concordance lines from a large representative body of authentic linguistic data. To this end, we used the function 'sample' of Sketch Engine, which created a random collection of concordances that involved the two target classifiers. We set the size of each subcorpus five hundred lines, which was more than sufficient to investigate the semantics of a common word.[8] After we input the extracted data to Excel, we went through the data manually to look for the collocating nouns and their frequencies in the sub-corpora. In addition, we looked up the frequencies of each of the collocating nouns in each of the sub-corpora. All the information acquired from the above steps was used to calculate the association measures and collocational strengths in the co-varying collexeme analysis. These association measures included: 1) T-score, which indicates the level of certainty with which one can argue for a clear association between the linguistic units analysed. A T-score higher than 2 is seen as statistically significant, which means that the co-occurrence of the two linguistic units is more than mere chance. 2) logDice, which is a measure of the typicality of the co-occurrence of the classifier and its collocating noun. The maximum logDice value is 14, which means the exclusive collocation between the linguistic units investigated (that all occurrences of X co-occur with Y and vice versa). A negative value means that the XY collocation is not statistically significant. 3) MI score, which stands for the extent to which words co-occur compared to the frequency of their separate appearance. An MI score higher than 3 is an indicator of a statistically significant collocation. The lower the MI score, the more likely the linguistic units co-occur only by chance.

7 https://www.sketchengine.eu.

8 Sinclair (2005) claims that it takes around 20 tokens to determine the meaning of a not particularly complicated lexeme and around 50 tokens for an average lexeme.

The three association measures may or may not converge, as we will show in the body of the analysis.

5 Results

In this section, we report the findings based on the data retrieved from Sketch Engine following the steps outlined above.

5.1 Nouns in [QUAN]-[*kē*]-[N]: Their T-Score and logDice

In the sub-corpus of 棵 *kē*, we found 38 different nouns that co-occurred with the classifier. Below, we discuss the association measures of T-score and logDice.

It is important to bear in mind that each of these measures takes a different approach in measuring the strength of the collocation. If we look at the most frequent noun collocating with 棵 *kē*, i.e 树 *shù* 'tree', its T-score and logDice are the highest among all collocating nouns, but its MI score is not. The reason is that the MI score is strongly influenced by the size of the corpus, hence it is usually considered subsidiary if compared to the T-score. As for the T-score, it promotes pairings that are frequently observed but does not concern the total frequencies of each of the linguistic units, hence the size of the corpus is irrelevant. For instance, if we look at the noun 木棉树 *mùmiánshù* 'cotton tree', the T-score is relatively low because there are only three tokens of its collocation with 棵 *kē*, but the MI score is quite high, as the MI score takes into account all the other occurrences of both of the words. As for the logDice, it is an important indicator of the typicality of a collocation.

Therefore, in this study, T-score and logDice are our main foci. Table 6 lists the first five nouns with the highest T-score and the highest logDice in the sub-corpus of 棵 *kē*.

Table 6 Top five collocations with 棵 *kē* in terms of T-score and logDice

Noun	T-score	Noun	LogDice
树 *shù* 'tree'	16.3200	树 *shù* 'tree'	5.7562
树木 *shùmù* 'tree-wood'	3.9987	杨树 *yángshù* 'poplar'	3.4276
杨树 *yángshù* 'poplar'	3.2991	树木 *shùmù* 'tree-wood'	3.2250
树苗 *shùmiáo* 'tree seedling'	3.1353	树苗 *shùmiáo* 'tree seedling'	3.2247
果树 *guǒshù* 'fruit tree'	3.0853	果树 *guǒshù* 'fruit tree'	2.8927

As we see in table 6, the two association measures largely overlap and jointly confirm the status of 树 *shù*, 杨树 *yángshù*, 树木 *shùmù*, 树

苗 *shùmiáo*, and 果树 *guǒshù* being statistically significant collocates of 棵 *kē*. 树 *shù* is the most significant lexeme attracted to [QUAN]-[*kē*]-[N], based on the T-score and the logDice.

5.2 Nouns in [QUAN]-[*zhū*]-[N]: Their T-Score and logDice

The same analysis was done with the nouns that co-occurred with 株 *zhū*. In the sub-corpus, there are 75 different nouns found to co-occur with 株 *zhū*. We also calculated the T-score and the logDice for each of the nouns, now listing the top five in terms of the T-score and the logDice in table 7.

Table 7 Top five collocations with 株 *zhū* in terms of T-score and logDice

Noun	T-score	Noun	LogDice
树 *shù* 'tree'	13.4313	苗 *miáo* 'seedling'	6.0427
苗 *miáo* 'seedling'	10.9546	树 *shù* 'tree'	5.1596
花 *huā* 'flower'	9.2243	植树 *zhíshù* 'plant-tree'	4.4780
植树 *zhíshù* 'plant-tree'	6.3984	菌 *jùn* 'bacteria'	4.4602
苗木 *miáomù* 'seedling'	6.0901	苗木 *miáomù* 'seedling'	4.4198

As we can see in table 7, the top five collocates in terms of each of the association measures still largely overlap, which confirms the status of 树 *shù*, 苗 *miáo*, 植树 *zhíshù*, and 苗木 *miáomù* as the most statistically significant lexemes that are attracted to [QUAN]-[*zhū*]-[N].

However, if we compare all the five most significant collocates between the two classifiers in the corpora, we see that 棵 *kē* generally collocates with nouns that contain 树 *shù* as part of it, whereas the significant collocates of 株 *zhū* are more diversified (that is, do not necessarily involve 树 *shù* as part of the lexeme). In addition, 株 *zhū* has collocates that invoke SMALL and VULNERABLE, such as 苗 *miáo*, 花 *huā*, and 菌 *jùn*. We will return to this point when we compare the results from this co-varying collexeme analysis with the results in Dosedlová and Lu (2019).

A comparison of tables 6 and 7 allows us to identify 树 *shù* as the lexeme that appears in both tables, meaning that it is the lexeme that has the highest T-score and logDice in both [QUAN]-[*kē*/*zhū*]-[N], indicating the strongest attraction between 树 *shù* and the two classifier constructions. Based on this fact, we may say that 树 *shù* is the prototypical lexical instantiation of PLANT that collocates with both 棵 *kē* and 株 *zhū* (but only within the particular construction of [QUAN]-[CLF]-[N] and only when it co-varies with 棵 *kē* and 株 *zhū*, rather than in Chinese in general). In addition to 树 *shù*, 苗 *miáo* is also a lexeme that has a very high T-score and logDice in [QUAN]-[*zhū*]-

[N], so is another prototypical lexical instantiation of PLANT in that classifier construction. We will return to this point in our discussion.

5.3 A Cluster Analysis of Nouns within [QUAN]-[*kē*/*zhū*]-[N]

After we obtained the association measures, we further submitted the numbers to a cluster analysis based on the Euclidean distance. In the analysis we used the same corpora, where we first identified the nouns that collocated with both of the classifiers. There are fourteen of such nouns, which includes 树 *shù* 'tree', 槐树 *huáishù* 'Chinese scholar tree', 果树 *guǒshù* 'fruit tree', 杨树 *yángshù* 'poplar tree', 植树 *zhíshù* 'plant-tree', 松树 *sōngshù* 'pine tree', 柳树 *liǔshù* 'willow', 树木 *shùmù* 'tree-wood', 林木 *línmù* 'forest', 银杏 *yínxìng* 'ginkgo', 柳杉 *liǔshān* 'Japanese cedar', 核桃 *hétáo* 'walnut', 樱花 *yīnghuā* 'cherry blossom', 玉米 *yùmǐ* 'corn', and 桂花 *guìhuā* 'osmanthus'.

Secondly, we calculated the Euclidean distance between the fourteen nouns that co-occurred with 棵 *kē* and 株 *zhū* within the construction [QUAN]-[CLF]-[N], following the formula introduced in § 3 and using the raw frequency, T-score, MI value and logDice of the fourteen lexemes as the possible variables. A summary of the Euclidean distances is given as table 9.

Table 9 Euclidean distances between pairs of the fourteen nouns co-occurring with 棵 *kē* and 株 *zhū* within [QUAN]-[CLF]-[N]

	shù	***huáishù***	***guǒshù***	***yángshù***	***zhíshù***	***sōngshù***	***liǔshù***	***shùmù***	***línmù***	***yínxìng***	***liǔshān***	***hétáo***	***yīngtáo***	***yùmǐ***	***guìhuā***
shù	0.0000	10.4612	6.4445	5.4078	3.3374	4.4432	11.7994	2.5257	8.3046	5.3465	8.7752	**14.0385**	**12.8982**	6.3567	8.4151
huáishù	10.4612	0.0000	5.3431	6.8035	9.5656	6.8840	3.0736	8.9606	6.9774	7.8378	10.7145	6.0012	2.4650	10.2923	9.9332
guǒshù	6.4445	5.3431	0.0000	1.4954	4.4740	2.0294	5.5865	4.2356	2.8454	2.4955	5.9007	7.6001	7.6979	4.9964	5.1410
yángshù	5.4078	6.8035	1.4954	0.0000	3.0277	1.1295	7.0205	2.9924	2.9934	1.0824	5.0565	8.8309	9.1825	3.6956	4.3665
zhíshù	3.3374	9.5656	4.4740	3.0277	0.0000	2.6831	10.0446	1.2045	5.4170	2.4293	5.4401	11.7935	12.0071	3.0495	5.0853
sōngshù	4.4432	6.8840	2.0294	1.1295	2.6831	0.0000	7.5671	2.2245	4.1106	1.8019	5.9873	9.6246	9.3303	4.2890	5.3452
liǔshù	11.7994	3.0736	5.5865	7.0205	10.0446	7.5671	0.0000	9.7916	5.8479	7.8344	9.5297	2.9601	3.6330	9.8310	8.7971
shùmù	2.5257	8.9606	4.2356	2.9924	1.2045	2.2245	9.7916	0.0000	5.7990	2.8214	6.4452	11.7939	11.4239	4.1497	6.0053
línmù	8.3046	6.9774	2.8454	2.9934	5.4170	4.1106	5.8479	5.7990	0.0000	3.0166	3.7724	6.7290	8.9456	4.1286	3.0044
yínxìng	5.3465	7.8378	2.4955	1.0824	2.4293	1.8019	7.8344	2.8214	3.0166	0.0000	4.1896	9.4007	10.1869	2.6200	3.5685
liǔshān	8.7752	10.7145	5.9007	5.0565	5.4401	5.9873	9.5297	6.4452	3.7724	4.1896	0.0000	9.8562	12.7176	2.4658	0.7830
hétáo	**14.0385**	6.0012	7.6001	8.8309	11.7935	9.6246	2.9601	11.7939	6.7290	9.4007	9.8562	0.0000	5.8824	10.8428	9.2455
yīnghuā	**12.8982**	2.4650	7.6979	9.1825	12.0071	9.3303	3.6330	11.4239	8.9456	10.1869	12.7176	5.8824	0.0000	12.5538	11.9492
yùmǐ	6.3567	10.2923	4.9964	3.6956	3.0495	4.2890	9.8310	4.1497	4.1286	2.6200	2.4658	10.8428	12.5538	0.0000	2.3161
guìhuā	8.4151	9.9332	5.1410	4.3665	5.0853	5.3452	8.7971	6.0053	3.0044	3.5685	0.7830	9.2455	11.9492	2.3161	0.0000

The summary in table 9 allows us to compare the Euclidean distance between all the nouns involved and the prototypical PLANT within the two particular grammatical constructions. Remember that 树 *shù* is the lexical prototype in both constructions. In table 9, we can see that

among the fourteen lexemes shared by the two classifier constructions, 核桃 *hétáo* and 樱花 *yīnghuā* are the two lexemes that have the highest Euclidean distance from 树 *shù*, with a Euclidean distance value of 14.0385 and 12.8982 (in bold), respectively. This means that the behaviours of these two lexemes are the most different from the prototype in the corpora. On the other hand, the two lexemes that have the smallest Euclidean distance with 树 *shù* are 树木 *shùmù* and 植树 *zhíshù*, having a Euclidean distance value of 2.5257 and 3.3374 (underlined), respectively, meaning that the two lexemes have the most similar behaviour with 树 *shù* in the corpora. Note that the two lexemes are also conceptually closer to 树 *shù* than the other lexemes, as they do not refer to any particular type of tree, so are at the same level with 树 *shù* in terms of taxonomy. Therefore, the similar behaviour between 树 *shù*, 树木 *shùmù* and 植树 *zhíshù* is natural.

6 Discussion and Concluding Remarks

The statistically informed analysis in the present paper largely confirms the results in Dosedlová and Lu's (2019) study based on raw lexical frequencies, but it also turns up meaningful patterns that were not reported in the previous study.

In particular, based on the T-score and the logDice, we firstly confirm that 树 *shù* is the lexeme that has the strongest association measures with both [QUAN]-[*kē*]-[N] and [QUAN]-[*zhū*]-[N]. This matches the fact that 树 *shù* is the most frequent noun that co-occurs both with 棵 *kē* and with 株 *zhū* (Dosedlová, Lu 2019, 123). Following on from that, we see that the raw frequency, T-score and logDice constitute pieces of converging evidence that jointly support the claim that 树 *shù* is the prototypical lexical instantiation of PLANT in [QUAN]-[*kē*/*zhū*]-[N]. Secondly, the statistically informed analysis allows us to confirm that [QUAN]-[*zhū*]-[N] does attract nouns that invoke SMALL and VULNERABLE, such as 苗 *miáo*, 花 *huā*, and 菌 *jùn* (Dosedlová, Lu 2019, 122). In the above two respects, the results obtained via a co-varying collexeme approach echo the findings based on raw lexical frequency.

However, a co-varying collexeme analysis can build on the previous analysis and can allow us to see patterns beyond an exclusively raw-frequency-based approach – first of all, it allows us to identify 苗 *miáo* as another lexeme that is strongly associated with [QUAN]-[*zhū*]-[N]. According to the list of token frequencies in Dosedlová and Lu (2019, 123), 苗 *miáo* accounts for 14.3% of the total usages in [QUAN]-[*zhū*]-[N], but that is only less than one third of the percentage of 树 *shù* (which is 47.3% in their table). Accordingly, a study merely based on the token frequency may not give the collocation between 苗 *miáo* and 株 *zhū* too much weight. But once the T-score and the logDice are included, that brings the lexeme back to our at-

tention. Secondly, another linguistic fact that is uncovered through the Euclidean distance is the similarity between each of the fourteen shared lexemes with the prototype 树 *shù*. For instance, the Euclidean distance analysis indicates 树木 *shùmù* and 植树 *zhíshù* to be the lexemes that are most similar to 树 *shù* in terms of the behavioural profile, which cannot be captured by a simple frequency count – that would only identify 木 *mù* and 植 *zhí* being infrequent lexical types in the corpus, about one eighth of 树 *shù* in [QUAN]-[*kē*]-[N] (Dosedlová, Lu 2019, 121) and one fourth of 树 *shù* in [QUAN]-[*zhū*]-[N] (Dosedlová, Lu 2019, 123). In addition, the cluster analysis has found the behavioural profiles of 核桃 *hétáo* and 樱花 *yīnghuā* to be the most distant from the prototype among the fourteen shared lexemes, meaning that the two lexemes behave most differently from 树 *shù* in [QUAN]-[*kē*/*zhū*]-[N], which is an observation that can be made only through a Euclidean distance analysis.

Despite of the advantages of a co-varying collexeme analysis and a cluster analysis mentioned above, we maintain and emphasise that an analysis based on type and token frequencies is still capable of uncovering linguistic facts about near-synonymy that cannot be seen through a collostructional analysis, and that the two approaches should be considered *complementary* to each other. An interesting part of the conceptual semantic difference between 棵 *kē* and 株 *zhū*, for instance, lies in the fact that [QUAN]-[*zhū*]-[N] has an extended group of usages that covers entities that do not invoke PLANT, such as MOLD, BACTERIUM, BIOLOGICAL SUBSTANCE and CHEMICAL SUBSTANCE (Dosedlová, Lu 2019, 122-3). These usages are peripheral members of the linguistic category (defined by the categorising structure [QUAN]-[*zhū*]-[N]) and are very low in lexical frequency. Such periphery of a linguistic category is typically difficult to observe given its low frequency, but may contain important conceptual information that helps define the linguistic category. Such information may become available only through an extensive type frequency analysis of the language sample.

Finally, we would like to conclude by proposing a synergy between different quantitative methods for analysing the near-synonymy of classifiers, similar to the advocacy for a methodological synthesis in Janda, Kudrnáčová and Lu (2019). As we have shown in this paper, each research method has its strengths and its limitations, so we consider it always advisable to try to obtain converging and consolidating evidence from different angles, or to try to obtain comprehensive results from complementary methodological approaches.

Bibliography

Aikvenhald, A.Y. (2003). *Classifiers. A Typology of Noun Categorization Devices*. Oxford: Oxford University Press.

Allan, K. (1977). "Classifiers". *Language*, 53(2), 285-311.

Desagulier, G. (2014). "Visualizing Distances in a Set of Near Synonyms: Rather, Quite, Fairly, and Pretty". Robinson, Glynn 2014, 145-78. https://doi.org/10.1075/hcp.43.06des.

Divjak, D. (2010). *Structuring the Lexicon. A Clustered Model for Near-Synonymy*. Berlin: De Gruyter Mouton.

Divjak, D.; Fieller, N. (2014). "Cluster Analysis. Finding Structure in Linguistic Data". Robinson, Glynn 2014, 405-41.

Divjak, D.; Gries, S.T. (2006). "Ways of Trying in Russian. Clustering Behavioral Profiles". *Corpus Linguistics and Linguistic Theory*, 2(1), 23-60. https://doi.org/10.1515/cllt.2006.002.

Dosedlová, A.; Lu, W. (2019). "The Near-Synonymy of Classifiers and Construal Operation. A Corpus-Based Study of 棵 *kē* and 株 *zhū* in Chinese". *Review of Cognitive Linguistics*, 17(1), 113-30. https://doi.org/10.1075/rcl.00028.dos.

Firth, J.R. (1957). "A Synopsis of Linguistic Theory 1930-1955". *Studies in Linguistic Analysis*. Oxford: Blackwell, 1-32.

Goldberg, A.E. (1995). *Constructions. A Construction Grammar Approach to Argument Structure*. Chicago: Chicago University Press.

Glynn, D. (2014). "Polysemy and Synonymy. Cognitive Theory and Corpus Method". Robinson, Glynn 2014, 7-38.

Janda, L.A.; Kudrnáčová, N.; Lu, W. (2019). "Deep Dives into Big Data. Best Practices for Synthesis of Quantitative and Qualitative Analysis in Cognitive Linguistics". *Review of Cognitive Linguistics*, 17(1), 1-6. https://doi.org/10.1075/rcl.00023.jan.

Jiang, S. (2017). *The Semantics of Chinese Classifiers and Linguistic Relativity*. London: Routledge.

Langacker, R.W. (2008). *Cognitive Grammar. A Basic Introduction*. Oxford: Oxford University Press.

Robinson, J.A.; Glynn, D. (eds) (2014). *Corpus Methods for Semantics. Quantitative Studies in Polysemy and Synonymy*. Amsterdam: John Benjamins.

Saalbach, H.; Imai, M. (2012). "The Relation between Linguistic Categories and Cognition. The Case of Numeral Classifiers". *Language and Cognition Processes*, 27(3), 381-428. https://doi.org/10.1080/01690965.2010.546585.

Schmid, H.-J. (2010). "Does Frequency in Text Instantiate Entrenchment in the Cognitive System?". Fischer, K.; Glynn, D. (eds), *Quantitative Methods in Cognitive Semantics. Corpus-Driven Approaches*. Berlin: De Gruyter Mouton, 101-32. https://doi.org/10.1515/9783110226423.101.

Schmid, H.-J.; Küchenhoff, H. (2013). "Collostructional Analysis and Other Ways of Measuring Lexico-Grammatical Attraction. Theoretical Premises, Practical Problems and Cognitive Underpinnings". *Cognitive Linguistics*, 24(3), 531-78. https://doi.org/10.1515/cog–2013–0018.

Sinclair, J. (2005). "Corpus and Text. Basic Principles". Wynne, M. (ed.), *Developing Linguistic Corpora. A Guide to Good Practice*. Oxford: Oxbow Books, 1-16.

Stefanowitsch, A.; Gries, S.T. (2003). "Collostructions. Investigating the Interaction of Words and Constructions". *International Journal of Corpus Linguistics*, 8(2), 209-43. https://doi.org/10.1075/ijcl.8.2.03ste.

Stefanowitsch, A.; Gries, S.T. (2005). "Covarying Collexemes". *Corpus Linguistics and Linguistic Theory*, 1(1), 1-43. https://doi.org/10.1515/cllt.2005.1.1.1.

Tai, J.H-Y. (1994). "Chinese Classifier Systems and Human Categorization". Chen, M.Y.; Tzeng, O.J. (eds), *Interdisciplinary Studies on Language and Language Change*. Taipei: Pyramid, 479-94.

Tang, X. (2016). "Lexeme-based Collexeme Analysis with DepCluster". *Corpus Lingustics and Linguistic Theory*, 13(1), 165-202. https://doi.org/10.1515/cllt-2015-0007.

Corpus-Based Research on Chinese Language and Linguistics
edited by Bianca Basciano, Franco Gatti, Anna Morbiato

Chinese Affixes in the Internet Era
A Corpus-Based Study of X-族 *zú*, X-党 *dǎng* and X-客 *kè* Neologisms

Bianca Basciano
Università Ca' Foscari Venezia, Italia

Sofia Bareato
Università Ca' Foscari Venezia, Italia

Abstract In the last few decades, under the influence of foreign languages and netspeak, many word-formation patterns emerged in the Chinese lexicon. This paper proposes a corpus-based investigation of three suffixes, i.e. 族 *-zú*, 党 *-dǎng*, and 客 *-kè*, which build words indicating persons with certain characteristics or behaviour, or doing a certain activity. The paper aims at describing and comparing the three word-formation patterns based on these suffixes. It also aims at describing their evolution over time and their grammaticalisation path. In addition, it discusses the diffusion of the three patterns in Chinese and compares their productivity.

Keywords Derivation. Affixes. Word formation. Neologisms. Productivity.

Edizioni Ca'Foscari

Sinica venetiana 6
e-ISSN 2610-9042 | ISSN 2610-9654
ISBN [ebook] 978-88-6969-406-6 | ISBN [print] 978-88-6969-407-3

Peer review | Open access
Submitted 2020-03-27 | Accepted 2020-12-10 | Published 2020-12-21

DOI 10.30687/978-88-6969-406-6/008

1 Introduction[1]

In the last few decades, under the influence of foreign languages and netspeak, many word-formation patterns emerged in the Chinese lexicon. In this paper, we will examine three formatives, i.e. 族 *zú* (orig. 'clan, ethnic group'), 党 *dǎng* (orig. 'party, clique'), and 客 *kè* (orig. 'guest'), used to form nouns indicating persons with certain characteristics or behaviour, or doing a certain activity, as in the following examples: [2]

1. a. 背包族
bèi-bāo-zú
back-pack-ZU
'backpackers'

b. 剁手党
duò-shǒu-dǎng
chop-hand-DANG
'online shopaholics [those who buy things online and then regret it, wanting to cut their own hands off]'

c. 换客
huàn-kè
exchange-KE
'one who sells/exchanges goods online'

Both 族 *zú* and 党 *dǎng* refer to groups of people with common characteristics or behaviour. X-族 *zú* is a quite established and widely studied word-formation pattern in Chinese. Neologisms containing the formative 族 *zú* were attested already in the Nineties and greatly increased in number over the years: between 1995 and 2006, 310 X-族 *zú* neologisms may be found in the 人民日报 *Renmin Ribao* (*People's Daily*) (Chen, Zhu 2010; 309 according to Cao 2007). Many X-族 *zú* words are now listed in dictionaries: the words 上班族 *shàng-bān-zú* 'go-work-ZU, office workers' and 工薪族 *gōngxīn-zú* 'salary-ZU, sal-

1 We are very grateful to two anonymous reviewers for their constructive comments and suggestions. We would also like to thank Giorgio Francesco Arcodia for carefully reading a draft of this paper.

The glosses follow the general guidelines of the Leipzig Glossing Rules, with the addition of SP = 'structural particle'. For academic purposes, Bianca Basciano is responsible for §§ 2, 3.4, 4, 5 and 6, and Sofia Bareato is responsible for §§ 1, 3.1, 3.2 and 3.3. X-族 *zú* and X-党 *dǎng* words were collected by Sofia Bareato in her MA dissertation (Bareato 2017).

2 In order to distinguish these formatives from the corresponding lexemes, we gloss them as ZU, DANG, and KE.

aried people' were included in the 现代汉语词典 *Xiandai Hanyu cidian* (The Contemporary Chinese Dictionary) in 2005 (Cao 2007).

In contrast, X-党 *dǎng* represents a quite novel pattern of word formation (Chen, Zhu 2010). Complex words containing the morpheme 党 *dǎng* as the right-hand constituent with the meaning 'group of people with common characteristics or behaviour' began to appear in the late 2000s; this pattern is mostly used on the web.

Finally, the morpheme 客 *kè* has been used as the right-hand constituent of complex words indicating 'a person doing a certain activity' or 'a person with certain characteristics', since the beginning of the twenty-first century, (again) mostly on the web.

In this paper we will examine a corpus of neologisms containing the three items at issue drawn from the following sources:

- the 新世纪新词语大词典 *Xin shiji xinciyu da cidian* (New Century Comprehensive Dictionary of Neologisms) (henceforth XCY), which collects about 5,400 neologisms coined between 2000 and 2015;
- the Leiden Weibo Corpus (henceforth: LWC),[3] an annotated 100-million-word corpus, consisting of 5,103,566 messages posted on Sina Weibo (China's premier Twitter-like microblogging service) in January 2012.
- the Buzzwords section of the *Shanghai Daily* (henceforth: SD). [4]

We collected 707 distinct words in total: specifically, 434 X-族 *zú* words, 189 X-党 *dǎng* words, and 84 X-客 *kè* words.

The aim of this paper is twofold. First, it aims at describing and comparing the three word-formation patterns at issue, highlighting their formal and semantic properties. Secondly, it aims at describing the evolution of the three formatives over time and their grammaticalisation path, as well as their diffusion in Chinese. To this end, we will also propose an analysis of productivity measures for the three word-formation patterns.

The paper is organised as follows: § 2 provides an overview of derivation in Mandarin Chinese, focusing on its status and on the charac-

3 http://lwc.daanvanesch.nl/index.php.

4 http://buzzword.shanghaidaily.com (2017-02-06). It is a weekly column of the *Shanghai Daily*, started in October 2005. It aims at recording and translating into English new words and phrases appearing in the press, online etc. According to the editor, the purposes of the column are "first, to provide a tentative English translation of new terms and phrases as a reference for our readers; second, to tell our readers what are the latest buzzwords used by locals in their work and daily life; and third, to invite readers to help us generate better English translations of such stylish or trendy Chinese words and phrases" (*Shanghai Daily* 2010). Unfortunately, the column is no longer available; presumably it ceased operations in 2017, when we last consulted it. The buzzwords appeared in the column up to mid-2009 have also been published as a book (*Shanghai Daily* 2010).

teristics of affixes. § 3 is devoted to the presentation of the word-formation patterns at issue, and of their formal and semantic properties. In § 4, we describe the grammaticalisation path of the three formatives, arguing for their affixal status, and we then propose a comparison of the word-formation patterns based on them. Then, in § 5 we compare their productivity in the Leiden Weibo Corpus. Lastly, in § 6 we present our conclusions.

2 Derivation in Mandarin Chinese

While compounding forms words made up of two or more units, be they words (Fabb 1998, 66; Katamba 1993, 54), base lexemes (Haspelmath 2002, 85), stems (Bauer 1998, 404), or roots (Katamba 1993, 54), depending on the morphological profile of the language at issue (Bauer 2006), derivation is a morphological process often involving an affix (Naumann, Vogel 2000, 933-4). Thus, in English, while a word like *zebrafish* is a compound, a word like *violinist* is a derived word. However, the distinction between compounding and derivation is not always clear-cut. In some cases, some elements have hybrid properties, which make it hard to classify them as words or as affixes (Bauer 2005, 106-7). For example, items like *monger*, *cade* or *scape* in English complex words such as *fishmonger*, *motorcade*, *seascape* are not words in Modern English but still retain some kind of full, lexical meaning. In some cases, an affix-like element co-exists with the word it originates from. For example, in Dutch the morpheme *boer* is a word meaning 'farmer'; however, it is also used as the right-hand (head) constituent in complex words with the meaning 'seller of X', as e.g. in *sigaren-boer* 'cigar-farmer, cigar seller', *kabel-boer* 'cable-farmer, provider of broadband cable services' (see Booij 2005). Therefore, we observe semantic differentiation: *boer* is considered an affixal element when, as a right-hand constituent, has the meaning of 'seller', which is not attested in its use as a free form (word).

It has been proposed by many to label these hybrid forms as pseudo-affixes or affixoids (see e.g. Naumann, Vogel 2000), a notion which has been employed in slightly different ways by different authors: as highlighted by Booij (2005), the notion of affixoid is not a theoretical notion, but a convenient descriptive label. These hybrid forms become affixes as soon as their connection with the corresponding lexeme is lost, either because of sound change or because of semantic change, following a process of grammaticalisation.

The issue of the distinction between compounding and derivation is much thornier in Chinese: the existence of derivation as a productive morphological process, distinct from compounding, is under debate (see e.g. Dong 2004). This is due to the characteristics of Mandarin Chinese, an isolating language, where words are generally formed

by the agglutination of morphemes, mostly lexical; compounding is generally regarded as the most productive means of word formation in this language (see Ceccagno, Basciano 2007, 208). In addition, the majority of lexical morphemes are bound (about 70% according to Packard 2000): this means that they cannot occupy a syntactic slot (i.e. they are not words) but have a full lexical meaning and are actively used to form complex words. However, they do not occupy a fixed position, differently from affixes: see e.g. 衣 *yī* 'clothes' (compare the corresponding free form 衣服 *yīfu*) in 大衣 *dà-yī* 'big-clothes, overcoat', 衣蛾 *yī-é* 'clothes-moth, clothes moth'. Furthermore, in Mandarin Chinese, as well as in other languages of East and South-East Asia, there is no regular formal distinction between lexical morphemes and grammatical morphemes (Bisang 1996): except for a few items, there is no phonological reduction nor other formal changes characterising grammaticalised forms.

Only a small number of items are commonly regarded as derivational affixes in the literature, in particular those items which became toneless and lost much of their meaning (and productivity), i.e. 子 *-zi* (< *zǐ* 'child'), as in 桌子 *zhuōzi* 'table', 儿 *-r* (< *ér* 'child'), as in 画儿 *huàr* 'painting', and 头 *-tou* (< *tóu* 'head'), as in 石头 *shítou* 'stone'. As a matter of fact, loss of tone and of lexical meaning seem to be the only criteria accepted by Chinese linguists to include an item among affixes (see Ma 1995).

Other formatives that are usually included among derivational affixes are 化 *-huà* '-ise, -ify' (< *huà* 'change'), as in 国际化 *guójìhuà* 'international-ise, internationalise, internationalisation', and 性 *-xìng* 'nature, -ity, -ness' (< *xìng* 'inherent nature'), as in 可能性 *kěnéng-xìng* 'possible-ity, possibility'. These two suffixes began to be productively used at the beginning of the 20th century due to the influence of Japanese, where they were used to render the equivalent of European suffixes (Masini 1993). The functional correspondence with suffixes in European languages probably favoured their inclusion among derivational affixes (Pan, Ye, Han 2004, 67). However, it must be noted that these word-formation patterns already existed in Chinese; for example, words containing the suffix 化 *-huà* are found in Premodern Chinese, even though this suffix could only be attached to monosyllabic bases (Arcodia, Basciano 2012). Thus, this pattern was somehow independent from the European model, but it strongly developed starting from the 20th century, due to foreign influence. After that, it started to be used independently, creating new words by analogy, thus not only to translate foreign words (Steffen Chung 2006). Therefore, the influence of foreign languages, in this case, gave impulse to an already existent, though not very productive, word-formation pattern.

Besides the cases mentioned above, there are many ambiguous formatives: how to deal with those lexical morphemes which appear

in a fixed position in a high number of complex words, thus showing a high degree of productivity, always conveying the same meaning? Are they to be analysed as compound constituents or as derivational affixes? Recall that, as mentioned earlier, generally there is no formal distinction between lexical morphemes and grammatical morphemes in Chinese. Take, for example, the root 人 *rén* 'person', which is used both as a word and as the right-hand bound constituent in complex nouns indicating a person from a country, town etc., as e.g. 上海人 *Shànghǎi-rén* 'Shanghai-person, Shanghaiese', 意大利人 *Yìdàlì-rén* 'Italy-person, Italian'. Given its "versatility", Yip (2000, 59-60) regards it as a suffix. A similar example is that of 店 *diàn* 'shop', typically used as a constituent in complex words, indicating any kind of shop, as e.g. 书店 *shū-diàn* 'book-shop, bookstore', 布店 *bù-diàn* 'cloth-shop, cloth store', 冷饮店 *lěng-yǐn-diàn* 'cold-drink-shop, cold-drink bar/shop'. Given the high productivity of the pattern, in which 店 *diàn* has a fixed position and a stable meaning, Lü (1941, quoted in Pan, Ye, Han 2004, 468) considers it as a quasi-affix (近似词缀 *jìnsì cízhuì*). However, it must be noted that the number of words built according to a morphological pattern is not normally used as a diagnostic test for affixhood, since compounding patterns too can be very productive. In addition, there is no semantic differentiation observed when these two formatives are used as right-hand constituents bearing a fixed meaning: the meaning of 人 *rén* as a bound right-hand constituent is not different from that of 人 *rén* when used as a free root (word); in the same way, 店 *diàn* retains its original meaning of 'shop', without any kind of bleaching. Also, we may remark that both formatives may be used as left-hand constituents in complex words, bearing the very same meaning: see e.g. 人产 *rén-chǎn* 'person-produce, production per person', 人堆 *rén-duī* 'person-pile, crowd', 店台 *diàn-tái* 'shop-platform, shop counter', 店员 *diàn-yuán* 'shop-member, shop assistant'.

Compare now the examples given above with the root 学 *xué*. In Modern Chinese, 学 *xué* is a free root, a verb meaning 'study'; however, it is also used as the right-hand bound constituent in complex nouns indicating a field of study, as e.g. 语言学 *yǔyán-xué* 'language-study, linguistics', 财政学 *cáizhèng-xué* 'finance-study, finance', 测地学 *cè-dì-xué* 'survey-hearth-study, geodesy'. This formative has two main characteristics: it can be used to build any word indicating a field of study, and it displays some semantic difference from the verb 学 *xué*. For these reasons, some authors have defined items like 学 *xué* as affixes (词缀 *cízhuì*) or affixoids (类词缀 *lèicízhuì* or 准词缀 *zhǔncízhuì*); however, in the literature on the topic, the criteria for the definition of affixes and affixoids, and thus what items should be included in these categories, vary greatly from author to author (see Pan, Ye, Han 2004). Ma (1995), for example, states that in Chinese it is possible to distinguish roots from affixes: affixes are never free, and they appear in a fixed position in complex words. Affixes may be

further divided in 'true affixes' (真词缀 *zhēn cízhuì*) and 'quasi-affixes' (准词缀 *zhǔn cízhuì*): the former are semantically empty, always bound, and are characterised by some sort of phonological reduction, typically loss of tone, as the above mentioned 子 *-zi*, 儿 *-r*, and 头 *-tou*, while the latter have some sort of categorial meaning, i.e. they assign the complex word to a lexical category and/or a semantic class (a taxonomical category), like in the case of 学 *-xué* mentioned above. In short, to be categorised as an affixoid, an item must be bound (independently from the fact that it has a corresponding free form) and must convey a meaning which is not its core meaning.

As highlighted by Arcodia (2011), if we were to regard as affixes only those items undergoing some sort of phonological reduction, we would ignore the features of grammaticalisation processes in the languages of East and South-East Asia, which, as we mentioned, generally do not display co-evolution of form and meaning. In addition, considering as affixes only items devoid of meaning would result in a definition of derivation which is cross-linguistically inconsistent, since typically derivational affixes carry some sort of meaning.

According to Sun (2000), the distinction between affixes and affixoids is not relevant in Chinese, since the system of derivational affixes in this language is still developing: those morphemes which behave as affixes but are phonologically (and orthographically) identical to their lexematic counterparts should be regarded as not fully grammaticalised. Thus, she holds a view according to which grammaticalisation necessarily involves some formal change. Arcodia (2011) too proposes to abandon the distinction between affixes and affixoids in Chinese but holds a different view: since in Mandarin we may have grammaticalisation of a sign without sound change, then the distinction between affixes and affixoids, which may be useful for European languages, is not relevant in Chinese. However, differently from Sun, Arcodia posits that the fundamental criterion to label a morpheme as a derivational affix in Mandarin Chinese is meaning differentiation. He claims that derivational affixes in Mandarin are the evolution of compound constituents, appearing in a fixed position, with a certain meaning, in a number of complex words. In order to become an affix, a lexeme must undergo a shift in meaning, which can either be more general than the meaning it has in other uses or be the extension of one of the possible non-core meanings of the lexeme. One example provided by Arcodia (2011) is 性 *-xìng*, whose development, as mentioned above, was favoured by the influence of European languages. This item was a word (a free form) in Classical Chinese, but it has turned into a bound root in the modern language, where it can be used to form complex words, as e.g. 性能 *xìng-néng* 'nature-capacity, natural capacity/function (of machine etc.)/property', 个性 *gè-xìng* 'personal-character, character/personality'. However, it also developed an affixal meaning, i.e. 'nature, -ity, -ness', as in 毒性 *dú-xìng*

'poison-nature, toxicity', 塑性 *sù-xìng* 'plastic-nature, plasticity' (see above). This meaning developed from the original meanings 'quality, intrinsic properties or characteristics of something' and 'inherent properties of the human being': through a process of generalising abstraction, 性 *-xìng* turned into a nominal suffix indicating just any property (not only intrinsic and everlasting properties), forming abstract nouns.[5]

In this paper, following Arcodia (2011), we dismiss the distinction between affixes and affixoids, since, as we have seen, in Mandarin grammaticalised signs often do not undergo any sound change. If a bound item used to build complex words appears in a fixed position with a fixed meaning, which (partially) departs from its original/core one and is more general or abstract than the meaning of the corresponding lexeme, then it can be regarded as a derivational affix, even if it is not formally different from the corresponding lexeme. Therefore, a formative like 学 *xué* 'field of study' above may be considered as a suffix; in other words, it is a grammaticalised item.

New affixes may emerge not only as a consequence of a grammaticalisation process inner to the language, but also due to the need of translating words containing affixes from foreign languages (see Shen 2015), as mentioned above. A possible example is 控 *kòng* 'buff, enthusiast, devotee', in words like 猫控 *māo-kòng* 'cat-enthusiast, cat lover', 长发控 *cháng-fà-kòng* 'long-hair-enthusiast, person extremely fond of long hair' (Ma 2016). This item is a phonetic adaptation of the Japanese suffix コン *-con*, which in turn originates from English *complex*, i.e. 'a group of attitudes and feelings that influence a person's behaviour, often in a negative way' (Cao, Mo 2012). In order to render affixes with no equivalent in Chinese, mostly lexical morphemes whose meaning is roughly similar or which are (quasi-)homophonous are chosen: when the number of words created by means of these morphemes increases, they gradually begin to assume a more general meaning.

In other cases, affixes may develop from a phonetic adaptation of a foreign word: (part of) a loanword may undergo a grammaticalisation process leading to an affix. One such example is 吧 *bā*, phonetic part of the hybrid 酒吧 *jiǔ-bā* 'alcohol-bar, bar', defined by *The Contemporary Chinese Dictionary* (2002) as "bar; counter at which alcoholic beverages are served in a Western-style restaurant or hotel". After the acceptance of this loanword, many complex words containing 吧 *bā* as the right-hand constituent have been created, as e.g 水

5 According to Arcodia (2011), the grammaticalisation process undergone by 性 *-xìng* is not fundamentally different from that leading to the English suffix *-hood* (< Old English *-hād*), as in e.g. *childhood*, *falsehood*. Originally a Germanic name meaning 'person, sex, condition, rank, quality', it has become a suffix forming nouns of condition or quality, or indicating a collection or group, from nouns and adjectives.

吧 *shuǐ-bā* 'water-bar' (a place where mostly soft drinks are served), 氧吧 *yǎng-bā* 'oxygen-bar' (a place where oxygen masks are available for customer usage), 网吧 *wǎng-bā* 'internet-bar, internet café' (see Arcodia 2011, 125-7). In the 新华新词语词典 *Xinhua xinciyu cidian* (Xinhua Dictionary of Neologisms, 2003), 吧 *bā* is listed with the following meaning: "broadly indicates an entertainment place with a particular function or supplied with some special equipment" (Arcodia 2011, 121). According to Arcodia (2011), 吧 *bā* underwent a further generalisation of meaning and does not indicate specifically an entertainment place, as e.g. 创意吧 *chuàngyì-bā* 'creativity-bar', a kind of enterprise in the field of business consulting, or 话吧 *huà-bā* 'talk-bar', basically a call shop. Thus, the starting point is a process of analogy, and then 吧 *bā* begins to be associated with more lexemes: drinks and food, other services (see e.g. 氧吧 *yǎng-bā* 'oxygen-bar' above), and then all sorts of meeting places (including virtual ones), where one can play games (e.g. 游戏吧 *yóuxì-bā* 'game-bar, amusement arcade'), exchange information on a topic (e.g. 贴吧 *tiē-bā* 'paste-bar, webpage where fans publish posts related to their idols', lit. 'post bar'), or even provide consulting or information for a charge (e.g. the above-mentioned 创意吧 *chuàngyì-bā* 'business consulting service'). According to Arcodia (2011, 126), metaphor is at work here: the meaning of 'bar' is extended to include any place which can be associated with the defining features of a bar. He also stresses the fact that this does not mean that 吧 *bā* has become a suffix with a pure locative meaning, since the connection with the original lexical meaning is always present somehow.[6]

In short, affixes in Chinese may develop through a grammaticalisation process inner to the language, due to the influence of foreign languages, or due to a combination of both: word-formation patterns already attested in the language may become productive due to the necessity to introduce foreign words; or, also, loanwords may develop an affixal use over time.

In what follows, after describing the three patterns, including their formal and semantic properties, we will focus on their development, and we will argue for their affixal status.

6 For further details on the development and meanings of 吧 *bā*, the reader is referred to Arcodia 2011.

3 Description of the Three Word-Formation Patterns

3.1 X-族 *zú* Words

As mentioned in the introduction, X-族 *zú* is a quite established and widely studied word-formation pattern in Chinese. The original meaning of 族 *zú* is 'clan, tribe, ethnic group', and it is still used with this meaning in compound words, as e.g. 大族 *dà-zú* 'big-clan, famous and influential clan', 白族 *bái-zú* 'Bai-group, Bai minority'. In the last decades, this root has also developed a more generic meaning, i.e. 'a category/group of people with common characteristics or behaviour' (see Zhao 2009), appearing in a fixed position (right-hand constituent) in complex words, as e.g. 星空族 *xīng-kōng-zú* 'star-sky-ZU, night workers' (XCY, LWC), 网购族 *wǎng-gòu-zú* 'net-purchase-ZU, those who love purchasing goods online' (XCY, LWC), 候鸟族 *hòuniǎo-zú* 'migratory.bird-ZU, the commuters' (LWC).

This use of 族 *zú* originates as a loan from Japanese 族 *zoku* 'a group of people with similar feelings or passions'; it was introduced to mainland China through Taiwan and Hong Kong (Cao 2007; Xiao 2009; Zhao 2009; Chen, Zhu 2010; Li 2013). According to Cao (2007), in Chinese it was originally used to indicate 'a category of things with shared characteristics or properties', as in 水族 *shuǐ-zú* 'water-ZU, aquatic animals', and later developed the above-mentioned meaning 'a category/group of people with common characteristics or behavior' (Cao 2007; Lu 2010). The first words containing 族 *zú* with this broader meaning, coined in the early Nineties, are 上班族 *shàng-bān-zú* 'go-work-ZU, office workers', 追星族 *zhuī-xīng-zú* 'follow-star-ZU, groupies', and 打工族 *dǎgōng-zú* 'have.a.temporary.job-ZU, those having a temporary or casual job' (Cao 2007; Yang, Chen 2012). Due to their use on the web and in the media, these words became widespread and increased in number over the years, as can be seen from the number of distinct X-族 *zú* words found in the newspaper 人民日报 *Renmin ribao* (*People's Daily*) between 1995 and 2005, shown in table 1 (Cao 2007, 151).

Table 1 X-族 *zú* words in the 人民日报 *Renmin ribao* between 1995 and 2005

1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
19	15	23	33	15	33	23	27	41	36	44

Thus, in short, X-族 *zú* developed into a word-formation pattern indicating different groups of people who have something in common: fans/people who love something (much like 控 *kòng* seen in § 2 above), as 朋克族 *péngkè-zú* 'punk-ZU, punk lovers' (LWC), 哈韩族 *hā-Hán-*

zú 'adore-[7]Korea-ZU, those who love Korean music, TV, clothes etc.' (LWC), or 爱车族 *ài-chē-zú* 'love-car-ZU, car lovers' (XCY, LWC); those who are addicted to something, as e.g. 爱邦族 *ài-bāng-zú* 'love-Lianbang.syrup-ZU, those who are addicted to the cough syrup Lianbang (联邦止咳露 *Liánbāng zhǐké lù*)' (XCY), 偷菜族 *tōu-cài-zú* 'steal-vegetables-ZU, those addicted to online games like Happy farm (开心农场 *Kāixīn nóngchǎng*) or Farmville',[8] 点赞族 *diǎn-zàn-zú* 'like-[9]ZU, like-clicking addicted', i.e. those who always click the 'like' button, e.g. on Facebook (XCY); users of various means of transport, such as 单车族 *dānchē-zú* 'bicycle-ZU, cyclists' (LWC), 地铁族 *dìtiě-zú* 'subway-ZU, those who use the subway' (LWC); workers, such as 办公族 *bàngōng-zú* 'work (in an office)-ZU, people who work in an office (LWC), 星空族 *xīng-kōng-zú* 'star-sky-ZU, night workers' (LWC), 陪逛族 *péi-guàng-zú* 'accompany-stroll-ZU, personal shoppers' (SD); those who share ideals/views/lifestyles etc., as e.g. 养生族 *yǎngshēng-zú* 'keep.in.good.health-ZU, health-conscious people' (LWC), 慢活族 *màn-huó-zú* 'slow-live-ZU, those who follow a slow living lifestyle' (XCY, SD), 素食族 *sù-shí-zú* 'vegetarian-food-ZU, the vegetarians' (XCY); people with a particular behaviour in common or engaged in certain activities (people who often do something or who like to do something), as 手机夜游族 *shǒujī-yè-yóu-zú* 'mobile.phone-night-travel-ZU, those who use mobile phones in bed before sleeping' (SD), 蹭网族 *cèng-wǎng-zú* 'freeload-net-ZU, Wi-Fi squatters' (SD),[10] 晒密族 *shài-mì-zú* 'show-secret-ZU, those who reveal their secrets on the web' (SD); people with some characteristics in common, as e.g. 肥腿族 *féi-tuǐ-zú* 'fat-leg-ZU, girls with fat legs' (LWC), 榴莲族 *liúlián-zú* 'durian-ZU, ill-tempered co-workers who have been working for many years and are hard to get along with, just like the smelly fruit with thick thorny skin' (SD),[11] 向日葵族 *xiàngrìkuí-zú* 'sunflower-ZU, people who, just like a sunflower, always look on the bright side of life and are resilient to pressure as they easily forget about unhappiness' (XCY, SD, LWC).[12]

7 The character 哈 *hā* is often used in Taiwan with the meaning of 'worship, adore': https://bit.ly/39kDjSa.

8 They are virtual farms, where you play the role of a farmer who plants and harvests crops. Players can sneak into their friends' farms and steal vegetables.

9 点赞 *diǎn-zàn*, lit. 'click-praise', in the Internet slang indicates the 'like' button, used by the users to express that they like, enjoy, or support something.

10 It refers to those who linger in a public location to use its Wi-Fi internet connection, or who use such a connection without authorisation. Definition from *China Daily*: https://language.chinadaily.com.cn/trans/2012-11/22/content_15951634.htm.

11 Definition from *China Daily*: http://www.chinadaily.com.cn/dfpd/2011-08/22/content_13162619.htm.

12 Definition from *China Daily*: https://language.chinadaily.com.cn/trans/2011-06/28/content_12793756.htm.

A few base words for X-族 *zú* neologisms are phonetic adaptations, like 辣奢族 *làshē-zú* 'luxury-ZU, fans of luxury goods' (XCY, SD), 飞特族 *fēitè-zú* 'freeter-[13]ZU, those who work only when they feel they need some money (having a work schedule more flexible than freelancers)' (SD). The base word may be also a phonetic adaptation of an acronym, as e.g. 丁克族 *dīngkè-zú* 'DINK-[14]ZU, young couples in big cities without children' (LWC). Sometimes the written form of the phonetic adaptation contains meaningful rather than neutral characters, as e.g. in 乐活族 *lè-huó-zú* 'happy-live-ZU, those following LOHAS',[15] where the characters chosen somehow convey the meaning of the acronym the phonetic adaptation refers to. In some cases, the base of such neologisms are direct loans, as e.g. *Emo*族 *zú* 'Emo people', including acronyms and initialisms, such as DIY族 *zú* 'DIY (do it yourself)-ZU, DIY lovers'. There are also a couple of words whose base is a single Latin letter, which stands for an acronym/initialism, as e.g. T族 *zú*, which refers to Chinese students who want to study abroad and must pass the TOEFL (Test of English as a Foreign Language). Sometimes X-族 *zú* words are calques from English, as e.g. 食男族 *shí-nán-zú* 'eat-male-ZU, maneaters' and 游族 *yóu-zú* 'game-[16]ZU, gamers' (LWC). There are also graphic loans from Japanese, as e.g. (御)宅族 (*yù)zhái-zú* 'nerd (< Jap. *otaku*)-ZU, nerds, geeks' (XCY, SD, LWC).

In addition, the variant 一族 *yīzú* (lit. 'one group') is attested as well, as e.g. 微博一族 *Wēibó-yīzú* 'Weibo users': we found 31 types in our corpus. Out of 31 neologisms ending in 一族 *yīzú*, 20 do not display the corresponding X-族 *zú* form, as e.g. 哈哈一族 *hāhā-yīzú* 'Harry Potter lovers' (XCY), while 11 appear both in the 一族 *yīzú* and in the 族 *zú* form, as e.g. 拇指族 *mǔzhǐ-zú* and 拇指一族 *mǔzhǐ-yīzú* 'thumb-(YI)ZU, young people who use text messages as main means of communication' (XCY, SD, LWC), or 上网族 / 上网一族 *shàng-wǎng-zú* / *shàng-wǎng-yīzú* 'go-web-ZU, web users', apparently with the same meaning (but see fn. 38).

3.2 X-党 *dǎng* Words

Complex words containing 党 *dǎng* as the right-hand constituent, indicating groups of people with common characteristics or behaviour, started to appear in the late 2000s, thus X-党 *dǎng* is a quite novel pattern of word formation. This pattern is typical of the web but is occasionally attested in other media as well (Chen, Zhu 2010).

13 From English *free* and German *Arbeiter* 'worker'.

14 DINK is the acronym of *dual income, no kids*.

15 LOHAS is the acronym of *Lifestyles of Health and Sustainability*.

16 游 *yóu* stands for 游戏 *yóuxì* 'game'.

The original meaning of 党 *dǎng* is 'political party, clique', and with this meaning it is found in complex words such as 党员 *dǎng-yuán* 'party-member, party member', 黑手党 *hēi-shǒu-dǎng* 'black-hand-clique, mafia, gang'. As the right-hand constituent in complex words, it also developed the meaning 'a group/category of people with common interests and characteristics or behaviour', much like the morpheme 族 *zú*: e.g. 剧透党 *jùtòu-dǎng* 'spoiler-DANG, people who like to spoil (films etc.)' (LWC). Chen and Zhu (2010) argue that this use of 党 *dǎng* derives from the meaning 'clique', which has a strong derogatory sense. However, they point out that after the morpheme acquired the meaning of 'political party' (from Japanese 党 *tō*, e.g. in 国民党 *kokumintō* 'Chinese Nationalist Party'), especially after the foundation of the Chinese Communist Party, it started to have a positive connotation: this new meaning contributed to 'lighten' the derogatory sense connected to 'clique'.[17]

Among neologisms containing 党 *dǎng*, we find words indicating different groups of people with something in common, as e.g.: people with a particular behaviour or habit in common, such as 自拍党 *zìpāi-dǎng* 'selfie-DANG, people who take a lot of selfies' (LWC), 游戏党 *yóuxì-dǎng* 'game-DANG, those who play online videogames' (LWC), 睡衣党 *shuìyī-dǎng* 'pyjamas-DANG, those who go out in pyjamas' (LWC), 早起党 *zǎo-qǐ-dǎng* 'early-wake.up-DANG, the early risers' (LWC), 格格党 *gégé-dǎng* 'princess(a loan from Manchu)-DANG, Chinese girls born after 1985 who do not take their work seriously, do not obey their superiors, are arrogant, pay too much attention to their own needs without understanding those of other people, thus being incompatible with traditional jobs' (XCY, SD, LWC); people addicted to something or who like something very much, be it a videogame, a sport, a musical genre, a dressing style, an instrument, or a brand, as e.g. 手机党 *shǒujī-dǎng* 'mobile.phone-DANG, mobile phone addicted' (LWC), 剁手党 *duò-shǒu-dǎng* 'chop-hand-DANG, online shopaholics (example (1b)), 甘党 *gān-dǎng* 'sweet-DANG, sweet lovers', 爱凤党 *àifèng-dǎng* 'Iphone-DANG, Iphone lovers' (LWC), where 爱凤 *àifèng* is a phonetic adaptation; people sharing some particular characteristics, such as 白意党 *bái-yì-dǎng* 'pure-intention-DANG, the sentimental' (LWC), 无聊党 *wúliáo-dǎng* 'bored-DANG, the bored' (LWC), 一见钟情党 *yī-jiàn-zhōngqíng-dǎng*, one-see-fall.in.love-DANG, those who fall

17 Chen and Zhu (2010) highlight that in Japanese 党 *tō* has also the meaning 'clique', just like in Chinese, as e.g. in 凶党 *kyō-tō* 'gang of partners in crime' (lit. 'evil/villain-clique'). Furthermore, it has also the meaning of 'a group/ category of people with common interests and characteristics', much like in Chinese, but it has very low productivity. Some of the few examples that can be found are 烟党 *kemuri-tō* 'smoke-TO, smokers' (compare Chinese 抽烟党 *chōuyān-dǎng* 'smoke-DANG'), and 甘党 *ama-tō* 'sweet-TO, sweet lovers' (compare Chinese 甘党 *gān-dǎng* 'sweet-DANG'; 甜食党 *tián-shí-dǎng* 'sweet-eat-DANG').

in love at first sight' (LWC), 美丽党 *měilì-dǎng* 'beautiful-DANG, beautiful people'(LWC), or 苍白党 *cāngbái-dǎng* 'pale-DANG, people with little vitality and energy' (LWC).

Among these words indicating different types of people, there are also words originating from online buzzwords, such as 寂寞党 *jìmò-dǎng* 'lonely-DANG', i.e. web users who often use the buzzword (哥)…的不是…, 是寂寞 *(gē)…de bù shì…, shì jìmò* 'what X is Y-ing is not Z, it is loneliness' (XCY, SD, LWC).[18]

In addition, just like 族 *zú*, 党 *dǎng* too can form neologisms which indicate certain types of workers, such as 上班党 *shàng-bān-dǎng* 'go-work-DANG, office workers' (LWC; compare the above-mentioned 上班族 *shàng-bān-zú* 'go-work-ZU, office workers'), and 配音党 *pèiyīn-dǎng* 'dub-DANG, dubbers' (LWC).

Chen and Zhu point out that 族 *zú* and 党 *dǎng* as the right-hand constituents of complex words indicating people with common characteristics or behaviour are actually interchangeable, i.e. they can attach to the same base without any apparent change in meaning: see e.g. 熬夜族 *áoyè-zú* 'stay.up.late-ZU' / 熬夜党 *áoyè-dǎng* 'stay.up.late-DANG', both indicating 'those who stay up late or all night'. However, Chen and Zhu (2010) observe that the oldest words containing 族 *zú* are generally not found in the corresponding X-党 *dǎng* form. In addition, after becoming an established pattern, X-族 *zú* words lost their novelty; at the same time, X-党 *dǎng* words started to appear on forums, becoming more and more widespread and replacing X-族 *zú* words as the most popular way to indicate groups of people with common interests, characteristics, or behaviour. Through a Baidu search, Chen and Zhu show that between 2008 and 2009 党 *dǎng* was the most used formative for words referring to groups of people: they considered the frequency of X-党 *dǎng* and X-族 *zú* words formed with the same base, showing that the X-党 *dǎng* pattern is the most frequently used for recent neologisms, while for older ('typical') words it is rarely used (Chen, Zhu 2010, 67). Therefore, apparently the difference between the two items is that 族 *zú* is more established, while 党 *dǎng* is more recent, popular and fashionable, and

18 This buzzword emerged in 2009 in the Chinese BBS community Baidu World of Warcraft forum: an user posted a low-resolution webcam image of a man eating noodles accompanied by the sentence 哥吃的不是面, 是寂寞 *gē chī de bù shì miàn, shì jìmò* 'what this brother is eating aren't noodles, but loneliness!'. Shortly after, other users on the forum began repeating this sentence with slight variations, giving rise to the template illustrated above, creating a series of parody images centred around the theme of loneliness, as e.g. 我呼吸的不是空气, 是寂寞 *wǒ hūxī de bù shì kōngqì, shì jìmò* 'what I am breathing is not air, is loneliness', 哥灌的不是水, 是寂寞 *gē guàn de bù shì shuǐ, shì jìmò* 'what (this brother) is pouring is not water, is loneliness', 我用的不是手机, 是寂寞 *wǒ yòng de bù shì shǒujī, shì jìmò* 'what I am using is not a mobile phone, is loneliness'. https://baike.baidu.com/item/%E5%AF%82%E5%AF%9E%E5%85%9A; https://knowyourmeme.com/memes/loneliness-party-%E5%AF%82%E5%AF%9E%E5%85%9A#fnr1.

it is mainly used on the web (we will return on this issue in §§ 4.4 and 5). A hint of the fact that 党 *dǎng* is perceived as more popular and fashionable is the significant presence in our corpus (53 out of 189 words, 28%) of X-党 *dǎng* words indicating fans of actors, singers, characters, books, TV series, comics etc., as e.g. 天使党 *tiānshǐ-dǎng* 'angel-DANG, fans of the Japanese anime television series *Angel beats!*' (LWC), 松井党 *Sōngjǐng-dǎng* 'Rena Matsui-DANG, fans of Rena Matsui (松井玲奈, Japanese actress and singer)'. In our corpus we did not find any X-族 *zú* words of this kind, with the exception of 哈哈一族 *hāhā-yīzú* 'Harry Potter lovers' (XCY), containing the variant 一族 *yīzú* (see § 3.1). Rather, among X-族 *zú* words we found some examples of fans/enthusiasts of a particular genre or category, like 朋克族 *péngkè-zú* 'punk-zu, punk lovers', 哈韩族 *hā-Hán-zú* 'adore-Korea-zu, those who love Korean music, TV, clothes etc.' (§ 3.1).

Furthermore, it must be noted that, among X-党 *dǎng* words, we find words referring to a series of illegal activities, which cannot be found among X-族 *zú* words, as e.g. 拎包党 *līnbāo-dǎng* 'bag-DANG, pickpockets' (SD), 撞车党 *zhuàng-chē-dǎng* 'collide-car-DANG, people who wilfully get hit by other cars to extort money from drivers' (LWC), 敲墙党 *qiāo-qiáng-dǎng* 'knock-wall-DANG, a mafia-style group that forces people to rely on their companies when they need to renovate their properties' (LWC), and 黄牛党 *huángniú-dǎng* 'scalper-DANG, scalpers'. This negative nuance is apparent in the word 摩托党 *mótuō-dǎng* 'motorcycle-DANG, the motorcyclists' as well, which usually refers to gangs of motorcyclist disturbing public security etc.,[19] and not simply to people who ride a motorcycle. Therefore, in some cases, 党 *dǎng* retains to an extent the negative nuance of its original meaning 'clique' (see Chen, Zhu 2010; we will return to this issue in § 4.2). We believe that this is the source of the ambiguity displayed by some neologisms, which can have two different meanings: e.g. 狗党 *gǒu-dǎng* 'dog-DANG' can refer either to 'close friends' or to 'spies' (LWC). The latter meaning retains the negative nuance of the term 'clique'.

3.3 X-客 *kè* Words

The original meaning of the morpheme 客 *kè* is 'guest, traveller', and with this meaning it is found in compound words as e.g. 旅客 *lǚkè* 'travel-guest, hotel guest/traveller', 请客 *qǐng-kè* 'invite-guest, invite/entertain guests', 客车 *kè-chē* 'guest-vehicle, passenger train'.

However, in recent years it started to appear as the right-hand constituent of complex words indicating 'a person doing a certain activity' or 'a person with certain characteristics'. Arguably, the most pop-

19 https://baike.baidu.com/item/%E6%91%A9%E6%89%98%E5%85%9A/18818561.

ular of these complex words is 黑客 *hēi-kè* 'black-KE, hacker', which entered the Chinese lexicon in the late Nineties, as a phonetic-semantic adaptation of the English word *hacker*: the Chinese word approximately recalls the pronunciation of the source word; in addition, the left-hand constituent, 黑 *hēi* 'black, shady, illegal', conveys the negative meaning of the word (compare 黑车 *hēi-chē* 'black-vehicle, illegal taxi, unlicensed motor vehicle'). This word-formation pattern has become popular starting from the beginning of the twenty-first century: according to Zhang and Xu (2008), with the spread and popularity of blogs (in Chinese 博客 *bókè* 'blog', also 'blogger') at the beginning of the 2000s, more and more X-客 *kè* words appeared, which, together with words already coined, like 黑客 *hēikè* 'hacker', contributed to form a word-formation pattern typical of the web.

Along with words indicating different kinds of 'hackers', such as 白客 *bái-kè* 'white-KE, online security guard; hacker-fighter', 红客 *hóng-kè* 'red-KE, patriotic hacker, defending the security of domestic networks and fending off attacks', 灰客 *huī-kè* 'grey-KE, unskilled hacker',[20] we find neologisms indicating persons engaged in different kinds of activities, such as 刷书客 *shuā-shū-kè* 'scan-book-KE, a person who record extracts from a book, either in a bookstore or in a library, with an electronic mini scanner, without any intention to buy it' (XCY), 换客 *huàn-kè* 'exchange-KE, one who sells/exchanges goods online'.

As Zhang and Xu (2008) point out, this word-formation pattern is typical of the web and was then extended to the media in general and to everyday language too, even though it is still mainly used by young people. Actually, many X-客 *kè* words belong to the domains of technology and the web, often indicating people doing some kind of activity online (38 out of 84 words in our corpus, almost half of the total); we will go back to this issue in § 4.3.

Among X-客 *kè* words, we find many neologisms which are phonetic adaptations: however, differently from what happens with 族 *zú* and 党 *dǎng*, generally speaking it is the whole complex word ending in 客 *kè* that is a phonetic adaptation (not just the base), as e.g. 极客 *jí-*

20 Following Arcodia and Basciano (2018), we excluded 'hacker' words from our analysis, since they do not indicate 'a person doing a certain activity' or 'a person with certain characteristics' related to the base. Rather, they are best analysed as analogical formations (see Booij 2010) from 黑客 *hēi-kè* 'black-KE, hacker', where the modifier is invariably a colour term, which is always understood in a metaphorical rather than in a literal sense. An anonymous reviewer pointed out that the semantic mechanism at work could be similar to reductions observed in English words such as *cheeseburger* or *fishburger*, where *burger* is the truncated form of *hamburger*, or also in Italian words like *auto-strada* 'car-road, motorway' or *auto-lavaggio* 'car-washing, car washing', where *auto* stands for *automobile* 'car'. Thus, in a word as 红客 *hóng-kè* 'red-KE, patriotic hacker (lit. red hacker)', 客 *kè* would be the truncated form of 黑客 *hēi-kè* 'hacker' (红(黑)客 *hóng-(hēi)-kè*). However, in our opinion analogy best explains these cases, since the modifier is always a colour term, which replaces 黑 *hēi* 'black' in 黑客 *hēi-kè* 'hacker', and is interpreted in a metaphorical sense, just like in 黑客 *hēi-kè*.

kè 'extremely-KE, geek', much like in the case of 黑客 *hēi-kè* 'hacker' (we will go back to this issue in § 4.4). Out of the 84 X-客 *kè* words collected from our sources, 15 (17.86%) are phonetic adaptations of this kind, i.e. the whole complex word is a phonetic adaptation. It must be noted, though, that 客 *kè* is not just a component of the phonetic adaptation, but is also the element which conveys the agentive meaning to the complex word. For example, 切客 *qiē-kè* 'cut-KE, fan of location-based services who regularly checks in to keep friends and relatives posted on her/his whereabouts' is a phonetic adaptation of English *check-in*; however, the word indicates a person, and this meaning is conveyed by the morpheme 客 *kè*. The same goes for the word 粉飞客 *fěn-fēi-kè* 'fan-[21]fly-KE, fanfictioner (fan who likes to write sequels or change plots of TV series to express her/his ideas, passions etc.)', which is a phonetic adaptation of English *fanfic*: besides recalling the pronunciation of the last part of the word, 客 *kè* also conveys the meaning of 'person'; as a matter of fact, the whole word means *fanfictioner*, not *fanfic*. Therefore, in these cases the X-客 *kè* word indicates a person involved in an activity connected to the meaning of the phonetic adaptation as a whole ('a person doing an activity connected to X-客 *kè*', not 'a person doing an activity connected to X'), where 客 *kè* is part of the phonetic adaptation but, at the same time, contributes the meaning of 'person'.

In addition to these cases, we also found 4 complex words (4.76%) the base of which is a phonetic adaptation, as e.g. 秀客 *xiù-kè* 'show-KE' (秀 *xiù* is the phonetic adaptation of *show*), which refers to those who share videos from the e-commerce platform 秀兜 *Xiùdōu* on their Weibo, among their friends (they receive a fee from the platform every time someone clicks on their sponsored links and then completes the purchase). All in all, we can observe that the proportion of phonetic adaptations among X-客 *kè* words is much higher than among X-族 *zú* (13 out of 434, about 3%) and X-党 *dǎng* words (10 out of 189, 5.29%). We will return to the possible motivations for this in § 4.3.

Besides phonetic adaptations, we also find calques and hybrid forms, as e.g.: 追客 *zhuī-kè* 'follow-KE, someone who regularly refreshes web pages to follow the latest updates of online series, TV series, bloggers, or podcasts', which looks like a calque of English *follower* (追 *zhuī* translates *follow*, and 客 *kè* is roughly equivalent to *-er*); 创客 *chuàng-kè* 'create-KE, maker', which can be regarded as a hybrid, where 创 *chuàng* translates *make*, while 客 *kè* acts as the equivalent of the suffix *-er* and, at the same time, recalls the pronunciation of the last part of the word *maker*.

However, X-客 *kè* words are not limited to loans and words connected to the Internet and new technologies; the X-客 *kè* pattern is

21 粉 *fěn* 'powder' stands for 粉丝 *fěnsī*, phonetic adaptation of the English word *fans*.

also used to coin words indicating persons involved in all sorts of different activities or having certain characteristics, as e.g. 必剩客 *bì-shèng-kè* 'certainly-remain-KE, a person above the typical marriage age but still single, considered to be doomed to remain unmarried', 代扫客 *dài-sǎo-kè* 'take.the.place.of-sweep-KE, a person who offers a service consisting in visiting tombs (sweeping and offering sacrifices) during the Qingming festival' (XCY), 排客 *pái-kè* 'line.up-KE, a person paid to stand in a queue for others', 帕客 *pà-kè* 'handkerchief-KE, a green consumer who prefers to use handkerchiefs instead of throw-away paper tissues in support of low-carbon life'[22] (LWC). However, even when X-客 *kè* words are not nouns connected to the Internet and new technologies, the role of the web in their creation and diffusion is apparent, at least for part of them. Take for example the word 帕客 *pà-kè* just mentioned above: it became popular after one of China's online messaging service providers launched a handkerchief design campaign in 2009 to encourage the use of handkerchiefs to protect the environment; the winner was called 帕客 *pà-kè* 'handkerchief-KE'.[23]

All in all, it can be stated that the morpheme 客 *kè* as the right-hand constituent of complex words has acquired a more general meaning, appearing in a fixed position, indicating various kinds of persons, with a function roughly comparable to that of English *-er* (Arcodia, Basciano 2018).

3.4 Are X-族 *zú* and X-党 *dǎng* Words Collective Nouns?

The three morphemes at issue, as we have seen, have apparently acquired a more general meaning, appearing in a fixed position (to the right of complex words), indicating various kinds of persons. At a first look, it would seem that 客 *kè* forms individual nouns, while 族 *zú* and 党 *dǎng* form collective nouns, thus preserving part of their original meaning, as suggested by the following examples:

2. a. 刷书客
shuā-shū-kè
scan-book-KE
'a person who scans with a mini-scanner the content from the books in a bookstore or a library'

22 http://language.chinadaily.com.cn/trans/2010-02/21/content_9480739.htm.
23 http://language.chinadaily.com.cn/trans/2010-02/21/content_9480739.htm.

b. 刷书族
shuā-shū-zú
scan-book-ZU
'people who scan with a mini-scanner the content from the books in a bookstore or a library'

In (2), we have two words differing only for the right-hand constituent used, i.e. 客 *kè* or 族 *zú*. The only difference in meaning between the two words seems to be individual *vs* collective. The X-族 *zú* term, thus, apparently denotes a collective whole, a (semantic) plurality ('more than one') obtained by grouping together a number of entities, which share a part-whole relation (see Gardelle 2019). This is further suggested by the fact that 客 *kè* and 族 *zú* may combine in the same word. See the following examples:

3. a. 换客
huàn-kè
exchange-KE
'one who sells/exchanges goods online'

b. 换客族
huàn-kè-zú
exchange-KE-ZU
'those who sell/exchange goods online'

4. a. 晒客
shài-kè
expose-KE
'a person who shares his experiences and thoughts with others on the Internet'

b. 晒客族
shài-kè-zú
expose-KE-ZU
'those who share their experiences and thoughts with others on the Internet'

However, a closer look at the data reveals a different picture: X-族 *zú* nouns can apparently refer to members of the group rather than the group as a whole, as in the following examples.[24]

24 In these examples, the plural classifier 些 *xiē* (i.e. the only plural classifier available in Chinese) is used. 些 *xiē* is never used in counting; it combines with the demonstratives 这 *zhè* 'this' or 那 *nà* 'that', resulting in 'these' and 'those', or with the numeral 一 *yī* 'one', leading to the indefinite meaning 'some' (cf. Eng. *a few*, *a couple of*, *a num-*

5. a. 现在, 人们的工作节奏较快, 对一些上班族来说, 下班之后想自己做顿像样的晚餐成了一种奢望 [...]
xiànzài rén-men de gōngzuò jiézuò jiào kuài duì ***yī***
now person-PL SP work rhythm quite fast for one
xiē shàng-bān-zú *láishuō xiàbān zhīhòu xiǎng*
CFL$_{PL}$ go-work-ZU concerning finish.work after want
zìjǐ zuò dùn xiàngyàng de wǎncān chéng le
oneself make CFL decent SP dinner become PFV
yī zhǒng shē-wàng
one CFL extravagant-hope
'Nowadays, the working rhythm of people is quite fast. For **some office workers**, preparing a decent meal for themselves after work has become an extravagant hope [...]'[25]

b. [...] 到了双休日那些爱运动的上班族都来了 [...]
dào le shuāng-xiū-rì ***nà xiē ài yùndòng de***
arrive PFV double-rest-day that CFL$_{PL}$ love sports SP
shàng-bān-zú *dōu lái le*
go-work-ZU all come PFV
'[...] In the weeks with two rest days, all **the/those office workers who love sports** came [...]'[26]

This is observed with X-党 *dǎng* nouns as well:

6. a. [...] 全国各地都有一些睡衣党出没。
quán-guó-gè-dì dōu yǒu ***yī xiē shuìyī-dǎng***
whole-country-each-place all have one CFL$_{PL}$ pyjamas-DANG
chūmò
come.and.go
'[...] everywhere in the country there are **some people who go out in pyjamas** [...]'.[27]

b. 对于这些熬夜党, 尤其是女性熬夜党来说, 护肤尤为重要。[28]
duìyú ***zhè xiē áoyè-dǎng*** *yóuqí shì nǚxìng*
for this CFL$_{PL}$ stay.up.late-DANG especially be woman

ber of; Sybesma 2017). According to Ilijc (1994), 些 *xiē* is a collective marker, referring to wholes, rather than a plural marker.

25 http://www.peopledailynews.eu/sp/20190417_57656.html.

26 https://hznews.hangzhou.com.cn/xinzheng/quxian/content/2010-06/23/content_3327660.htm.

27 https://kknews.cc/zh-my/news/ebbe42z.html.

28 https://k.sina.com.cn/article_7026285403_1a2cc9b5b00100saxq.html?from=fashion.

áoyè-dǎng *láishuō* *hùfū* *yóuwéi* *zhòngyào*
stay.up.late-DANG concerning skincare particularly important
'For **these people who stay up late**, especially for women, skincare is particularly important'.

Besides, it must be noted that both X-族 *zú* words and X-党 *dǎng* words may be followed by the plural / collective suffix 们 *-men*:

7. a. 眼看着本月底地铁4号线就将推行"禁食令", 本市不少"地铁快餐族"们同样提出了自己的质疑
yǎnkànzhe *běn* *yuè-dǐ* *dìtiě* *sì* *hào* *xiàn*
watch.helplessly this month-end subway 4 number line
jiù *jiāng* *tuīxíng* *jìn* *shí* *lìng* *běn* *shì*
then will carry.out forbid eat decree this city
bùshǎo ***dìtiě-kuài-cān-zú-men*** *tóngyàng* *tíchū* *le* *zìjǐ*
many subway-fast-food-ZU-PL same pose PFV oneself
de *zhíyí*
SP call.into.question
'While watching helplessly that by the end of this month Line 4 of the subway will implement a 'no eating decree', many **"subway fast-food eaters"** in town called it into question'.

b. […] 酱油党们也因为在片中露脸而找到了狂欢的理由 (LWC)
jiàngyóu-dǎng-men *yě* *yīnwèi* *zài* *piàn* *zhōng* *lùliǎn*
soy.sauce-DANG-PL also because at film in appear
ér *zhǎodào* *le* *kuánghuān* *de* *lǐyóu*
and find PFV revel SP reason
'**Those who feign ignorance**[29] too found a reason to revel because they appeared in the film'.

According to Li and Thompson (1981, 40), the suffix 们 *-men* is generally used only when there is some reason to emphasise the plurality of the noun. According to others (e.g. Iljic 1994; Cheng, Sybesma 1999), it is a collective rather than a plural marker. Iljic (1994, 96), for example, points out that "[t]he speaker resorts to *men* whenever he has grounds to view several persons as a group, either relative to himself or relative to a third party". The function of this suffix, then, would be to group different units, to construct a group from several elements. According to

29 From 打酱油 *dǎ jiàngyóu* 'it's none of my business; it has nothing to do with me' (orig. 'buy soy sauce'). This meaning developed from a buzzword: in 2008 the Guangzhou Broadcasting Network interviewed a local man about the Edison Chen (a celebrity from Hong Kong) photo scandal, who answered: "关我鸟事，我出来打酱油的 *guān wǒ niǎo shì, wǒ chūlái dǎ jiàngyóu de*" (it's none of my business / what the f**k does it have to do with me? I was just out buying soy sauce). This answer then became a meme, applicable to any context: https://chinadigitaltimes.net/space/Get_soy_sauce.

Cheung (2016), count nouns suffixed with 们 *-men* can be used to refer to a group of people that are known to both speakers and hearers. As a matter of fact, they are regularly used as a term of address in gatherings, as e.g. 女士们、先生们 *nǚshìmen, xiānshēngmen* 'ladies and gentlemen'.

Therefore, the co-occurrence of 族 *zú* and 党 *dǎng* with the suffix 们 *-men* would be unexpected if they were simply used to form collective nouns (which involve the gathering of a plurality of entities, specifically a group), unless 们 *-men* is seen just an emphatic marker (i.e. if it is used to emphasise collectivity). If the function of 们 *-men* is to group several entities, we should then conclude that X-族 *zú* and X-党 *dǎng* nouns in these contexts refer to members of the group, rather than to the collective whole.

In addition, individuation may be observed in yet other contexts: X-族 *zú* and X-党 *dǎng* nouns can combine with sortal classifiers[30] (or individual classifiers, Peyraube 1998) used for humans, and individual members can be counted. See the following examples, where X-族 *zú* nouns clearly indicate single entities, and not the collective whole:[31]

8. a. 背上旅行包, 带上相机, 做个背包族, 继续我的浪子情怀。 (LWC)
bèi-shàng lǚxíng-bāo dài-shàng xiàngjī ***zuò ge bèi-bāo-zú***
back-on travel-bag bring camera be CLF back-pack-ZU
jìxù wǒ de làngzǐ qínghuái
continue 1SG SP wastrel mood
'Carrying my luggage on the shoulder, taking the camera with me, **being a backpacker**, carrying on my nomad spirit'.

b. 粗略统计, 3分钟内竟出现40个"车缝族"。 (XCY)
cūlüè tǒngjì sān fēnzhōng nèi jìng chūxiàn
rough statistics 3 minute inside actually appear
sìshí ge chē-fèng-zú
40 CLF vehicle-crack-ZU
'With a rough estimate, in 3 minutes **40 jaywalkers** appeared'

30 As pointed out by Croft (1994), sortals simply name the unit that is already present in the semantic denotation of the noun, while measures create a unit by which we can count or measure; they include real measures (kilo, mile), containers (cup, spoon), and collectors (group, mass). Measures carry their own, noun-independent semantics, as confirmed by the fact that they can be used with count nouns and mass nouns alike (Sybesma 2017). Chinese sortal classifiers represent a closed class, and each classifier combines with a set of nouns that can be seen to belong to one and the same class. Classifiers are compulsory with numerals, i.e. there is no counting without a classifier, so that they are often referred to as numeral classifiers (Sybesma 2017).

31 In the examples, we observe the use of the classifiers 个 *ge*, used for all humans (regardless of sex, age, social status, occupation etc.), and the honorific classifier for people 位 *wèi*. Actually, 个 *ge* is also used as a generic classifier for nouns lacking more specific sortals, or even as a 'default' – speakers often use it with nouns that combine with another sortal according to prescriptive grammar (see Sybesma 2017).

c. 我只是一个朝九晚五的上班族 (Zhao 2009, 36)
wǒ zhǐ shì ***yī ge zhāo jiǔ wǎn wǔ de shàng-bān-zú***
1SG only be one CLF morning nine evening five SP go-work-ZU
'I am only **a 9- to 5-er**' (translation provided by the source)

d. 一位27岁的上班族写完一首诗后跳楼自杀了[...][32]
yī wèi èrshíqī suì de shàng-bān-zú *xiě-wán yī shǒu shī hòu tiào-lóu zì-shā le*
one CLF 27 year SP go-work-ZU write-finish one CLF poem after jump-building self-kill PFV
'**A 27-year-old office worker** jumped to his death from a building after finish writing a poem [...]'

X-党 *dǎng* words too are attested in numeral-classifier constructions like the ones above, as in the following examples (see also Chen, Zhu 2010):

9. a. 群里面就我一个电脑党 (LWC)
qún lǐmiàn jiù wǒ ***yī ge diànnǎo-dǎng***
group inside only 1SG one CLF computer-DANG
'I am the **only computer expert** of the group'

b. [...] 这份调查报告研究了南京1840位"剁手党" [...]
zhè fèn diàochá bàogào yánjiū le Nánjīng ***yīqiānbābǎisìshí wèi duò-shǒu-dǎng***
this CLF survey report study PFV Nanjing 1840 CLF cut-hand-DANG
'[...] this survey studied 1840 online shopaholics in Nanjing [...]'[33]

Further examples where X-族 *zú* and X-党 *dǎng* nouns are used to indicate individuals rather than groups are the following ones, where a member-class/category relationship is displayed: the X-族 *zú* and X-党 *dǎng* nouns represent a class/category indicating the nature of the individuals (see § 4.1):

10. a. 你是御宅族吗? (LWC)
nǐ shì yùzhái-zú ma
2SG be otaku-ZU Q
'Are you an otaku (nerd)?'

32 https://3g.163.com/dy/article_cambrian/EIU2S9G10544809Y.html.
33 http://china.cnr.cn/qqhygbw/20160123/t20160123_521212278.shtml.

b. 可怜我这熬夜党每晚只睡三小时 (LWC)
kělián *wǒ* *zhè* *áoyè-dǎng* *měi* *wǎn* *zhǐ*
poor 1SG this stay.up.late-DANG each evening only
shuì *sān* *xiǎoshí*
sleep three hour
'Poor me, this night owl who sleeps only three hours per night'

But what about cases like those in (3) and (4), where 客 *kè* and 族 *zú* may combine in the same word, so that both the X-客 *kè* and the X-客族 *kèzú* version of a word are attested? Generally speaking, in those cases it seems that actually the X-族 *zú* word is not used to refer to individuals. As a matter of fact, X-客族 *kèzú* words, differently from X-客 *kè* words, are not generally used with a sortal classifier in numeral-classifier constructions:

11. a. 一个/位换客
yī *ge/wèi* *huàn-kè*
one CLF exchange-KE
'an exchanger'

b.?? 一个/位换客族
yī *ge/wèi* *huàn-kè-zú*
one CLF exchange-KE-ZU

However, both of them are apparently allowed with a measure numeral classifier, as e.g. the collector 群 *qún* 'group':

12. a. 在惠州, 也有一群"换客"[34]
zài *Huìzhōu* *yě* *yǒu* ***yī*** ***qún*** ***huàn-kè***
in Huizhou also have one CLF_{group} exchange-KE
'There is a **group of "exchangers"** in Huizhou too'

b. […] 并涌现出一群"换客族"[35]
bìng *yǒngxiàn-chū* ***yī*** ***qún***
and emerge.in.large.numbers-come.out one CLF_{group}
huàn-kè-zú
exchange-KE-ZU
'[…] and a large **group of "exchangers"** emerged'

34 http://news.ifeng.com/gundong/detail_2013_11/19/31364520_0.shtml.
35 http://news.sina.com.cn/c/2011-05-16/112422472490.shtml.

c. 福州的这群“换客”们, 带来的“宝贝”都不太多 […][36]
*Fúzhōu de **zhè** **qún** **huàn-kè-men** dàilái de bǎobèi*
Fuzhou SP this CLF_{group} exchange-KE-PL bring SP treasure
dōu bù tài duō
all not too many
‘**This group of “exchangers”** in Fuzhou did not bring many ‘treasures […]’

d. 这可不是在做梦,而是一群“换客族”们在网络交换平台上发出的召唤。[37]
zhè kě bù shì zài zuòmèng ér shì
this actually NEG be PROG dream but be
***yī qún huàn-kè-zú-men** zài wǎngluò jiāohuàn*
one CLF_{group} exchange-KE-ZU-PL at Internet exchange
píngtái shàng fāchū de zhàohuàn
platform on issue SP call
‘This is not a dream, but the call issued by **a group of “exchangers”** in an exchange platform on the web’.

This is possibly due to the fact that for X-客族 *kèzú* nouns a less degree of individuation is licensed, and thus they can be used to refer to the members of the group but not to indicate a single entity; accordingly, they imply plurality. This issue requires further investigation.

In a nutshell, both 族 *zú* and 党 *dǎng* have undergone further extension of meaning, departing more from their original meaning – indicating a group –, and at present they can be used to refer to individuals.[38] As pointed out by an anonymous reviewer, similar cases of collective > individual metonymic shift are observed in different languages, as e.g. Spanish *policia* ‘police’: *un policia* ‘a policeman’ (lit. ‘a police’). We will go back to this issue in § 4.1.

4 On the Development of 族 *zú*, 党 *dǎng* and 客 *kè*

In the preceding section, we have shown that 族 *zú*, 党 *dǎng* and 客 *kè* appear in a fixed position, with a fixed meaning, building families of words indicating people doing certain activities or with shared characteristics or behaviour. Can they be labelled as suffixes then? In order to answer this question, in this section we will focus on the evolution of the three items at issue.

36 http://news.sina.com.cn/s/2006-11-21/010010551237s.shtml.

37 https://baike.baidu.com/item/%E6%8D%A2%E5%AE%A2%E6%97%8F.

38 We may remark that 一族 *yīzú* (see § 3.1), despite bearing the same meaning as 族 *zú*, cannot refer to individuals (see Cao 2007; Lu 2010).

4.1 The Evolution of 族 *zú*

As we have mentioned in § 3.1, 族 *zú* as an affix-like item originates from Japanese. Its original meaning of 'clan, tribe, group' developed into the affixal 族 *zoku* 'a group of people with similar feelings or passions', which was then imported in Taiwan, Hong Kong, and later in Mainland China, as we have seen. It then acquired the more generic meaning of 'a category/group of people with common characteristics or behaviour'.

13. 族 *zú* 'clan/ethnic group' > a group of people with similar feelings or passions > a category/group of people with common characteristics or behaviour

Thus, it is evident that this item underwent a process of generalising abstraction, which involves taking a lexeme to a higher taxonomical level (Heine, Claudi, Hünnemeyer 1991; Arcodia 2011). This is confirmed by the fact that 族 *zú* can be used to indicate a variety of referents (see § 3.1): fans/people who love something, workers, people with a particular behaviour in common or engaged in certain activities, people with some characteristics in common etc. This can be seen as a process of grammaticalisation through metaphorical extension, with increased lexical generality and contextual expansion (see Arcodia 2011, 126-7); we argue that the different meanings conveyed by this item may all be subsumed under the meaning 'a category/group of people with common characteristics or behaviour'. Given these characteristics, we maintain that 族 *zú* can be classified as a proper suffix: as pointed out by Arcodia (2011, 125-6),

> [s]ince the meaning expressed by a derivational affix, a grammaticalised sign, may be very general, it is not surprising that it can be used to design a huge variety of referents, provided that it is still possible to identify the commonalities among the various instances.

Recall that in Chinese grammaticalisation usually does not display co-evolution of form and meaning, i.e. affixes are generally characterised by meaning generalisation but not by phonological reduction (see § 2).

In addition, we have shown that 族 *zú* underwent further meaning extension, and it is now used to indicate single entities as well (§ 3.4). This appears to be similar to the development of the suffix 家 *jiā*, which was first used to indicate a group ('school of thought'), as e.g. 法家 *fǎjiā* 'Legalists' (*一个法家 *yī ge fǎjiā* 'one Legalist'), and then started to form individual nouns, with the meaning of 'expert', as e.g. 艺术家 *yìshùjiā* 'artist', 语言学家 *yǔyánxuéjiā* 'linguist' (一个艺术家 / 语言学家 *yī ge yìshùjiā* / *yǔyánxuéjiā* 'an artist / a linguist'); see Wang ([1980] 2002, 230).

As we mentioned in § 3.4, this kind of metonymical semantic shift is not uncommon in the world's languages. Specifically, this metonymical pattern can be seen as an extension of the part-whole relationship to the domain of collections, i.e. sets of roughly equal members: for example, a swarm of bees is made up only of bees, thus it is a collection, because its parts are largely identical (Peirsman, Geeraerts 2006, 302). In collections, entities are conceived as relatively independent but still closely associated. Through this kind of metonymical pattern, a collective term can be used for one entity only, as e.g. in the case of German *Imme* 'bee' (single entity), which developed from Middle High German 'swarm of bees' (collection; Peirsman, Geeraerts 2006, 304). This phenomenon, as we mentioned in § 3.4, is observable in the polysemy displayed by some nouns in synchrony as well. Gardelle (2019, 112-19) observes that in English originally collective nouns, as e.g. *crew*, may come to mean "more than one member in a group", as in *these crew* (uninflected plural), or even, for some of them, "a member in a group" (*one crew*, in the sense of 'one member of a crew'). Another example is *police* used in the sense of 'policeman', as in *those police* 'those policemen', *two police* 'two policemen'.[39]

For these nouns of collective origin, Gardelle argues that the mechanism at work is 'type coercion', i.e. a rather unusual use of a word as regards its grammatical features (in this case, use as uninflected plural instead of singular count) (see Audring and Booij 2016). Gardelle (2019, 115-16) hypothesises that this kind of type coercion goes through three stages: 1) the noun has collective sense and takes grammatical agreement (*this crew has…*); 2) the noun, still having a collective sense, licenses semantic override agreement (foregrounding of the individuals) outside the NP, in the verb and in pronouns (*this crew have…they…*) – non-additivity is lost, and the predicates and anaphors only apply to the individuals; 3) uninflected lexical plural use (*these crew have…*). This plural denotes units, not a collective whole, "though they are expected to belong to a group of the kind denoted by the collective sense of the noun" (Gardelle 2019, 115). This is considered a type of coercion by Gardelle, since this sense is not freely accessible with all collective nouns that denote humans (**these*

39 Gardelle (2019, 109-10) notes that the uninflected plural, meaning "more than one member in a group", is less individuated than the noun that names the separate units: she points out that *those police* is found in cases in which the police officers act together, react together, without any differentiation, while *those policemen* may be used in the same contexts or where there is individuation, as in "[i]t was directed to those policemen who kill and mistreat Blacks". Similarly, *two police* is found only in contexts of professional activities (arrests, or to count victims) – what matters is that they belong to the same socio-professional category –, while *two policemen* are found either in the same contexts or with a higher degree of individuation.

committee) and does not allow for free combination with determiners. As a matter of fact, Gardelle shows that the only determiners licensed by all the uninflected plurals of collective origin are plural demonstratives (these/those), while quantities (one, two, several) are acceptable only with a few nouns (e.g. *crew*, *police*, *faculty*). Gardelle observes that, semantically, conceptualisation with a demonstrative determiner only requires a very low degree of individuation of the units, if compared e.g. with quantities. This could explain why numerals are found only with some of these nouns. She further stresses that actual numerals seem to stand one step further in the evolution of these uses, since they are available only for some of the nouns examined: as for 'one' ('one member'), it is restricted to very few nouns (*clergy*, *crew*, *faculty*, *police*, *staff*), possibly due to potential referential ambiguity. Finally, the use of the indefinite article *a* is very rare. Gardelle argues that stage 3 is reached through plural uses (*these*/*those*); only at this point, for some nouns and to some speakers, more individuation may be licensed, including, ultimately, the singular.

Gardelle (2019, 116) points out that type coercion is accompanied by semantic coercion, from group to members, "as the loss of the /count/ feature entails a loss of boundedness at lexical level"; the noun becomes polysemous. The shift from the collective sense to the uninflected plural sense takes place at a notional level, from the notion of group to that of members; the uninflected plural denotes a class, a socio-professional category, "albeit one in which people are expected to be members of groups".

Gardelle (2019, 117) concludes that these uninflected plural nouns are not collective: the units do not stand in a part-whole relationship with the plurality (**crew are composed of crew*/*members*/*members of crew*). These nouns do not denote a collective whole but a class, indicating the nature of the individuals. Thus, they are characterised by a member-class relation (e.g. *she is crew*).

Let us now go back to X-族 *zú* nouns in Chinese. Given the characteristics displayed by these nouns observed in § 3.4, we argue that 族 *zú* underwent a metonymical semantic shift from 'group' to 'members' (see examples 5a-b, 7a), and then more individuation has been licensed, as shown by the compatibility of X-族 *zú* nouns with quantities (one, two, several): these nouns can refer to single entities as well (examples (8a)-(8d)). In this 'member' sense, the X-族 *zú* noun does not denote a collective whole, but rather a class/category, indicating the nature of the individuals (see the discussion above on English nouns of collective origin). This is confirmed by examples like (10a): sentences like 我是上班族/背包族/爱车族 *wǒ shì shàngbānzú*/*bèibāozú*/*àichēzú* 'I am an office worker/a backpacker/a car lover' express belonging to a category (member-class relation), rather than being part of a group (part-whole relation).

The meaning shift from collective to individual underwent by 族 *-zú* can thus be described as follows: group > members of a category/class > individual (a member of the category/class).

4.2 The Evolution of 党 *dǎng*

In § 3.2 we have seen that the meaning of 党 *dǎng* as the right-hand constituent of complex words indicating groups of people with common characteristics or behaviour probably originates from the meaning 'clique', though the meaning of 'political party' contributed to the development of this new sense as well (Chen, Zhu 2010). As we have shown, X-党 *dǎng* words can indicate a variety of referents: people with a particular behaviour or habit in common; people addicted to something or who love something; people with some characteristics in common. We argue that all these meanings can be subsumed under the meaning 'category/group of people with common characteristics or behaviour'. Given this generalisation of meaning, and the variety of referents it can designate, we conclude that 党 *dǎng* underwent a grammaticalisation process and should be then considered as a suffix. However, as we have pointed out in § 3.2, some X-党 *dǎng* words retain the negative nuance of the original meaning 'clique', arguably reflecting an earlier stage in the semantic evolution of this formative.

Furthermore, we argue that 党 *dǎng*, much like 族 *zú*, also underwent a semantic shift from collective to individual. As a matter of fact, we pointed out in § 3.4, that 党 *dǎng* can be used to refer to 'members' rather than to a collective whole (see examples (6a)-(6b) and (7b)), and to single entities as well (see examples (9a)-(9b)). We can conclude that, like 族 *zú*, it refers to a class/category, indicating the nature of the individuals, as emerges from examples like the one in (10b). See also the following example:

14. 因为我是游戏党,所以当初买这部手机时最先看中的就是它的性能配置。[40]

yīnwèi	***wǒ***	***shì***	***yóuxì-dǎng***	*suǒyǐ*	*dāngchū*	*mǎi*	*zhè*
because	1SG	be	game-DANG	so	at.first	buy	this
bù	*shǒujī*		*shí*	*zuìxiān*	*kànzhōng*	*de*	*jiùshì*
CLF	mobile.phone		time	very.first	take.a.fancy.to	SP	be
tā	*de*	*xìngnéng*	*pèizhì*				
3SG.N	SP	function	configuration				

'Since **I am an online videogame player**, when I bought this mobile phone, what I first considered was its performance settings'.

40 https://c.m.163.com/news/a/FNA6NOCJ05318V7C.html.

The meaning shift from collective to individual underwent by 党 *dǎng* is similar to the one underwent by 族 *zú*: group > members of a category/class > individual (a member of the category/class).

Therefore, given the meaning generalisation and semantic shift underwent by 党 *dǎng*, we conclude that it can be included among affixes.

4.3 The Evolution of 客 *kè*

As highlighted by Wu (2010), Basciano (2017), and Arcodia and Basciano (2018), the pattern X-客 *kè* already existed in previous stages of the language. The basic meaning of 客 *kè*, as we have seen (see § 3.3), is 'guest, visitor'; however, if we look at its meaning in Classical Chinese, we also find 'person specialised in a certain activity', 'person engaged in a particular pursuit' (see 古汉语大词典 *Gu Hanyu da cidian* 'Great Dictionary of Classical Chinese', 1999), as it is evident e.g. in words like 侠客 *xiá-kè* 'chivalrous-KE, knight errant', 掮客 *qián-kè* 'serve.as.broker-KE, broker', 剑客 *jiàn-kè* 'sword-KE, swordsman'. Thus, it can be argued that the use of 客 *kè* as the right-hand constituent of complex words indicating 'a person doing a certain activity', or 'a person with certain characteristics', developed from this meaning.

Wu (2010) argues that the meaning 'guest, visitor' is the oldest one, which is attested since the pre-Qin period (before 221 BC); it then underwent extension of meaning, and its scope widened, beginning to indicate not only home visitors, but also travellers, people travelling or residing away from home, and even emissaries and invaders or aggressors. Later, 客 *kè*, while preserving the original meaning of 'guest', also developed other meanings: for example, Wu observes that for 水客 *shuǐ-kè* 'water-KE' the meaning 'boatman' emerged in the Wei-Jin period (220-420). Then, it underwent further extension of meaning: in the Tang period (618-907), for example, the word 瘦客 *shòu-kè* 'thin-KE, emaciated' emerged. Therefore, this morpheme underwent gradual generalisation of meaning, departing from its original meaning and starting to indicate 'a person involved in some activity' (e.g. 刺客 *cì-kè* 'assassinate-KE, assassin', 说客 *shuō-kè* 'speak-KE, persuasive talker') or a 'person with certain characteristics' (e.g. 醉客 *zuì-kè* 'drunk-KE, drunkard').

Thus, apparently the influence of English and netspeak gave an impulse to the development of an already existing pattern, rather than leading to the creation of a new one. Arcodia and Basciano (2018, 248) even speculate that the choice of 客 *kè* as a phonetic adaptation of the second syllable of English *hacker*, among many other morphemes which are commonly used in Modern Chinese for phonetic adaptations in loanwords (e.g. 克 *kè* 'overcome', 科 *kē* 'department' etc.), could have been motivated also by the meaning which 客 *kè* already had in word formation.

At present, 客 *kè* as the right-hand constituent of complex words can form nouns indicating different types of persons doing any kind of activity (not only on the web) or having certain characteristics. According to Arcodia and Basciano (2018), the general word-formation schema for these words is 'person related to X' ('person doing X' or 'person characterised by X'). Given the gradual extension of meaning underwent by this item, we consider 客 *kè* as an affix.

However, Arcodia and Basciano (2018) point out that those neologisms where the whole X-客 *kè* word is a phonetic adaptation of an English word not indicating a person, as e.g. 切客 *qiē-kè* 'cut-KE, fan of location-based services who regularly checks in to keep friends and relatives posted on her/his whereabouts', do not fit well this schema. As we have seen in § 3.3, the whole word is a phonetic adaptation of English *check-in*, but it indicates a person involved in an activity connected to the semantic of the phonetic adaptation as a whole ('person doing X-客 *kè*', rather than 'person doing X'). The role of 客 *kè*, thus, is not only phonetic: as we mentioned earlier, it contributes the meaning of 'person' as well. Therefore, Arcodia and Basciano (2018) consider these words as a special case of the X-客 *kè* construction.

As suggested by an anonymous reviewer, an alternative explanation could be that there are two routes of generalisation of 客 *kè*: one is the native route; the other one is the loan route, possibly resulting from the introduction of 黑客 *hēikè* 'hacker'. The native route ('person doing X' or 'person characterised by X') may be argued to have developed from a gradual extension of the meaning 'person specialised in a certain activity' and contributes to form words as e.g. 排客 *pái-kè* 'line.up-KE, a person paid to stand in a queue for others', 必胜客 *bì-shèng-kè* 'certainly-remain-KE, person doomed to remain single'. The loan route ('person doing X-客 *kè*'), instead, can be argued to have developed from the 黑客 *hēi-kè* 'hacker' model. As we have seen, while 黑客 *hēi-kè* 'hacker' is a phonetic adaptation of an English word indicating a kind of person (see also e.g. 极客 *jí-kè* 'extremely-KE, geek'), in many cases X-客 *kè* is not a phonetic adaptation of a word indicating a person; the meaning 'person' is rather conveyed by 客 *kè*, which is not only a phonetic component. We may hypothesise that 客 *kè*, originally part of a loanword, over time developed as an affix, whose meaning ('a person doing a certain activity') is somehow connected to the one of the loanword it was part of (i.e. *hacker*, a person engaged in a particular kind of activity). The development from the meaning 'hacker' could also explain the quite high number of X-客 *kè* words indicating persons doing activities on the web, or anyway using computers or new technologies (see § 3.3): a *hacker* is someone who does a particular activity online, i.e. someone who uses computers to get access to data in somebody else's computer or phone system without permission.

Even in this scenario, though, it cannot be excluded that the choice of 客 *kè* for the phonetic adaptation and its development into an agen-

tive suffix in these words have been influenced by the meaning this item already had in word formation, as mentioned above, and that the influence of English simply gave a new impulse to its development: thus, 黑客 *hēi-kè* 'hacker' and other words indicating different kinds of hackers may have had a role in reinforcing the word-formation schema at issue, given their basic agentive meaning, rather than being the source of it (Arcodia, Basciano 2018). Needless to say, the issue requires further investigation.

4.4 A Comparison of the Three Word-Formation Patterns

In the previous sections, we described the evolution of 族 *-zú*, 党 *-dǎng* and 客 *-kè*, and we argued for their affixal status, since they underwent a gradual generalisation of meaning and can now be used to indicate a wider variety of referents. In addition, we have shown that 族 *-zú* and 党 *-dǎng* also underwent a semantic shift from collective to individual and can be currently used to refer to single individuals as well.

The development of these three affixes also shows the different mechanisms at work in grammaticalisation processes and the interplay between native patterns and foreign models. As for 族 *zú*, its affixal use was apparently imported from Japanese, a source language for many neologisms, as well as for new word-formation patterns, especially in the period between the end of 19th and the beginning of the 20th century (Masini 1993). As for 客 *-kè*, we pointed out that English had a key role in its development, as suggested also by the high proportion of phonetic adaptations of English words among X-客 *kè* neologisms. At the same time, though, this word-formation pattern was already present in Chinese and developed through a grammaticalisation process inner to the language; thus, it may be argued that English favoured the development of an existing pattern, rather than creating a new one.

The grammaticalisation paths followed by 族 *-zú* and 党 *-dǎng* are very similar, and actually the two affixes are very close in meaning; they may be found attached to the same bases without apparent changes in meaning (see Chen, Zhu 2010, and § 4.2). However, we pointed out that 党 *-dǎng* appeared later as a suffix, and conveys a more modern flavour; in addition, in some words it still retains the negative nuance of the original meaning 'clique' (see Chen, Zhu 2010; § 4.2). This suffix is not as established as 族 *-zú*, and apparently its use is typical of user-generated content (i.e. created by the users of an online system). This is quite clear if we compare the number of types (i.e. the number of different words created by a word-formation process) found in the dictionary of neologisms (XCY) with those found in the corpus (LWC) for X-族 *zú* and X-党 *dǎng* words:

Table 2 X-族 *zú* and X-党 *dǎng* neologisms in XCY and LWC[41]

	XCY	**LWC**
X-族 ***zú* words**	215	142
X-党 ***dǎng* words**	3	184

As may be seen in table 2, we find an abundance of X-族 *zú* words in the dictionary of neologisms (XCY), which is a hint of the fact that this affix has been consistently and continuously used over the last thirty years, and its use is widespread in society. This word formation pattern is now established in the Chinese lexicon, and many X-族 *zú* words have been 'institutionalised', i.e. after having been widely employed for a reasonable amount of time, they have started to be accepted and recognised by language users as items of their regular vocabulary (see Bauer 1983; Fernández-Domínguez 2010). In contrast, only 3 X-党 *dǎng* words are listed in the XCY, which is in line with the relatively young age of this suffix: this word-formation pattern is not as established as X-族 *zú*, and most of these neologisms are not 'institutionalised'. Coinages may be produced and used for some time, and then disappear: these words are known by the speakers who coined them, and perhaps to the speaking community around, but remain unnoticed for most language users (Hohenhaus 2005; Fernández-Domínguez 2010). Blocking could avoid the institutionalisation among speakers of part of X-党 *dǎng* words: it is possible that some X-党 *dǎng* words appear in the language, are used for a short period of time, and then disappear in favour of the previously existing X-族 *zú* words which are already widely used in the community (e.g. 上班党 *shàngbāndǎng vs* 上班族 *shàngbānzú* 'office workers'; see Fernández-Domínguez, Díaz-Negrillo and Štekauer 2007). Blocking, indeed, does not avoid the coinage of words, but rather their institutionalisation, i.e. their wide usage in the community (Bauer [1988] 2003, 80-1).

However, if we look at the XCY and LWC columns in table 2, we can see that, despite the much greater number of distinct X-族 *zú* words in the XCY, in the LWC there are actually more X-党 *dǎng* words than X-族 *zú* words. This suggests that 党 *-dǎng* as a suffix is particularly frequent in user-generated texts. The preference for X-党 *dǎng* words can be stylistically motivated, since it represents a more fashionable pattern (see § 3.2). According to Plag (2006b, 550), productivity is also influenced by fashion, regardless of any need to name things (social factors or pragmatic needs can motivate new word creation; see

41 We excluded from the count words in which 党 *dǎng* and 族 *zú* bear their original meaning, as e.g. 政党 *zhèngdǎng* 'political party' or 藏族 *zàngzú* 'Tibetan ethnic group'.

Dal, Namer 2016). We will go back to this issue in the next section.

As for X-客 *kè*, we showed that it followed a peculiar grammaticalisation path, in which a native pattern interacted with a foreign model. The agentive meaning it acquired ('person doing a certain activity' or 'person with certain characteristics') is close to that conveyed by 党 *-dǎng* and 族 *-zú* (compare 刷书客 *shuā-shū-kè* 'scan-book-KE' and 刷书族 *shuā-shū-zú* 'scan-book-ZU', both referring to someone who scans with a mini-scanner the content from the books in a bookstore or a library); in addition, we remarked that the two suffixes may combine in the same word (see § 3.4). Nevertheless, we have stressed the fact that, in our corpus, many X-客 *kè* words indicate persons involved in online activities, or anyway activities connected to technology, and that there is a high proportion of loanwords among them, differently from X-党 *dǎng* and X-族 *zú* words, highlighting the role of English and of the word 黑客 *hēikè* 'hacker' in the development of this pattern (§ 4.3). What about the diffusion of this word-formation pattern? Judging from the number of X-客 *kè* types found in the XCY and in the LWC **[tab. 3]**, this pattern is not particularly established and widespread in the language, neither it is particularly common in netspeak, if compared to X-族 *zú* and X-党 *dǎng* **[tab. 2]**.

Table 3 X-客 *kè* words in XCY and LWC[42]

	XCY	LWC
X-客 ***kè* words**	21	56

The figures in table 3 shows that the number of X-客 *kè* 'institutionalised' words is not high: the words listed in the XCY are more than those listed for X-党 *dǎng*, which is quite expected, since X-党 *dǎng* is the newest word-formation pattern among those considered; however, they are very few if compared to X-族 *zú* words. In addition, the number of types of X-客 *kè* found in the LWC is quite low compared to the other suffixes at issue, suggesting that the number of new words coined by means of this process is relatively limited: as pointed out by Fernández-Domínguez,

> accepting the assumption that corpora are reliable reflections of language (Bauer 2001: 47; Plag 2003: 52), *V* [type frequency] should be a good indicator of the number of words coined by a pro-

42 We excluded those words in which 客 *kè* bears the meaning of 'guest' or 'client', as e.g. 顾客 *gùkè* 'customer', and compounds in which the right hand constituent is a X-客 *kè* word, as 心理黑客 *xīnlǐ-hēikè* 'psychology-hacker, a person who helps others solve psychological issues'. Also, we decided to exclude all words indicating different kinds of 'hackers', for the reasons explained in fn. 20.

> cess, so that the higher the figure of types, the more units a process has formed. (2010, 198)

We will return to this issue in the next section.

All in all, the morphological processes involving the three suffixes at issue are all productive, in the sense that they are 'available', i.e. they can be used in the present stage of the language to build new words (Bauer 2001, 205-11). But to what extent is their availability exploited in language use, i.e. to what extent are they 'profitable' (Bauer 2001, 205-11)? In the next section, we will compare their productivity by assessing their 'profitability' in the LWC: while availability is a qualitative notion (a process is either available or not), profitability is a quantitative notion because it deals with how many lexemes an available process coins, thus one process may be more profitable than another (Fernández-Domínguez 2010; for an overview on qualitative and quantitative approaches to productivity, see Dal, Namer 2016).

As pointed out by Plag (2006a, 124), "it is well known that certain affixes are more commonly found in certain types of texts than in others": given the characteristics of the three affixes illustrated here, LWC is best suited to assess their profitability, since it is quite recent and is made up of user-generated content. As the LWC collects all the posts by Weibo users within a certain period of time, it reflects how words are actually, spontaneously and creatively used, and consequently the vitality of the three suffixes. The use of corpora rather than dictionaries as a source of data is motivated by the fact that in a corpus we may find productively formed derivatives which are not listed in dictionaries, and thus "corpus-based descriptions of productivity reflect how words are actually used" (Nishimoto 2003, 51).

5 A Comparison of the Productivity of 族 *-zú*, 党 *-dǎng* and 客 *-kè*

Several methods have been proposed in the literature to measure the profitability of a given process (for an overview, see Plag 2006a, 2006b). The same affix may score differently for different measures, thus yielding different productivity rankings, depending on the method used (for a summary, see Plag 2006b, 544-6). This is because each measure "highlights a special aspect of productivity" (Plag 2006a, 123).

As we have already shown in the previous section, if we look at type frequency, widely used as a productivity measure in the literature (see Fernández-Domínguez 2013), then 党 *-dǎng* is the most productive suffix, while 客 *-kè* is the least productive one.

Table 4 X-族 *zú*, X-党 *dǎng* and X-客 *kè*: type frequency in the LWC

	Types
X-党 ***dǎng* words**	184
X-族 ***zú* words**	142
X-客 ***kè* words**	56

However, as observed by Fernández-Domínguez (2013), this measure may tell us something about the degree of generalisation (the degree to which a process has spread its derivatives in language) of a process but does not say anything about its availability, ignoring the synchronic status of word-formation processes: it focuses on the attestation of lexemes. This measure describes past productivity, i.e. the productivity of a process up to the present, and it is independent of its actual use (see also Dal, Namer 2016).

Other approaches to productivity look at this notion from a probabilistic-statistical perspective and focus on the likeliness of a given pattern to coin new words in the future (see the overview in Fernández-Domínguez 2013). Here we adopt Baayen's hapax-based index of productivity (P-index),[43] which is based on the number of *hapax legomena* (Baayen 1992): if an affix is very productive, we expect to find many *hapax legomena* containing that affix in a large text corpus, since it is typically among hapaxes that we find the higher proportion of neologisms (Renouf, Baayen 1996). Therefore, the crucial assumption behind this method is that the number of hapaxes of a given morphological category correlates with the number of neologisms of that category. In this sense, the number of hapaxes can be seen as an indicator of productivity.

Baayen's P-index is obtained by dividing the number of *hapax legomena* with a given affix (n1) by the number of tokens containing that affix (N) in the corpus considered:

15. P = n1 / N

If all of the words found in a text sample are hapaxes, the P-index will be 1 (maximal productivity), while many high frequency words increase the value of N, leading to a low productivity index.[44] Thus, high token frequency is connected with a high degree of lexicalisa-

43 Baayen's models have undergone a number of modifications over the years, but in all of them *hapaxes* occupy a central position (for an overview, see Fernández-Domínguez 2013; Dal, Namer 2016).

44 Several shortcomings of this hapax-based measure of productivity have been pointed out (see e.g. Bauer 2001; Fernández-Domínguez 2013; Dal, Namer 2016). Generally speaking, larger corpora lead to increased accuracy in calculating the P-index.

tion (storage in the lexicon) and low productivity, while low token frequency is connected with a low degree of lexicalisation and high productivity: as observed by Plag (2006a, 123), the presence of a large number of low-frequency words keeps the rule alive, since they force speakers to segment the derivatives, strengthening the existence of the affix. *Hapax legomena* are often unfamiliar words, but they are understandable for the hearer or reader if the process which created them is still 'active'.

Table 5 shows the P-index of 族 *-zú*, 党 *-dǎng* and 客 *-kè* in the LWC.[45]

Table 5 P-index of X-族 *zú*, X-党 *dǎng* and X-客 *kè* in the LWC

	Tokens (N)	***Hapax legomena* (n1)**	**P-index**
X-党 ***dǎng* words**	342	137	0,400
X-族 ***zú* words**	1335	80	0.059
X-客 ***kè* words**	469	23	0,049

As we can see from the figures in table 5, the P-index of 党 *-dǎng* ranks the highest, while that of 客 *-kè* ranks the lowest, in line with the productivity ranking obtained by calculating type frequency **[tab. 4]**. 党 *-dǎng* has the highest number of hapaxes but the lowest number of tokens, meaning that among X-党 *dǎng* words there are not many high frequency words, leading to a very high P-index: this means that this pattern has a high potential to be used for the coinage of new forms, if needed. In contrast, 族 *-zú* displays a number of tokens significantly higher than that of the other two suffixes, meaning that many X-族 *zú* words are quite frequently used, leading to a large number of tokens and, consequently, an overall decrease of the P-index. As for 客 *-kè*, it is characterised by a low number of hapaxes (the lowest among the three suffixes) but a relatively high number of tokens (higher than 党 *-dǎng*), meaning that some of these words are frequently used; this leads to a low P-index.

These data confirm what already emerged from the discussion in the previous sections, i.e. that the X-族 *zú* pattern is quite established, and that many X-族 *zú* words are widespread in the language

45 We must remark that the Leiden Weibo corpus has one major problem, namely that many messages are simply reposted from other users, and thus there are many cases of duplicated messages. This leads to an increase in the number of tokens; we thus manually removed the repeated messages, in order to get a more reliable picture.

Furthermore, we had to exclude the word 博客 *bókè* from the count. The overall number of tokens in the LWC is 3,006, but it includes both the meaning of 'blog' and that of 'blogger'. Since it was not feasible to separate manually 'blogger' from 'blog', given the high number of tokens, we decided to exclude it. However, at a cursory look, we noticed that the meaning of 'blog' is predominant.

and have become 'institutionalised'. Also, the high productivity displayed by 党 *-dǎng* is in line with the young age of this pattern and with its current popularity among netizens. The X-族 *zú* pattern was widely used for a certain period of time, producing many words which eventually became accepted as part of the common language and 'institutionalised' (as confirmed by the high number of types in the XCY; § 4.4), but it has apparently been superseded by the newly popular X-党 *dǎng* pattern, confirming what observed by Chen and Zhu (2010) on the two patterns. Its P-index predicts a high potential to build new words in future, much higher than that of X-族 *zú*.

As for the X-客 *kè* pattern, it is not as established as X-族 *zú*, but, at the same time, it displays limited productivity. The reasons of its low productivity should be investigated in depth: what are the factors restricting its productivity? Since many affixal elements indicating a type of person are currently found in Chinese, especially in user-generated texts, pragmatic factors, sociological factors, and blocking phenomena should be probably taken into account in order to get a clearer picture.

6 Conclusions

The influence of foreign languages and netspeak in the past few years not only led to the creation of a large number of neologisms, but also to the development of new word-formation patterns in Chinese, with the creation of many derivational affixes. Some of these items may be widely used at a given time but are then superseded after a while by a newer word-formation pattern. In this paper, we examined three suffixes emerged in the last thirty years, i.e. 族 *-zú*, 党 *-dǎng* and 客 *-kè*, all forming nouns referring to persons. After describing the three word-formation patterns, we focused on the evolution of the three formatives, characterised by meaning generalisation, arguing that at present they can all be considered as suffixes, based on their fixed position in complex words (to the right) and on the meaning generalisation observed.

The suffixes 族 *-zú* and 党 *-dǎng* both form nouns indicating a variety of referents, which we argued can all be subsumed under the general meaning 'people with common characteristics or behaviour'. These two suffixes can also be attached to the same base without any change in meaning. In addition, we also remarked that both of them underwent a meaning shift from collective to individual, and thus they can be used to refer to single entities as well. However, from the point of view of meaning, the two suffixes do not exactly overlap: differently from X-族 *zú* words, some X-党 *dǎng* words retain the negative nuance of the original meaning 'clique', indicating a series of illegal activities. In addition, we pointed out that X-党 *dǎng* is

currently a more popular and fashionable pattern, thus possessing a more modern flavour.

Through the analysis of productivity based on the data of the LWC, we also showed that, while both word-formation patterns are 'available' to form new words at the present stage of the language, the degree of profitability of the X-党 *dǎng* pattern is much higher, meaning that it has a high potential to build new words: the X-族 *zú* pattern was widely used for a certain period of time, producing many words which eventually became accepted as part of the common language, but it has been apparently superseded by the newly popular X-党 *dǎng* pattern.

As for 客 *kè*, a number of words containing this suffix emerged starting from the 2000s in user-generated content. We argued that the influence of English and netspeak gave impulse to an already existent, though limitedly productive, word-formation pattern. From the analysis of the data in our corpus, the X-客 *kè* word-formation pattern is not particularly established and widespread in the language, neither it is particularly frequent in user-generated content: in the LWC it ranks the lowest for type frequency, number of hapaxes, and P-index, while ranks higher than 党 *-dǎng* in terms of token frequency, meaning that some X-客 *kè* words are frequently used. Even though this pattern is available for the creation of neologisms, its potential to create new words is quite limited.

As we mentioned, besides those investigated in this study, at present we can find a number of emerging suffixes indicating people in Chinese. One may wonder why so many different affixes are needed to create words referring to persons. A broader investigation comparing the properties and usage differences of different suffixes would be welcome. Since the creation of neologisms is not always meant to satisfy naming needs, it would be worth investigating the role of social factors, pragmatic needs, as well as language trends, in the development of these suffixes.

Bibliography

Arcodia, G.F. (2011). "A Construction Morphology Account of Derivation in Mandarin Chinese". *Morphology*, 21, 89-130. https://doi.org/10.1007/s11525-010-9173-2.

Arcodia, G.F.; Basciano, B. (2012). "On the Productivity of the Chinese Affixes *–r*, –化 *–huà* and –頭 *–tou*". *Taiwan Journal of Linguistics*, 10(2), 89-117. http://dx.doi.org/10.6519/TJL.2012.10(2).3.

Arcodia, G.F.; Basciano, B. (2018). "The Construction Morphology Analysis of Chinese Word Formation". Booij, G. (ed.), *The Construction of Words. Advances in Construction Morphology*. Berlin: Springer, 219-53. https://doi.org/10.1007/978-3-319-74394-3_9.

Audring, J.; Booij, G. (2016). "Cooperation and Coercion". *Linguistics*, 54(4), 617-37. https://doi.org/10.1515/ling-2016-0012.

Baayen, R.H. (1992). "Quantitative Aspects of Morphological Productivity". Booij, G.; Van Marle, J. (eds), *Yearbook of Morphology 1991*. Dordrecht; London: Kluwer Academic Publishers, 109-49. https://doi.org/10.1007/978-94-011-2516-1_8.

Bareato, S. (2017). *La derivazione in cinese. Uno studio su corpora dei formanti* 族 *zú e* 党 *dǎng* [MA dissertation]. Venice: Ca' Foscari University of Venice.

Basciano, B. (2017). "黑客 *hēikè*, 白客 *báikè*, 红客 *hóngkè*. Hacker e altri 'ospiti' tra i neologismi del cinese moderno". Bulfoni, C. et al. (eds), 文心 *Wenxin. L'essenza della scrittura. Contributi in onore di Alessandra Cristina Lavagnino*. Milano: FrancoAngeli, 384-95.

Bauer, L. (1983). *English Word-Formation*. Cambridge: Cambridge University Press.

Bauer, L. (1998). "Is there a Class of Neoclassical Compounds, and if so Is It Productive?". *Linguistics*, 36(3), 403-22. https://doi.org/10.1515/ling.1998.36.3.403.

Bauer, L. (2001). *Morphological Productivity*. Cambridge: Cambridge University Press.

Bauer, L. [1988] (2003). *Introducing Linguistic Morphology*. 2nd ed. Edinburgh: Edinburgh University Press.

Bauer, L. (2005). "The Borderline between Derivation and Compounding". Dressler, W. et al. (eds), *Morphology and Its Demarcations*. Amsterdam; Philadelphia: John Benjamins, 97-108. https://doi.org/10.1075/cilt.264.07bau.

Bauer, L. (2006). "Compound". Brown, K. (ed.), *Encyclopedia of Language and Linguistics*, vol. 2. Oxford: Elsevier, 719-26.

Bisang, W. (1996). "Areal Typology and Grammaticalization. Processes of Grammaticalization Based on Nouns and Verbs in East and Mainland South East Asian Languages". *Studies in Language*, 20(3), 519-97. https://doi.org/10.1075/sl.20.3.03bis.

Booij, G. (2005). "Compounding and Derivation. Evidence for Construction Morphology". Dressler, W. et al. (eds), *Morphology and Its Demarcations*. Amsterdam; Philadelphia: John Benjamins, 109-32. https://doi.org/10.1075/cilt.264.08boo.

Booij, G. (2010). *Construction Morphology*. Oxford: Oxford University Press.

Cao D. 曹大为 (2007). "'Zu' de leicizhuihua shiyong fenxi" "族" 的类词缀化使用分析 (Analysis of the Pseudo-Affixal Use of 族 *zú*). *Shandong shehui kexue*, 5, 150-2.

Cao T. 曹铁根; Mo W. 莫伟勇 (2012). "Wangluo xinciyu 'X-*kong*' yuyi jiexi" 网络新词语"X 控"语义解析 (Analysis of the Meaning of X-控 *kòng* Neologisms on the Web). *Hunan keji daxue xuebao*, 15(1), 122-4.

Ceccagno, A.; Basciano, B. (2007). "Compound Headedness in Chinese. An Analysis of Neologisms". *Morphology*, 17, 207-31. https://doi.org/10.1007/s11525-008-9119-0.

Chen C. 陈昌来; Zhu Y. 朱艳霞 (2010). "Shuo liuxingyu 'X-dang' —— yi jian lun zhi ren yusu de leicizhuihua" 说流行语 "X党"——兼论指人语素的类词缀化 (On the Buzzword 'X-党 *dǎng*'. A Discussion on the Affixal Status of Morphemes Referring to Persons). *Dangdai xiucixue*, 159, 64-70.

Cheng, L.L.-S.; Sybesma, R. (1999). "Bare and Not-so-bare Nouns and the Structure of NP". *Linguistic Inquiry*, 30(4), 509-42. https://doi.org/10.1162/002438999554192.

Cheung, C.C.-H. (2016). *Parts of Speech in Mandarin. The State of the Art*. Singapore: Springer.

Croft, W. (1994). "Semantic Universals in Classifier Systems". *Word*, 45(2), 145-71. https://doi.org/10.1080/00437956.1994.11435922.

Dal, G.; Namer, F. (2016). "Productivity". Hippisley, A.; Stump, G. (eds), *The Cambridge Handbook of Morphology.* Cambridge: Cambridge University Press, 70-89. https://doi.org/10.1017/9781139814720.004.

Dong X. 董秀芳 (2004). *Hanyu de ciku yu cifa* 汉语的词库与词法 (Chinese Lexicon and Morphology). Beijing: Peking University Press.

Fabb, N. (1998). "Compounding". Spencer, A.; Zwicky, A.M. (eds), *Handbook of Morphology*. Oxford: Blackwell, 66-83. https://www.blackwellpublishing.com/content/BPL_Images/Content_Store/WWW_Content/9780631226949/01Prelim.pdf.

Fernández-Domínguez, J. (2010). "Productivity vs. Lexicalisation. Frequency-Based Hypotheses on Word-Formation". *Poznan Studies in Contemporary Linguistics*, 46(2), 193-219. https://doi.org/10.2478/v10010-010-0010-x.

Fernández-Domínguez, J. (2013). "Morphological Productivity Measurement. Exploring Qualitative versus Quantitative Approaches". *English Studies*, 94(4), 422-47. https://doi.org/10.1080/0013838x.2013.780823.

Fernández-Domínguez, J.; Díaz-Negrillo, A.; Štekauer, P. (2007). "How Is Low Morphological Productivity Measured?". *Atlantis*, 29, 29-54. https://www.jstor.org/stable/41055264.

Gardelle, L. (2019). *Semantic Plurality. English Collective Nouns and Other Ways of Denoting Pluralities of Entities*. Amsterdam; Philadelphia: John Benjamins.

Haspelmath, M. (2002). *Understanding Morphology*. Oxford: Oxford University Press.

Heine, B.; Claudi, U.; Hünnemeyer, F. (1991). *Grammaticalization. A Conceptual Framework*. Chicago: University of Chicago Press.

Hohenhaus, P. (2005). "Lexicalization and Institutionalization". Štekauer, P.; Lieber, R. (eds), *Handbook of Word-Formation*. Dordrecht: Springer, 353-73. https://doi.org/10.1007/1-4020-3596-9_15.

Iljic, R. (1994). "Quantification in Mandarin Chinese. Two Markers of Plurality". *Linguistics*, 32(1), 91-116. https://doi.org/10.1515/ling.1994.32.1.91.

Katamba, F. (1993). *Morphology*. New York: Martin's Press.

Li J. 李杰 (2013). "'Zu' leici zhong cizhuihua qingxiang xianxiang" "族" 类词中词缀化倾向现象 (The Development of 族 *zú* Words into Derivates). *Xi'an hangkong xueyuan xuebao*, 2(31), 21-3.

Li, C.; Thompson, S.A. (1981). *Mandarin Chinese. A Functional Reference Grammar.* Berkeley: University of California Press.

Lu Y. 鲁瑛 (2010). "'xx-*zu*' cilei de yuyanxue yanjiu" "××族" 词类的语言学研究 (Linguistic Research on "xx-*zu*" Neologisms). *Waiguo yuwen*, 2(26), 71-5.

Lü S. 吕叔湘 (1941). *Zhongguo wenfa yaolüe* 中国文法要略 (An Outline of Chinese Grammar). Shanghai: The Commercial Press.

Ma M. 马渺沙 (2016). "Cong ci de xingcheng ji qi jiegou de jiaodu qianxi wangluo liuxing leicizhui de xianxiang —— yi "X-*zu*", "X-*kong*" wei lie" 从词的形成及其结构的角度浅析网络流行类词缀的现象——以"X族""X控"为例 (An

Analysis of the Phenomenon of Popular Affixoids on the Web from the Perspective of the Form and Structure of Words. The Case of X-族 *zú* and X-控 *kòng*). *Meijie yu wenhua yanjiu*, 12, 99-101.

Ma Q. 马庆株 (1995). "Xiandai Hanyu cizhui de xingzhi, fanwei he fenlei" 现代汉语词缀的性质, 范围和分类 (Nature, Domain and Classification of Affixes in Modern Chinese). *Zhongguo yuyan xuebao*, 6, 101-37.

Masini, F. (1993), *The Formation of Modern Chinese Lexicon and Its Evolution Towards a National Language. The Period from 1840 to 1898*. Berkeley: University of California.

Naumann, B.; Vogel, P.M. (2000). "Derivation". Booij, G.; Lehmann, C.; Mugdan, J. (eds), *Morphologie-Morphology.* Berlin; New York: Mouton de Gruyter, 929-43.

Nishimoto E. (2003). "Measuring and Comparing the Productivity of Mandarin Chinese Suffixes". *Computational Linguistics and Chinese Language Processing*, 8(1), 49-76.

Packard, J. (2000). *The Morphology of Chinese*. Cambridge: Cambridge University Press, 71-3.

Pan W. 潘文国; Ye B. 叶步青; Han Y. 韩洋 (2004). *Hanyu de goucifa yanjiu* 汉语的构词法研究 (The Research on Word Formation in Chinese). Shanghai: Huadong Shifan Daxue Chubanshe.

Peirsman, Y.; Geeraerts, D. (2006). "Metonymy as a Prototypical Category". *Cognitive Linguistics*, 17(3), 269-316. https://doi.org/10.1515/cog.2006.007.

Peyraube, A. (1998). "On the History of Classifiers in Archaic and Medieval Chinese". T'sou, Benjamin (ed.), *Studia Linguistica Serica*. Hong Kong: City University of Hong Kong, 39-68.

Plag, I. (2003). *Word-Formation in English*. Cambridge: Cambridge University Press.

Plag, I. (2006a). "Productivity". Brown, K. (ed.), *Encyclopedia of Language and Linguistics*, vol. 10. 2nd ed. Oxford: Elsevier, 121-8. https://doi.org/10.1016/B0-08-044854-2/00125-5.

Plag, I. (2006b). "Productivity". Aarts, B.; McMahon, A. (eds), *The Handbook of English Linguistics*. Oxford: Blackwell Publishing, 537-56.

Renouf, A.; Baayen, R.H. (1996). "Aviating among the Hapax Legomena. Morphological Grammaticalisation in Current British Newspaper English". Renouf, A. (ed.), *Explorations in Corpus Linguistics*. Amsterdam; Atlanta, GA: Rodopi, 181-9.

Shanghai Daily (2010). *Chinese Buzzwords. With English Explanations*. 2nd ed. Shanghai: Shanghai Press.

Shen G. 沈光浩 (2015). *Hanyu paishengshi xinciyu yanjiu* 汉语派生式新词语研究 (Research on Derived Neologisms in Chinese). Beijing: Zhongguo Shehui Kexue Chubanshe.

Sproat, R.; Shih, C. (1996). "A Corpus-Based Analysis of Mandarin Nominal Root Compound". *Journal of East Asian Linguistics*, 5, 49-71. https://doi.org/10.1007/bf00129805.

Steffen Chung, K. (2006). *Mandarin Compound Verbs*. Taipei: Crane.

Sun Y. 孙艳 (2000). "Xiandai Hanyu cizhui wenti tantao" 现代汉语词缀问题探讨 (A Study on Some Problems Related to Modern Chinese Affixes). *Hebei Shifan Daxue xuebao*, 23, 55-8.

Sybesma, R. (2017). "Classifiers, Nominal". Sybesma, R. et al. (eds), *Encyclopedia of Chinese Language and Lingustics*, vol. 1. Leiden: Brill, 620-7. http://dx.doi.org/10.1163/2210-7363_ecll_COM_00000091.

Wang L. 王力 [1980] (2002). *Hanyu shi gao* 汉语史稿 (History of Chinese). Beijing: Zhonghua Shuju.

Wu L. 吴琅琅 (2010). "Xiao xi Wenzhou fangyan 'ke' zu ci —— cong gu Hanyu jiaodu zhengming 'ke' houzhui cunzai keneng" 小析温州方言 "客" 族词——从古汉语角度证明 "客" 后缀存在可能 (A Brief Analysis of the 客 *kè* Family of Words in Wenzhou Dialect. Proving the Existence of the 客 *-kè* Suffix from the Perspective of Old Chinese). *Yuyan wenxue yanjiu*, 3, 27-8.

Xiao Y. 肖遥遥 (2009). "'Zu' lei Hanyu xinci yufahua qianxi" "族" 类汉语新词语法化浅析 (Analysis of the Grammaticalization of 族 *zú* Neologisms in Chinese). *Zhongzhou Daxue xuebao*, 2(26), 75-7.

Yang H. 杨海明; Chen Q. 陈倩仪 (2012). "'~zu' ci yu xinwen yuyan de shidai gan" "～族"词与新闻语言的时代感 (～族 *zú* Words and the Sense of Time in the Language of the News). *Xinwen aihaozhe*, 92-3.

Yip P.-C. (2000). *The Chinese Lexicon. A Comprehensive Survey*. London; New York: Routledge.

Zhang Y. 张谊生; Xu X. 许歆媛 (2008). "Qian xi 'X ke' ci zu —— cihuihua he yufahua de guanxi xin tan" 浅析"X 客"词族 —— 词汇化和语法化的关系新探 (A Preliminary Analysis of the "X-客 *kè*" Family of Words. A New Look at the Relation between Lexicalization and Grammaticalization). *Yuyan wenzi yingyong*, 4, 77-82.

Zhao A. 赵爱萍 (2009). "Shimao 'zuci' jiqi shehui, xinli toushi" 时髦 "族词" 及其社会、心理透视 (The Fashionable 族 *zú* Words. A Sociological and Psychological Perspective). *Qiqihar Shifan Gaodeng Zhuanke Xuexiao xuebao*, 4, 35-6.

Dictionaries

Gu Hanyu da cidian 古汉语大词典 (Great Dictionary of Classical Chinese) (1999). Shanghai: Shanghai Cishu Chubanshe.

The Contemporary Chinese Dictionary (Chinese-English edition) (2002). Beijing: Foreign Language Teaching and Research Press.

Xiandai Hanyu cidian 现代汉语词典 (The Contemporary Chinese Dictionary) (2005). Beijing: The Commercial Press.

Xin shiji xinciyu da cidian 新世纪新词语大词典 (New Century Comprehensive Dictionary of Neologisms) (2015). Shanghai: Shanghai Cishu Chubanshe.

Xinhua xinciyu cidian 新华新词语词典 (The Xinhua Dictionary of New Words) (2003). Beijing: The Commercial Press.

Sociolinguistics

Corpus-Based Research on Chinese Language and Linguistics
edited by Bianca Basciano, Franco Gatti, Anna Morbiato

What Can the Corpus of Mid-20th Century Hong Kong Cantonese Tell Us About Hong Kong Society of Half a Century Ago?

Andy Chin
The Education University of Hong Kong

Abstract This paper reports on a corpus-based sociolinguistic study of terms of address with a special focus on kinship terms found in *The Corpus of Mid-20th Century Hong Kong Cantonese*, which has a size of about one million Chinese character tokens. The corpus data was collected by transcribing the speech dialogues of 81 black-and-white movies produced in Hong Kong between 1940 and 1970. The kinship terms extracted from the corpus can tell us about the family structure and marital life of Hong Kong six decades ago.

Keywords Corpus-based sociolinguistic study. Cantonese corpus. Early Hong Kong society. Terms of address. Family culture.

Edizioni Ca'Foscari

Sinica venetiana 6
e-ISSN 2610-9042 | ISSN 2610-9654
ISBN [ebook] 978-88-6969-406-6 | ISBN [print] 978-88-6969-407-3

Peer review | Open access
Submitted 2020-03-06 | Accepted 2020-03-29 | Published 2020-12-21

DOI 10.30687/978-88-6969-406-6/009

1 Introduction[1]

Baker (2010) commented that cross-fertilisation between two seemingly unrelated disciplines, namely corpus linguistics and sociolinguistics, has been done very little although the two disciplines have established their traditions in the field of linguistics for a long time. Baker explained that this may be due to the fact that corpus linguistics sometimes gives the impression that it "has made only a relatively small impact on sociolinguistics" (2010, 1). In spite of this, Baker (2010, 8-9) showed that the two disciplines share a lot of common features: a) analysing naturally occurring and empirical language data; b) emphasising on language-in-use or social context; c) making use of quantitative methodologies; d) examining and comparing variations and changes; e) providing explanations for the findings. All these common features demonstrate that these two disciplines can produce cluster research. One notable example is Davies' study of "issues related to culture and society, either in terms of change over time or variation between [English] dialects" (2017, 19) by means of various gigantic English corpora.[2] For example, Davies (2017, 27) found that, with data from GloWbe, the word 'terrorism' appears more in the varieties of English spoken in South Asian countries, such as Pakistan and Sri Lanka, than in British English and American English. Furthermore, he found that Australian English has more word types with the suffix *-ies* than other varieties of English in the Inner Circle *à la* Braj Kachru's model of World Englishes.

One research area in sociolinguistics seeks to examine language variations and changes either in diachronic or synchronic dimensions. Adopting a corpus-based approach to study linguistic variations from a diachronic perspective entails that one has to look for

1 Earlier versions of this paper were presented in the BK21PLUS Conference organised by The Hankuk University of Foreign Studies, South Korea (Co-Author: Ou Lili, 27-30 October 2017), and in the 2019 Annual Conference of Society for Hong Kong Studies (22 June 2019). The Author would like to acknowledge the following funding support for the construction of the corpus reported in this paper: (a) *Spoken Corpus Construction and Linguistic Analysis of Mid-Twentieth-Century Cantonese* (Internal Research Grant, The Hong Kong Institute of Education, Project No.: RG41/2010-2011); (b) *A Preliminary Linguistic Analysis of Mid-Twentieth-Century Cantonese from a Corpus-based Approach* (Internal Research Grant, The Hong Kong Institute of Education, Project No.: RG62/12-13R); (c) *Linguistic Analysis of Mid-Twentieth-Century Hong Kong Cantonese by Constructing an Annotated Spoken Corpus* (Early Career Scheme, Research Grants Council, Hong Kong SAR Government, Project No.: ECS859713); (d) *Initiatives in Digital Humanities* (Central Reserve for Strategic Development, The Education University of Hong Kong).

2 These corpora include the *Corpus of Contemporary American English* (COCA), the *Corpus of Historical American English* (COHA), the *Google Books* corpus, *Global Web-based English* (GloWbE), and *News on the Web* (the NOW corpus). These corpora can be accessed at `https://www.english-corpora.org/`.

historical data or to construct a historical corpus. This is not an easy task when one wants to collect real-time language data produced from the past. As McEnery and Hardie put it,

> for these and other extinct languages there is a fixed "corpus" of surviving texts which will never grow any further, except in the rare circumstance that hitherto unknown texts are discovered. An electronic corpus composed of all of these surviving texts (or a sampled subset of them) is thus the ideal tool for taking into account as much data on these historical forms as possible in an analysis of how language has changed. (2012, 94-5)

A corpus-based study of the diachronic development of a language will become fruitful and illustrative only when we manage to collect and process language data produced in the period we want to examine. At the same time, we also need to ensure that the corpus data we collect is "representative", "balanced" and "comparable" (McEnery, Hardie 2012, 10), although it is always not easy to have a corpus that perfectly meets all these three attributes.

2 The Corpus of Mid-20th Century Hong Kong Cantonese

This paper introduces a corpus-based sociolinguistic study of kinship terms in Hong Kong Cantonese, a language spoken as a home language by nearly 90% of the population in Hong Kong.[3] The data comes from *The Corpus of Mid-20th Century Hong Kong Cantonese* (hereafter HKCC) developed at The Education University of Hong Kong since 2011.[4] The data of HKCC was collected by transcribing the speech dialogues of 81 black-and-white movies produced in Hong Kong between 1940 and 1970. There are two phases of corpus development, at different stages and with different sources of funding.[5] The two phases of HKCC have processed spoken Cantonese data with a size of nearly one million Chinese characters.[6] The transcribed data of both phases in HKCC was tokenised and assigned with Cantonese pronunciations. The data in the second phase of HKCC was also annotated with parts-of-speech.

3 See table 3.12 of CSD 2016. For the sociolinguistic situation of Hong Kong, see Tsou 1997 and Bacon-Shone, Bolton, Luke 2015.

4 The URL of HKCC is `http://hkcc.eduhk.hk`.

5 Dialogues of 21 and 60 movies were transcribed in the first and second phases respectively. HKCC is now available online for searching.

6 Dialogues of three genres of movies were transcribed in HKCC: a) melodramas with themes on family and romance; b) detective and suspense; c) comedy.

Chin (2013; 2019a) provided detailed descriptions of the two phases of HKCC, including the data source and the rationales behind the construction of the corpus. The primary aim of HKCC is to provide real time language data for conducting diachronic studies on Cantonese and comparing the Cantonese language spoken in Hong Kong in the contemporary period and that of half a century ago. The HKCC data also bridges the gap of Cantonese linguistic research on early Cantonese (back to early 19th century) and contemporary Cantonese. Specifically, the mid-20th century is a transitional period in which some critical linguistic changes took place in Cantonese: the corpus data can thus provide authentic language data to examine the switchover from the old features to the new features.[7]

Another important feature of HKCC is that it can supply quantitative and qualitative information for examining the characteristics of the Cantonese language. HKCC can generate lists of segmented tokens according to their parts-of-speech and usage frequency, which can provide useful data for selecting items for compiling learning and teaching materials. Furthermore, the sample sentences based on the movie dialogues can allow users to have a better understanding of the use of language in context. Although one many argue that the data of HKCC comes from half a century ago and may be considered outdated and unsuitable for language teaching and learning, HKCC is still valuable because some of the usages and sentence patterns had not changed significantly since mid-20th century. This is especially the case for function words such as aspect markers, which have exceptionally high occurrences in HKCC. For example, the perfective aspect marker 咗 *zo2*[8] has a frequency of 3,300 in HKCC, which is far more than its occurrence (869 tokens) in HKCanCor.[9] To our best understanding, no existing learning and teaching resources can provide comparable amount of data and sample sentences for illustration. In addition, the search functions of the second phase of HKCC have been significantly enhanced so that users can incorporate flexible search criteria such as 'Numeral + Classifier + Noun' to retrieve more results for analysis and comparison.[10]

7 Some examples include the development of neutral questions (also known as Yes-No questions) and indirect object markers (also known as dative markers). For details, see, for example, Cheung 2001 and Chin 2011 respectively.

8 Cantonese examples are transcribed with the Jyutping Romanisation scheme developed by The Linguistic Society of Hong Kong. For details, see `https://www.lshk.org/jyutping`.

9 HKCanCor (*The Hong Kong Cantonese Corpus*) was developed by Professor Luke Kang Kwong at the University of Hong Kong in the late 1990s. The corpus has 869 occurrences of 咗 *zo2* out of 180,000 word tokens. The corpus data can be downloaded from `http://compling.hss.ntu.edu.sg/hkcancor`. For details of HKCanCor, see Luke, Wong 2015.

10 For the search functions in the second phase of HKCC, see Chin (forthcoming).

While there are Cantonese corpora developed in the past two decades, none of them is comparable to HKCC in terms of size and data source.[11] In spite of the availability of Cantonese corpora, linguistic research with Cantonese corpus data mainly focuses on the internal system such as syntax, lexicon, and phonology. This can be seen from a search of the keywords 'corpus' and 'Cantonese' in Google Scholar. Some of the research outputs include, for example, loanword truncation in Cantonese (Luke, Lau 2008), comparisons of temporal and tonal aspects in Mandarin and Cantonese (Peng 2006), the GIVE-construction in Mandarin and Cantonese (Wong 2009), the analysis of type and token frequencies of phonological units in Hong Kong Cantonese (Leung, Law, Fung 2004), the verbal suffix 着 *zoek6* (Lai, Chin 2018). These sample studies show how corpus data can enhance our understanding of the linguistic properties of Cantonese. However, they are still limited to language internal features. There are in fact many extra-linguistic issues that can be pursued with corpus data. One of the merits of HKCC is the dialogic and highly interactive nature of its data. It is thus useful for studying issues on discourse, pragmatics and sociolinguistics, which are relatively under-explored in Cantonese linguistic research. The author and his research team have conducted a number of studies on Cantonese discourse with data from HKCC. For example, Tse and Chin (2015) examined the features of co-referential noun phrases such as 你個衰人 *nei5 go3 seoi1jan4* 'you CLF bad guy, you the bad guy', that have the same surface structure as the possessive noun phrase with a classifier used as possessive marker, such as 你個公仔 *nei5 go3 gung1zai2* 'you CLF doll, your doll'. Chin (2018a) explored discourse markers including the tag questions 好唔好 *hou2 m4 hou2* 'is it alright' and sentence final particles. Chin (2018b) compared the two Cantonese prohibitive markers 唔好 *m4hou2* and 咪 *mai5*, which are usually treated as synonyms in Cantonese dictionaries and textbooks. The study examined the verbs these two prohibitive markers take, as well as the length of the verb phrases. It is interesting to see that each marker shows some distinct features which are not found in the other marker.

11 For details on the nature and data source of other Cantonese corpora, see Chin 2013; 2019a.

3 Applications of HKCC: Tracking Changes of Society

HKCC is important and useful for studying variations and development of Hong Kong Cantonese over time. There are lexical items and syntactic structures in HKCC which are no longer active in contemporary Cantonese. Examples include 霎氣 *saap3hei3* 'having an argument with someone', 蘇蝦 *sou1haa1* 'baby'. As for syntactic structures, we can find both old and new patterns co-existing in the same sentence, i.e. hybrid forms.[12] Besides linguistic analysis, we can also make use of the data from HKCC to examine sociocultural issues, because the content of the movies can reflect the popular and key social issues of Hong Kong society of the period concerned. Lui (1988) studied the housing issue of Hong Kong in the 1950s with reference to two melodrama movies, namely *In the Face of Demolition* (危樓春曉, 1953) and *The Kid* (細路祥, 1950).[13] Specifically, Lui argued that

> these films do provide corroborative evidence in understanding the decade of the 1950s. The feeling among Hong Kong people that the government should play a leading role in solving their housing problem grew only in the past ten to twenty years. (1988, 90)

In his study of Cantonese melodrama with the theme of familial relationships in the 1950s and 1960s, Law observed that the disappearance of Cantonese melodrama after the 1960s could be due to "rapid modernisation of Hong Kong" and "the spread of the nuclear family as the basic social unit and its accompanying individualism". These changes of social life and interpersonal relations "outstripped the development of the form and content of Cantonese melodrama" (Law 1986, 19).

The above two studies of Hong Kong society through early Cantonese movies show that movies can act as a telescope allowing us to look at some deeper issues of the community in which they are depicted. As language is argued to be the carrier of culture, we can thus observe, through the movie dialogues, what was being practised by people, as well as the characteristics of the social life and culture in the community concerned.

Mid-20th century saw the booming of Hong Kong's movie industry. According to Chung (2004), more than 1,500 movies, literally known as 'Cantonese long movies' (粵語長片 *jyut6jyu5 coeng4pin2*), were produced between 1950 and 1960. The dialogues in these movies can be claimed to have faithfully recorded the Cantonese language

12 One example is neutral questions produced in the movies included in HKCC. For details, see Chin 2019b.

13 These two movies were also included in HKCC.

spoken in Hong Kong at that time. Some of these Cantonese movies have their stories centring on the social situation of Hong Kong of that time. Some of the themes include familial relationships, especially conflict of interest among family members, romance among young people, and tragedies arising from social issues such as poverty and humanity. We thus believe that the data from HKCC can serve as a good resource for conducting a corpus-based sociolinguistic study.

In the following, based on the data extracted from HKCC, we will examine the kinship terms and lexical items related to family and marriage with an aim to explore the family culture and family organisation in Hong Kong half a century ago.

4 Kinship Terms and Family Culture

Terms of address are lexical items used to address a person in conversations. For kinship terms which are used to refer to family members, the amount and complexity are highly correlated with the concepts of family structure in the respective speech community. There have been numerous studies comparing the kinship term systems between the Chinese language and other languages such as English. It is generally acknowledged that kinship terms in Chinese have a "finely grained semantic structure" (Qian, Piao 2009, 190), which can be associated with the complex family structure of Chinese society. For example, Chinese families reflect the patrilineal character (Wu 1927) and this is rendered in the kinship terms referring to grandparents. Kinship terms for maternal grandparents carry the prefix 外 *ngoi6*, literally 'external, outside', such as 外公 *ngoi6gung1* 'maternal grandfather' and 外婆 *ngoi6po4* 'maternal grandmother'. Furthermore, Chinese kinship terms make distinction in terms of age and gender, while English in some cases uses one single kinship term instead.[14] Typical examples are *uncle*, *aunt* and *cousin*. All these differences between kinship terms in Chinese and English can reflect the family structures of the two cultural traditions.

We can also have a look at the family structure of early Hong Kong by examining the kinship terms found in HKCC. As we discussed in § 2, the movies we selected to transcribe cover three genres, namely melodrama, detective and suspense, and comedy. Many of these movies have their stories and plots centring on family members. For example, in some suspense movies, the stories were about disputes among family members, such as brothers and sisters fighting for the

14 Taking all these attributes into consideration, kinship terms in Chinese (including its dialects) can be examined by means of componential analysis. See, for example, Chao 1956; McCoy 1970; Cheung 1990; Qian, Piao 2009.

property left by their parents. Sometimes members of extended families such as uncles and aunts were also involved in the story.

Furthermore, it is noted that "propositional synonyms" referring to "a single kinship concept" always exist (Qian, Piao 2009, 193). These are also interesting terms that we can examine as they may signify different styles or degrees of solidarity between the addresser and the addressee. This will be discussed in § 5.3.

Besides kinship terms, we will also examine words related to the concept of marriage. Kinship relationships are built upon marriage between a man and a woman although, in modern society, families with single-parent, single-child, same-sex couples or heterosexual cohabiting partners give rise to many new kinship terms, as illustrated by Qian and Piao (2009). In other words, the examination of kinship terms of different time periods can allow us to observe the development of society in terms of marital life and family organisation.

5 Terms of Address in HKCC

5.1 Terms of Marriage

Before examining the kinship terms in HKCC, let us start with the concept of *marriage*, which is the foundation for family organisation. Besides core terms like 婚姻 *fan1jan1* 'marriage' and 結婚 *git3fan1* 'getting married', we also searched for words describing different stages in the marital journey. These lexical items and their frequencies in HKCC are shown in table 1.[15]

Table 1 Lexical items related to the concept of 'marriage' in HKCC

Term related to marriage*	Meaning	Frequency in HKCC
相睇 *soeng1tai2*	blind date	8
婚姻 *fan1jan1*	marriage	47
拍拖 *paak3to1*	dating	29
求婚 *kau4fan1*	proposal	29
訂婚 *ding6fan1*	engagement	58
結婚 *git3fan1*	getting married	379
離婚 *lei4fan1*	divorce	44
嫁 *gaa3*	marry a man	376
娶 *ceoi2*	take a wife	206

15 Unless stated otherwise, the data of HKCC are based on the second phase, which has about 800,000 Chinese character tokens.

媒人 *mui4jan2*	matchmaker	30

* The frequency also includes items such as 離咗婚 *lei4 zo2 fan1* 'having divorced', which was segmented in HKCC into three tokens: 離, 咗, 婚.

Among the terms associated with marriage, 結婚 *git3fan1* 'get married' has the highest frequency, suggesting that this is one of the major events in movies with plots on romance and familial relationships.

In traditional Chinese families, children's marriage is always arranged by their parents, possibly through a matchmaker and blind-dates. The relevant words 媒人 *mui4jan2* 'matchmaker' and 相睇 *soeng1tai2* 'blind date' appear 30 times and 8 times respectively in HKCC, as shown in table 1 above. This kind of marital arrangement received a lot of criticism as young people tended to bargain for more freedom and autonomy in their own marriage. In the following dialogues, we can see the pre-arrangement of marriage by senior family members.

1. *Your Infinitive Kindness* (恩義難忘, 1965)
   婚姻大事都係由老人家作主好啲嘅
   fan1jan1 daai6si6 dou1hai6 jau4 lou5jan4gaa1 zok3zyu2 hou2 di1 ge2
   'It is better for the elderly to decide on the marriage'.

2. *Love Burst* (難為了嬌妻, 1966)
   婚姻大事係要聽父母之命媒酌之言
   fan1jan1 daai6si6 hai6 jiu3 teng1 fu6mou5 zi1 ming6 mui4zoek3 zi1 jin4
   'Marriage has to be based on parents' order and matchmaker's word'.

We also see how young people feel against the tradition of having marriage arranged by their parents or other senior members such as grandparents in the family. The following dialogue shows an argument between a father and his daughter.

3. *Foster-Daddy's Romantic Affairs* (契爺艷史, 1952)
   Father: 你嘅婚姻事爸爸會同你揸主意㗎。
   nei5 ge3 fan1jan1si6 baa4baa1 wui5 tung6 nei5 zaa1 zyu2ji3 gaa3
   'Daddy will take care of your marriage'.
   Daughter: 爸爸, 婚姻嘅事情我哋自己會理㗎啦。
   baa4baa1, fan1jan1 ge3 si6cing4 ngo5dei6 zi6gei2 wui5 lei5 gaa3laa3
   'Daddy, we can take care of our marriage'.

The following dialogue illustrates how young people feel dissatisfied toward pre-arranged marriage and ask for freedom on the decision of their marriage.

4. *Stubborn Love* (癡兒女, 1943)
   取消呢種封建嘅婚姻制度。
   ceoi2siu1 ni1 zung2 fung1gin3 ge3 fan1jan1 zai3dou6
   'We need to abolish this kind of feudal style of marriage system'.
   而且婚姻要自由呀。
   ji4ce2 fan1jan1 jiu3 zi6jau4 aa3
   'Furthermore, we need to have freedom in marriage'.
   阿媽點都唔能夠強迫我婚姻自由。
   aa3maa1 dim2 dou1 m4 nang4gau3 koeng4bik1 ngo5 fan1jan1 zi6jau4
   'Mother cannot take away my freedom of marriage'.

It is also common for parents (especially those of a daughter) to have business partners as their potential in-laws. There is one proverb in Chinese, namely 門當戶對 *mun4dong1wu6deoi3* 'families of equal rank', advocating for marriage between people with similar backgrounds. In spite of this old-fashioned mindset, there were sometimes parents who were open-minded and willing to allow their children to choose their lifelong partners. Dialogue (5) below is an utterance made by a mother to her daughter, whose marriage was arranged by her father.

5. *When Girls are in Love* (女生外向, 1965)
   Mother: 我時時都唔贊成你爸爸將佢嘅生意
   ngo5 si4si4 dou1 m4 zaan3sing4 nei5 baa4baa1 zoeng1 keoi5 ge3 saang1ji3
   同埋你嘅婚姻拉埋一齊。
   tung4maai4 nei5 ge3 fan1jan1 laai1maai4 jat1cai4
   'I have never agreed with your father in linking his business with your marriage'.

What the above dialogues extracted from HKCC show is that marriage in the old days was not necessarily built upon love and could be arranged by parents without the consent of the children. In a survey conducted by Podmore and Chaney with 1,123 respondents aged between 15 and 30 in the 1970s, 91% indicated that "love was the appropriate basis for marriage" (1974, 403), while 94% of the respondents were "against the idea of arranged marriage" (404). In this connection, it is relevant to examine the verb 娶 *ceoi2* 'to marry a woman' as it can take two different objects: 老婆 *lou5po4* 'wife' and 新抱 *san1pou5* 'daughter-in-law'. The two verb-object phrases capture different perspectives on 'marrying a woman'.[16] The former takes the perspective of the son, while the latter that of the parents. In HKCC, the two phrases have 83 and 14 occurrences in HKCC respectively. Interestingly, among the 83 phrases of 娶老婆 *ceoi2 lou5po4* 'taking

16 It is interesting to note that the verb 嫁 *gaa3* 'to marry a man' does not have such a dual usage. This verb can only be used to mean 'marrying a man to be his wife'.

a wife', 28 contain a prepositional phrase headed by 同 *tung4* 'for', carrying the meaning of *for*. Two examples are given below.

6. *She's so Neat* (彩鳳引金龍, 1957)
   而家同你娶老婆嘅
   ji4gaa1 tung4 nei5 ceoi2 lou5po4 bo3
   'We are now going to take a wife for you'.

7. *Standard Husband* (標準丈夫, 1965)
   你快啲話畀佢聽同佢娶老婆喺咪得囉
   nei5 faai3di1 waa6 bei2 keoi5 teng1 tung4 keoi5 ceoi2 lou5po4 gam2 mai6 dak1 lo1
   'You'd better tell him that we are going to take a wife for him'.

The adjunct phrase headed by 同 *tung4* 'for' shows that the act of taking a wife is not necessarily initiated by the son himself, but by someone in his family, such as parents or even grandparents. For the verb phrase 娶新抱 *ceoi2 san1pou5*, the subject is always the parents, and we do not find the adjunct phrase headed by 同 *tung4* (see the three examples below), which re-affirms that the act of marrying a woman as one's wife could be done sometimes by the family. From example (10), we can even see that in some families, getting a daughter-in-law (i.e. 娶新抱 *ceoi2 san1pou5*) is more important than marrying off the daughter (i.e. 嫁女 *gaa3 neoi5*).

8. *Lovesick* (為情顛倒, 1952)
   我阿媽成日都想娶新抱。
   ngo5 aa3maa1 seng4jat6 dou1 soeng2 ceoi2 san1pou5
   'My mother always wants to get a daughter-in-law'.

9. *The Merry Matrimony* (喜結良緣, 1966)
   阿強媽想快啲娶新抱呀嘛。
   aa3 koeng4 maa1 soeng2 faai3di1 ceoi2 san1pou5 aa1maa3
   'Ah Keung's mother wants to get a daughter-in-law as soon as possible'.

10. *Foster-Daddy's Romantic Affairs* (契爺艷史, 1952)
    噉呀梗係娶咗新抱先至嫁女㗎嘅。
    gam2 aa6 gang2hai6 ceoi2 zo2 san1pou5 sin1zi3 gaa3 neoi5 go3bo3
    'Then, we certainly take a daughter-in-law before we marry off the daughter'.

The above HKCC dialogues containing words related to 'marriage' show the family structure and the arrangement of marriage in mid-20th century Hong Kong. Generally speaking, it was considered a normal practice for someone to get married when they become adults. If the children did not have any intention to form their own families,

their parents would do that for them by all means. In other words, the concept of family is somewhat important in the old days of Hong Kong, as the majority of the population in Hong Kong were Chinese who follow the tradition that men and women form their own families through marriage (Wu 1927; Baker 1979). In the next section, we will examine the kinship terms found in HKCC.

5.2 Terms of Kinship

Since the data in HKCC was only tagged with parts-of-speech, it is not easy to extract kinship terms as a semantic notion directly from HKCC. However, as Qian and Piao (2009) show, there are some unique morphemes referring to kinship. We thus compiled a list of Cantonese kinship morphemes, plotted on a simplified family tree according to the generations they belong to in a traditional Cantonese family **[fig. 1]**.

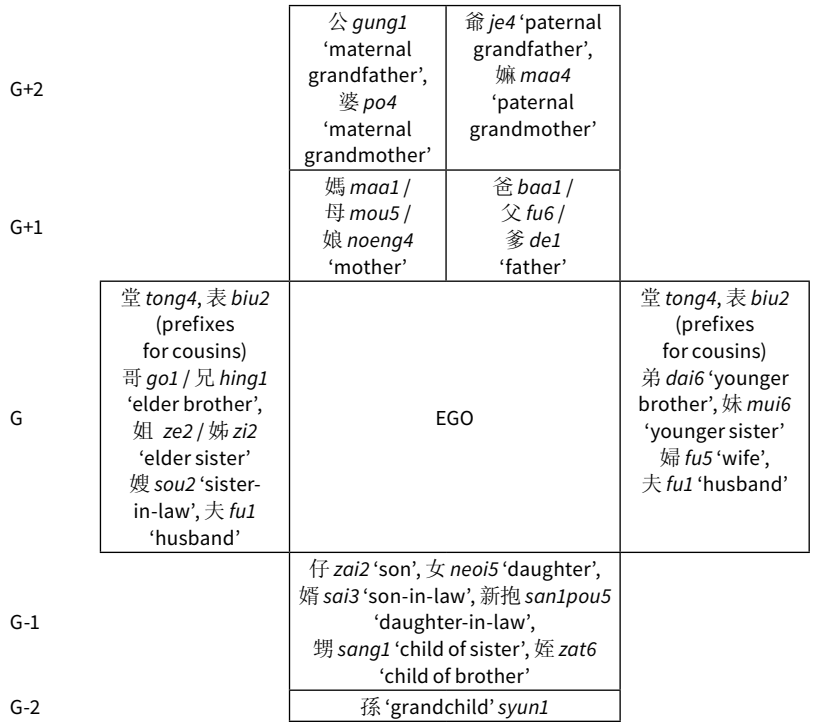


Figure 1 Cantonese kinship morphemes

The above is not an exhaustive list but these morphemes cover the basic kinship that a traditional Hong Kong family might have.[17] With these kinship morphemes, we were able to retrieve about 100 kinship terms from HKCC. Among these 100 items, some are core and common kinship terms such as *father*, *mother*, *brother*, and *sister*, which are listed in table 2.

In addition, there are a few items referring to members of extended families in the grandparents' generation: 叔公 *suk1gung1* 'the younger brother of the paternal grandfather' (i.e. father's paternal uncle); 姑婆 *gu1po4* 'the sister of one's paternal or maternal grandfather' (i.e. father or mother's paternal aunt); 姨婆 *ji4po4* 'the sister of the maternal grandmother' (i.e. mother's maternal aunt). There are also terms that are used by a wife to address the relatives of her husband: 姑奶奶 *gu1naai4naai2* and 舅老爺 *kau5lou5je4*.[18] The former is used to refer to the husband's paternal aunt, while the latter to the husband's maternal uncle. These kinship terms of grandparents' generation demonstrate the scale of the family of old Hong Kong.

Table 2 Kinship terms of core family members and their frequencies in HKCC

Father		**Mother**		**Elder sister**		**Elder brother**	
爸爸 *baa4baa1*	1551	(阿)媽 *aa3maa1*	1080	家姐 *gaa1ze1*	349	大哥 *daai6go1*	175
老豆 *lou5dau6*	122	媽媽 *maa4maa1*	142	(阿)姐 (*aa3*)*ze1*	52	(阿)哥 *aa3go1*	213
(阿)爹 (*aa3*)*de1*	95	媽咪 *maa1mi4*	83	大家姐 *daai6gaa1ze1*	33	哥哥 *go4go1*	35
(阿)爸 (*aa3*)*baa4*	19	老母 *lou5mou5*	48	姐姐 *ze4ze1*	25	二哥 *ji6go1*	17
家父 *gaa1fu6*	10	母親 *mou5can1*	8	大姊 *daai6zi2*	2	大佬 *daai6lou2*	15
父親 *fu5can1*	7	家母 *gaa1mou5*	7	姊姊 *zi2zi2*	1		

Younger sister		**Younger brother**	
(阿)妹 *aa3mui2*	92	細佬 *sai3lou2*	53
妹妹 *mui6mui2*	24		
三妹 *saam1mui2*	19		
二妹 *ji6mui2*	5		
細妹 *sai3mui2*	2		

17 The tree only provides the general meaning of the kinship morphemes. Some of these morphemes can have more than one meaning depending on the kinship terms they form. For example, the morpheme 公 *gung1* is usually understood as 'maternal grandfather', as in the kinship term 公公 *gung1gung1* or 外公 *ngoi6gung1*. However, 公 *gung1* can also appear in the term 老公 *lou5gung1*, meaning 'husband'.

18 The above five kinship terms 叔公 *suk1gung1*, 姑婆 *gu1po4*, 姨婆 *ji4po4*, 姑奶奶 *gu1naai4naai2* and 舅老爺 *kau5lou5je4* appear 2 times, 4 times, 3 times, 10 times, and 4 times respectively in HKCC.

In addition, there are a few items referring to members of extended families in the grandparents' generation: 叔公 *suk1gung1* 'younger brother of paternal grandfather' (i.e. father's paternal uncle); 姑婆 *gu1po4* 'sister of one's paternal or maternal grandfather' (i.e. father or mother's paternal aunt); 姨婆 *ji4po4* 'sister of maternal grandmother' (i.e. mother's maternal aunt). There are also terms that are used by a wife to address the relatives of her husband: 姑奶奶 *gu1naai4naai2* and 舅老爺 *kau5lou5je4*.[19] The former is used to refer to the husband's paternal aunt while the latter the husband's maternal uncle. These kinship terms of grandparents' generation demonstrate the scale of the family of old Hong Kong.

5.3 Other Terms of Address for Family Members

It is common to have more than one item addressing the same person, as shown in table 2 above. Sometimes, the choice among the different items depends on extra-linguistic factors such as solidarity and politeness (Wardhaugh 1992; Gu 1990). Some of these terms are used to show the respect of the addresser towards the addressee, and these terms are usually called honorific terms. In HKCC, there are a number of honorific terms referring to the core family members of the addressee. These honorific forms carry the prefix 令 *ling6*. Interestingly, the kinship terms following the prefix are not the same as the common forms.[20] Table 3 lists the honorific terms and their frequencies in HKCC.

Table 3 Honorific terms in HKCC

Meaning	Honorific term	Frequency in HKCC
your father	令尊翁 *ling6zyun1jung1*	6
	令尊 *ling6zyun1*	15
your mother	令壽堂 *ling6sau6tong2*	6
	令堂 *ling6tong2*	1
your brother	令兄 *ling6hing1*	1
your sister	令妹 *ling6mui2*	1
your son	令郎 *ling6long2*	7
your daughter	令千金 *ling6cin1gam1*	14

19 The above five kinship terms, 叔公 *suk1gung1*, 姑婆 *gu1po4*, 姨婆 *ji4po4*, 姑奶奶 *gu1naai4naai2*, and 舅老爺 *kau5lou5je4*, appear 2 times, 4 times, 3 times, 10 times, and 4 times respectively in HKCC.

20 For example, the honorific form for 'your father' is 令尊 *ling6zyun1* or 令尊翁 *ling6zyun1jung1*, but not 令爸 *ling6baa4*.

These terms are seldom used in modern Cantonese, and only in some very traditional settings.[21]

Another feature of the family structure of mid-20th century Hong Kong society is polygamy. It was quite common for men to take more than one wife, especially when the first wife could not bring any children to the family. There are several terms found in HKCC addressing the concubine or second wife of a man, and the stepmothers.

Table 4 Terms for concubines and stepmothers

Term	Meaning	Frequency
妾侍 *cip3si6*	Concubine	16
細姐 *sai3ze2*	Vocative for father's concubine	51
細婆 *sai3po4*	Vocative for grandfather's concubine	4
後底嫗 *hau6dai2naa2*	Stepmother	2
填房 *tin4fong4*	Stepmother	10

The practice of polygamy ended in 1971 as a result of the changes in the marriage law (Liu 1999; Sullivan 2005; Ip 2014). Therefore, we can see that terms addressing second wives and stepmothers were still quite common in mid-20th century movies.

Many families keep house workers, generally known as servants or maids. As Watson stated, maids were "purchased" (1991, 240), suggesting that the masters were usually wealthy and in the higher socioeconomic class. As for those maids who were bought to the family when they were very young, they were referred to as 妹仔 *mui1zai2* 'little maid'. There were also some servants who helped the mistresses of the family to take care of the children in activities such as breast-feeding. They were called 奶媽 *naai5maa1* 'wet nurse'. Below are some dialogues containing these terms. In dialogue (11), we can see that maids and servants were usually badly treated by the master and his family members.

11. *A Ready Lover* (十月芥菜, 1952)
阿爸爸呀, 你唔好因佢係妹仔睇低佢喎!
aa3 baa4baa1 aa3, nei5 m4hou2 jan1 keoi5 hai6 ***mui1zai2*** *tai2dai1 keoi5 wo3*
'Daddy, you should not look down on her just because she is a little maid'.

21 These terms are not found in HKCanCor, whose data were collected from speakers in their '20s and '30s in 1997 and 1998 (Luke, Wong 2015).

12. *The Joyful Matrimony* (龍鳳合歡花, 1960)
    邊個養大個女嚟做妹仔服侍佢呀, 吓?
    bin1go3 joeng5 daai6 go3 neoi5 lai4 zou6 ***mui1zai2*** *fuk6si6 keoi5 aa3, haa2*
    'Who is willing to raise a daughter to be a little maid to serve him?'

13. *Midnight Werewolf* (夜半人狼, 1963)
    喫, 肚餓叫奶媽攞嘢食啦!
    je1, tou5ngo6 giu3 ***naai5maa1*** *lo5 je5 sik6 laa1*
    'Yeah, if you are hungry, ask wet-nurse for food'.

14. *The Millionaire's Daughter* (千金之女, 1963)
    TOMMY, 奶媽頭先話佢唔精神。
    Tommy, ***naai5maa1*** *tau4sin1 waa6 keoi5 m4 zing1san4*
    'Tommy, wet nurse just said she did not feel well'.

6 Concluding Remarks

In this paper, we made use of the data from *The Corpus of Mid-20th Century Hong Kong Cantonese* to examine how Hong Kong society looked like half a century ago. Our focus was on kinship terms and terms related to marriage. Through these terms, we were able to see the family structure of the old Hong Kong, which was significantly different from contemporary Hong Kong. This could be due to changes in the concept of family and also in the lifestyle, such as working habits. Since the 1970s, Hong Kong people were strongly advised to have serious family planning and many families had only one or two children; this subsequently reduced the size of families.[22] There were no more 'big families' (大家族 *daai6gaa1zuk6*), which led to the reduced use of many kinship terms.[23]

This paper also demonstrates how HKCC can be used to conduct corpus-based sociolinguistic studies in Cantonese which had not been extensively and systematically explored. The corpus data is highly relevant in terms of time (i.e. mid-20th century) and nature (movies with their themes on daily life situations). It is hoped that more corpus-based sociolinguistic studies can be carried out in future with the development of more Cantonese corpora covering a broader variety of language data.

22 Wong discussed how the family planning campaign of Hong Kong in the 1970s challenged "traditional Chinese values in the areas of family size and gender dominance [...] that reshaped society in Hong Kong" (2018, 123).

23 There are some kinship terms showing the traditional big family structure. For example, 舅父仔 *kau5fu2zai2* 'little maternal uncle' is used to refer to the maternal uncle whose age is close or even smaller than the addresser. Other terms include 七妹 *cat1mui2* 'the seventh sister' and 四姨 *sei3ji1* 'the fourth maternal aunt'.

Bibliography

Bacon-Shone, J.; Bolton, K.; Luke, K.K. (2015). *Language Use, Proficiency and Attitudes in Hong Kong*. Hong Kong: Social Sciences Research Centre; University of Hong Kong.

Baker, H.D.R. (1979). *Chinese Family and Kinship*. New York: Columbia University Press.

Baker, P. (2010). *Sociolinguistics and Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Chao, Y.R. (1956). "Chinese Terms of Address". *Language*, 32(1), 217-41. https://doi.org/10.2307/410666.

Cheung, S.H. (1990). "Terms of Address in Cantonese". *Journal of Chinese Linguistics*, 18(1), 1-43. https://www.jstor.org/stable/23767129.

Cheung, S.H. (2001). "The Interrogative Construction. (Re)constructing Early Cantonese Grammar". Chappell, H. (ed.), *Sinitic Grammar. Synchronic and Diachronic Perspectives*. Oxford: Oxford University Press, 191-231.

Chin, A.C. (2011). "Grammaticalization of the Cantonese Double Object Verb [pei35] 畀 in Typological and Areal Perspectives". *Language and Linguistics*, 12(3), 529-63. http://www.ling.sinica.edu.tw/files/publication/j2011_3_02_0726.pdf.

Chin A.C. 錢志安 (2013). "Yueyu yanjiu xin ziyuan. Xianggang ershi shiji zhongqi yueyu yuliaoku" 粵語研究新資源——《香港二十世紀中期粵語語料庫》(New Resources for Cantonese Language Studies. A Linguistic Corpus of Mid-20th Century Hong Kong Cantonese). *Zhongguo yuwen tongxun*, 92(1), 7-16.

Chin, A.C. (2018a). "Discourse Markers in Cantonese". Paper presented at *The 30th North American Conference on Chinese Linguistics* (Columbus, Ohio, 9-11 March 2018). Ohio State University.

Chin, A.C. (2018b). "唔好客氣 vs. 咪走寶. A Corpus-Based Study of Cantonese Prohibitive Markers". Paper presented at *The 18th Workshop on Cantonese* (Hong Kong, 21 April 2018). The Chinese University of Hong Kong.

Chin, A.C. (2019a). "Initiatives of Digital Humanities in Cantonese Studies. A Corpus of Mid-Twentieth-Century Hong Kong Cantonese". Tso, W.B.A. (ed.), *Digital Humanities and New Ways of Teaching*. Singapore: Springer, 71-88.

Chin A.C. 錢志安 (2019b). "Yanbian zhong de yuyan. Yi yueyu zhongxing wenju weili" 演變中的語言——以粵語中性問句為例 (Linguistic Change in Progress. A Case Study of Cantonese Neutral Questions). Paper presented at 海外珍藏漢語文獻與漢語研究高端論壇 *Symposium of Chinese Texts from Overseas and Chinese Linguistic Studies* (Guangzhou, 29 October 2019). San Yat-sen University.

Chin A.C. 錢志安 (forthcoming). "Hanyu fangyan yuliaoku de jiangou he yingyong. Yi Xianggang ershi shiji zhongqi yueyu yuliaoku weili" 漢語方言語料庫的建構和應用——以《香港二十世紀中期粵語語料庫》為例 (The Construction and Application of the Corpus of Chinese Dialects. A Case of the Corpus of Mid-20th Century Hong Kong Cantonese). *Hanyu yuyanxue jikan*.

Chung P.Y. 鍾寶賢 (2004). *Xianggang yingye bai nian* 香港影業百年 (The Movie and Television Industry of Hong Kong over the Past Hundred Years). Hong Kong: Joint Publishing Limited.

CSD. Census and Statistics Department (2016). *Summary Results of 2016 Population By-census*. https://www.bycensus2016.gov.hk.

Davies, M. (2017). "Using Large Online Corpora to Examine Lexical, Semantic, and Cultural Variation in Different Dialects and Time Periods". Friginal, E. (ed.), *Studies in Corpus-Based Sociolinguistics*. London: Routledge, 19-82.

Gu, Y. (1990). "Politeness Phenomena in Modern Chinese". *Journal of Pragmatics*, 14, 237-57. https://doi.org/10.1016/0378-2166(90)90082-o.

Ip, K.Y. (2014). *The Abolishment of Concubinage in Hong Kong. An Analysis of Its Process and Opinions on the Issue* (*1948-1971*) [PhD Dissertation]. Hong Kong: The Chinese University of Hong Kong.

Lai Y.P. 黎奕葆; Chin A.C. 錢志安. (2018). "Yueyu de dongci houzhui 'zhe'" 粵語的動詞後綴'着' (The Cantonese Verbal Suffix *zoek6*). Ho, D.-A. et al. (eds), *Hanyu yu Hanzangyu qianyan yanjiu. Ding Bangxin xiansheng badie shouqing lunwenji*《漢語與漢藏語前沿研究——丁邦新先生八秩壽慶論文集》(Frontiers in Sinitic and Sino-Tibetan Linguistics: Studies in the Languages of China: Festschrift in Honour of Professor Ting Pang-Hsin on His 80th Birthday). Beijing: Social Sciences Academic Press, 697-710.

Law, K. (1986). "Archetype and Variations". *Cantonese Melodrama 1950-1969*. Hong Kong: The Urban Council, 10-20.

Leung, M.; Law, S.; Fung, S. (2004). "Type and Token Frequencies of Phonological Units in Hong Kong Cantonese". *Behavior Research Methods, Instruments, & Computers*, 36(3), 500-5. https://doi.org/10.3758/bf03195596.

Liu, A.N.C. (1999). *Family Law for the Hong Kong SAR*. Hong Kong: Hong Kong University Press.

Lui, T.L. (1988). "Home at Hongkong". *Changes in Hong Kong Society through Cinema*. Hong Kong: The Urban Council of Hong Kong, 83-92.

Luke, K.K.; Lau, C. (2008). "On Loanword Truncation in Cantonese". *Journal of East Asian Linguistics*, 17, 347-62. https://doi.org/10.1007/s10831-008-9032-x.

Luke, K.K.; Wong, M.L.Y. (2015). "The Hong Kong Cantonese Corpus. Design and Uses", in Tsou, B.; Kwong, O.O. (eds), "Linguistic Corpus and Corpus Linguistics in the Chinese Context", monogr. no., *Journal of Chinese Linguistics*, 25, 312-33. https://www.jstor.org/stable/26455290.

McCoy, J. (1970). "Chinese Kin Terms of Reference and Address". Freedman, M. (ed.), *Family and Kinship in Chinese Society.* Stanford: Stanford University Press, 209-26.

McEnery, T.; Hardie, A. (2012). *Corpus Linguistics*. Cambridge: Cambridge University Press.

Peng, G. (2006). "Temporal and Tonal Aspects of Chinese Syllables. A Corpus-Based Comparative Study of Mandarin and Cantonese". *Journal of Chinese Linguistics*, 34(1), 134-54. https://www.jstor.org/stable/23754151.

Podmore, D.; Chaney, D. (1974). "Family Norms in a Rapidly Industrializing Society. Hong Kong". *Journal of Marriage and Family*, 36(2), 400-7. https://doi.org/10.2307/351167.

Qian, Y.; Piao, S. (2009). "The Development of a Semantic Annotation Scheme for Chinese Kinship". *Corpora*, 4(2), 189-208. https://doi.org/10.3366/e1749503209000306.

Sullivan, P.L. (2005). "Culture, Divorce, and Family Mediation in Hong Kong". *Family Court Review*, 43(1), 109-23. https://doi.org/10.1111/j.1744-1617.2005.00011.x.

Tse C.M. 謝明燊; Chin, A.C. 錢志安 (2015). "Yueyu 'ming-liang-ming' jiegou de tongzhi yongfa" 粵語「名-量-名」結構的同指用法 (The Co-Referential Usage of "Noun-Classifier-Noun" in Cantonese). Paper presented at *The Fifteenth*

LSHK Workshop on Cantonese (Hong Kong, 11 April 2015). The University of Hong Kong. https://www.jstor.org/stable/23756692.

Tsou B. 鄒嘉彥 (1997). "San yan, liang yu shuo Xianggang" 三言兩語說香港 (Three Spoken Languages and Two Written Languages in Hong Kong). *Journal of Chinese Linguistics*, 25(2), 290-307.

Wardhaugh, R. (1992). *An Introduction to Sociolinguistics*. Oxford: Blackwell.

Watson, R. (1991). "Wives, Concubines, and Maids. Servitude and Kinship in the Hong Kong Region, 1900-1940". Watson, R.S.; Ebrey, P.B. (eds), *Marriage and Inequality in Chinese Society*. Berkeley; Los Angeles: University of California Press, 231-55.

Wong, M.L.Y. (2009). "*Gei* Constructions in Mandarin Chinese and *Bei* Constructions in Cantonese. A Corpus-Driven Contrastive Study". *International Journal of Corpus Linguistics*, 14(1), 60-80. https://doi.org/10.1075/ijcl.14.1.04won.

Wong, W.S. (2018). "Reconfiguring a New Tradition of Ideal Family Size. A Case Study of the Family Planning Association of Hong Kong, 1977-1982". Wong, W.S. (ed.), *The Disappearance of Hong Kong in Comics. Advertising and Graphic Design*. Cham: Palgrave Macmillan, 123-40. https://doi.org/10.1007/978-3-319-92096-2_6.

Wu, C.-C. (1927). "The Chinese Family. Organization, Names, and Kinship Terms". *American Anthropologist*, 29(3), 316-25. https://doi.org/10.1525/aa.1927.29.3.02a00100.

Corpus and Database Building

Corpus-Based Research on Chinese Language and Linguistics
edited by Bianca Basciano, Franco Gatti, Anna Morbiato

Form and Meaning Representation of Chinese Constructions
Fundamental Issues on Constructicography

Weidong Zhan (Peking University, China)

Jiajun Wang (Peking University, China)

Long Chen (Peking University, China)

Haibin Huang (Peking University, China)

Abstract This paper introduces a Chinese constructicon (CCL-CxnBank) and a corpus annotation platform for the description of actual usages of constructions in contexts. CCL-CxnBank is an online repository that contains more than 1,000 constructions, as well as the linguistic descriptions of their various features. Based on our practice of constructicography, we hold that constructions differ from phrases in that they are not recursive. We propose that the formal representation of a given construction should be linear, while its meaning should be represented through paraphrase templates and semantic frames. In the future, contextual features will be integrated to analyse the semantics of constructions.

Keywords Chinese constructicon. Constructicography. Construction grammar. Form and meaning representation. Principle of compositionality. Language engineering.

Edizioni Ca'Foscari

Sinica venetiana 6
e-ISSN 2610-9042 | ISSN 2610-9654
ISBN [ebook] 978-88-6969-406-6 | ISBN [print] 978-88-6969-407-3

Peer review | Open access **307**
Submitted 2020-10-02 | Accepted 2020-11-30 | Published 2020-12-21

DOI 10.30687/978-88-6969-406-6/010

1 Introduction

This paper introduces the work on knowledge representation of Chinese constructions done in recent years by the Centre for Chinese Linguistics (CCL) of Peking University. Our work includes two parts: the development of a Chinese constructicon (provisionally named as CCL-CxnBank)[1] and the annotation of a corpus consisting of sentences that display various usages of construction instances.[2] Our work stems from the belief that linguistic knowledge resources can better support natural language processing and language teaching if they are well organised, analysed, and digitised into databases and annotated corpora.

In the past 30 years, the construction approach to language has thrived among Chinese linguistic studies and has brought rich knowledge to both case studies and systematic studies (Zhang B. 2008, 2018; Zhang J. 2013). Against this background, since 2015 CCL has been running a project on the development of a Chinese constructicon database, which is the first Chinese constructicon project comprising both a construction knowledge database and an annotated corpus. CCL-CxnBank serves as a supplement to the current natural language engineering practice that in mainstream computational linguistics is based on commonly-used grammatical units, such as words and phrases. Up to now, this project has already collected over 1,000 Chinese constructions and recorded their syntactic, semantic, and pragmatic information. Moreover, relationships among constructions, such as synonymy, antonymy, and hyponymy/hyperonymy relations, have also been included, in order to provide a more systematic and coherent knowledge representation scheme for Chinese constructions. Finally, an online corpus annotation platform has been developed to annotate the internal structure and the subjective attitude meaning of each construct that occurs in real texts, with the aim of providing a comprehensive description of the actual usages of constructions in real contexts.[3]

This paper presents our work in progress and some of the major challenges we encountered in the development of CCL-CxnBank. § 2 presents our definition and understanding of the term 'construction' by comparing it with the conventional grammatical unit notion of 'phrase', which is commonly used to refer to a formal representation

1 The website of CCL-CxnBank is http://ccl.pku.edu.cn/ccgd.

2 We have also set up a website as a working platform for annotating the corpus, which is currently only accessible to authorised annotators. The website is http://162.105.161.162:8088/cclannotator/public/index.php.

3 'Construction' and 'construct' in this paper are used to refer to construction type and token respectively. 'Constructicon' refers to the construction database in which construction entries and their linguistic attributes are systematically organised and recorded.

scheme in syntactic structures in the knowledge engineering practices for computer. § 3 discusses issues in the representation of the forms and meanings of constructions. § 4 gives an overview of CCL-CxnBank and discusses the methodology adopted in its development. § 5 presents our work on corpus annotation, including an introduction of the online platform for annotation and some related challenges. The last section concludes by presenting the significance of our work and the future direction of development of construction resources.

2 The Properties of Constructions. Comparing Constructions with Phrases

From the viewpoint of language resources development, Zhan (2017) analysed the relationship and differences existing between constructions and conventional grammatical units, i.e. words, phrases etc. This work adopts Zhan's (2017) perspective: below, we discuss some major tenets and propose some further considerations.

Unlike some constructionists who maintain that all units of a grammatical system are constructions (Croft 2001), we treat constructions as complements to common phrases: in our view, constructions complement words and phrases rather than totally replacing them.[4] This is based on our understanding of constructions and conventional language units. Conventional language units can be classified into words and phrases. Words have fixed internal structures and cannot be recursively composed of smaller grammatical units. Phrases have expandable internal structures and can be recursively composed of smaller phrases. This classification allows greater efficiency and convenience in developing and maintaining language resource databases. In a language resource database, a limited (but large) number of words are listed entry by entry, while an infinite number of phrases can be described with a finite number of syntactic rules based on a finite number of grammatical categories such as noun, verb, noun phrase, verb phrase etc. However, in a linguistic system, other types of linguistic units can be identified (that we call 'constructions', Zhan 2017), which differ in the following respects.

First, constructions emerge from common phrases, which are formed by words. Therefore, constructions are different from words,

4 Treating words as constructions is merely a theoretical or labelling issue. Words *can* be treated as constructions from a 'form-meaning' pair perspective, but it makes little difference in the knowledge engineering practice. For languages with little or no inflection such as Chinese, knowledge in a dictionary is stored in exactly the same way as in constructions' description: each entry is a 'word form-word meaning' pair. In other words, referring to words as 'word constructions' or 'words' makes no difference in the knowledge engineering practice.

which are not composed of smaller grammar units. From the point of view of formal grammar, a word can even be regarded as the smallest grammatical unit or atomic unit and there is no need to analyse its internal components.

Second, constructions are different from phrases. In traditional linguistics, phrases are treated as core grammatical units. The formalisation of phrases includes four elements: relationships, heads, categories and hierarchies. These four elements jointly display a syntagmatic and recursive nature within phrases: (1) the syntagmatic relations between constituents within phrases, (2) the head roles in the phrases, (3) the grammatical categories the phrases and their constituents belong to, and (4) the hierarchical (tree) structures in which the phrases are internally organised. The syntactic description of these four aspects is the foundation for the computation of the meaning of phrases (Jurafsky, Martin 2000, chs. 15.1, 15.2). On the contrary, typical constructions have weak relationships between the constituents, no prominent head roles, only limited variations in their de-categorised components, and a linear internal structure rather than a hierarchical one. From the perspective of meaning, the acquisition of the meanings of phrases generally follows the so-called 'principle of compositionality', stating that the meaning of a whole sentence is acquired by the semantic combination of its constituent parts (Partee 2004). As for constructions, the meaning of a construct is the combination of the meanings of its constituents and the meaning of the construction in which these words occur. Therefore, constructions are not conventional phrases.

Third, we can either refer to constructions as phrases or refer to phrases as constructions (Croft 2001). If we refer to constructions as phrases, constructions are unique phrases; if we refer to phrases as constructions, phrases are schematised constructions (Zhan 2017). It is theoretically reasonable to refer to phrases as constructions; however, categorising them as the same grammatical unit does not mean they have identical grammatical properties. Constructions and conventional phrases still differ in many basic grammatical properties such as recursiveness and compositionality. For example, constructions can usually be embedded in conventional phrases, while only a limited number of phrases can be embedded into constructions. Example (1) illustrates two sentences with the same pattern: [不是 *búshì* + N_1 + 的 *de* + N_2].[5] N_1 differs from N_2 in (1a), while in (1b) N_1 and N_2

5 The glosses follow the general guidelines of the Leipzig Glossing Rules. Additional glosses include: BEI = 'Chinese 被 *bèi* marker', often labelled as a passive marker; DE = 'Chinese particle 的 *de*', functioning as modification marker or nominaliser; MP = 'mood particle' (in Chinese they are used to add various moods, including interrogation, request, command, emphasis and exclamation, to an utterance); SFP = 'sentence final particle'. In-text abbreviations are as follows: N = 'noun'; NP = 'noun phrase'; V = 'verb';

are identical (repetition of nouns with the same form): (1b) includes a construct of the construction [不是 *búshì* + N + 的 *de* + N], meaning 'N that is not N'.

1. a. 怎么解决这不是正品的问题?
zěnme jiějué zhè bú shì zhèng-pǐn de wèntí
how solve this not COP genuine-product DE problem
'How to solve the problem that this commodity is not genuine?'

   b. 怎么解决这不是问题的问题?
zěnme jiějué zhè bú shì wèntí de wèntí
how solve this not COP problem DE problem
'How to solve this problem which is not a problem?'

By comparing the examples above, it is obvious that the instance of the linear pattern [不是 + N1 + 的 de + N2] in (1a) has a different internal hierarchical structure, which can be expanded into a different form. (2) is the expansion of (1a), which maintains the original hierarchical structure.

2. 怎么解决这个商品不是厂家正品的严重失信问题?
zěnme jiějué zhè ge shāngpǐn bú shì chǎngjiā
how solve this CLF commodity not COP manufacturer
zhèng-pǐn de yánzhòng shīxìn wèntí
genuine-product DE serious dishonesty problem
'How to solve the problem that this commodity is not a genuine product of the manufacturer, which indicates a serious dishonest conduct?'

However, the instance of the pattern [不是 *bú shì* + N + 的 *de* + N] 'N that is not N' in (1b) cannot be expanded as that in (1a). [不是 *bú shì* + 问题 *wèntí* + 的 *de* + 问题 *wèntí*] 'a problem which is not a problem' is a fixed language unit: 问题 *wèntí* 'problem' can only be substituted with a limited number of nouns such as 办法 *bànfǎ* 'method', 理由 *lǐyóu* 'reason', 机会 *jīhuì* 'chance', 结局 *jiéjú* 'outcome', 妈妈 *māma* 'mother' etc. The generative capacity of this pattern is limited if compared with that of phrase patterns shown in example (1a) and (2). Furthermore, it carries an additional inherent meaning that goes beyond the meaning of 不是 *bú shì* and 问题 *wèntí*, which could be paraphrased as 'it is only a titular N' or 'it is not a typical N, but, nonetheless, we can grudgingly treat it as one' etc. The specific meaning is determined by the context in which the pattern occurs.

VP = 'verb phrase'; A = 'adjective'; AP = 'adjective phrase'; CLP = 'numeral plus classifier phrase'; X, Y,... = 'constituents with arbitrary syntactic category'.

Above all, constructions are different from conventional grammatical units, i.e. words and phrases, in a major respect: in language engineering, mapping between forms and meanings of constructions need to be listed entry by entry, just like those of words; combinatorial properties of constructions, on the other hand, need to be described like those of phrases.

3 The Form and Meaning Representation of Constructions

According to Zhan (2017) and following the considerations above, the forms of constructions should be described as linear patterns with specific lexical elements (which we call 'constants') and schematic elements (which we call 'variables'). Within a construction, constants are specific words, and variables are represented by part-of-speech tags or syntactic categories of phrases (N, V, NP, VP etc.). Some variables in certain constructions can be instantiated with elements of different phrase categories, which is mentioned above as 'de-categorisation'. The following examples illustrate the variables instantiated by word categories, phrase categories and cross-category elements.

Table 1 Some examples of constructions combined with constants and variables

Constructions	Constructs	Constants (Words)	Variables (Categories)
V + 一 *yī* 'one' + CLF + 是 *shì* 'COP' + 一 *yī* 'one' + CLF 'Every behaviour which V indicates counts'	说一句是一句 *shuō yī jù shì yī jù* 'Every word counts'	一 *yī* 'one' 是 *shì* 'COP'	V, CLF
说 *shuō* 'speak' + VP + 就 *jiù* 'immediately' + VP 'Carry out the behaviour indicated by VP immediately after promising to VP'	说给钱就给钱 *shuō gěi qián jiù gěi qián* 'Pay immediately after promising to pay'	说 *shuō* 'speak' 就 *jiù* 'immediately'	VP
除了*chúle* 'besides' + X + 还是 *háishì* 'still' + X 'There is nothing but X'	除了下雨还是下雨 *chúle xià yǔ háishì xià yǔ* 'It rains endlessly' 除了馒头还是馒头 *chúle mántou háishì mántou* 'There is nothing to eat but steamed buns'	除了 *chúle* 'besides' 还是 *háishì* 'still'	X (N, V, A etc.)

Constructions share semantic properties both with words and with phrases. On the one hand, the meanings of constructions have to be listed entry by entry just like words, in order to describe fixed relations between form and meaning. On the other hand, the meaning of constructions has to be computed by combining the meanings of the constituents following the 'principle of compositionality', just like phrases. The following two sections present and discuss issues in the representation of construction forms and meanings.

3.1 The Representation of Forms. Variations and Extensions of Constructs in Actual Use

The internal structure of a construction is represented as a linear pattern consisting of several constants and variables. While it is generally not necessary to consider recursiveness in the structural representation of a construction (typically, a construct cannot be embedded into a construct of the same construction), some constructions display a limited expansion capacity. Zhan (2017) analysed the basic forms of constructions, which are considered to be stable and fixed. Here we further discuss the form variations of constructions, which can be distinguished into three types.

3.1.1 The Variation of Lexically Specified Elements of a Construction

Let us consider the following examples:

3. a. 有什么大惊小怪的
yǒu shénme dàjīngxiǎoguài de
have what fuss DE
'There is nothing to fuss about'

b. 有什么可大惊小怪的
yǒu shénme kě dàjīngxiǎoguài de
have what may fuss DE
'There is nothing to fuss about'

c. 有什么好大惊小怪
yǒu shénme hǎo dàjīngxiǎoguài
have what worth fuss
'There is nothing to fuss about'

d. 有什么好大惊小怪的
yǒu shénme hǎo dàjīngxiǎoguài de
have what worth fuss DE
'There is nothing to fuss about'

a'. 没有什么大惊小怪的
méi yǒu shénme dàjīngxiǎoguài de
not have what fuss DE
'There is nothing to fuss about'

b'. 没有什么可大惊小怪的
méi yǒu shénme kě dàjīngxiǎoguài de
not have what may may fuss DE
'There is nothing to fuss about'
没有什么好大惊小怪

c'. *méi yǒu shénme hǎo dàjīngxiǎoguài*
not have what worth fuss
'There is nothing to fuss about'

d'. 没有什么好大惊小怪的
méi yǒu shénme hǎo dàjīngxiǎoguài de
not have what worth fuss DE
'There is nothing to fuss about'

The form of the construction in example (3) is [有 *yǒu* + 什么 *shénme* + VP + 的 *de*] 'there is no need to VP', as in (3a). (3b)-(3d) are variations of this construction with other constants added, such as 可 *kě* 'may', 好 *hǎo* 'worth', or with the constant 的 *de* omitted. 有 *yǒu* 'have' in these constructs may also appear in its negated form, 没有 *méi yǒu*, as in (3a')-(3d'), meaning 'there is no need to VP', 'it is worthless to VP' etc. The variations of a construction form can be either exhaustively listed in the constructicon or captured by regular expressions. The construction form in example (3) can be represented as [(有 *yǒu* | 没有*méi yǒu*) 什么 *shénme* (好*hǎo* | 可 *kě*)? (VP) (的*de*)?], where '?' indicates zero or one leftward character, and '|' indicates disjunction, matching either left or right character. Regular expressions can be represented by the finite state transition network (FSTN). The FSTN of the construction in example (3) is illustrated in figure 1 below (Chomsky 1956).

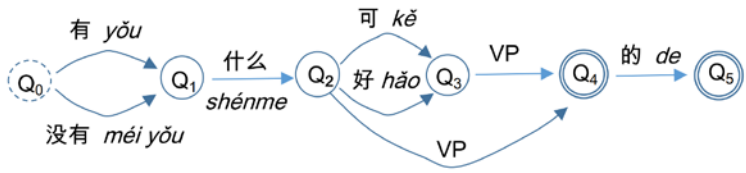


Figure 1 The FSTN recognising form variations of the construction [有 *yǒu* + 什么 *shénme* + VP + 的 *de*]

Among the 1,066 entries in CCL-CxnBank, 816 are marked as not having form variations, and 250 entries are marked as having some (about 23.45%). Constructions vary both in the number and the degree of form variations. The basic form of the construction in example (4) is [A + 就 *jiù* 'exactly' + A + 在 *zài* 'on' + X] 'it is indeed X which makes it A', with more complicated instantiations than example (3): (4a) is an instance that can match the construction form exactly; in (4b), the auxiliary 可能 *kěnéng* 'may' is inserted before 就 *jiù* 'indeed' as a constant of the construction; in (4c) and (4d), 就 *jiù* 'indeed' is replaced by 就是 *jiùshì* 'exactly' and 也就 *yě jiù* 'also exactly', respectively. Besides, in (4c) and (4d), the first variable is separated from the rest by a comma.

3. a. 个人申请贷款, 麻烦就麻烦在担保和抵押。
gèrén shēnqǐng dàikuǎn máfan jiù máfan
individual apply.for loan troublesome indeed troublesome
zài dānbǎo hé dǐyā
on guarantee and mortgage
'The troubles of individual application of loans lie exactly on guarantees and mortgages'.

b. 他的主子大人将来倒霉可能就倒霉在狗的身上。
tā de zhǔzi dàrén jiānglái dǎoméi kěnéng jiù
he DE master lord future unfortunate may indeed
dǎoméi zài gǒu de shēn shàng
unfortunate on dog DE body on
'Something unfortunate may happen to his lord master exactly because of the dog'.

c. 很多学生觉得文言文难, 就是难在一些实词和虚词上。
hěnduō xuéshēng juéde wényánwén nán jiùshì
many student think Classical.Chinese difficult exactly
nán zài yìxiē shící hé xūcí shàng
difficult on some content.word and function.word above
'Many students think that the difficulties of Classical Chinese lie exactly on some content words and function words'.

d. 处方的'含金量'高, 也就高在用进口药和合资企业药的比重猛增。
chùfāng de hánjīnliàng gāo yě jiù gāo
prescription DE gold.content high also indeed high
zài yòng jìnkǒu yào hé hézī qǐyè
on use imported medicine and joint.venture enterprise
yào de bǐzhòng měngzēng
medicine DE ratio soar
'The 'true value' (price) of the medical prescriptions is high exactly because of the soaring of the ratio of the medicines used, which are produced by foreign and joint venture enterprises'.

The form variations in (4a) and (4b) are complete grammatical units, while in (4c-d) the construction variations may not be grammatical constituents. In (4c), 难, 就是难在一些实词和虚词上 *nán jiùshì nán zài yìxiē shící hé xūcí shàng* 'difficulties lie on some content words and function words' can be treated either as a complete constituent or as two clauses separated by a comma, with each clause acting as a constituent. Thus, (4c) is no longer appropriate to be treated as a construct instantiated from the form variation of the construction [A + 就 *jiù* + A + 在 *zài* + X], at least not the same as that instantiated by examples (4a) and (4b), even though they almost share the same meaning.

This construction has even more form variations, such as (4a') and (4c') below, which are expanded from (4a) and (4c).

4. a'. 个人申请贷款的麻烦, 最主要就麻烦在担保和抵押。
gèrén shēnqǐng dàikuǎn de ***máfan*** *zuì zhǔyào*
individual apply.for loan DE trouble most main
jiù máfan zài dānbǎo hé dǐyā
indeed troublesome on guarantee and mortgage
'The troubles of individual application of loans lie exactly on guarantees and mortgages'.

c'. 文言文难, 很多学生觉得就是难在一些实词和虚词上。
wényánwén ***nán*** *hěnduō xuéshēng juéde* ***jiù shì***
Classical.Chinese difficult many student think indeed COP
nán zài yìxiē shící hé xūcí shàng
difficult on some content.word and function.word above
'The difficulties of Classical Chinese, some students think, lie exactly on some content words and function words'.

In (4a') and (4c'), if the bold parts are treated as the form variations of the construction [A + 就 *jiù* + A + 在 *zài* + X], some problems arise when trying to represent the form of the construction variations, because regular expressions will capture chunks with no linguistic significance when trying to match the constructs in the sentences. The chunks 最主要 *zuì zhǔyào* 'the most important' in (4a') and 很多学生觉得 *hěnduō xuéshēng juéde* 'many students think that...' in (4c') appear between a constant and a variable. A module needs to be specifically designed to handle these strings appropriately.

Examples in (3) and (4) show that, while the constants in [有 *yǒu* + 什么 *shénme* + VP + 的 *de*] have limited form variations which can be captured rather precisely and exhaustively by regular expressions, the relation between the first variable 'A' and the constant 就 *jiù* in [A + 就 *jiù* + A + 在 *zài* + X] is relatively loose. In real texts, language chunks of various categories can be inserted between the constant and the variable in the constructs, displaying great variability. Although these constructs express the same basic meaning, their forms cannot be exhaustively and appropriately described. The internal structure of the construction in (4) requires further examination. In other words, the construction [A + 就 *jiù* + A + 在 *zài* + X] is not a monolithic whole. The chunk responsible for the explanation is [A + 在 *zài* + X], occurring after 就 *jiù*. [就 *jiù* + A + 在 *zài* + X] is a relatively independent chunk, which can be used separately from the preceding variable 'A', as in (4a') and (4c'). It would be more reasonable to include [A + 在 *zài* + X] as a separate construction entry, specifying that it is a synonym of [A + 就 *jiù* + A + 在 *zài* + X]. When processing sentences with such constructs, the construct [A + 就 *jiù* + A + 在 *zài* + X] has the priority over the others, according to

the greedy matching principle. If [A + 就 *jiù* + A + 在 *zài* + X] fails to match any construct, [A + 在 *zài* + X] will be called in for matching.[6]

3.1.2 Expansion by Juxtaposition of Constructions in the Form of Chunks

Let us consider the following examples:

5. 新就新在财政部门认真对待全国人大、政协、‘两会’代表、委员的意见上，新在他们转变作风、行动迅速上。
xīn jiù xīn zài cáizhèng bùmén rènzhēn duìdài
new exactly new on financial department seriously treat
quánguóréndà zhèngxié liǎng-huì dàibiǎo
N.P.C. N.P.P.C.C. Two-Sessions representatives
wěiyuán de yìjiàn shàng xīn zài tāmen zhuǎnbiàn
committee DE advice on new be.at they change
zuòfēng xíngdòng xùnsù shàng
style act rapidly above
‘The novelty lies in the fact that the Financial Department took very seriously the advice given by N.P.C., N.P.P.C.C., and the representatives and committees of the Two Sessions, regarding their change in working style and action speed’.

6. 一整天在湖上晃呀晃、拐呀拐的也是一种度日方式吧。
yì zhěng tiān zài hú shàng huàng ya huàng
one whole day on lake on waggle MP waggle
ya guǎi de yě shì yì zhǒng dù rì fāngshì ba
MP turn DE also COP one CLF spend day way MP
‘Waggling and turning around all day long on the lake is also a way to spend the day’.

7. 喜剧不是喜剧, 闹剧不是闹剧, 丑角不是丑角, 痞子不是痞子, 简直滑稽至极。
xǐjù bú shì xǐjù nàojù bú shì nàojù
comedy not COP comedy farce not COP farce
chǒujué bú shì chǒujué pǐzi bú shì pǐzi
clown not COP clown ruffian not COP ruffian
jiǎnzhí huájī zhì jí
simply ridiculous to utmost

6 Another method is to treat examples (4a’) and (4c’) as separable usages of a construction, which requires form matching of a discontinuous string, thus making the matching process more complicated.

'It is neither a comedy nor a farce, the clown is not a clown and the ruffian is not a ruffian: it's ridiculous'.

The basic form of the construct in (5) is [A + 就 *jiù* + A + 在 *zài* + X] (same as in example (4)), with [A + 在 *zài* + X] partially expanding, appearing twice in the sentence. Similarly, the pattern [V + 呀 *ya* + V] 'V again and again', whose basic form is [V + 呀 *ya* + V + 的 *de*] 'V-*ing* again and again', expands and appears twice in (6). The basic form of the construction in (7) is [N_1 + 不是 *bú shì* N_1, N_2 + 不是 *bú shì* + N_2] 'it is neither N_1 nor N_2', which already includes two juxtaposed chunks. In (7), the whole construct expands, differently from (5) and (6), where the constructs only partially expand.

The 'Expandable' (是否可扩展 *shìfǒu kě kuòzhǎn*) feature in CCL-CxnBank is used to describe the constructs illustrated above. Its default value is 'true', which allows expansion by juxtaposition. For constructions which cannot expand juxtapositionally, the value will be 'false'.

8. a. 一个不留神, 摔了个大跟头。
yí ge bù liúshén shuāi le ge dà gēntou
one CLF no caution fall PFV CLF big somersault
'Without caution, (someone) fell heavily'.

b. 一个愿打, 一个愿挨。
yí ge yuàn dǎ yí ge yuàn āi
one CLF willing beat one CLF willing endure
'One is willing to beat, the other is willing to be beaten'.

c. 一个使劲骂一个偷东西的孩子, 还有一个 […]
yí ge shǐjìn mà yí ge tōu dōngxi de
one CLF continuously scold one CLF steal thing DE
háizi hái yǒu yí ge
child also have one CLF
'One keeps on scolding a child who steals, the other […]'

一个不留神 *yí ge bù liúshén* in (8a) is an instantiation of the construction [一 *yí* + 个 *ge* + VP] 'one moment of VP (leads to)...', which is also shared by instantiations such as [一 *yí* 'one' 个 *ge* 'CLF' 没 *méi* 'not' 站稳 *zhàn-wěn* 'stand-steady'] 'one moment of instability...', [一 *yí* 'one' 个 *ge* 'CLF' 手 *shǒu* 'hand' 软 *ruǎn* 'soft'] 'one moment of loosened grip...' etc., all conveying the happening of unexpected events which bring about undesirable results. However, although the sentences in (8b) and (8c) formally display the [一 *yí* + 个 *ge* + VP] pattern, they are not instantiations of this construction. Rather, 一个 *yí ge* 'one' acts as the subject (with the head noun omitted) of the following predicate. In (8c), there is also a second 一个 *yí ge*, which is

part of the modifier of the NP's head noun 孩子 *háizi* 'child', together with the relative clause 偷东西的 *tōu dōngxi de* 'who steals', altogether meaning 'the child who steals'. In CCL-CxnBank, the 'Expandable' feature of [一 *yí* + 个 *ge* + VP] is thus set to 'false', therefore preventing the chunks like those in (8b) and (8c) from being recognised as constructs of the construction [一 *yí* + 个 *ge* + VP] in automatic syntactic parsing.

3.1.3 Schematic Elements (Variables) Which May not Form a Constituent as a Whole

Let us consider the following examples:

9. a. 这批货要多少有多少。
zhè pī huò yào duōshǎo yǒu duōshǎo
this CLF goods require how.many have how.many
'As for this batch of goods, you can have as many as you need'.

b. 接下来不作解释了, 能理解多少理解多少。
jiēxiàlái bú zuò jiěshì le néng lǐjiě duōshǎo
next not conduct explain SF can comprehend how.much
lǐjiě duōshǎo
comprehend how.much
'(I) shall explain no more. Try to comprehend as much as (you) can'.

c. 观众爱给多少给多少, 不给也无妨。
guānzhòng ài gěi duōshǎo gěi duōshǎo
audience like give how.much give how.much
bù gěi yě wúfáng
not give also acceptable
'Audience may give as much as they like, even nothing'.

d. 有多少根发梢便会传递多少缕柔情蜜意。
yǒu duōshǎo gēn fà-shāo biàn huì chuándì
have how.many CLF hair-end therefore can convey
duōshǎo lǚ róuqíng-mìyì
how.many CLF tender-affection
'Men will be fascinated by her thick hair'.

(9a) displays the construction [V_1 + 多少 *duōshǎo* + V_2 + 多少 *duōshǎo*] 'the amount of V_1 leads to the same amount of V_2'. Chunks with similar patterns also appear in (9b)-(9d), which convey similar meanings, indicating that the quantity involved in the latter event is dependent on the quantity involved in the former one. However, chunks in (9b-d) cannot be treated as true instantiations of the construction

[V_1 + 多少 *duōshǎo* + V_2 + 多少 *duōshǎo*] 'the amount of V_1 leads to the same amount of V_2', in that the chunks after 多少 *duōshǎo* 'how many', such as [有 *yǒu* 'have'... 根 *gēn* 'CLF' 发梢 *fàshāo* 'hair end'] in (9d), do not form complete constituents. To account for this, sentences in (9) can be first treated with common phrase structure rules. Each sentence consists of two juxtaposed phrase structures and interrogative chunks with the same form generally occur at the same syntactic position. The whole structure expresses a dependency correlation, which can be instantiated by any number of event pairs with conditional relation. The quantity included in the second event corresponds to the quantity included in the first event.

Similar phenomena are more common in compound sentences. Take the construction [再 *zài* 'again' + VP_1 + 也 *yě* 'still' + VP_2] 'no matter how much one VP_1, VP_2 still occurs' as an example. Simple constructs can be decomposed into the constants 再 *zài* and 也 *yě*, and two predicative variables. However, for more complicated constructs, the pattern […再 *zài*…也 *yě*…] establishes a long-distance relation which connects two clauses, as happens in (10):

10. 你奉献得再多, 那些人也觉得不够

nǐ	*fèngxiàn*	*de*	*zài*	*duō*	*nàxiē*
you	give	COMP	again	much	those
rén	*yě*	*juéde*	*bú*	*gòu*	
people	still	think	not	enough	

'No matter how much you give, they will always think it is not enough'.

The meaning of the whole sentence can be decomposed into the basic propositional meanings of the two clauses with an adversative relation, which is represented by the two function words 再 *zài* and 也 *yě*. Describing the adversative relation using the linear pattern [再*zài* 'again' + VP_1 + 也 *yě* 'still' + VP_2] 'no matter how much one VP_1, VP_2 still occurs' is an over-simplification. In fact, the variables between 再 *zài* and 也 *yě* may not form a constituent, but separately belong to the two clauses as shown in example (10). In addition, the constant 也 *yě* can be replaced by other tokens, such as 都 *dōu*, 总 *zǒng*, 还 *hái* etc. (all roughly with the meaning 'still', when used here).

Constructions such as those in (9) and (10) require similar analyses: they are first processed using phrase structure rules and then marked as constructs with specific relations according to construction evoking elements such as [再 *zài*…也…*yě*], [多少 *duōshǎo*…多少 *duōshǎo*] etc.

3.2 The Representation of Meanings. A Strategy Combining Paraphrase Template and Semantic Frame

The semantics of common sentences follows the principle of compositionality: the meanings of words are combined according to the structural meanings of the sentences where these words occur, as is the case in (11).

11. 北大中文系培养计算语言学本科生
Běidà Zhōngwén-xì *péiyǎng* *jìsuàn-yǔyánxué* *běnkē-shēng*
PKU Chinese-department train computational-linguistics undergraduate-student
'The department of Chinese language and literature of PKU has an undergraduate program in computational linguistics'.

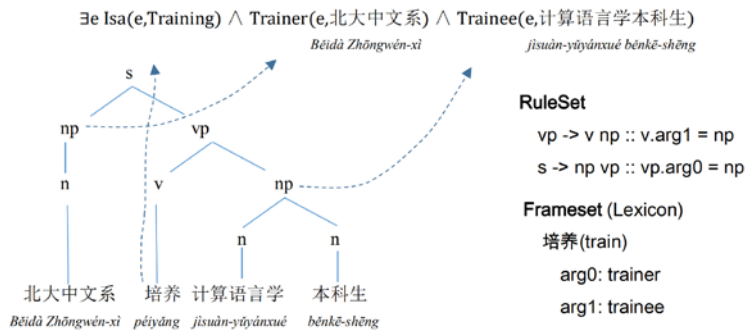


Figure 2 The semantic composition of sentence (11)

The syntactic structure derived from the syntactic rules set in (11) allows identification of the semantic roles of the NPs, where 北大中文系 *Běidà Zhōngwén-xì* 'the department of Chinese language and literature of PKU' plays the role of a 'trainer', which is often annotated as 'arg0' in propbank-style corpus, and 计算语言学本科生 *jìsuàn-yǔyánxué běnkē-shēng* 'undergraduates in computational linguistics' plays the role of a 'trainee', which is often annotated as 'arg1'.

One way to compute the meaning of a construct is to paraphrase it into a structure which can be handled by general phrase structure rules. The paraphrased sentence can then be processed by a semantic analyser, where a semantic representation can be computed according to the 'principle of compositionality'. See the example below:

12. 贝多芬十一岁时, 就已经显露了他的音乐天才, 被认为是莫扎特第二。
Bèiduōfēn shíyī suì shí jiù yǐjīng xiǎnlù le
Beethoven eleven year.old time then already show PFV
tā de yīnyuè tiāncái bèi rènwéi shì
he DE music talent BEI regard COP
Mòzhātè dì-èr
Mozart second
'Beethoven showed his music talent quite early, at the age of eleven. At that time, he was regarded as a second Mozart'.

莫扎特第二 *Mòzhātè dì-èr* 'a second Mozart' in example (12) is an instantiation of [N + 第二 *dì-èr*]. In the CCL-CxnBank, the 'Paraphrase Template'(释义模板 *shìyì múbǎn*) of this construction is set as either [像 *xiàng* + N + 一样 *yíyàng*] or [很 *hěn* + 像 *xiàng* + N], both meaning 'like N'. Thus, (12) can be paraphrased as 贝多芬十一岁时, 就已经显露了他的音乐天才, 被认为是很像莫扎特 *Bèiduōfēn shíyī-suì shí, jiù yǐjīng xiǎnlù-le tā de yīnyuè tiāncái, bèi rènwéi shì* ***hěn xiàng Mòzhātè*** 'Beethoven showed his music talent early at the age of eleven. At that time he was believed to be very much like Mozart', where 很像莫扎特 *hěn xiàng Mòzhātè* 'very much like Mozart' is an ordinary phrase structure, whose meaning can be computed by the semantic analyser designed for processing ordinary phrase structures.

The paraphrasing method encounters difficulties when dealing with complicated meanings of constructs, at least in the following two aspects. First, paraphrase templates fail in the constructs where there is a variable that does not form a constituent. The constructs illustrated in 3.1.3 with the pattern […再 *zài*…也 *yě*…], for example, display variables that do not form complete constituents. In this case, it is more appropriate to determine their meanings by first analysing the structure of the compound clauses where the construct appears, and then representing such meanings separately, rather than applying the paraphrasing method as in (12). Suppose there are two clauses S1 and S2, where S1 includes 再 *zài* and S2 includes 也 *yě*. The propositional meanings of S1 and S2 are separately represented as P1 and P2. The meaning of the whole sentence is represented with two predicate formulas 'AND(P1, P2)' and 'INEVITABLY(P2)', where the former represents the basic propositional meaning of the whole sentence, and the latter represents the subjective attitude brought by the constants 再 *zài* and 也 *yě*, expressing the speaker's attitude that P2 will inevitably happen.

Paraphrase templates also fail to process construction meanings when the acquisition of the meanings depend on the context rather than on the construction itself, such as [用 *yòng* 'use' + N + 说话 *shuō-huà* 'speak'] 'speak with N'. While the paraphrase templates of this construction is given in CCL-CxnBank, such as [凭借 *píngjiè* 'rely on' + N + 获得 *huòdé* 'gain' + 优势 *yōushì* 'advantage' / 认同 *rèntóng* 'approval' / 权力 *quánlì* 'power] 'gain advantage / approval / power

with N', the specific meaning of certain constructs has to be fully determined in the specific context.

More specifically, 说话 *shuō-huà* 'speak' may either have a literal meaning, as in [用智慧 *yòng zhìhuì* 'use wisdom' 说话 *shuō-huà* 'speak'], meaning 'speak with wisdom', or display a metaphoric reading, as in [用行动说话 *yòng xíngdòng* 'use action' + 说话 *shuō-huà* 'speak'], meaning 'speak with action', [用拳头说话 *yòng quántou* 'use fist' + 说话 *shuō-huà* 'speak'], meaning 'speak with fists'. The constant 说话 *shuō-huà* 'speak' in this construction can have different meanings in different contexts. Therefore, the meanings of the instances of this construction cannot be easily formalised through paraphrase templates, which can only provide abstract and general meaning descriptions. Some other representation schemas that try to represent construction meanings by paraphrasing also fail in this construction, e.g. AMR for constructions (Bonial et. al. 2018).

Construction meanings that are determined by context are more suitable to be formalised by frame representations, where constructional meanings can be included through attributes in the frame: specific meanings implied in certain contexts can be specified as values of the attributes. For example, the meaning of the construction [用 *yòng* 'use' + N + 说话 *shuō-huà* 'speak'] 'speak with N' can be represented with the frame in figure 3.

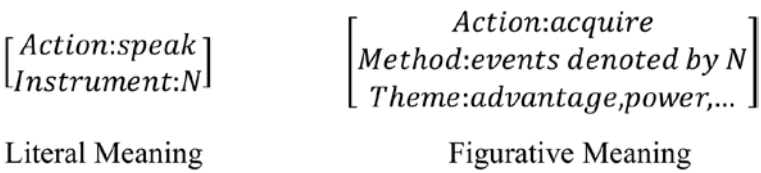


Figure 3 The frames representing the literal and figurative meanings of [用 *yòng* 'use' + N + 说话 *shuō-huà* 'speak'] 'speak with N'

The frames below represent the meaning of two instances of the construction: 用数据说话 *yòng shùjù shuō-huà* 'use figures speak, gain approval with data', and 用拳头说话 *yòng quántou shuō-huà* 'use fist speak, acquire power by beating others, assert one's authority through force'.

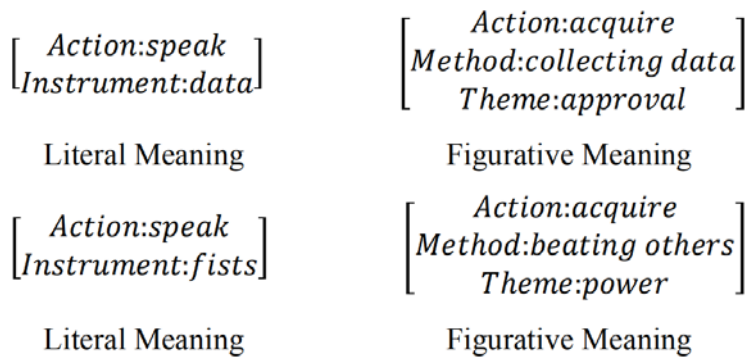


Figure 4 The frames representing the literal and figurative meanings of 用数据说话 *yòng shùjù shuō-huà* 'gain approval with data', and 用拳头说话 *yòng quántou shuō-huà* 'assert one's authority through force'

In conclusion, the meaning representation of constructions can be decomposed into two layers, including:

1. Paraphrase Template: the meanings which can be expressed by simply manipulating symbols linearly.
2. Semantic Frame: the implicit meaning of a construction often need to be represented with frames.

The semantic frame can be further divided into two types:

a. the Implied Meaning: additional semantic relations are acquired after the structural analysis of variables and constants, which are discontinuous and remotely related, expressing meanings such as the dependency meaning expressed by the repetition of interrogatives on the same syntactic position, or the adversative relation expressed by [再 *zài* + VP_1 + 也 *yě* + VP_2] 'no matter how much one VP_1, VP_2 still occurs' (see above), which further indicates a subjective attitude of necessity etc.

b. The Contextual Meaning: the meaning of a specific construct has to be clarified in the context where it occurs. For instance, in the construction [用 *yòng* 'use' + N + 说话 *shuō-huà* 'speak'] 'speak with N', the variable N expresses the means by which certain actions take place. The whole construction uses means as a metaphor of the purpose of an action, such as 用数据说话 *yòng shùjù shuō-huà* 'use data speak, support an idea' (lit. 'speak with data'), 用成绩说话 *yòng chéngjì shuō-huà* 'use grades speak, prove some ability' (lit. 'speak with grades'), and 用拳头说话 *yòng quántou shuō-huà* 'use fist speak, to defeat one's opponent' (lit. 'speak with fists') etc. The abstract

aspects of the construction's derived meanings (such as the abstract 'objective' meaning highlighted in certain constructions) can be described in the CCL-CxnBank, and the specific aspects, including subjective attitudes such as evaluations, standpoints and emotions etc., can be analysed and added according to the context while annotating the corpus. This aspect will be elaborated in § 5 below.

4 The Framework and Current Status of CCL-CxnBank

Some constructicon projects on several languages are described in Lyngfelt et al. (2018). The quantity of data included in these projects so far is not very large. A brief survey on these constructicons is listed in Appendix I. This section introduces the design framework and the current status of our project on the basis of Zhan (2017), where the basic issues of developing CCL-CxnBank were briefly introduced and discussed.

The descriptive framework of the construction knowledge is a core issue for the development of constructicons. Using the English rate-construction as an example, Fillmore, Lee-Goldman and Rhodes (2012) summarised six types of construction knowledge: (1) a bracketing formula with syntactic and semantic information attached to mother and daughter nodes; (2) a mnemonic name (used to address the constructions); (3) syntactic categories of the mother and daughter nodes, sometimes followed by informal descriptions of their syntagmatic distributions; (4) (optional) informal descriptions of the semantic information of the mother and daughter nodes; (5) an informal interpretation of the meaning of the construction as a whole (similar to traditional dictionary explanations); (6) annotated sentences containing the construction.

The German constructicography project described in Lyngfelt et al. (2018) concluded that, in order to appropriately describe the idiosyncratic characteristics of constructions of a specific language, the design of the description framework has to suit the grammatical characteristics of the specific target language, rather than trying to stipulate a universal grammatical framework for constructions of all the languages around the world. The design of the framework of CCL-CxnBank is in accordance with this view, implementing Yu (2003) and Zhan (1999; 2000) as the fundamental grammatical framework for the description of constructions, which have their origins in Zhu (1982; 1985).

Compared with Fillmore, Lee-Goldman and Rhodes (2012), we have developed a framework which allows to describe a richer amount of information in a more fine-grained manner (see Appendix II). The framework includes seven parts: (1) basic information,

(2) constants and variables, (3) relations between constants and variables, (4) syntactic information, (5) semantic information, (6) pragmatic information, (7) references. Each part describes a specific aspect of a construction entry. Due to space limitation, only the first part is explained and illustrated in detail below:[7]

- **Form Variations** (构式变体 *gòushì biàntǐ*) of a given construction carry the same variable(s) but different constant(s). For example, the construction [有 *yǒu* + 什么 *shénme* + VP + 的 *de*] 'what is the worth of VP' has form variations such as [有 *yǒu* + 什么 *shénme* + 可 *kě* + VP + 的 *de*], [有 *yǒu* + 什么 *shénme* + 好*hǎo* + VP + 的 *de*] etc., as shown in § 3.1.1.
- **Construction Type** (构式类型 *gòushì lèixíng*) may either be fixed (凝固型 *nínggù xíng*), semi-fixed (半凝固型 *bàn-nínggù xíng*), phrasal (短语型 *duǎnyǔ xíng*) or compound (复句型 *fùjù xíng*). For example, [用 *yòng* 'use' + 脚 *jiǎo* 'foot' + 投票 *tóupiào* 'vote'] 'vote with feet' is a fixed construction, which contains only constants and no variable components. [人 *rén* 'person' + 见 *jiàn* 'meet' + 人 *rén* 'person' + V] 'whoever meets him/her V', on the other hand, is a semi-fixed construction: this type of construction usually has a fixed length, mostly four syllables, and contains one or two variants. [NP + 倒 *dào* 'but' 是 *shì* 'be' + NP] 'although NP is NP,...' is an example of phrasal constructions, which have variable length and contain more than one variable component; the variants are to some extent replaceable. Finally, [NP_1 + X_1, NP_2 + 还 *hái* 'still' + X_2 + 呢 *ne*] 'NP_1 X_1 is trivial compared with NP_2 X_2' is a compound construction, with variable length and more variants, which have higher replaceability. The definitions of these four types are elaborated in Zhan (2017).
- **Features** (构式特征 *gòu shì tèzhēng*): tags indicate construction features related to their syntax, semantics or other aspects. This set of tags is open, i.e. new tags can be added. For example, the tags attached to the construction [NP + 不 *bù* 'not' + VP + 谁 *shéi* 'who' + VP] 'NP do not VP, who else is supposed to do so' are: (i) 省略 *shěnglüè* 'ellipsis', as the original form of this construction is [如果 *rúguǒ* 'if' + NP + 不 *bù* 'not' + VP + 那么 *nàme* 'then' + 谁 *shéi* 'who' + VP], in which the conditional connectives 如果 *rúguǒ* 'if' and 那么 *nàme* 'then' marking the logical relation between the two clauses in the construction are omitted; (ii) 复现 *fùxiàn* 'recurrence', since there are two perfectly identical VPs in the construction; (iii) 含否定成分 *hán fǒudìng chéngfèn* 'containing negation markers', since there is a negative word 不 *bù* 'not' in the construction; (iv) 含疑问成分

7 For more details on the remaining six parts, please visit the website of CCL-CxnBank.

hán yíwèn chéngfèn 'containing question markers', since there is a question word 谁 *shéi* 'who'; and (v) 修辞 *xiūcí* 'rhetoric', since this construction is a rhetorical question.

- **Number of Syllables** (构式音节数 *gòu shì yīnjié shù*) captures the length of a construction or the number of syllables allowed for the construction. For fixed constructions it is a fixed number (e.g. (2) for [甩 *shuǎi* 'throw' + 锅 *guō* 'pot'] 'pass the buck'), while for other types of constructions it is a range of possible numbers (e.g. (4)-(10) for [有 *yǒu* 'have' + 什么 *shénme* 'what' + NP] 'there is no NP').
- **Number of Chunks** (组块数 *zǔ kuài shù*): the number of chunks of the construction that describes whether the construction is separated into two parts by a comma. For example, [没有*méi yǒu* 'not have' + NP_1 + 就 *jiù* 'then' + 等于 *děngyú* 'be equal to' + 没有 *méi yǒu* 'not have' + NP_2] 'the loss of NP_1 leads to the loss of NP_2' and [别 *bié* 'don't' + 说 *shuō* 'speak' + X, 连 *lián* 'even' + Y + 都 *dōu* 'all' + Z] 'even Y Z, let alone X' are both compound constructions, but the former has one chunk, while the latter has two chunks.
- **Expandable** (是否可扩展 *shìfǒu kě kuòzhǎn*) refers to the property of whether the construction can be expanded by juxtaposition. For example, [A + 得 *de* 'COMP' + 够呛 *gòuqiāng* 'terribly/extremely'] 'extremely A' can be expanded by juxtaposition, in sentences such as 他累得够呛，困得够呛，倒头就睡 *tā lèi de gòuqiàng, kùn de gòuqiàng, dǎotóu jiù shuì* 'he fell asleep immediately, as he was extremely tired and sleepy'.
- **Sense Number** (义项编号 *yìxiàng biānhào*) indicates the number of the meanings of a construction; not all constructions with the same form share the same meaning. If a construction form has only one meaning, the sense number of this entry is recorded as 0. Constructions with the same form but different meanings are listed in CCL-CxnBank as different entries, with sense numbers recorded as 1, 2, 3,... etc.
- **Paraphrase Templates** (释义模板 *shìyì múbǎn*) specifies the ordinary phrase that is synonymous to a construction. This column records phrases which can replace constructs of the entry in language use. For example, the paraphrase templates of [NP + 不 *bù* 'not' + VP + 谁 *shéi* 'who' + VP] 'NP do not VP, who else is supposed to VP' are: [如果 *rúguǒ* 'if' + NP + 不 *bù* 'not' + VP + 那么 *nàme* 'then' + 谁 *shéi* 'who' + VP] 'if NP does not VP, then who else is supposed to do so', [NP + 一定 *yídìng* 'surely' + 会 *huì* 'be likely' + VP] 'NP is surely to VP', [NP + 就 *jiù* 'then' + 该 *gāi* 'should' + VP] 'NP is obliged to VP' etc.
- **Samples** (构式实例 *gòushì shílì*) specify at least 3 samples of the actual usages of the construction, either in contexts or not. The

samples are collected from the CCL corpus[8] or built by the lexicographer according to her/his intuition. For example, the samples of [NP + 不 *bú* 'not' + VP + 谁 *shéi* 'who' + VP] 'NP does not VP, who else is supposed to VP' are: 劳模不干谁干 *láomó bù gān shéi gàn* 'if the model worker doesn't do it, who else is supposed to do it', 你不失败谁失败 *nǐ bù shībài shéi shībài* 'if you do not fail, who else is supposed to fail', and 我不入地狱谁入地狱 *wǒ bú rù dìyù shéi rù dìyù* 'if I do not step into hell, who else is supposed to do so'.

- **Synonym Constructions** (同义构式 *tóngyì gòushì*):[9] construction entries which share the same meaning and the same constants of the construction. For example, the synonym construction of [N_1 + 多 *duō* 'much, many' + N_2 + 少 *shǎo* 'few, little'] 'N_1 is abundant while N_2 is lacking' is [V_1 + 多 *duō* 'much' + V_2 + 少 *shǎo* 'little'] 'always V_1 but seldom V_2', and vice versa. The two constructions share the same template of interpretation.
- **Antonym Constructions** (反义构式 *fǎnyì gòushì*): construction entries which have the opposite meaning of the construction. For example, the antonym construction of [NP + 倒 *dào* 'but' 是 *shì* 'be' + NP] 'although NP is NP,...' is [NP + 倒 *dào* 'but' 不是 *bú shì* 'not be' + NP] 'although NP is not NP', and vice versa.
- **Hyperonym Constructions** (上位构式 *shàngwèi gòushì*) specifies the more general construction entry which subsumes (both syntactically and semantically) the construction.[10]
- **Hyponym Constructions** (下位构式 *xiàwèi gòushì*): the more specific construction entries which are subsumed (both syntactically and semantically) by the construction.
- **Negated Forms** (否定形式 *fǒudìng xíngshì*): the constructions collected in the CCL-CxnBank are idiosyncratic patterns which cannot be further decomposed with phrase structure rules. Therefore, their negated forms have to be manually recorded rather than deduced with phrase structure rules. The same goes for **Interrogative Forms** (疑问形式 *yíwèn xíngshì*). See examples (13) and (14) below for a comparison between a common sentence that has a corresponding interrogative form and a construction that has no corresponding interrogative form. For constructions which do not have negated forms or inter-

8 http://ccl.pku.edu.cn:8080/ccl_corpus.

9 Ideally, the information content of this field, including synonym, antonym, hypernym and hyponym, can help establish hierarchical network relationships between constructs. But, in fact, there are only some local relationships of parts of constructs at present, and no network relationships covering all the constructions has been established.

10 There is nothing to fill in the field 'Hyperonym Constructions' in the current database, since there is no schematic construction recorded in CCL-CxnBank at the current stage. The same goes for 'Hyponym Constructions'.

rogative forms, or which are already negated or interrogative, these two columns are recorded as 'none'.

- **Origin** (形成机制 *xíngchéng jīzhì*): describes how a construction emerges, or the process of grammatical constructionalisation of a construction. An academic paper is usually required to explain the origin of a construction.
- **Notes** (备注 *bèizhù*): the place where the lexicographer may keep notes on issues related to an entry, which need to be logged in detail for further investigation.

The goal of CCL-CxnBank is to accurately describe all the syntactic distribution information of each construction, which is illustrated in the following examples.

13. a. 张三也买了那本书。
Zhāngsān yě mǎi le nà běn shū
Zhangsan also buy PFV that CLF book
'Zhangsan also bought that book'.

b. 谁也买了那本书?
shéi yě mǎi le nà běn shū
who also buy PFV that CLF book
'Who also bought that book?'

c. 张三也买了哪本书?
Zhāngsān yě mǎi le nǎ běn shū
Zhangsan also buy PFV which CLF book
'Which book did Zhangsan also buy?'

14. a. 连张三也买了那本书。
lián Zhāngsān yě mǎi le nà běn shū
even Zhangsan also buy PFV that CLF book
'Even Zhangsan bought that book'.

b. *连谁也买了那本书?
lián shéi yě mǎi le nà běn shū
even who also buy PFV that CLF book
*'Even who also bought that book?'

c. *连张三也买了哪本书?
lián Zhāngsān yě mǎi le nǎ běn shū
even Zhangsan also buy PFV which CLF book
*'Even which book did Zhangsan also buy?'

(13a) is a sentence whose internal structure is subject-predicate, while (13b) and (13c) are its interrogative forms. In general, sentences consisting of regular phrases have both a declarative form and a corresponding interrogative form. However, (14a) is an instance of the [连 *lián* 'even' + X + 也 *yě* 'also' + Y] construction, which does not have interrogative forms like those of (13a). Both (14b) and (14c), which contain the question words 谁 *shéi* 'who' and 哪 *nǎ* 'which', respectively, are ungrammatical.

Based on the detailed description of each construction, a variety of statistical information on all entries in CCL-CxnBank is available now. There is a web page that displays the frequency of occurring constants, variables, and features, including both single features and combinations of features, which can be extracted from all the constructions or just only from a selected type of constructions. For example, figure 5 shows the 8 most frequently occurring constants in CCL-CxnBank. They are 不 *bù* 'not', 一 *yī* 'one, a', 的 *de* 'DE', '是 *shì* 'be', 个 *ge* 'CLF', 有 *yǒu* 'have', 了 *le* 'PFV', 也 *yě* 'also', in descending order of frequency. Obviously, high frequency function words and verbs with more abstract meanings are more common in constructions.

Figure 5 The webpage that displays statistics of CCL-CxnBank

The left side of figure 5 shows the statistical results, i.e. the frequency list of items being counted. The right side of figure 5 shows a menu for the user to select 'Items that need to be counted', 'Scope of statistics', which have been explained above, and 'Sort criteria' (the statistical result can be presented both in order of frequency or in alphabetical order).

Based on the statistics of variable components and features in current CCL-CxnBank, we can sketch an overview of common features of Chinese constructions: (1) the top three variable categories (ignoring the category X which matches all the categories) are V (verb), A (adjective) and AP (adjective phrase), indicating that predicative con-

stituents are more likely to fill the slots of constructions than nominal constituents; (2) the top three construction features are recurrence (复现 *fùxiàn*), grammatical mismatch (语法错配 *yǔfǎ cuòpèi*) and ellipsis (省略 *shěnglüè*), which conforms to our expectation that, according to phrase-based rules of grammar, Chinese constructions usually have grammatical mismatches to some extent, which are often caused by recurrence or ellipsis of certain constituents.

5 Building a Syntactically and Semantically Annotated Corpus of Chinese Constructions

5.1 An Online Platform for the Annotation of Constructs

As a hand-built knowledge base, CCL-CxnBank alone cannot fully reflect the constructs' overall usages in real texts, especially their form and meaning variations. Just as lexicons and phrase structure rule bases have to be accompanied by treebanks to reflect the overall usages of linguistic units, constructicons too have to be accompanied by annotated corpora, in which each construction entry is complemented with a collection of sentences where the corresponding constructs occur.

The English FrameNet constructicon described in Lyngfelt et al. (2018) contains 73 constructions and 1,471 annotated sentences. The constructs in the sentences are annotated with linguistic information, including construction elements (CE), construction-evoking elements (CEE), words in the sentence and their syntactic categories etc. The linguistic information annotated on the constructs are mainly concerned with the constituents of the constructs, and the direct analysis of the meaning of the constructs is lacking.

In order to fill this gap, i.e. to fully reflect the uses of constructions in real texts and to investigate the sentiment information carried by constructions (Huang, Zhan 2018), we have selected from CCL-CxnBank 50 constructions that have subjective attitudinal meanings. These constructions are tagged with construction features such as negative evaluation (负面评价 *fùmiàn píngjià*), subjective large amount (主观大量 *zhǔguān dàliàng*), and subjective little amount (主观小量 *zhǔguān shǎoliàng*) in the database table that describes the basic information of the construction. For each of the 50 constructions, about 100 sentences from the CCL corpus are extracted, resulting in a total of 4,777 sentences.

For constructs within sentences, three types of information are annotated: the construct's boundary, constituents, and the subjective attitudinal meaning. A construct's boundary serves to separate a construct from its surrounding context. Within the boundary, con-

stituents are respectively annotated as constants and variables, according to the pattern of the construction. In figure 7, the coloured tiles highlight the construct 别说干事业，连吃饭走道都打不起精神 *bié shuō gàn shìyè, lián chīfàn zǒudào dōu dǎ bù qǐ jīngshén* 'be spiritless even when walking and eating, let alone working' in its context 一个人要是没有奋斗目标 *yí ge rén yàoshi méiyǒu fèndòu mùbiāo* 'if a person does not have a goal to strive for', with black tiles indicating the constants and red tiles indicating the variables.

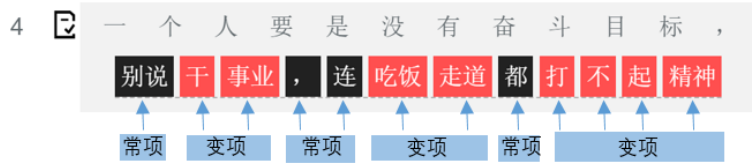


Figure 6 The annotation of the constituents of a construct. The constituents in black tiles and red tiles are constants and variables, respectively. The check mark on the top left corner indicates that this sentence's annotation has been proof-read

As for the subjective attitudinal meaning, four dimensions are designed to describe it: evaluation (评价 *píngjià*), standpoint (立场 *lìchǎng*), emotion (情感 *qínggǎn*), and intensity (强度 *qiángdù*). As for evaluation, there are three options: positive (正面 *zhèngmiàn*), negative (负面 *fùmiàn*), or neutral (中立 *zhōnglì*). Standpoint also has three options to choose from: accept (接受 *jiēshòu*), refuse (拒绝 *jùjué*), or noncommittal (不置可否 *bù zhìxìn kěfǒu*). The value of emotion can be defined by the annotator according to her/his judgement on the emotion the specific construct expresses in the context. As for intensity, four values are given to choose from: none (缺省 *juéduì*),[11] very high (极 *jí*), high (很 *hěn*), or not high (不很 *bù hěn*). Below is the subjective attitudinal meaning of the construct 别说干事业，连吃饭走道都打不起精神 *bié shuō gàn shìyè, lián chīfàn zǒudào dōu dǎ bù qǐ jīngshén* 'be spiritless even when walking and eating, let alone working'.

Statistics of subjective attitudinal meanings are shown in table 2 below. Among the 4,777 sentences of 50 constructions, about 70% of them are concerned with evaluations and standpoints; about 25% of the sentences express emotions; about half of the sentences have a relatively high intensity of subjective attitudes.

11 'None' is the default option. It is used to check automatically whether the intensity of a sentence is marked or not by the platform.

Figure 7 The subjective attitudinal meaning of the construct 别说干事业, 连吃饭走道都打不起精神 *bié shuō gàn shìyè, lián chīfàn zǒudào dōu dǎ bù qǐ jīngshén* 'be spiritless even when walking and eating, let alone working', with evaluation set as 'negative', standpoint set as 'refuse', emotion set as 'detest', and intensity set as 'not high'

Table 2 Statistics of subjective attitudinal meanings in the annotated construction corpus

Evaluation		Standpoint		Emotion	Intensity			Total
positive	negative	accept	refuse		very high	high	not high	
1,141	2,223	1,204	2,111	1,238	785	1,731	1,022	4,777
23.89%	46.53%	25.20%	44.19%	25.92%	16.43%	36.24%	21.39%	
3,364		3,315			3,538			
70.42%		69.40%			74.06%			

The subjective attitudinal meaning, as its name implies, is subjective, and it is up to the annotator's language intuition to determine the value of the four dimensions, given a specific construct and its context. In this project, each construct is annotated by one annotator and checked by another annotator to ensure the internal consistency of annotation results, in order to control the quality of the annotation.

5.2 Some Challenges in the Annotation of the Form and Meaning of Constructs

The annotation of constructs is a challenging task in language resource development. There are several issues in annotating the forms and meanings of constructs. This is shown in the following example of the annotation of the [连 *lián* 'even' + X + 都 *dōu* 'all' + Y] 'even X do/be Y' construction.

15. 连他离京, 做妹妹的都不知道。
lián tā lí Jīng zuò mèimei de dōu bù zhīdào
even he leave Beijing do sister DE all not know
'Even his sister does not know his departure from Beijing'.

In (15), the text string between the constants 连 *lián* and 都 *dōu* does not form a constituent, but stretches across two clauses. Therefore, the form [连 *lián* 'even' + X + 都 *dōu* 'all' + Y] does not precisely match the construct in (15), which requires a more flexible representation of the form of the construction [连 *lián* 'even' + X + 都 *dōu* 'all' + Y]. It is the same situation as the one we have shown in example (10) for the pattern […再 *zài* 'again' …也 *yě* 'still'…]: the pattern [连 *lián* 'even' + X + 都 *dōu* 'all' + Y] too establishes a long-distance relation which connects two clauses. In (15), the two clauses are 他离京 *tā lí Jīng* 'he leaves Beijing' and 做妹妹的不知道 *zuò mèimei de dōu bù zhīdào* 'his sister does not know', respectively, and are separated by a comma. The internal components of sentence (15) are analysed in the same way as sentence (10) in § 3.1.3.

16. 别说放弃了棋类的爱好, 连一般人天天都看的电视都没空看。
bié-shuō fàngqì le qílèi de àihào lián yìbān
not-say give.up PFV chess DE hobby even ordinary
rén tiāntiān dōu kàn de diànshì dōu méi-kòng kàn
person every.day all watch DE TV all not-time watch
'(He) does not even have time for TV programs that ordinary people watch, let alone having time for hobbies like playing chess'.

In (16), the second 都 *dōu* 'all' is a constant of the construction, but the first 都 *dōu* 'all' is used as a common adverb. This gives rise to difficulties when we try to design algorithms to automatically identify the construct's boundary.

17. 下雨天, 别说打(不)到车, 连地铁都会挤爆。
xiàyǔ tiān bié-shuō dǎ (bú) dào chē lián dìtiě dōu
rain day not-say call not able taxi even subway also
huì jǐbào
will overcrowded
'On rainy days, the subways will be crowded, not to mention that you cannot find a taxi'.

In (17), the speaker means that the hearer cannot find a taxi, and public transportation is not a solution, no matter whether the negative 不 *bù* appears in the clause introduced by 别说 *bié shuō* 'not to say' or not. This meaning is inferred from the literal meaning of the [连 *lián* 'even' + X + 都 *dōu* 'all' + Y] construction. The mechanism of how a construct interacts with constituents outside of its boundary is a challenging problem and is still under investigation.

18. 这是连天气预报都可以放假的日子。
zhè shì lián tiānqì-yùbào dōu kěyǐ fàngjià de rìzi
this COP even weather-forecast all can have.a.day.off DE day
'The weather is so good that even the weather forecast can have a day off'.

In (18), the literal meaning 'the weather forecast being able to have a day off' is an improbable event. The occurrence of this improbable event is caused by the fact that the weather is extremely pleasant, so there is no reason to worry about changes in weather. The construction [连 *lián* 'even' + X + 都 *dōu* 'all' + Y] invites listeners to discover the reason for the occurrence of an improbable event. The mechanism by which this inference is carried out also needs further investigation.

The current construct annotation project is still in the early stages of exploration. Our goal is to annotate construction information based on treebanks and propbanks, where basic syntactic and semantic information has already been annotated. In this way, further investigation on the interaction between the constructs and the contexts can be carried out, where pragmatic information (such as inferences) shall be elicited and added into CCL-CxnBank.

6 Conclusions

As Ronald Langacker said in his book, "language is a mixture of regularity and idiosyncrasy" (1987, 411). During the development of Peking University Treebank (Zhan 2016), we already realised that constructions are necessary complements to common phrase structures, and common phrases are well suited to describe their internal constructs in terms of recursive tree structures defined by a formal grammar. However, for the constructions discussed in this paper, it is not suitable to describe their internal structures with hierarchical tree structures. As already pointed out in the analysis above, it is more suitable to describe the internal composition patterns of constructions as flat linear sequences.

The practical work of developing CCL-CxnBank taught us that constructicons and annotated construction corpora should be compatible with existing language resources, make full use of the work under the theory of phrase structure grammar, and integrate their annotation guidelines into systems of language resources such as treebanks, propbanks and FrameNet etc. The new language resources developed in this way will be more valuable from the perspective of language engineering.

As to the meaning representation of constructions, we recognise that, although constructional approaches to language emphasise the

integrity of constructions and neglect the combinatorial semantic analysis of the constituents of constructions to some extent, the principle of compositionality holds in the analysis of construction meanings. In order to correlate the form and meaning of a construction, it is still necessary to decompose the construction form and combine the meanings of the constituents. This principle deserves much consideration in the design of the annotation of construction constituents and meanings. On the other hand, another principle of semantic analysis, i.e. the contextuality principle, should also be considered in the analysis of construction meanings in our future research. The analysis of construction meanings needs to be combined with the annotation of contextual features of constructions.

Bibliography

Bonial, C. et al. (2018). "Abstract Meaning Representation of Constructions. The More We Include, the Better the Representation". *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (LREC, May 2018, Miyazaki, Japan). Luxembourg; Paris: European Language Resource Association, 1677-84. http://www.lrec-conf.org/proceedings/lrec2018/pdf/856.pdf.

Chomsky, N. (1956). "Three Models for the Description of Language". *Transactions on Information Theory*, 2(3), 113-24. https://doi.org/10.1109/tit.1956.1056813.

Croft, W.; Cruse, D.A. (2004). *Cognitive linguistics*. Cambridge: Cambridge University Press.

Croft, W. (2001). *Radical Construction Grammar. Syntactic Theory in Typological Perspective*. Oxford: Oxford University Press.

Fillmore, C.J.; Kay, P.; O'Connor, M.C. (1988). "Regularity and Idiomaticity in Grammatical Constructions. The Case of Let Alone". *Language*, 64(3), 501-38. https://doi.org/10.2307/414531.

Fillmore, C.J.; Lee-Goldman, R.R.; Rhodes, R. (2012). "The FrameNet Constructicon". Boas, H.C.; Sag, I.A. (eds), *Sign-Based Construction Grammar*. Stanford, CA: CSLI Publications, 309-72.

Goldberg, A.E. (1995). *A Construction Grammar Approach to Argument Structure*. Chicago: The University of Chicago Press.

Goldberg, A.E. (2013). "Constructionist Approaches". Hoffmann, T.: Trousdale, G. (eds), *The Oxford Handbook of Construction Grammar*. Oxford: Oxford University Press, 15-31.

Hoffmann, T. (2017). "The Renaissance of Constructions. From Constructions to Construction Grammars". Dancygier, B. (ed.), *The Cambridge Handbook of Cognitive Linguistics*. Cambridge: Cambridge University Press, 284-309.

Huang S. 黄思思; Zhan W. 詹卫东 (2018). "Mianxiang qinggan fenxi de goushi zhuguan taidu yi chutan" 面向情感分析的构式主观态度义初探 (A Rudimentary Investigation of the Subjective Attitudinal Meaning of Construction Towards Sentiment Analysis). *Waiyu Jiaoxue*, 39(6), 27-33.

Jurafsky, D.; Martin, J.H. (2000). *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. New Jersey: Pearson Education.

Kay, P.; Fillmore, C.J. (1999). "Construction Grammar and Linguistic Generalizations. The What's X Doing Y? Construction". *Language*, 75(1), 1-33. https://doi.org/10.1353/lan.1999.0033.

Kay, P.; Michaelis, L.A. (2012). "Constructional Meaning and Compositionality". Maienborn, C.; von Heusinger, K.; Portner, P. (eds), *Semantics. An International Handbook of Natural Language Meaning*, vol. 3. Berlin: Mouton de Gruyter, 2271-96.

Langacker, R.W. (1987). *Foundations of Cognitive Grammar. Theoretical Prerequisites*, vol. 1. Stanford: California: Stanford University Press.

Li C.N.; Thompson, S.A. (1981). *Mandarin Chinese A Functional Reference Grammar*. Berkeley; Los Angeles: University of California Press.

Lyngfelt, B. et al. (eds) (2018). *Constructicography. Constructicon Development Across Languages*. Amsterdam: John Benjamins. https://doi.org/10.1075/cal.22.

Partee, B.H. (2004). "Compositionality". Partee, B.H. (ed.), *Compositionality in Formal Semantics. Selected Papers*. Malden (MA): Blackwell, 153-81.

Yu S. 俞士汶 (2003). *Xiandai Hanyu yufa xinxi cidian xiangjie* 现代汉语语法信息词典详解 (The Grammatical Knowledge Base of Contemporary Chinese. A Detailed Explanation). Beijing: Tsinghua University Press.

Zhan W. 詹卫东 (1999). "Yi ge Hanyu yuyi zhishi biaoda kuangjia. Guangyi peijia moshi" 一个汉语语义知识表达框架: 广义配价模式 (A Framework of Chinese Semantic Representation. Generalised Valence Mode). *Proceedings of Joint Symposium on Computational Linguistics* (Beijing, 1-3 November 1999). Beijing: Tsinghua University Press, 1-7.

Zhan W. 詹卫东 (2000). *Mianxiang Zhongwen xinxi chuli de xiandai Hanyu duanyu jiegou guize yanjiu* 面向中文信息处理的现代汉语短语结构规则研究 (A Study of Constructing Rules of Phrases in Contemporary Chinese for Information Processing). Beijing: Tsinghua University Press.

Zhan W. 詹卫东 (2017). "Cong duanyu dao goushi. Goushi zhishiku jianshe de ruogan lilun wenti tanxi" 从短语到构式: 构式知识库建设的若干理论问题探析 (On Theoretical Issues in Building a Knowledge Database of Chinese Constructions). *Zhongwen Xinxi Xuebao*, 1, 230-8.

Zhang B. 张伯江 (2008). "Jushi yufa lilun yu Hanyu jushi yanjiu" 句式语法理论与汉语句式研究 (Constructional Approaches to Grammar and Research on Chinese Constructions). Sheng Y. 沈阳; Feng S. 冯胜利 (eds), *Dangdai yuyanxue lilun he Hanyu yanjiu* 当代语言学理论和汉语研究 (Contemporary Linguistic Theories and Chinese Linguistic Studies). Beijing: The Commercial Press, 497-507.

Zhang B. 张伯江 (2018). "Goushi yufa yingyong yu Hanyu yanjiu de ruogan sikao" 构式语法应用于汉语研究的若干思考 (Some Reflections on the Application of Construction Grammar in Chinese Studies). *Yuyan jiaoxue yu yanjiu*, 192(4), 2-11.

Zhang J. 张娟 (2013). "Guonei Hanyu goushi yufa yanjiu shi nian" 国内汉语构式语法研究十年 (Ten Years of Construction Grammar Research in China). *Hanyu Xuexi*, 2, 65-77.

Zhu D. 朱德熙 (1982). *Yufa jiangyi* 语法讲义 (Lecture Notes on Grammar). Beijing: The Commercial Press.

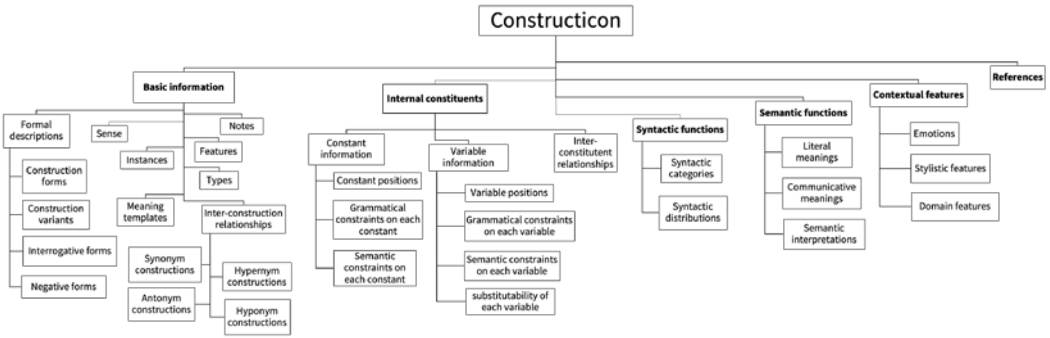
Zhu D. 朱德熙 (1985). *Yufa dawen* 语法答问 (Questions and Answers on Grammar). Beijing: The Commercial Press.

Appendix I. Constructicon Development across Languages

The table below is summarised from the content of each chapter in Lyngfelt et al. 2018.

Language	Name	Statistics	Website	Resources Dependent on
English	FrameNet Constructicon	73 constructions	http://www1.icsi.berkeley.edu/~hsato/cxn00/21colorTag/index.html	FrameNet Lexicon
Swedish	SweCcn	400 constructions	https://spraakbanken.gu.se/konstruktikon	Språkbanken SweFN++ Karp/Korp
Brazilian Portuguese	FN-Br	289 constructions	https://www.ufjf.br/framenetbr-eng/projects/frames-and-constructions http://webtool.framenetbr.ufjf.br/index.php/webtool/report/cxn/main	FrameNet
Japanese	JFN	N/A	http://jfn.st.hc.keio.ac.jp	FrameNet
Russian	FrameBank	Including 2700 high frequency verbs and 600 constructions which contain them	http://framebank.ru https://github.com/olesar/framebank	Språkbanken
German	GCon	39 constructions	http://gsw.phil.uni-duesseldorf.de https://gsw.phil.hhu.de https://gsw.phil.hhu.de/constructicon/constructionindex	FrameNet

Appendix II: The Framework of CCL-CxnBank



Corpus-Based Research on Chinese Language and Linguistics
edited by Bianca Basciano, Franco Gatti, Anna Morbiato

Some Reflections on the *Database of Medieval Chinese Texts* as a Multi-Purpose Tool for Research, Teaching, and International Collaboration

Christoph Anderl
Ghent University, Belgium

Abstract This paper gives an introduction to a Digital Humanities project at the Department of Languages and Cultures (Ghent University), the *Database of Medieval Chinese Texts* (DMCT), a collaborative project with several international partners. The structure of the DB is multi-modular, consisting of reference modules in the form of XML marked-up medieval non-canonical Chinese Buddhist texts, as well as analytical modules such as the Variants, Syntax, and Sentence Analysis modules. The architecture is 'open' and modules can be added, modified, and interlinked based on specific research requirements. The DB is multifunctional and not only provides information on key texts and their linguistic features, but also constitutes a research tool (featuring sophisticated online input masks and analytical tools) with which researchers can input and process data. In addition to its function in a research environment, it is also used in advanced master classes, in the framework of master thesis and PhD projects, as well as for internships. The DB has also an important 'socio-institutional' function, being situated at the intersection of Buddhological and historical linguistic studies, two of the main fields of research at the department.

Keywords Digital humanities. Linguistic database. XML mark-up. Medieval Chinese. Chinese syntax. Chinese character variants.

Edizioni Ca'Foscari

Sinica venetiana 6
e-ISSN 2610-9042 | ISSN 2610-9654
ISBN [ebook] 978-88-6969-406-6 | ISBN [print] 978-88-6969-407-3

Peer review | Open access
Submitted 2020-09-29 | Accepted 2020-12-06 | Published 2020-12-21

DOI 10.30687/978-88-6969-406-6/011

1 Introduction

The digitisation of premodern Chinese texts and the availability of an increasing number of huge text corpora have revolutionised many aspects of Sinological research during the last decades. Nowadays, the tracing of the source of a specific text passage, a term, a name, or a grammatical marker can ideally be performed within a very short period, whereas previously one frequently had to consult multiple indices or dictionaries, or even read through entire texts in order to retrieve the information. In addition, statistical material concerning the frequency of semantic items or syntactic function words can be collected much more speedily as compared to pre-digitisation times.

During my participation in projects involving text corpora and databases during the last 25 years, I have been observing a variety of approaches concerning the use and integration of the swiftly developing digital collections of texts, as well as a variety of continuously changing database and programming environments, which often entailed numerous problems and often rendered certain technical frameworks obsolete after a relatively short period. Naturally, the 'fall-out' rate in this field of research is significant; on the other hand, various projects have proven to become stable digital platforms and are continuously maintained and improved, greatly facilitating the work of the targeted research community. The reasons why certain database/digitisation projects have been successful – while others have not – are manifold and will not be discussed in detail in this paper.[1]

Considering the above, initiating a new database (DB) project is a risky task, since the initial technical framework will have a great impact on the future development of the DB. Therefore, when we first started designing the Database of Medieval Chinese Texts (DMCT)[2] in 2014, we decided to take a 'hybrid' approach, i.e. a project which could

1 Based on my experience with database projects, I have observed that successful projects seem to be often driven by the vision *of one person* or a small group of people, capable of motivating others to participate and contribute (as well as attracting the necessary funding). Among the databases I personally use most frequently, I want to mention the *Digital Dictionary of Buddhism* (DDB; ed. in chief: Charles Muller), which has developed immensely during the last years, with dozens of researchers contributing their research results, as well as the huge and ever-expanding digital collections of Buddhist texts in the form of the Chinese Buddhist Electronic Text Association (CBETA) and the SAT Daizōkyō Text databases. The collections of East Asian digital Buddhist corpora have expanded and improved at a very fast pace, one of the reasons being the work of innumerable anonymous contributors who input and proofread a vast number of texts. Another successful and innovative DB project I want to mention is *Thesaurus Linguae Sericae* (TLS, initiated more than 20 years ago by Christoph Harbsmeier), which has become an indispensable analytical tool for research on premodern Chinese texts.

2 Concerning the editors of and contributors to the DB project, please see `https://www.database-of-medieval-chinese-texts.be`.

develop in a multi-functional, multi-purpose and flexible way, and a DB which could 'grow' organically according to varying research and teaching requirements (for further elaborations, please see below).

From the beginning, the DMCT has been an international and collaborative project, drawing on the expertise of specialists in various fields, the main partners being Ghent University (Department of Languages and Cultures; Ghent Centre for Buddhist Studies) and Dharma Drum Institute of Liberal Arts (DILA, New Taipei City),[3] one of the leading Asian research centres concerning the digitisation of premodern Chinese texts. In addition, we have been collaborating with specialists in digitisation and Chinese text mark-up, most importantly, with Marcus Bingenheimer (formerly DILA; now Temple University).

2 The Technical Framework

When initiating the project in 2014, we were using eXist, a platform I had used in previous projects and which is very suitable for dealing with files in XML format (i.e. the mark-up language we use for the digitised texts), but for technical reasons we migrated to MySQL ca. three years ago.[4] MySQL is a relational DB, which is organised in tables. It can use different storage engines and, depending on the specific table, we use InnoDB[5] or MyISAM. MyISAM is specifically used for all tables which are designed for full-text searches, whereas InnoDB is used for all other tables, such as the user management tables.

The programme logic is implemented in PHP,[6] using object-oriented programming (OOP) and other interfaces, like PDOs (i.e. PHP Data Objects) combined with the Open Source PHP User Management Framework UserSpice.[7]

The view of the DB is designed with Cascading Style Sheets (CSS) and further languages are HTML5 and JavaScript. Since the encoded

3 These two institutions, in addition to the Research Foundation Flanders (FWO), have been the main sponsors of the DB. We also received financial support from the Tianzhu Foundation for the programming work. Administrative support and expert advice have been provided by members of the Dunhuang Academy, as well as by the international project *From the Ground Up. Buddhism and East Asian Religions* at the University of British Columbia.

4 The technical work on the DB has been primarily performed by the programming specialists Christian Bell (Bell Internet Design) and Jan Schrupp.

5 InnoDB is a product of the Oracle Corporation and is distributed under the GNU General Public Lincence. For an introduction to InnoDB storage engine, see `https://dev.mysql.com/doc/refman/8.0/en/innodb-introduction.html`. On MyIsam, see `https://dev.mysql.com/doc/refman/8.0/en/myisam-storage-engine.html`.

6 PHP is a programming language used especially for web development.

7 See `https://userspice.com`.

texts are XML files but the InnoDB itself is not suitable for storing XML files (unlike eXist), a XML import/export function was implemented.

Since recently, we have been using OpenProject[8] for the communication between editors/contributors and programmers, in order to improve the management of the work packages. All modules of the DB have commentary functions integrated, in order to add an interactive element in the communication with the (registered) users. The DB also features an advanced system of user management,[9] as well as sophisticated input interfaces for each module.

The DB consists of several modules whose data can be cross-referenced to each other. Currently, only some of the modules are public (the Text module, the Variant module, and the Bibliography), while some are currently for internal use only. A module for defining user rights makes it possible to assign permission to 'view' and/or 'edit' to each registered user/editor of the site, which has proven very useful in teaching environments (i.e. the students learn how to directly input data) and in the context of internships (see § 8). Unregistered visitors can fully access the public parts of the DB. By 2020, the public parts comprise all marked-up texts in two viewing modes ('diplomatic' and 'regularised'; see § 6 for more details), the module of Variant Chinese Characters ('Variant DB'; see § 7.1), and a bibliography. The internal modules are the Module of Medieval Chinese Syntactic Markers ('Syntax DB'; see § 7.2), the 'Sentence Analysis' module (see § 7.3), and the DB of 禪 *Chan* idiomatic phrases (see § 7.4). Currently, work on an additional module on Phonetic Loan Characters (通假字 *tōngjiǎzì*) is under construction.[10]

8 OpenProject is an open-source management software which we use for the assignment and coordination of work packages in the maintenance and development of the DB (for more information on this app, see `https://www.openproject.org`). This software has proved to be very useful for enhancing the communication and workflow efficiency among the participants.

9 I.e. 'editing'/'new entry'/'delete entry' functions can be assigned very specifically for each module of the DB. This is especially important when granting user rights to master students in the context of their internships (in order to limit the possibility of 'accidental damage' to the DB).

10 This module will collect references concerning character substitutions in manuscript texts, including phonetic loan characters, characters exchanged based on their structural similarities in handwriting, and other types of substitutions. Since the analysis of substitutions in handwritten manuscripts is highly complex, it was not included into the standard mark-up procedures. However, substitutions were systematically marked with 'sic' in the XML files, and can thus be extracted and compiled in lists, awaiting further analysis. The editors of the DB have also initiated collaboration with Fudan University, which hosts a large project on medieval Chinese phonetic loan characters. Within the framework of a PhD project on medieval Chinese writing (main researcher: Suzanne Burdorf), we also work on the visualisation of the 'social network' of Chinese characters/variant forms, i.e. visualising the various relations a given character form has with other forms, based on phonetic substitutions and/or word family relations, graphic variations, or structural similarities (structural similarities in hand-

3 Workflow and Technical Challenges

The maintenance and development of the DB is time- and resource-intensive, since it has to be periodically updated, adjusted and programmed to include data from current research activities, and the participants of the project have to be coordinated. However, as an international project, work processes and costs are shared between several institutions, and funding has been relatively stable so far. In addition, the DB profits from the work invested in the course of specific PhD and MA projects, and a system of 3-month internships in the framework of the Ghent University MA program.

4 Stable and Flexible Aspects of the Data

Digital tools and web-based DBs are often relatively short-lived, since they have to be continuously hosted and maintained. As such, data management and preservation has become an important issue and has been addressed from the beginning of the project. The project is therefore construed so as to ensure the *long-term preservation of the raw data* in the form of digitised and high-quality marked-up texts in XML format[11] and in accordance with the guidelines of the Text Encoding Initiative (TEI). Once produced, the format of the documents allows easy storage and maintenance and can be universally decoded beyond the limitations of specific research projects.[12] In the further development of the DB we will collaborate with the Ghent Centre of Digital Humanities in order to insure long-term preservation and universal accessibility of the raw data. All textual raw data are made accessible as open-source files.

By contrast, the transformations of these raw data into specific formats and technical environments are by nature more short-lived, based on the need of continuity in the maintenance and – related to

written forms of Chinese characters are one of the main reasons of 'erroneous' substitutions in copying processes).

11 Extensive Markup Language (XML) is an open standard for encoding documents, providing marked-up raw data (in this case textual documents) which can be conveniently transformed into a variety of applications, e.g. into XHTML for web pages, into versions suitable for printing etc. The production of XML documents is a very time-intensive process for the encoder, since the documents have to be well-formed in order to be validated. In order to facilitate the encoding to a certain degree, we use an XML editor (concretely, oXygen). The project generally follows the guidelines of the Text Encoding Initiative (the last version of the manual, TEI P5, consists of 1934 pages! For the mark-up of manuscripts, see especially pages 320-424).

12 All marked-up manuscript texts are freely downloadable and can be used in accordance with the Creative Commons Attribution 3.0 Unported Licence (`https://creativecommons.org/licenses/by/3.0`).

that – in funding. As such, the integration and publication of the raw data as the web-based DCMT is aimed at more short-term goals, based on local research projects, publication strategies, international collaboration, and pedagogical aspects.

5 The Reference Data Collections

The core of the DB project is the collection of texts, consisting of meticulously marked-up manuscript texts, with a focus on the period between ca. 700 and 1000 CE. The late Tang (618-907), Five Dynasties (907-960) and early Song (960-1279) periods are crucial for the study of the development of grammatical markers and semantic items typical for early Mandarin/early 白話 *báihuà* literature. As such, non-canonical texts preserved in the Dunhuang corpus[13] dating from this period are of great significance for reconstructing the early phase of the development of many important features of Mandarin and other Chinese dialects. In the project, we collect a corpus of medieval Chinese texts which is relevant from *various angles of research*. Since the great majority of pre-Song Medieval Chinese texts containing colloquial elements were composed in the context of Buddhism, the DB mainly constitutes a repository of editions of non-canonical Buddhist texts. In addition, several important semi-vernacular literary genres are represented, such as early Chan doctrinal[14] and appraisal texts, 'Transformation texts' (變文 *biànwén*), Avadāna (緣起 *yuánqǐ* / 因緣 *yīnyuán*, i.e. popular versions of narratives concerning the Buddha's life), and Sūtra Lecture texts (講經文 *jiǎngjīng wén*; i.e. vernacular sermons on Buddhist scriptures).[15] All of these text types had an important impact

13 Dunhuang texts are spread in collections around the world (for the main holdings, see Rong 2013). However, a great number of manuscripts have been made publicly available in the form of facsimiles by the International Dunhuang Project (IDP, `http://idp.bl.uk`, London, with mirror sites in Paris and Beijing).

14 Many early Chan texts (especially those attributed to the so-called "Northern School") were contributed to the DB by Marcus Bingenheimer, based on the project *Four Early Chan Texts from Dunhuang. A TEI-Based Edition* (2014-17). The results of this project were also published in a printed form (Bingenheimer, Chang 2018). Although early Chan texts show a lesser degree of vernacularisation as compared to other late Tang genres, they are still of great importance for the study of the colloquial features of the Chinese varieties spoken during the Tang period. Some manuscripts are of special interest, e.g. S.735v, S.2503, S.7961, Beijing 1351v, S.2058, P.2270 etc., which are a treasure grove for researching the earliest predecessors of Modern Mandarin interrogative pronoun 什麼 *shénme* 'what'. In addition, some early Chan texts also show features typical for Northwestern Medieval Chinese (for an overview of scholarship on this historical dialect, see Osterkamp, Anderl 2017).

15 For a short overview of Dunhuang popular literature, see Rong 2013, 398-412. The above genres constitute our most important sources for the study of the spoken language of the late Tang, Five Dynasties and early Song periods. Particularly the Trans-

on the development of various literary genres during the Song period. As the project progresses, we will also try to include other relevant material, such as Tang poems preserved in Dunhuang containing colloquial elements, colloquial (and sometimes bilingual) phrasebooks, schooling texts, lexicographical material etc. This corpus of texts is of great importance for research on early colloquial grammatical markers and syntactic constructions, as well as the development of lexical items. In the current version of the DB, ca. 140 texts are included (representing ca. nine years of work for an experienced encoder) with a rate of ca. fifteen new texts added every year.

6 The Digitisation of the Texts and Their Embedding in the DMCT

The manuscripts are encoded following the guidelines established by Marcus Bingenheimer (in collaboration with DILA), based on the mark-up conventions formulated by the Text Encoding Initiative (TEI). The mark-up focus is on textual features such as variant characters, loan characters and character substitutions (通假字 *tōngjiǎzì*), damaged and unclear passages, added/deleted/repeated characters, punctuation and diacritic markers, abbreviations, notes in the text etc.[16] Mark-up work is very time-consuming and difficult and one professional encoder completes in average ca. 15 manuscript texts per year, depending on the length and difficulties of the texts. After the completion of the mark-up, the texts are sent to Ghent University in XML format, transformed into HTML form and embedded in the DMCT by the project programmers. In DMCT, all texts are visualised in two ways (based on the same XML file), as a 'diplomatic' version (including references to variant characters which are projected as images on the upper right side of the screen, when the cursor moves over a character with a var-

formation texts have received considerable scholarly attention (for the genre features, see for example Mair 1983). Since recently, in the framework of a PhD project, also the variant characters of 祖堂集 *Zutang ji* (ZTJ; 10th century) are in the progress of being integrated in the DB, based on a digitised version of an original print preserved at Kyōto University (see below for more information). Currently, ca. 1,300 variants from the initial fascicles of ZTJ have been input and analysed by Laurent Van Cutsem. For a full list of marked-up texts currently publicly available in the DB, see `https://www.database-of-medieval-chinese-texts.be/views/texts/mcgbd_project/showText.php` and `https://www.database-of-medieval-chinese-texts.be/views/texts/chan_dunhuang/showText.php`.

16 For a full list of features and how they are expressed in the mark-up, see `http://wiki.dila.edu.tw/pages/%E6%95%A6%E7%85%8C%E6%BC%A2%E6%96%87%E4%BD%9B%E6%95%99%E5%AF%AB%E5%8D%B7%E9%BB%9E%E6%A0%A1%E6%9C%AC%E5%B7%A5%E4%BD%9C%E6%89%8B%E5%86%8A`. Variant characters are also cross-checked with the large Taiwanese variant DB, *Dictionary of Chinese Character Variants* (`https://dict.variants.moe.edu.tw/variants/rbt/home.do`).

iant form, in addition to displaying other manuscript features), and a 'regularised'[17] version in which characters are represented in their standard forms and other textual features are resolved into a 'readable' text (frequently, annotations are added in the footnotes, including parallel passages from other manuscripts/texts, as well as references to dictionaries and secondary literature).

The flexibility of the XML format does not only allow various HTML transformations, but can also be used as the basis for a printed edition of a text. Below, I provide a schematic figure of the workflow from manuscript facsimile to TEI-compatible mark-up, and the transformations of the XML file to two HTML visualisations.

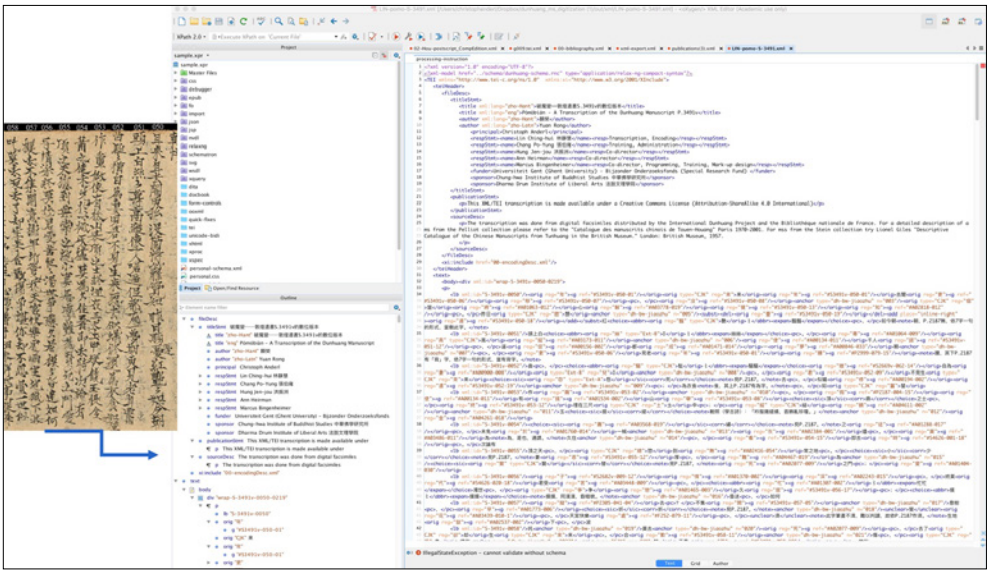


Figure 1.1 Based on the digitised facsimile of the manuscript, the text is encoded in oXygen by a specialist encoder (during the last six years, this work has been performed by Dr. Lin Ching-hui 林靜慧, DILA), following the TEI conventions for manuscript encoding with some adaptations. In addition to basic information (line number, missing/unreadable characters etc.; notes are integrated through an <anchor> element), the focus is on the identification and recording of variant characters. Phonetic loans and other substituted characters are presently only marked with <sic>, awaiting further analysis at a later date (currently, they are integrated in a <choice> element structure, 'X' being a substitution and 'Y' the assumed regularised form), the typical structure being: <choice> <sic>X</sic><corr>Y</corr></choice>. The screenshot shows the mark-up of several lines of the 破魔變 *Pò Mó Biàn* (Transformation [Text] on the Destruction of [Demon King] Māra), lines 50-58 of the manuscript Stein 3491*v*., a Dunhuang manuscript stored at the British Library and a digitised facsimile provided by IDP

17 On details concerning the 'regularisation' of variants, please see the link above (fn. 16).

Figure 1.2 Screenshot exemplifying a typical workflow:the passage encoded in 1.1 is transformed into two types of HTML visualisations in the DMCT. On the left side is a 'diplomatic transcription' with information on many original features of the manuscript preserved (including the projection of variants, here referred to as "non-Unicode characters", on the right upper corner when moving the cursor over passages in light orange). To the right side, a 'regularised transcription' is visualised, with problematic passages resolved into a readable text and including annotations. Note that the ID number of the image of the variant visualised on the right corner indicates its exact positioning in the manuscript, concretely, being character 13 of the column ('line') 50 of Stein 3491 (S3491-50-13). This type of referencing helps us to interlink the graphical variants stored in the Variants Module directly with the corresponding line number of the text in which they appear

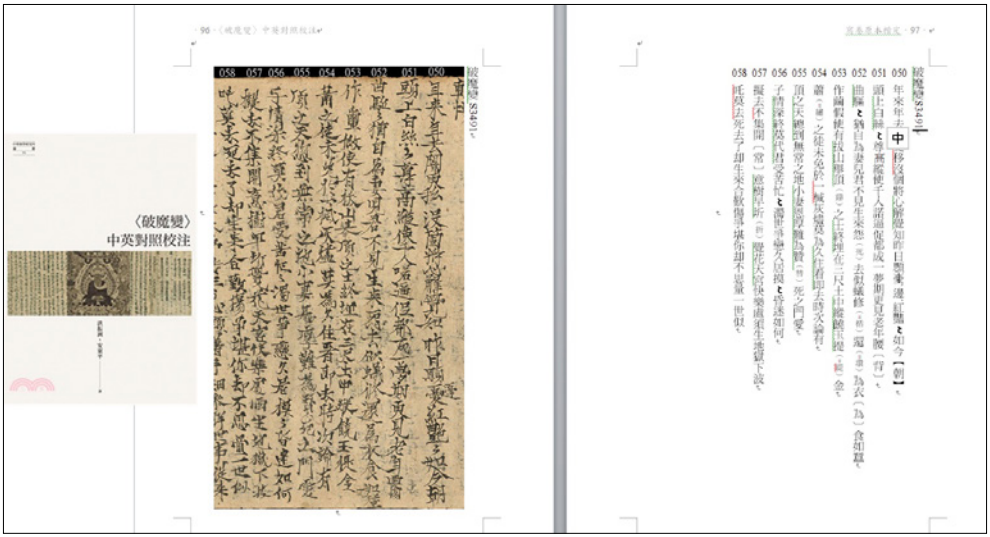


Figure 1.3 Occasionally, in the project, the marked-up XML file of a text will 're-materialise' in the physical form of a printed edition. As such, the circle of a text from the (physical) manuscript to a digitised facsimile, and then to digital versions in XML and HTML formats, returns to the material world in printed form. The figure here shows the same passage discussed in 1.1 and 1.2 as edited text in Lin, Anderl, Hung 2017, 97[18]

7 The Modules of the DB

7.1 The Variants DB Module

Since several research projects at the department deal with graphical variant forms of Chinese characters as encountered in medieval manuscript texts, the mark-up of the variants has become one of the priorities of the DMCT project. The mark-up is not quite homogenous in this respect, based on the fact that it combines the materials of two projects (i.e. the collaborative project with DILA, and prof. Bingenheimer's previous mark-up of early Chan texts). During the latter project, variants were, whenever possible, cross-checked with the *Dictionary of Chinese Character Variants* (DCCV), and the drawings of those graphs extracted and used in the mark-up (using the unique labels of the graphical forms in DCCV). Variants which were neither found in Unicode nor in the DCCV were newly created as drawings (many of these forms are pending to be included in future versions of Unicode fonts).

18 For a detailed description of the process of transforming the XML file into a printed edition, please see `https://bit.ly/3sMQpPF`.

The DMCT project has continued to use those drawings whenever possible, however, every 'new' variant is extracted from the manuscript as an *image*, and integrated as such in the DB. In addition to the Text module, the Variants module is the most developed part of the project, currently featuring ca. 37,000 variant–text passage relations.[19]

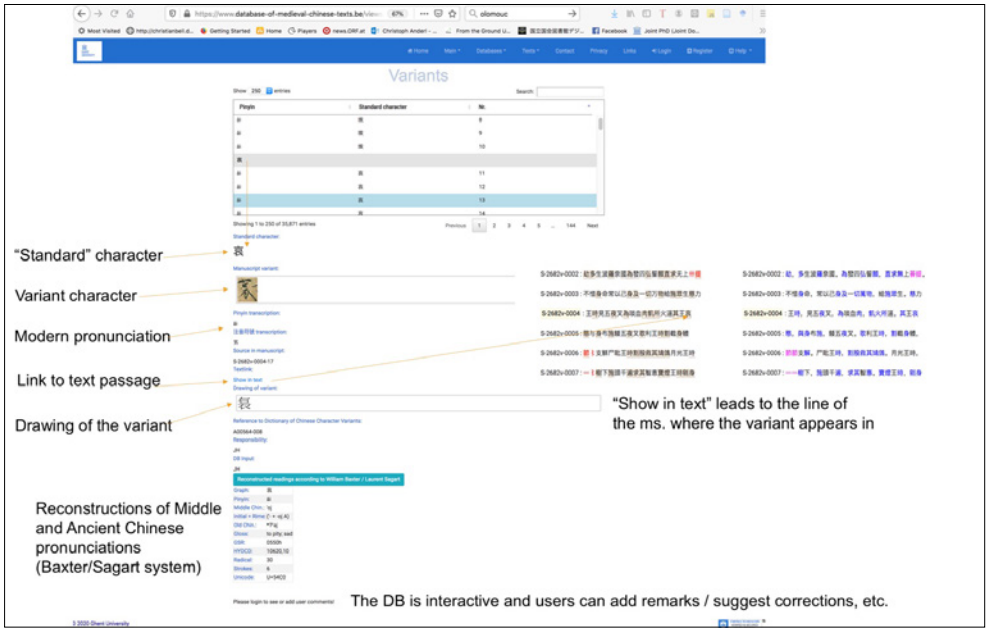


Figure 2.1 This is a screenshot of an entry in the Variants module (a variant of character 哀 *āi*), with explanations of the various fields. The "Source in manuscript" field leads directly to the line of the manuscript the variant appears in (exemplified by the text passage to the right). Since recently, the reconstructed readings of Old and Medieval Chinese, based on the system of Baxter and Sagart, are integrated into the Variants Module

19 In general, we only include variants extracted from Dunhuang manuscripts. However, in the framework of a research project on the ZTJ (a text of crucial importance for studying the vernacular language of the Late Tang and Five Dynasties periods), ca. 1,300 variants were recently input by Laurent van Cutsem (covering the first fascicle of this 20-fascicle work). As a collaborative project with Kyōto University (Zinbun kenkyūjo, Research Institute for Humanistic Studies), the variants are extracted from a digitised version of a unique print of the woodblocks of ZTJ, housed at Haein-sa in Korea (as supplement to the second carving project of the Korean Buddhist Canon in the middle of the 13th century). The textual history of ZTJ – the early parts of which were probably compiled in the middle of the 10th century – is highly complicated. In addition, van Cutsem has recently produced heavily annotated marked-up versions of the two prefaces to the ZTJ (Van Cutsem 2020b, 2020c), as well as to an extensive table and visualisation in Gephi of the lineage system promoted in the text (currently integrated into the DB; see Van Cutsem 2020a).

A very useful feature that enables users to simultaneously view *all registered variants* of a given character was added recently:

Figure 2.2 Screenshot of the function of the DB to collect and visualise all the variants of a specific character registered in the Variants module, here illustrated by the variants of the character 棄 *qì* (clicking on the link in the "Source in manuscript" column, the specific variant can also be viewed as part of the text it appears in). The systematic study of variant forms is of great importance for our understanding of medieval writing practices. Whereas in more 'formal' genres (e.g. copies of canonical Confucian texts, Buddhist *sūtras*, official administrative documents etc.) the character forms are frequently adjusted to contemporary 'standard' (正 *zhèng*) forms, semi-vernacular genres are an important source for actual everyday writing practices, often using popular non-standard forms (俗字 *súzì*). From our example here, showing variants of 棄 *qì* from the 8th to the 10th century, it can be deduced that the dominant popular form for this character during that period was actually very similar to its modern abbreviated counterpart (弃)

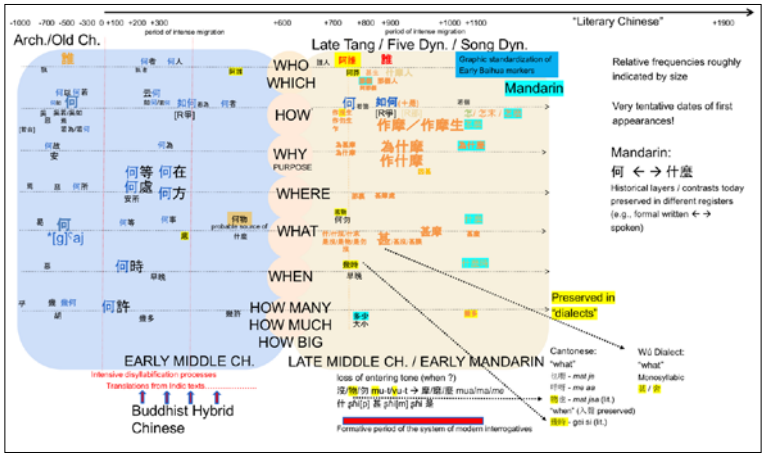


Figure 3 Highly schematic figure of the development of the 'modern' Chinese interrogatives, many of them having their source in the period between 700-1100 (marked with orange colour). The visualisation is based on information extracted from vernacular Dunhuang manuscripts, supplemented with other primary and secondary sources. As can be deduced from the data, a new set of interrogatives started to replace the 何-type system (which appeared frequently in compound form from the early medieval period onward, as evidenced especially in Buddhist texts; the 何 interrogatives are marked with blue colour; the light blue 'box' covers the period of Ancient and Early Medieval Chinese (EMC), before the appearance of the 'modern' interrogatives). By the 10th century, the system of early Mandarin pronouns and their 'standard' orthography had been nearly completely established (marked in light green shading; the beige 'box' covers the period from ca. 700 to 1100, Late Medieval Chinese). Other pronouns evidenced by medieval manuscript material survived in other Chinese dialects (marked with yellow shading). In the figure it is also shown how external features influenced the development and spread of interrogatives, e.g. disyllabification processes since the beginning of EMC, as well as the development of 'Buddhist Hybrid Chinese', a new type of Literary Chinese mixed with vernacular elements and 'Sanscritisms' heavily influenced by translation processes from Indic languages into Chinese. Other external factors include intensive migration events between ca. the 2nd and the 4th century, and then again between the 8th and the 10th century

7.2 Syntax Module

In this part of the DB, information on syntactic markers of Late Medieval Chinese (LMC) are collected. The information on these markers is extracted from texts collected in the Text Module, external text corpora (such as SAT and CBETA), additional Dunhuang manuscript material, as well as relevant secondary literature. The module aims at functioning as a *reference tool*, providing information on the use of LMC function words, their historical development, their orthography as encountered in manuscripts, their relation to other function words etc.[20] The use of the markers is illustrated by example sen-

20 The fields in the input interface also include information on (historical) pronunciations, notes on variants and phonetic loans used for the marker, dictionary references, as well as references of occurrences in primary and secondary sources. Since the information provided on the function words is still fragmentary, this part of the DB has not yet been opened to the public.

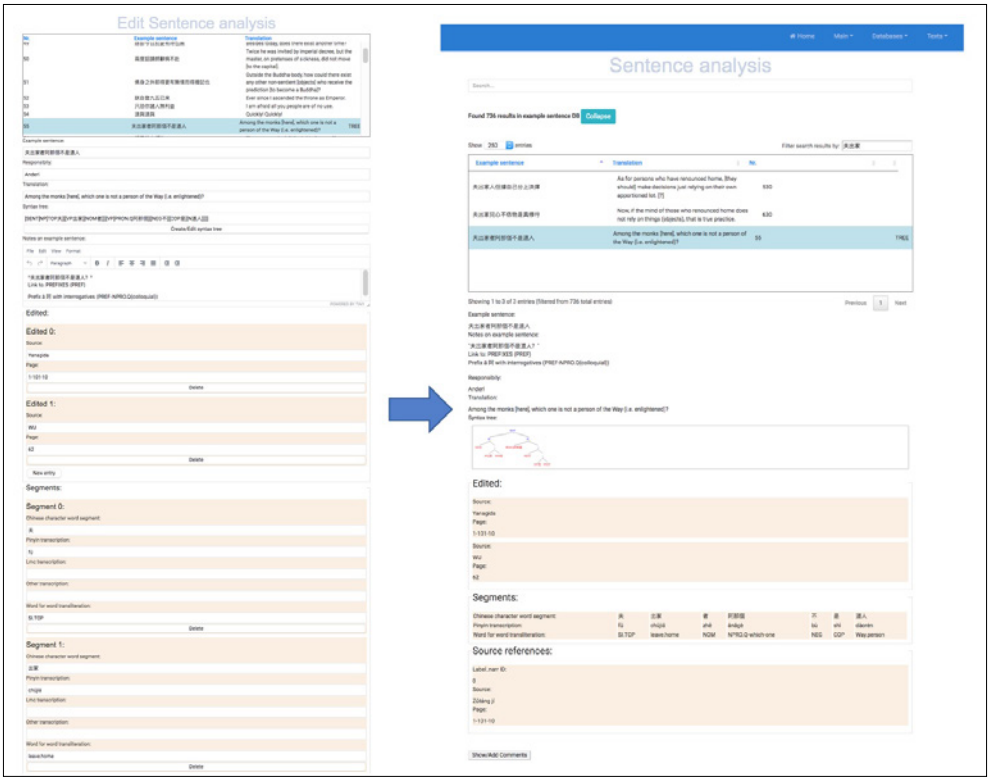


Figure 4 The left side shows the input mask of the Sentence Analysis Module, featuring a segmentation tool (each segment has fields for the word in Chinese characters, the *pinyin* reading, reconstructed LMC readings, as well as word-for-word transliterations), a tree generator, in addition to several fields for various references (e.g. translation, notes, editions etc.). On the right side of the figure, the HTML transformation of the interface entry into a page of the Sentence Analysis Module of the DMCT is shown. The entries in this module can be linked to the respective entries in the Syntax Module (in the example above, to the entry on prefix [jiā])

tences (collected in the Example Sentence Module and linked to the respective entries in the Syntax Module), as well as links to the line where they appear in the digitised manuscripts of the DB. The individual entries (currently ca. 700) can also be arranged to form 'chapters' (e.g. on classifiers, or interrogatives etc.), and we aim at developing this feature in our future work on the DB (ideally, this module can eventually be used as a 'reference grammar'). The Syntax Module plays an important role in the department's research on Chinese historical syntax (for an example, see **fig. 3**).

7.3 Sentence Analysis Module

This module is interrelated with the Syntax Module (which is descriptive in nature and records the basic functions and the historical development of a marker) and serves the purpose of illustrating and analysing the functional realms of syntactic markers by presenting examples of their usage in phrases/sentences. The interface contains fields for the example sentence and its translation, notes on the phrase/sentence, a segmentation tool, and the possibility to include a tree analysis **[fig. 4]**.

7.4 Chan Phrases Module

This DB module has been recently added in order to accommodate the results of an ongoing PhD project[21] on the syntax and semantics of 4-character Chan phrases of the Song dynasty, which are often contextually and pragmatically encoded, and the meaning of which is frequently very difficult to retrieve.[22] In addition, these phrases often contain dialect and local vernacular expressions (some of them still preserved in modern dialects), and are as such important sources for the historical development of lexical items.[23] The module aims at collecting these 4-character phrases which play an important role in the rhetorical structure of colloquial Chan texts of the Song and thereafter, register the source texts they appear in, collect referenc-

21 The material of this module has been mainly collected by Zeng Chen 曾辰 (researcher of Sichuan and Ghent Universities in the framework of a Joint PhD project). Currently, most data are collected in spread sheets, including thousands of Chan phrases with references to their sources. In the further work processes, these data sets will be imported into the Chan Phrases Module. As a sub-project concerning this part of the DB and the research related to it, we will focus on the identification of dialect elements in Chan phrases, as well as try to trace their development from their historical sources to Modern Chinese dialects (the results of this work will be also presented in the form of a joint research paper, currently in production).

22 In addition, these phrases were often alluded to and commented on in later works, as well as re-embedded in new contexts.

23 Some of these semantic items spread even 'internationally'; a famous example is 挨拶 *āizā* 'come close and squeeze > to check; to probe' (in the Chan context, often concretely referring to engaging in an exchange of questions and answers about the Buddhist teaching), which first appeared in a Song Dynasty Chan text in the phrase 一挨一拶 *yī ái yī zā* (圓悟佛果禪師語錄 *Yuanwu Foguo chanshi yulu* 'The Recorded Sayings of Chan Master Yuanwu Foguo'; CBETA, T.47, no. 1997, p. 756, b20-5; for another example, see CBETA, T.47, no. 1998A, p. 915, b18-24). After Chan (Jap. Zen) was introduced in Japan during the 12th/13th century, the word 挨拶 *āizā* started spreading beyond the confines of the monastic communities, eventually becoming a high-frequency word with the meaning 'to greet sb. (formally)' (Jap. あいさつ *aisatsu*). In this meaning the word was probably re-introduced to China and is preserved as loanword in the Minnan dialect (*ai*^{35}*sat*^{5}*tsuh*^{3}).

es from historical and contemporary secondary material, analyse their syntactic structure and provide tentative English translations, as well as trace their path of development **[fig. 5]**.

Figure 5 Screenshot of an entry in the Chan Phrases Module (the phrase 鼻孔累垂 *bíkǒng léichuí*). The entry provides a description of the phrase, a tree analysis, sources in primary texts and references in secondary literature, links to related phrases etc. In addition, occasionally the path of development of semantic items is traced (i.e. the usage in modern Chinese dialects). Here 累垂 *léichuí* is traced to Cantonese *lœy^{21}-sœy^{21}*, which has preserved the original semantic ('to hang; dangle') of the word

8 The DB as a Pedagogical Tool

The above description focused on the DB as a tool for research on medieval Chinese texts. An additional important aspect is the integration of the DB into the teaching environment of advanced master student courses at the Department of Languages and Cultures, Ghent University. The materials provided by the DB are regularly used in classes on Chinese Buddhist texts and culture, as well as for training the students in manuscript decipherment, historical Chinese writing conventions, medieval Chinese syntax and semantics. The materials are also used to compare the Dunhuang Buddhist narratives edited in the DB to their 'canonical' versions, in order to demonstrate how

key narratives have been adapted in terms of contents, language, and genre features to specific audiences (e.g. the vernacularisation processes one can observe in many manuscript versions, in order to adapt a narrative to a Chinese general audience).[24] In the master course, students also have to produce annotated translations of selected parts of the specific Dunhuang text discussed during the term. For the future development of the DB, we plan to feed the results of the master courses back into the DB, for example as revised and edited versions of the translations jointly produced by the students.

In addition to training master students in a classroom environment, the DB has also served as the basis for several master theses on Chinese Buddhist texts and/or Medieval Chinese linguistics.[25] Another aspect, which has become increasingly important during the last years, is the possibility to work on the DB in the framework of obligatory internships which master students have to perform as part of their master education (ca. 240 work hours). Most of the work is performed online (e.g. collection of materials, input of the materials into specific modules, analytical work etc.), in addition to regular meetings with the supervisor. This aspect related to the education of master students in the framework of the writing of their theses, as well as the internships,[26] have proven very promising in the development of the DB, and provides the students with an efficient training platform for working with (manuscript) texts; at the same time, it generates manpower for refining and expanding the DB.

24 As a concrete example, the master course *Buddhism. Texts and Material Culture* (MA, Spring 2020) dealt with the conversion story of Nanda (who figures as one of the main disciples of Śākyamuni in Buddhist scriptures), comparing canonical versions with the 因緣 *yīnyuán* genre version preserved among the Dunhuang manuscripts. The students gained reading practice in both Buddhist Hybrid Chinese (i.e. the language of Buddhist translation literature), as well as the semi-vernacular of the Dunhuang manuscript version. In addition to the philological/linguistic aspects, the students would become familiar with various genre features and would analyse the literary structure of the various versions (which emphasise different aspects of the story).

25 In the most recent master thesis, a student analysed the structure of prepositional phrases based on the data provided by DMCT (Dewaele 2019). Methodologically, the candidate extracted all prepositional phrases from the texts published in the DB, and analysed them comparatively and diachronically, as well as sorted by genre. Another recent master thesis dealing with vernacular Dunhuang materials is van Rentergem 2019, analysing the Buddha biographies of the so-called 八相變 *bāxiàng biàn* genre (transformation of the eight [main] events [of the Buddha's life]).

26 Internship assignments of 2020-21 will focus on the input of character variants of the earliest period of Dunhuang manuscripts, dating from the mid-fifth and early sixth centuries (see Silk, Galambos 2017), and the comparison of several Dunhuang version of the 搜神記 *Soushen ji* (Records of the Search for the Supernatural).

9 Final Reflections

DBs and digital collections of textual materials have become indispensable tools in the field of corpus linguistics. While typical corpora are repositories of text samples reflecting natural languages, collections of premodern texts necessarily will feature a number of particularities in terms of the selection, gathering, and the preparation of texts, as well as concerning the 'mining' and analysis of linguistically meaningful data. While there are a variety of large digital DBs available for premodern Chinese texts,[27] specialised DBs on non-canonical manuscript materials (which are of paramount importance for research in the culture and language of the Late Medieval period) are still very rare and the information they provide is rather limited. Establishing the DMCT is an attempt to fill this gap, by providing high-quality digital editions of LMC key texts, and develop an analytical apparatus dealing with this type of manuscript material. As described above, the DB also has a 'socio-institutional' function, trying to address the specific research constellation at our department, and providing material for both more Buddhologically oriented, and linguistic studies.

In addition to fulfilling its main function of producing and providing high-quality marked-up medieval text versions, the DB project is driven by specific research interests and topics, and is as such in a permanent state of change and evolution. Accordingly, the DMCT is built as a system of interconnected modules, each module fulfilling a certain function and being embedded in a specific research context (predominantly PhD research projects and international collaborative projects).

In order to widen its significance – justifying the considerable investment of work power and financial resources – the DB has also become an important element in the training of advanced master students, exchange students from China, in addition to being used in the framework of internships. The work invested in the DB in the framework of these pedagogical contexts is also an important source for expanding the scope of the DB by feeding the produced data and research results back into the DB.

27 In addition to those already mentioned, large DBs suitable for research in Chinese historical linguistics include: `www.cncorpus.org`, provided by Peking University and including both Chinese modern and premodern text collections; a variety of large text DBs offered by Academia Sinica, Taiwan (`http://www2.ihp.sinica.edu.tw/index.php`), including the Scripta Sinica Database (which comprises ancient and medieval Chinese texts, consisting of more than 700 million characters); and the huge number of premodern texts provided by CTEXT (`https://ctext.org`).

Bibliography

Anderl, C. (2018a). "Linking Khotan and Dunhuang. Buddhist Narratives in Text and Image". *Entangled Religions*, 5, 250-311.

Anderl, C. (2018b). "Metaphors of 'Sickness and Remedy' in Early Chán Texts from Dunhuang". Edzard, Borgland, Hüsken 2018, 27-46.

Anderl C.; Sørensen, H. (2020-21). "Northern Chán and the Siddhaṁ Songs". Anderl, C.; Wittern, C. (eds), *Chan Buddhism in Dunhuang and Beyond. A Study of Manuscripts, Texts and Contexts in Memory of John R. McRae*. Leiden: Brill, 99-139.

Bingenheimer, M.; Chang P.-Y. (eds) (2018). *Four Early Chan Texts from Dunhuang. A TEI-Based Edition*. Taipei: Shin Wen Feng.

Chen, J. (project director). *From the Ground Up. Buddhism and East Asian Religions*. Vancouver: University of British Columbia. https://frogbear.org.

CBETA = *Zhonghua dianzi fodian xiehui* 中華電子佛典協會 (Chinese Buddhist Electronic Text Association). https://www.cbeta.org.

DDB = *Digital Dictionary of Buddhism*. Ed. in chief: Charles Muller. Tokyo University. http://www.buddhism-dict.net/ddb/.

DCCV = *Dictionary of Chinese Character Variants* (2017). Taipei: Ministry of Education. https://dict.variants.moe.edu.tw/variants/rbt/home.do.

DMCT = *Database of Medieval Chinese Texts*. Ghent University, Belgium and Dharma Drum Institute of Liberal Arts, Taiwan. https://www.database-of-medieval-chinese-texts.be.

Dewaele, J. (2019). *On Coverbs and Prepositions in Late Medieval Chinese. A 'Field' Study and Diachronic Perspective Based on Early Chan and Dunhuang Avadāna Texts* [MA thesis]. Ghent: Ghent University.

Edzard, L.; Borgland, J.W.; Hüsken, U. (eds) (2018). *Reading Slowly. A Festschrift for Jens E. Braarvig*. Wiesbaden: Harrassowitz.

IDP = *The International Dunhuang Project*. http://idp.bl.uk.

Lin C.-H. 林靜慧; Anderl, C.; Hung C.-C. 洪振洲 (2017). "*Po Mo bian*" *zhong-ying duizhao jiaozhu*《破魔變》中英對照校注 ["Po Mo Bian" Critical Edition with Annotated Translations into Modern Chinese and English]. Taipei: Fagu wenhua.

Mair, V. (1983). *Tun-huang Popular Narratives*. Cambridge: Cambridge University Press.

Osterkamp, S.; Anderl, C. (2017). "Northwestern Medieval Chinese". Sybesma, R. et al. (eds), *The Encyclopedia of Chinese Language and Linguistics*, vol. 3. Leiden: Brill, 218-29.

Rong X. (2013). *Eighteen Lectures on Dunhuang*. Translated by I. Galambos. Leiden: Brill.

SAT = *The SAT Daizōkyō Text* Database. https://21dzk.l.u-tokyo.ac.jp/SAT/index_en.html.

Silk, J.; Galambos, I. (2017). "An Early Manuscript Fragment of Dharmarakṣa's Translation of the **Ajātaśatrukaukṛtyavinodana*". Edzard, Borgland, Hüsken 2018, 409-31.

Sūn C.-W. 孫昌武; Kinugawa K. 衣川賢次; Nishiguchi Y. 西口芳男 (eds) (2007). *Zutang ji* 祖堂集 (Collection from the Patriarchs' Hall). 2 vols. Beijing: Zhonghua shuju.

TEI P5 = *Guidelines for Electronic Text Encoding and Interchange*. Edited by the Technical Council of the TEI Consortium. Text Encoding Initiative Consortium, July 2019. https://tei-c.org/guidelines/p5.

The Dunhuang Research Academy. Dunhuang yanjiuyuan 敦煌研究院. http://public.dha.ac.cn/index.html.

TLS = *Thesaurus Linguae Sericae*. https://hxwd.org/index.html.

Van Cutsem, L. (2020a). *The Zutang ji* 祖堂集 *(K. 1503; B25, No. 0144). A Comprehensive.xlsx Table on its Contents and Structure*. Draft Version. Ghent: Ghent University and Database of Medieval Chinese Texts.

van Cutsem, L. (2020b) "Chán Master Jìngxiū's 淨修禪師 Preface to the Zǔtáng jí 祖堂集 (K.1503): A TEI/XML-Based Edition". Database of Medieval Chinese Texts. Ghent University and Dharma Drum Institute of Liberal Arts 法鼓文理學院. https://www.database-of-medieval-chinese-texts.be/views/texts/zutang_ji/showText.php.

van Cutsem, L. (2020c). "The Goryeo 高麗 Preface to the Zǔtáng jí 祖堂集 (K.1503): A TEI/XML-Based Edition". Database of Medieval Chinese Texts. Ghent University and Dharma Drum Institute of Liberal Arts 法鼓文理學院. https://www.database-of-medieval-chinese-texts.be/views/texts/zutang_ji/showText.php.

Van Rentergem, S. (2019). *A Study of the Dunhuang Baxiangbian. With Annotated Translations* [MA thesis]. Ghent: Ghent University.

Zutang ji 祖堂集 (Collection from the Patriarchs' Hall). Scanned Copy of an Original Print of the second *Goryeo Daejanggyeong* 高麗大藏經 Woodblock Edition (1245) of the *Zutang ji* Stored at the Library of the Institute for Research in Humanities of Kyōto University, Japan.

Zutang ji 祖堂集 (Collection from the Patriarchs' Hall). Digital versions of the text: https://raw.githubusercontent.com/cbeta-org/xml-p5/master/B/B25/B25n0144.xml and https://cbetaonline.dila.edu.tw/zh/B0144.

Bio-bibliographies

Christoph Anderl Christoph Anderl holds a PhD in Chinese linguistics (Oslo 2005) and has been Professor of Chinese Language and Culture at Ghent University since 2015. He is currently also a Research Cluster leader in the interdisciplinary project *From the Ground Up: Buddhism and East Asian Religions* (UBC), investigating text-image relations at Medieval Chinese Buddhist sites, and editor in chief of the Database of Medieval Chinese Texts. His research focuses on Late Medieval Chinese, Buddhist Chinese, Dūnhuáng manuscripts, aspects of Chinese Buddhism (Chán), and text-image relations in the transmission of Buddhist narratives. Recent publications include the monograph 【破魔變】中英對照校注 - *Pò Mó biàn* critical edition with annotated translations into Modern Chinese and English (with Lin Ching-hui 林靜慧 and Hung Chen-chou 洪振洲, Taipei 2017), the edited volumes *Chán Buddhism in Dūnhuáng and Beyond: A Study of Manuscripts, Texts and Contexts in Memory of John R. McRae* (with C. Wittern. Brill, 2020-21) and *Buddhist Encounters and Identities Across East Asia* (with C. Meinert and A. Heirman. Brill, 2018), as well as several papers on linguistics published in the *Journal of Chinese Linguistics*, the *Cahiers de Linguistique Asie Orientale*, and the Brill *Encyclopedia of Chinese Language and Linguistics*. For further publications, please consult `https://ugent.academia.edu/ChristophAnderl`.

Sofia Bareato Sofia Bareato holds a double degree title: Master's Degree in Languages and Civilisations of Asia and North-Africa from Ca' Foscari University of Venice, with a thesis on derivation in Mandarin Chinese, and Master's Degree of Teaching Chinese to Speakers of Other Languages from Capital Normal University, Beijing. She also obtained a Master's degree in Teaching Italian to foreigners from the University for Foreigners of Perugia. She is currently a secondary school teacher of Chinese language and culture in Milan.

Bianca Basciano Bianca Basciano is Associate Professor of Chinese at Ca' Foscari University of Venice. She obtained a PhD in Linguistics from the University of Verona with a thesis entitled *Verbal Compounding and Causativity in Mandarin Chinese*. Her research focuses on Chinese morphology and the syntax-semantics interface, especially on compounding, reduplication, resultatives, and causative constructions. She wrote a number of research papers on these topics. She also authored several entries of the Brill *Encyclopedia of Chinese Language and Linguistics* and co-authored the entry on Morphology in Sino-Tibetan languages in the *Oxford Research Encyclopedia of Linguis-*

tics. She is co-author of the book *Chinese Linguistics: An Introduction* (Oxford University Press, forthcoming).

Adriano Boaretto Adriano Boaretto is research fellow in Chinese language at Ca' Foscari University of Venice His research interests concern the grammar of contemporary Mandarin Chinese, with a focus on the syntax of relative clauses and the aspect system of Chinese (e.g. "Corrispondenti Funzionali Cinesi della Frase Relativa Italiana: alcune implicazioni dal punto di vista pedagogico". *Anna Maria Palermo, Atti del IX Convegno dell'Associazione Italiana di Studi Cinesi, "La Cina e l'Altro"*. Il Torcoliere, 311-21). He has also researched the differences between the variety of Chinese spoken in Mainland China and that of Taiwan (e.g. "Alcune osservazioni sulle differenze tra il cinese parlato nella Repubblica Popolare Cinese e quello parlato nella Repubblica di Cina". *La lingua cinese: variazioni sul tema*. Edizioni Ca' Foscari, 2015).

Erik Castello Erik Castello is Associate Professor of English Language and Linguistics at the University of Padua. His research interests include (learner) corpus linguistics, discourse analysis, and English language teaching and testing. He has recently published several articles on these topics (e.g. "Holding Up One's End of the Conversation in Spoken English: Lexical Backchannels in L2 Examination Discourse". *International Journal of Learner Corpus Research*, 5(2), 2019); "Pope Francis's *Laudato Si'*: A Corpus Study of Environmental and Religious Discourse", with S. Gesuato. *Lingue e Linguaggi*, 29, 2019). He has also co-edited a volume on Learner Corpus Research (*Studies in Learner Corpus Linguistics: Research and Applications for Foreign Language Teaching and Assessment*, with K. Ackerley and F. Coccetta. Peter Lang, 2015) and the special issue of *Lodz Papers in Pragmatics*, "Assessing Pragmatic Aspects of L2 Communication: Reflections and Practices" (16(1), 2020, with S. Gesuato, 2020).

Long Chen Long Chen received his Bachelor's degree in Applied Linguistics from Peking University in 2018. He is currently a graduate student in the Department of Chinese Linguistics and Literature, Peking University. His research interests include Chinese information processing, language knowledge engineering, applied linguistics, and computational linguistics.

Andy Chin Andy Chin is currently Head of the Department of Linguistics and Modern Language Studies, The Education University of Hong Kong. His research interests include Chinese linguistics, linguistic typology, sociolinguistics, corpus linguistics, discourse analysis. He received a number of awards in research such as the Young Scholar Award of The International Association of Chinese Linguistics (2009) and the LFK Young Scholar conferred by The Li Fang-Kuei Society for Chinese Linguistics (2013). In 2012, he started the construction of the Corpus of Mid-20th Century Hong Kong Cantonese, with an aim to provide authentic language data for Cantonese linguistic research, especially in the diachronic and discourse dimensions. This corpus won the Gold Medal and Special Award in the Silicon Valley International Invention Festival in 2019. He has published in *Journal of Chinese Linguistics*, *Language and Linguistics*, *Bulletin of Chinese Linguistics*, *Bulletin of the School of Oriental and African Studies*, *Minzu yuwen* 民族語文, *Yuyanxue luncong* 語言學論叢.

Aneta Dosedlová Aneta Dosedlová received her Master's degree from the Chinese Department of the Faculty of Arts, Masaryk University. Her research interest is corpus-cognitive linguistics.

Franco Gatti Franco Gatti is Associate Professor of Chinese Language and Literature at Ca' Foscari University of Venice. He obtained a PhD in Chinese Language, Literature and History from the Sapienza University of Rome. His research interests include Chinese language, Chinese linguistics, and Chinese literature of the Tang period. He is currently working on an annotated translation of the *Xuanshi zhi* 宣室志 by Zhang Du 張讀 (fl. late 9th century).

Haibin Huang Haibin Huang is now working as a web developer in Bytedance Inc., Beijing. He received his Master's degree from Peking University in 2020. He is interested in Natural Language Processing and in exploring the mystery of language with the algorithms of machine learning and deep learning.

Hong Gang Jin Hong Gang Jin is currently William R. Kenan Professor of Chinese Language and Culture Emeritus at Hamilton College in the US. She was Chair Professor of Applied Linguistics at the University of Macau for 5 years. With her PhD in Educational psychology and second language acquisition from the University of Illinois, Jin researches in areas of cognition and second language learning, second language processing, and second language teacher development. She has published 7 books and textbooks and over 60 book chapters and articles in refereed journals in the US and China.

Zhuo Jing-Schmidt Zhuo Jing-Schmidt is professor of Chinese Linguistics at the Department of East Asian Languages and Literatures, University of Oregon. She holds a PhD from the University of Cologne, Germany, and publishes in English, German, and Chinese on topics related to language and cognition, emotion, gender, digital media and language, linguistic typology, historical linguistics, and acquisition of Chinese as a second language.

Sophia Xiaoyu Liu Sophia Xiaoyu Liu is a doctoral student at the department of East Asian Languages and Literatures, University of Oregon. She is interested in perceptions of dialects, corpus linguistics, quantitative methods using R, and Chinese as a second language pedagogy.

Wei-lun Lu Wei-lun Lu, PhD, is Assistant Professor at the Department of Chinese Studies and the Language Center of Masaryk University. Dr. Lu has research interests in cognitive-oriented contrastive analysis that involves Chinese, with an emphasis on the cultural, stylistic, and poetic ramification of the linguistic tool. He is Special Assistant to Director for Strategic Development (Language Center) and Language Program Coordinator (Department of Chinese Studies). Dr. Lu is currently a Council Member of European Association for Chinese Teaching (2019-21) and is also involved in the following professional organisations: Czech Association for Language and Cognition, Association for Researching and Applying Metaphor, and Linguistic Society of Taiwan. He is currently a Review Editor of *Frontiers in Psychology* (Language Sciences) and serves on the editorial board of *Asian-Pacific Journal of Second and Foreign Language Education* (Springer), *Studia Orientalia Slovaca* (the only Sinological journal in Slovakia), and the book series "Cultural Linguistics" (Springer).

Anna Morbiato Anna Morbiato is Assistant Professor (RTD/A) at Ca' Foscari University of Venice and Research Affiliate at the University of Sydney. She holds a PhD in Linguistics from the University of Sydney and a PhD in Asian and African studies from Ca' Foscari University of Venice. She publishes in English, Italian, and Chinese on topics related to language and cognition, syntax-semantics-pragmatics interface, contrastive linguistics,

and second language acquisition, with a focus on Mandarin Chinese, English, and Italian. She also conducts research in frame semantics and NLU.

Heidi Hui Shi Heidi Hui Shi is a PhD candidate at the Department of East Asian Languages and Literatures, University of Oregon. Her research interests include gender and language, gender socialisation, cognitive linguistics, corpus linguistics, digital media and language, Chinese as a second language pedagogy, and quantitative methods using R. Her language areas include Chinese, Korean, and English.

Carlotta Sparvoli Carlotta Sparvoli is Associate Professor at the University of Bologna. In 2012, she was awarded a PhD from Ca' Foscari University of Venice and in the same year she won a six-month research grant within the Taiwan Fellowship Program. From 2012 to 2015, she conducted a post-doc research at the University of Parma. Between 2016 and 2019, she served as Director of the MA programme in Teaching Chinese to Speakers of Other Languages at the School of Asian Studies of University College Cork (Ireland). She published one monograph and numerous research papers on peer reviewed journals and edited volumes. Her most recent publication is "Modality in the general linguistic investigations carried out in China before 1949" (in Meisterernst, B. (ed.) *New Perspectives on Aspect and Modality in Chinese Historical Linguistics*. Springer, 2019). She is currently contributing to the *Oxford Research Encyclopedia of Linguistics*, serves in the editorial board of *Chinese as a Second Language Research* (De Gruyter Mouton), and is also active as reviewer and external examiner for several academic journals and institutions.

Vittorio Tantucci Vittorio Tantucci is Lecturer of Chinese and Linguistics at Lancaster University, UK. His publications focus on usage-based intersections of pragmatics and cognition. These issues are addressed typologically and cross-culturally, from both a synchronic and a diachronic perspective. His recent major publications include *Language and Social Minds: The Semantics and Pragmatics of Intersubjectivity* (Cambridge University Press, forthcoming); "Diachronic Change of Rapport Orientation and Sentence-Periphery in Mandarin" (*Discourse Studies*, 22(2), 2020; authored with A. Wang), "From Co-Actionality to Extended Intersubjectivity: Drawing on Language Change and Ontogenetic Development" (*Applied Linguistics*, 41(2), 2020), "From Co-actions to Intersubjectivity Throughout Chinese Ontogeny: A Usage-Based Analysis of Knowledge Ascription and Expected Agreement" (*Journal of Pragmatics*, 167, 2020).

Hongyin Tao Hongyin Tao is Professor of Chinese language and linguistics and applied linguistics at UCLA; he also holds a honorary Distinguished Chair Professor position at the National Taiwan Normal University. His research areas include corpus linguistics, Chinese discourse and grammar, and applications of linguistic research to language teaching and learning. Among his over 130 publications are the recent books *Chinese for Specific/Professional Purposes* (Springer, 2019), *Integrating Linguistics Research with Chinese Language Teaching and Learning* (John Benjamin, 2016), *Chinese under Globalization* (World Scientific, 2011), and *Working with Spoken Chinese* (Penn State University, 2011). He serves on over 20 editorial boards of journals and book series, and was the 2014 President of the Chinese Language Teachers Association, USA.

Aiqing Wang Aiqing Wang is a Senior Teaching Associate in Chinese at the Department of Languages and Cultures, Lancaster University. Her PhD project investigates clause-internal preposing in Late Archaic Chinese. Apart from syntax and pragmatics, her research areas also include historical linguistics and cultural studies.

Jiajun Wang Jiajun Wang is a PhD student in the Department of Chinese Language and Literature at Peking University. He received his Master's degree from Shanghai International Studies University in 2017. His research interests include feature- and unification-based grammatical theory, language resource development, statistical machine learning, and natural language processing.

Weidong Zhan Weidong Zhan is Professor at the Department of Chinese Language and Literature, Peking University. His main research areas are modern Chinese formal grammar, language knowledge engineering, and Chinese information processing. His PhD dissertation, *A Study of Constructing Rules of Phrases in Contemporary Chinese for Chinese Information Processing*, was published in 2000. He participated in the compilation of two textbooks, *Modern Chinese* (Higher Education Press, 2014) and A*n Introduction to Computational Linguistics* (The Commercial Press, 2003). He is the first author of the amendment of national standard, titled as "General Rules for Writing Numerals in Publishing Texts" (GB/T 15835-2011). He also compiled and published a book as a user guide of the standard in 2012. He published dozens of articles in leading academic journals of China. He was awarded as "New Century Outstanding Scholar" in 2012 and "Changjiang Outstanding Young Scholar" in 2017 by the Ministry of Education of the People's Republic of China.

Jie Zhang Jie Zhang is Associate Professor of Chinese Pedagogy and Applied Linguistics in the Department of Modern Languages, Literatures, and Linguistics at the University of Oklahoma, USA. She received her PhD in Applied Linguistics from the Pennsylvania State University. Her research interests are second language acquisition, foreign language pedagogy, and Chinese as a second language. She has published in the *Modern Language Journal*, *Language Testing*, *Language Teaching Research*, *Chinese as a Second Language*, *Teaching Chinese in the World*, among others. She is co-editor of the volume *Chinese Language Education in the United States* (Springer, 2016).

This volume collects papers presenting corpus-based research on Chinese language and linguistics, from both a synchronic and a diachronic perspective.
The contributions cover different fields of linguistics, including syntax and pragmatics, semantics, morphology and the lexicon, sociolinguistics, and corpus building.
There is now considerable emphasis on the reliability of linguistic data: the studies presented here are all grounded in the tenet that corpora, intended as collections of naturally occurring texts produced by a variety of speakers/writers, provide a more robust, statistically significant foundation for linguistic analysis.
The volume explores not only the potential of using corpora as tools allowing access to authentic language material, but also the challenges involved in corpus interrogation, analysis, and building.

Università
Ca'Foscari
Venezia