A distinctive collexeme analysis of near-synonym constructions “*ying-dang/ying-gai* + verb”

Zhuo Zhang
Department of Linguistics
and Translation
City University of
Hong Kong
83 Tat Chee Avenue,
Kowloon, Hong Kong, China
jessyzh@pku.edu.cn

Meichun Liu
Department of Linguistics
and Translation,
City University of
Hong Kong,
83 Tat Chee Avenue,
Kowloon, Hong Kong, China
meichliu@cityu.edu.hk

Dingxuan Zhou
School of Mathematics
and Statistics,
University of Sydney,
Sydney NSW 2006,
Australia
dingxuan.zhou@
sydney.edu.au

Abstract

This paper aims to differentiate the two modal auxiliary near-synonyms *ying-dang* ‘should; ought to; shall’ and ying-gai ‘should; ought to’ through investigating their distinctive collexemes of the near-synonym constructions of “*ying-dang*/ *ying-gai* + verb”. We employed the distinctive collexeme analysis (Gries and Stefanowitsch, 2004) to find the distinctive verbal collexemes and then categorized them into different semantic types based on the notion of frames (Fillmore, 1982). The sample was extracted from a news corpus of 2.95 billion Chinese characters. The results found that *ying-dang* and *ying-gai* display critical differences in attracted collexemes regarding modality types, usage patterns, and semantic types of verbs. Specifically, *ying-dang* prefers to exhibit a robust deontic meaning, whereas *ying-gai* can display both deontic and epistemic senses. *Ying-dang* also tends to take two-character verbs and appears on formal occasions, whereas *ying-gai* likes single-character verbs better and shows up in various informal contexts. In terms of semantic frames, *ying-dang* is inclined to take verbs with purposive efforts, whereas *ying-gai* attracts verbs of self-motion and emotion.

1 Introduction

Modal auxiliary verbs and adverbs are crucial in understanding the intention of a speaker. The Chinese modals pair *ying-dang* ‘shall; ought to; should’ and *ying-gai* ‘should; shall’ is considered as should-modals (similar to should) with overlaps in semantics and functions. Despite native speakers themselves may not be able to explain why they choose one word over another, their choices between near-synonym modals are found to display a distinctive trend in distribution (Hilpert and Flach, 2021). The distribution hypothesis that the semantics of a word is reflected by the company it keeps (Firth, 1957; Harris, 1970: 785; Turney and Pantel, 2010) may offer a potential theoretical foundation for this phenomenon. The collocational preferences can be distinguished between near-synonym constructions (Gries and Stefanowitsch, 2004; Hilpert, 2008) as illustrated by usage-based Construction Grammar (Goldberg, 1995; Diessel, 2019). Also, the divergence of collocational profiles associated with different words could offer vital clues to their differentiated semantics (Hilpert and Flach, 2021).

The subtle usage and semantic differences of modals can also be found through the pair-wise comparison of near-synonym constructions. Hilpert and Flach (2021) found that English modals could be differentiated by their collocational profiles, and near-synonymous pairs like may and might or must and have to differ in distribution. Urunbaevna (2022) employed Distinctive Collexeme Analysis (DCA) to analyze the strong lexemes of construction “should + verb” and “have to + verb” in 3000 sentences of each structure from the British National Corpus. The study found that have to is inclined to verbs with dynamic senses, whereas should prefers stative verbs that tend to appear in the written register and exhibits a stronger obligation sense. However, although “modal + verb” is one of the most representative constructions of modals, few studies have been conducted on differentiating the verbal collexemes between the Chinese near-synonym pairs. Thus, this study intends to compare the distinctive verbal collexemes of the Chinese should modals *ying-dang*/*ying-gai* in the construction of "modal + verb."

Methodologically, this study adopted distinctive collexeme analysis (DCA), one of the emerging quantitative corpus-linguistic methods, to compare the distinctive collexemes between the near-synonym constructions (Gries and Stefanowitsch, 2004; Gries, 2012). As a type of collostructional analysis (Stefanowitsch and Gries, 2003), DCA focuses on usage-based and pattern-specific properties through objective and systematic statistical investigation (Stefanowitsch and Flach, 2020). The method can go beyond the raw frequency to find positively and negatively associated collexemes (Flach, 2020). The application of DCA to the Chinese near-synonymous pair *ying-dang* and *ying-gai* is expected to exhibit a promising result and reflect the subtle semantic features and usage patterns associated with each modal.

Organization of this paper. The rest of this paper is organized as follows. Section 2 presents research aims and questions. Section 3 focuses on previous works. Section 4 illustrates the corpus and method applied in this study. Section 5 analyzes the distinctive verbs associated with the construction of "*ying-dang*/*ying-gai* + verb," and then discusses the limitations and. The last section concludes this paper with the differences between *ying-dang* and *ying-gai*.

2 Research aims and questions

This paper intends to find out the collexeme variances of the Chinese two-character *should* modal pair *ying-dang* and *ying-gai* in the representative construction of "modal + verb" and hopes to illustrate how speakers choose between the two modals. The research questions are:

1) What are the distinctive collexemes of the two near-synonym constructions?
2) How to explain the collexeme preferences of *ying-dang* and *ying-gai* in terms of modality, usage patterns, and semantic types?

3 Previous works

This section presents the previous works on the Chinese modals and introduces the recent advances in distinctive collexeme analysis.

3.1 A comparison of *ying-dang* and *ying-gai*

Ying-dang and *ying-gai* are two frequent two-character modals expressing a similar meaning to *should* in Chinese. They are usually categorized as *gai*-modals in Chinese, which can be translated as modals of *should*', which could both express possibility and necessity (Ding, 1961; Zhu, 1982; Xu, 1990; Chen, 2006; Guo, 2011; Wu, 2021), corresponding to epistemic and deontic modality (Peng, 2007; Pan, 2010). Cao (1999) found that 必须 *bi-xu* 'shall,' *ying- dang* 'should or ought to', 可以 *ke-yi* 'may', and 不得 *bu-de* 'shall not' were often used in legal performatives to impose obligations, conferred rights, permission, and prohibition. The study also found that 必须 *bi-xu* and 应当 *ying-dang* are employed to impose illocutionary force of obligations, whereas *ying-gai* seldom appears in the legislation. Zhou (2008) proposed that *ying- gai* is easier to be adapted to different occasions compared with other modals such as *ying-dang* and *bi-xu*. In comparison, the usage of *ying-dang* is always discussed in legislative texts (Zhao, 2009). Li et al. (2016) studied three pairs of modals with deontic meaning in Hong Kong legislation, including 1) obligation (similar to *shall/must*) 须 *xu*/ 必须 *bi-xu*; 2) no-obligation (similar to *needn't*) *wu-xu*/*bu-bi*; 3) obligation (similar to *should*) *ying-gai*/*ying-dang*. The study found that when lexemes like *ying* and *xu* are encoded in the linguistic context, the semantics must fit in to match the occurrence of *ying* and *xu*, which explains their similarities in meaning. The authors also mentioned that such classification ignored the subtle inner-group differences between *ying-gai* and *ying-dang*. Yao (2017) systematically compared the differences between *ying-dang* and *ying-gai*. The author discovered that 1) *ying-gai* is more flexible in expressing both epistemic and deontic modality while *ying- dang* tends to express deontic modality; 2) when expressing deontic modality, *ying-dang* expresses deontic necessity whereas *ying-gai* deontic obligation; 3) *ying-gai* is more informal than *ying-dang* and can be used in daily conversation; 4) in terms of the constructions of "*ying-gai*/*ying-dang* + 说 *shuo* 'say'," the former is frequently used in the front of a sentence as a summary with a relatively weaker subjectivity. *Ying-gai* and *ying-dang* also displayed distinctively different attractions to prepositions and adverbs (Yao, 2017).

In summary, various studies touched upon the usage patterns of *should* modals in multiple aspects, but few studies have focused on the distinctive collexemes attracted to a specific construction.

3.2 Recent advances in Distinctive Collexeme Analyses

Distinctive collexeme analysis (Gries and Stefanowitsch, 2004) aims to find the more attractive words of one construction as opposed to the other (Hilpert, 2014). One of the examples in this study is the comparison between the near-synonymous construction "*will* + verb" and "*be going to* + verb" denoting the meaning of future. The underlying theoretical foundation is that a verb is only suitable on the condition that its argument matches the construction under the principle of Semantic Coherence (Goldberg 1995: 50). The study found that *will* prefers relatively non-agentive or low-dynamicity actions (*find*, *receive*, *hold*, *finish*, *reach*), including perception/cognition events (*see*, *know*, *want*, *consider*, *notice*, *need*, *accept*), or states (*depend*, *remain*, *become*) whereas the opposite trend was found for *be going to*. DCA is often used to compare collexemes between two or more near- synonym constructions to draw implications on differentiating the near-synonyms (cf. S + *can* + V vs. S + *be able to* + V, Lojanica, 2021)

However, collostructional analysis is not without disputes. Schmid and Küchenhoff (2013) argued that collostructional analysis, including DCA, inherited potential problems in terms of statistical measures used to calculate collocational strengths (mostly Fisher-Yates exact, abbreviated as FYE). Gries (2015, 2019) defended that 1) many association measures, such as G^2 (the log-likelihood ratio), *chi-square*, Mutual Information (MI), and a logged odds ratio, could be used other than the FYE; 2) the FYE was correlated with the most proposed statistical tests and did not influence the top collexemes which were used to report the results; conflation due to large corpus size was itself a feature that describes the frequency of (co-) occurrence, which is important to usage-based grammar.

In a nutshell, DCA is an efficient method for differentiating near-synonymous constructions but rarely used in studying Chinese near- synonyms, and the technique is still developing with continuous trials and improvement.

4 Method

This section presents the corpus and major procedures of this study.

4.1 The corpus

The corpus adopted in this study is *news2016zh* (version 1.0), an open-access corpus with more than 2.95 billion Chinese characters in the large- scale natural language processing Chinese corpus (Xu, 2019). The data was downloaded on March 3, 2022. It contains 2.5 million news articles collected from 63 thousand media platforms between 2014 and 2016. All sentences that contain the structure "*ying-dang* + verb" (CNX1) and "*ying-gai* + verb" (CNX 2) in the corpus were extracted and utilized as samples in this study.

4.2 Major procedures

The major steps are presented as follows:

1. Build the corpus, segment words and perform the POS tagging.
2. Find and extract all near-synonym constructions and use them as the samples of *ying-dang* and *ying-gai*.
3. Count the frequency of each verb respectively in the two samples.
4. Perform the distinctive collexeme analysis.
5. Analyze the top distinctive collexemes by groups and report the results.

The Stanford CoreNLP Natural Language Processing Toolkit (Manning et al. 2014) was used for word segmentation and POS tagging. The collexemes of near-synonym constructions "ying-dang + verb" (CNX1) and "ying-gai + verb" (CNX 2) were counted via self-established programs in the programming language of Java. The frequency was conducted via existing R codes (Flach, 2021) with minor adjustments. FYE was adopted as the association measure as it generally exhibits good performance and correlates with other statistical measures. Also, our focus was on the report of top collexemes rather than the values of the statistical values (Gries, 2015; 2019). The threshold of the statistic measure was decided via multiple trials based on the number of distinctive collexemes with statistical significance. To exclude the low-frequency verbs, we set the minimum sum of CNX1 and CNX2 for each collexeme (referred to as threshold in DCA) as 100 to be considered for DCA, and collexemes that were not verbs were excluded or analyzed during the manual analysis. The analysis of verbal collexemes was based on the notion of frames (Fillmore, 1982) and the grammatical functions of the collexemes. Specifically, we referred to Mandarin VerbNet (Liu, 2017) and FrameNet (Fillmore and Baker, 2010) for deciding the semantic frames of verbs and reported the attracted frames along with some other superficial features associated with the distinctive collexemes.

5 Result and discussion

This section illustrates the descriptive statistics, distinctive collexemes of the two constructions, together with limitations and future implications.

5.1 Descriptive statistics

The corpus contains 161,602 sentences with *ying-dang*, and 548,303 with *ying-gai*, about 3.39times of *ying-dang*. After POS tagging, 41,693 sentences were found tohave the construction of "*ying-dang* + verb" involving 3,057 individual verbs, among which 1,621 verbs appeared over once and 727 more than five times (See Table 1). The corpus also includes 226,309 sentences of "*ying-gai* + verb" (5.43 times of *ying-dang*), which involved 8,876 unique verbs, with 4707 showing up not less than twice and 1,974 over five times. Also, *ying-dang* prefers to take verbs directly than *ying-dang* because the count of *ying-gai* is 3.39 times of *ying-dang* while the usage of "modal + verb" is 4.43 times of *ying-dang*.

Total	*Ying-dang*	*ying-gai*	*ying-gai/ ying-dang*
sent	161,602	548,303	3.39
CNX	41,693	226,309	5.43
verbs	3,057	8876	2.90

Table 1: Descriptive statistics

5.2 Most distinctive collexemes in CNX 1

After the DCA, 281 out of 3,027 individual verbs met the threshold, 110 verbs are distinctive to *ying-dang* with statistical significance in CNX 1, and 61 verbs are not distinctive to *ying-dang* or *ying-dang*. The top 20 distinctive collexemes associated with CNX1 are presented in Table 2 on the next page, together with observed and expected frequencies of the two CNXs, collocation strength, and significance level of $p < 0.0001$. *Ying-dang* is more frequently used in formal occasions associated with deontic needs in an obligation or requirement scenario, such as laws, regulations, official announcements, policy, obligations, and requirements, as shown in Example (1)-(6), in which the typical verbs are 遵守 *zun-shou* 'obey', 要求 *yao-qiu* 'require,' 允许 *yun-xu* 'permit,' 承认 *cheng-ren* 'admit,' 负责 *fu-ze* 'account for', and 禁止 *jin-zhi* 'forbid.' *Ying-dang* are inclined to take two-character verbs from various semantic types (frames) with a strong inclination of purposive efforts (See Table 3).

Semantic frames	Typical verbs	E.g.
Purposive efforts	遵循 *zun-xun* 'obey' 履行 *lv-xing* 'perform (duties)'	(1)
Existence	具备 *jv-bei* 'possess' 具有 *jv-you* 'possess'	(2)
Cognition	坚持 *jian-chi* 'insist' 视为 *shi-wei* 'consider (as)'	(3)
Communication	说明 *shuo-ming* 'state' 告知 *gao-zhi* 'tell'	(4)
Transference	支付 *zhi-fu* 'pay'	(5)
Attribute of manner	真实 *zhen-shi* 'real' 谨慎 *jin-shen* 'cautious'	(6)

Table 3: Representative verbs and frames in the construction of "ying-dang + verb" (CNX1)

No	LEX	*pinyin*	Eng	O.CXN1	E.CXN1	O.CXN2	E.CXN2	COLL	SIGNIF
1.	承担	*cheng-dan*	undertake	1485	381.6	968	2071.4	Inf	*****
2.	符合	*fu-he*	conform	1022	186.5	177	1012.5	Inf	*****
3.	认定	*ren-ding*	identify	574	95.4	39	517.6	Inf	*****
4.	提交	*ti-jiao*	submit	492	80.1	23	434.9	Inf	*****
5.	予以	*yu-yi*	give	599	120.1	173	651.9	Inf	*****
6.	遵守	*zun-shou*	obey	578	125.5	229	681.5	277.5758	*****
7.	建立	*jian-li*	Establish	697	179.8	459	976.2	262.8857	*****
8.	具备	*jv-bei*	have	880	321.6	1187	1745.4	189.5545	*****
9.	履行	*lv-xing*	fulfill	326	62.7	77	340.3	185.423	*****
10.	取得	*qu-de*	get	253	45	36	244	161.308	*****
11.	提供	*ti-gong*	supply	346	92.6	249	502.4	124.0307	*****
12.	遵循	*zun-xun*	follow	422	135.2	447	733.8	114.3772	*****
13.	组织	*zu-zhi*	organize	165	29.4	24	159.6	105.0159	*****
14.	采取	*cai-qu*	take	504	195.1	750	1058.9	97.49452	*****
15.	说明	*shuo-ming*	illustrate	168	32.8	43	178.2	93.84047	*****
16.	包括	*bao-kuo*	include	420	153.2	565	831.8	90.92773	*****
17.	载明	*zai-ming*	state	112	17.4	0	94.6	90.55912	*****
18.	真实	*zhen-shi*	real	116	18.5	3	100.5	88.57565	*****
19.	加强	*jia-qiang*	amplify	439	169.3	649	918.7	85.66282	*****
20.	披露	*pi-lou*	reveal	129	23	19	125	82.06816	*****

Note: O.CXN = Observed Construction Frequency; E.CXN = Expected Construction Frequency

Table 2: The most distinctive verbs in the construction of “ying-dang + verb

Example (1) 制定和实施城乡规划**应当遵循**…的原则。

The formulation and implementation of urban and rural planning **shall follow** the principles of …

Example (2) 申请认定小学教师资格，**应当具备**高等院校专科毕业及其以上学历。

To apply for the qualification of primary school teachers, one **should have** a college degree or above from an institution of higher learning.

Example (3) 行政机关依申请公开政府信息**应当坚持**公正、公平、便民、及时的原则

When an administrative organ discloses government information upon application, it**shall adhere to** the principles of impartiality, fairness, convenience, and timeliness.

Example (4) 报告期内发生重大会计差错更正需追溯重述的，公司**应当说明**情况、更正金额、原因及其影响。

Suppose a significant accounting error correction occurs during the reporting period thatneeds to be retrospectively restated, the company**shall explain** the situation, correct the payment amount, and indicate the reason and its impact.

Example (5) 劳动合同法第八十五条进一步明确了法律责任，劳动报酬低于当地最低工资标准的，**应当支付**其差额部分;

Article 85 of the Labor Contract Law further clarifies the legal responsibility. If the labor remuneration is lower than the local minimum wage, the difference part **shall be** paid;

Example (6) 报考人员提交的报考申请材料**应当真实**、准确

The application materials submitted by the candidates **should be true** and accurate.

Ying-dang also prefers two-character light verbs (see Table 4 on the next page), which add little semantic meaning to the sentences other than bring or amplify the following verbs.

Apart from light verbs, the construction of “verb of attaining/achieving/adopting + goal” is also frequently occurred with *ying-dang*, in which the meaning is largely reliant on their objects (See Table 5 on the next page).

Example (7) 该通告针对蓬江、江海、新会三区有效，要求其范围内销售燃气具的单位**应当取得**营业执照并注明相应的经营范围。

Example (7) This circular is valid for the three districts of Pengjiang, Jianghai, and Xinhui. and it requires that units selling gas appliances in these areas should obtain a business license and indicate the corresponding business scope.

Semantics	Light verb	Literal meaning	E.g.
To bring the following verbs	予以 *yu-yi* 给予 *ji-yu*	give	予以赔偿 'to be compensated'
Existence	作出 *zuo-chu* 进行 *jin-xing* 实行 *shi-xing* 加以 jia-yi	do/ perform	作出无罪判决 'to be acquittal'
Cognition	引起 *yin-qi*	cause	引起警惕 'to arouse vigilance'
To amplify the following verbs	加强 *ji-qiang* 加大 *jia-da*	amplify/ strengthen	加强中外交流 'to strengthen Sino-foreign exchanges'

Table 3: The distinctive light verbs in "*ying-dang* + verb

Semantic frames	Typical verbs	E.g.
attaining	取得 *qu-de* 'obtain' 获得 *huo-de* 'achieve'	(7)
achieving	达到 *dao-da* 'reach' 发挥 *fa-hui* 'play a role' 满足 *man-zu* 'satisfy'	(8)
adopting	采用 *cai-yong* 'use' 采取 *cai-qu* 'take'	(9)
Communication	说明 *shuo-ming* 'state' 告知 *gao-zhi* 'tell'	(4)

Table 4: Typical verbs belong to verbs of attaining/achieving/adopting

Example (8) 普通话水平**应当达到**国家语言文字工作委员会颁布的普通话水平测试等级标准二级乙等以上**标准**。

Example (8) The proficiency of Putonghua **should meet the standard of** Level 2 or above of the Putonghua Proficiency Test issued by the State Language and Character Work Committee.

Example (9) **应当采用**反担保等必要**措施**防范风;

Example (9) Necessary **measures,** such as counter-guarantee, **should be adopted** to prevent risks

5.3 Most distinctive collexemes in CNX 2

Compared with ying-dang, ying-gai tends to take a wider semantic range of verbs with no bias on single-characters verbs. The top 20 distinctive collexemes are presented in Table 7 on the next page. Compared with ying-dang, verbs of motion transference, perception, and emotion are more likely to co-occur with ying-gai. Some of the representative verbs are summarized in Example (10)-(15) in Table 6.

Semantic frames	Typical verbs	E.g.
Self-Motion	去 *qu* 'go' 来 *lai* 'come'	(10)
Communication	说 *shuo* 'say'	(11)
Cognition	懂 *dong* 'understand' 想 *xiang* 'think' 知道 *zhi-dao* 'know' 明白 *ming-bai* 'understand'	(12)
Existence	有 *you* 'have/ possess'	(13)
Transference	买 *mai* 'buy' 花 *hua* 'spen 学 *xue* 'learn' 换 *huan*'exchange'	(14)
Emotion	高兴 *gao-xing* 'happy'	(15)

Table 5: Representative verbs and frames in the construction of "ying-gai + verb" (CNX2)

Example (10) 你若问我中国哪里最**应该去**，我会告诉你新疆!

If you ask me which is the best place people **should go** to in China, I will tell you Xinjiang!

Example (11) 这句话似乎**应该说**成我们现在都是老兵了。(Literal meaning: should say)

It seems this sentence should be put in this way: we are all veterans now.

Example (12) 他爱我，就应该懂我、满足我。

He loves me, so he should understand me and satisfy my needs.

Example (13) 他是第一个听我哭的男人，他教我男人就**应该有**强壮的臂弯。

He is the first man to hear me cry, and he teaches me that men **should have** a strong arm.

Example (14) 在跑步一开始，**应该买**些最好的鞋子和衣服吗?

Should we **buy** the best shoes and clothes as soon as we begin to run?

Example (15) 你见到我**应该高兴**呀。

You **should be happy** when seeing me.

It was also observed when the typical perception verb 看 kan 'see/look at' is used after ying-gai, it tends to express abstract cognitive activities, such as focusing on, checking, valuing, reading something or the indicated meaning of seeing a doctor. The respective examples (extracted sentence clips) can be found in Example (16)-(20).

Example (16) **应该看的是**它的潜藏价值

'**should focus on** its potential value'

Example (17) 应该**看两证**

'**should check** two certificates

Example (18) **应该看得**比生命还重要

'should value (it) over than life'

Example (19) 应该看过精益创业 'should read *The Lean Startup'*

Example (20) 应**该看**什么科?

Word-for-word: Should – see – what – section?

'Which section should (a potential patient) go?'

Among the top 50 verbs distinctive to ying-gai, more a half are single-character verbs, whereas those to ying-dang are all two-character verbs. Ying-gai can be considered more colloquial than ying-dang, evidenced by the frequent uses of single-character verbs and informal expressions like trifling in daily chat. For example, informal verbs like 叫 jiao 'call … as' and 算 suan 'is barely considered as' were found distinctive to ying-gai as shown in Example (20)-(21).

Example (21) 职称药师考试其实**应该叫**药学职称考试

The Professional Pharmacist Qualification Exam **should** actually **be called** the Pharmacy Job Title Exam.

No	LEX	*pinyin*	Eng	O.CXN1	E.CXN1	O.CXN2	E.CXN2	COLL	SIGNIF
1.	说	*shuo*	say	144	1122	7068	6090	Inf	*****
2.	是	*shi*	be	2644	9185.6	56401	49859.4	Inf	*****
3.	会	*hui*	can/will	131	1029.1	6484	5585.9	309.0393	*****
4.	做	*zuo*	do	117	746.6	4682	4052.4	203.3121	*****
5.	有	*you*	have	882	1822.3	10832	9891.7	156.5101	*****
6.	要	*yao*	will	124	564.6	3505	3064.4	125.4005	*****
7.	去	*qu*	go	96	438.2	2721	2378.8	97.55963	*****
8.	算是	*suan-shi*	be considered	31	244.4	1540	1326.6	73.33712	*****
9.	可以	*ke-yi*	can	67	318.9	1983	1731.1	73.2193	*****
10.	能	*neng*	can	59	287.8	1791	1562.2	67.22511	*****
11.	怎么办	*zen-me-ban*	how to do	25	198	1248	1075	59.65883	*****
12.	让	*rang*	let	171	433.1	2613	2350.9	52.67851	*****
13.	没有	*meiyou*	no	22	159.9	1006	868.1	46.60287	*****
14.	没	*mei*	no	5	107.3	685	582.7	42.28269	*****
15.	给	*gei*	give	110	291.2	1762	1580.8	37.9745	*****
16.	叫	*jiao*	call	13	105.3	664	571.7	32.31799	*****
17.	知道	*zhi-dao*	know	306	511.5	2982	2776.5	25.73305	*****
18.	算	*suan*	count	15	92.7	581	503.3	25.32514	*****
19.	明白	*ming-bai*	clear	45	146.5	897	795.5	24.87238	*****
20.	买	*mai*	Buy	2	57.6	368	312.4	23.81272	*****

Table 6: The most distinctive verbs in the construction of "*ying-gai* + verb"

Example (22) 如果孩子写的是 make, 那**应该算**对还是错呢?

If children write ‘make,’ **should it be considered as** right or wrong?

Ying-gai also prefers highly transitive constructions over *ying-dang*, evidenced by 让 *rang* /给 *gei* construction ‘enable …’ as shown in Example (23) and frequent usage of caused-motion/position verb 放 *fang* ‘put’ as presentedin Example (24).

Example (23) 我们**应该让它恢复**到自然的规律当中去。

We **should enable it to restore** the law of nature.

Example (24) 主要精力**应该放**在教育教学管理上。

The primary energy **should be put** intoeducation and teaching management.

Apart from these examples, *Ying-gai* also tends to appear in front of modal auxiliary verbs. Among the top 10 distinctive collexemes, fourare modals, including 会 *hui* ‘can/will’, 要 *yao* ‘want’, 可以 *keyi* ‘can’ and 能 *neng* ‘can’.Compared with those of *ying-dang*, the distinctive collexemes of *ying-gai* could already display epistemic probabilities as shown inExample (23) and (24).

Example (23) 志玲姐姐的导航声音嗲嗲的，**男司机应该会喜欢**。

Female actor Zhiling's navigation voice is squeaking, so **male drivers should like it.**

Example (24) 但是范冰冰私生子的谣言**应该可以不攻自破了**。

However, the rumor of Bingbing Fan's illegitimate child **should be self-defeating**.

Similarly, the distinctive collexemes of *ying- gai* also include interrogative phrase 怎么办 *zen-me-ban* ‘how to deal with it’, negation adverb negation 没 / 没有 *mei*/*mei-you* ‘not/haven’t’though these words are not considered as verbs.

5.4 Limitations and future implications

In terms of the research scope, this study only focuses on the distinctive collexemes associated with the near-synonym constructions “*ying-dang*/*ying-gai* + verb”. To fully address their differences, a future study may explore more constructions, such as “modal + adv,” “adv + modal,” and “modal + prepositions”. In the corpus, we also found that “Degree markers + modal” is much less frequently collocated with *ying-dang* rather than *ying-gai* (Yao, 2017). As shown in Example (10), degree marker 最 *zui* ‘the most’ is used before *ying-gai*. Flach (2020) employed Co-Varying Collexeme Analysis, also a type of collostructional analysis, to study gradient idiomaticity in MOD + ADV collocations, and revealed that collocational behaviors of modal auxiliaries could serve as a cue for measuring the scope of adverbial modification.

Besides, this study mainly relies on POS tagging to extract the target constructions, which was hardly 100% accurate. However, some improper tags may still be meaningful as the same POS model was applied to the samples. It is suggested to keep the original data as much as possible while pointing out the potential improper tags, especially when such tagging could also reveal differences in usage patterns. The point is that loyalty to the actual data is essential, and ‘wrong’ or unexpected results should not be ignored. Ignoring the ‘wrong’ or unexpected tags may miss critical aspects in the findings. In the case of this study, modal auxiliary verbs, such as 会 *hui* ‘can/will’, 要 yao ‘want’, 可以 *ke-yi* ‘can’, interrogative phrase 怎么办 ‘how to deal with it’, and negation adverb negation 没/没有 ‘haven’t’, were found to be distinctive to *ying-gai*. To the authors’ knowledge, such preferences to different modals were first proposed to differentiate *ying-dang* and *ying-gai*.

Also, when analyzing the collexemes, it is suggested to draw some implications on the existing theories like semantic domains (Biber, 1999:365-371) as illustrated by Deshors (2017) and the notion of frames (Fillmore, 1982) evidenced by Wiliński (2019) to provide a relatively objective and linguistically-wise interpretations on semantic types. In fact, the uses of semantic types and frames are not new but seldomly cited in terms of specific theories (Gries and Stefanowitsch, 2004). In the case of this project, we were largely reliant on the frame- based interpretations of verbs.

Last, methodologically, DCA is not just limited to the constructions associated with near- synonyms and can potentially be applied to more construction-based studies. For instance, Newman (2021a) applied this approach to study the differences between singular *child* and plural form *children* by comparing three typical constructions, such as (a) adjective + child/children, (b) child’s/children’s +

noun, and (c) child/children + present participle of a verb. In the same year, Newman (2021b) also employed this approach to study the collexemes distinctive to singular and plural forms of animal nouns *dog* and *cat*. The studies on singular and plural forms are hardly considered as near-synonyms; however, the purpose of this approach is to find the more attractive or repelled collexemes distinctive to the typical constructions of the studied words so as to reveal their usage variances and semantic differences

6 Conclusion

In sum, this paper presented the collexemes respectively distinctive to *ying-dang* and *ying-gai* in the construction of "modal + verb". In terms of modality, the study finds *ying-dang* prefers to take collexemes with a stronger obligation sense, whereas *ying-dang* prefers common verbs with no obvious indication of deontic senses and is able to be used together with other modal verbs to express epistemic meaning. As for usage patterns, *ying-dang* is frequently appeared in a formal context where obligations or requirements are involved and likes to take two-character verbs, whereas *ying-gai* attracts single-character verbs, and appears in various occasions including but not limited to daily conversation, gossips, and forums. Regarding semantic preferences, *ying-dang* prefers verbs associated with purposive efforts whereas *ying-gai* enjoys collocating with verbs of self-motion and emotion. *Ying-gai* is also more likely to take interrogative phrase and negation adverb negation. Methodologically, this study offers some practical suggestions for applying DCA to study Chinese modal auxiliary verbs.

Acknowledgement

The research is supported by the CityU of HK grant for Project #7020010 - Developing a Linguistically-wise and Computationally Feasible Digital System of Semantic Representation for NLP Applications.

References

Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman grammar of spoken and written English*. Longman, Harlow, UK.

Deborah Cao. 1999. 'Ought to' as a Chinese Legal Performative. *International Journal for the Semiotics of Law*, vol. 12, no. 2, , pp. 151–167.

Sandra C Deshors. 2017. Zooming in on Verbs in the Progressive: A Collostructional and Correspondence Analysis Approach. *Journal of English Linguistics*, vol. 45, no. 3, pp. 260–290.

Holger Diessel. 2019. *The Grammar Network: How Linguistic Structure is Shaped by Language Use*. Cambridge University Press, Cambridge, UK.

Charles J. Fillmore and Collin F. Baker. 2010. A Frames Approach to Semantic Analysis. In B. Heine and H. Narrog (Eds.) *The Oxford Handbook of Linguistic Analysis. Oxford University Press*. Oxford, UK/New York, New York.

John R. Firth. 1957. A synopsis of linguistic theory 1930–55. Reprinted in Frank. R. Palmer (Ed.), (1968). *Selected Papers of J.R.* Firth 1952–1959. Longman, London, UK.

Susanne Flach. 2021 Collostructions: An R implementation for the family of collostructional methods. Package version v.0.2.0, URL: https://sfla.ch/collostructions/. ED: 3 March. 2022.

Susanne Flach. 2021. Beyond Modal Idioms and Modal Harmony: a Corpus-Based Analysis of Gradient Idiomaticity in Mod Adv Collocations. *English Language and Linguistics*, vol. 25, no. 4, pp. 743–765., doi:10.1017/S1360674320000301.

Adele E. Goldberg. 1995. *Constructions. A Construction Grammar Approach to Argument Structure*. Chicago University Press, Chicago, USA

Stefan Th. Gries 2012. Corpus linguistics: Quantitative methods. In Carol A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics*, 1380-1385. Oxford: Wiley-Blackwell.

Stefan Th. Gries and Anatol Stefanowitsch. 2004. Extending collostructional analysis: A corpus-based perspectives on 'alternations'. *International Journal of Corpus Linguistics*, 9.1: 97-129.

Martin Hilpert. 2008. *Germanic Future Constructions. A Usage-based Approach to Language Change*. John Benjamins, Amsterdam, the Netherlands.

Martin Hilpert. 2014. Collostructional analysis: Measuring associations between constructions abd lexical elements. In D. Glynn and J. A. Robinson (Eds.), *Corpus Methods for Semantics: Quantitative studies in polysemy and synonymy* (Human Cognitive Processing, vol. 43) (pp.391– 404). John Benjamins, Amsterdam, the Netherlands.

Martin Hilpert and Susanne Flach. 2021. Disentangling modal meanings with distributional semantics. *Digital Scholarship in the Humanities*, 36(2): 307–321.

Jian Li, Le Cheng, and Winnie Cheng. 2016. Deontic meaning making in legislative discourse. *Semiotica*, 209 (2016): 323-340.

Meichun Liu. 2017. A frame-based morpho-constructional approach to verbal semantics. In Chunyu Kit and Meichun Liu (Eds.) *Empirical and Corpus Linguistic Frontiers*, China Social Sciences Press, BJ.

Tošić T. Lojanica. 2021. Exploring present ability: A collostructional approach. *Nasledje Kragujevac* 18.48:105-115.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55–60. Association for Computational Linguistics, Baltimore, Maryland.

John Newman. 2021a. Child and children in a corpus of American fiction: Contrasting semantic preferences and their experiential motivations. *Cognitive Semantics*, 7.1: 1-30.

John Newman. 2021b. Singular and plural preferences among adjectival collocates of CAT and DOG. *LaMiCuS,* 5: 12-32.

Anatol Stefanowitsch and Susanne Flach. 2020. Too big to fail but big enough to pay for their mistakes:A collostructional analysis of the patterns [too ADJ to V] and [ADJ enough to V]. In Gloria Corpas and Jean Pierre Colson (eds.), *Computational and corpus-based phraseology*, (pp. 248–272). John Benjamins, Amsterdam, the Netherlands.

Anatol Stefanowitsch and Stefan Th. Gries. 2003. Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics,* 8.2: 209-243.

Peter. D. Turney and Patrick Pantel (2010). From frequency to meaning: vector space models of semantics. *Journal of Artificial Intelligence Research*, 37: 141–88.

Bright Xu. 2019 Sep. "NLP Chinese corpus:Largescale Chinese corpus for NLP." Corpus version 1.0, URL: https://doi.org/10.5281/zenodo.3402023. ED: 10 March. 2022.

Jiajia Chen. 2006. Multi-perspective research on "*ying-gai*". Hunan Normal University. Ph.D. thesis.

Shengshu Ding. 1961. *Speeches on Modern Chinese Grammar*. Commercial Press. BJ.

Zhaojun Guo. 2011. Two modal types of “gai-type” auxiliary verbs and their selection factors. *Nankai Journal of Linguistics*, (11):132-149.

Lizhen Peng. 2007. *Research on modality of modern Chinese*. China Social Science Press. BJ.

Tiantian Wu. 2021. *A Comparative Study of the Modal Verbs "ying," "gai," "dang," "ying-gai," and "ying."* Shanghai International Studies University, MA thesis.

Heping Xu. 1990. A Preliminary Study on Chinese Modal Verb Semantics and Syntax. World Society for Chinese Language Teaching, *Selected Papers of the 3rd International Conference on Chinese Language Teaching*, Beijing Language Institute Press, BJ.

Wenbiao Yao. 2017. A Comparative Study of "*ying-gai*" and "*ying-dang*". Central China Normal University, MA thesis.

Yancai Pan. 2010. Analysis of the meaning of "*ying-gai*". *Chinese Knowledge*, (04):59-61.

Wei Zhao. 2009. Language Modal Expressions and Its Standardization. *Rhetorical Learning*, (02):30-36.

Youbin Zhou. 2008. Standards of Parts of Speech in Chinese and Determination of Auxiliary Verbs. *Journal of Huaibei Vocational and Technical College*, (06):50-52.

Dexi Zhu. 1982. *Lecture Notes on Grammar*. Commercial Press. BJ.