ISSN 1798-4769
Journal of Language Teaching and Research, Vol. 11, No. 6, pp. 995-1004, November 2020
DOI: http://dx.doi.org/10.17507/jltr.1106.18

On the Establishment of Database of Morpheme Meaning Annotations for Modern Chinese Compounds

Yu Xu
Beijing International Studies University, Beijing, China

Weiqi Chen
Beijing International Studies University, Beijing, China

Fei Song*
Beijing International Studies University, Beijing, China

Abstract—**This research establishes Database of Morpheme Meaning Annotations for Compounds, and statistically analyses the using frequency of morphemes and their meanings in the database. The research shows that the using frequency of "zi" (子) and "er" (儿) is the highest, and their meanings with the highest using frequency are applied when the morphemes are affixes. Among various meanings of one morpheme, only one meaning would be applied with high frequency, with simpler and specific implications. By overviewing all morphemes, it is observed that those with higher using frequency do not necessarily have meanings with high frequency of use. By analyzing internal structures of compounds, this paper finds that noun-modification compounds are the most, followed with verb-object compounds and affixation compounds. The semantic logic relations between morphemes and compound words can be: word definitions can be the abstraction of objects signified by morpheme meanings, or of senses signified by morpheme meanings, or of spatial meanings signified by morpheme meanings.**

Index Terms—**compound words, morpheme meaning, database, word formation**

I. Introduction

Searching on CNKI, there are 214 academic papers researching on morpheme meanings of modern Chinese words. By analyzing the existing studies in this area, it is discovered that they mainly focus on the Chinese language itself. For example, by investigating on new word selection and new meaning compilation in *Modern Chinese Dictionary*, Cheng (2017) discusses the relationship between them and other related issues. In addition, most of the existing studies are on the annotation of certain vocabulary in a specific corpus. For example, Wang, Yang et al. (2017) research on the annotation of polysemous words in *Modern Chinese Dictionary*, but it is short of studying on database with all compounds' morphemes and their meaning annotations. Moreover, from the perspective of natural language processing, the current academia lacks a database of all morphemes to construct compounds and morpheme meaning annotations. To solve these issues, this research will establish a Database of Morpheme Meaning Annotations for Modern Chinese Compounds.

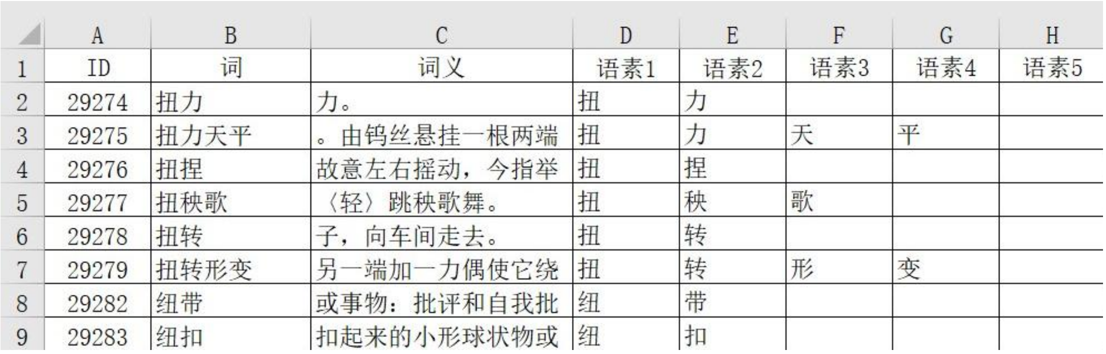
II. Establishment of Database of Morpheme Meaning Annotations for Compounds

A. Design of Database Structure

In the fifth edition of Modern Chinese Dictionary (herein referred to as MCD 5), a compound entry includes Pinyin, part of speech, word definition and sample sentence. Database of Morpheme Meaning Annotations for Compounds (hereinafter referred to as Database of Morpheme Annotations) annotates morphemes in compounds. As pronunciation and part of speech are not directly relevant to morpheme annotation, they are omitted in the database, whereas only word definition and sample sentence are kept since they contribute to an accurate morpheme annotation. In addition, every compound in the database has an ID (see column A in picture 2-1). As the longest compound in the database has eight morphemes, eight morpheme fields are set up, with each morpheme distributed in its corresponding field.

In summary, Database of Morpheme Annotations includes several fields, such as ID, word, word definition, morpheme 1, morpheme 2, morpheme 8, and so on (see figure 2-1).

[*] Corresponding author

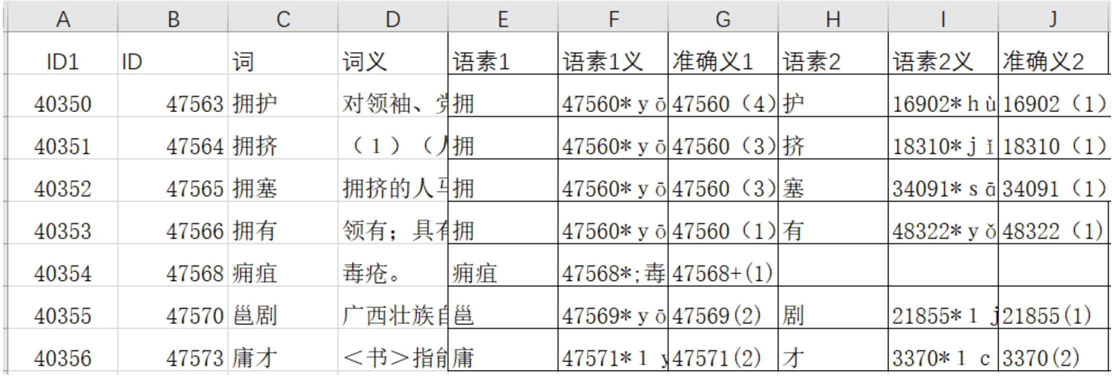


	A	B	C	D	E	F	G	H
1	ID	词	词义	语素1	语素2	语素3	语素4	语素5
2	29274	扭力	力。	扭	力			
3	29275	扭力天平	。由钨丝悬挂一根两端	扭	力	天	平	
4	29276	扭捏	故意左右摇动，今指举	扭	捏			
5	29277	扭秧歌	〈轻〉跳秧歌舞。	扭	秧	歌		
6	29278	扭转	子，向车间走去。	扭	转			
7	29279	扭转形变	另一端加一力偶使它绕	扭	转	形	变	
8	29282	纽带	或事物：批评和自我批	纽	带			
9	29283	纽扣	扣起来的小形球状物或	纽	扣			

Figure 2-1 The structure of Database of Morpheme Meaning Annotations for Compounds

Meaning Annotation Database is based on Morpheme Annotation Database. “Morpheme meaning” field is added after “morpheme” field, and “morpheme meaning” contains full meanings of a certain morpheme, which is associated with word definition under a single-character entry. Also, “accurate meaning” field is added to indicate the exact meaning of a morpheme in a compound. At the same time, computer automatically generates a new ID (see “ID” bar in Figure 2-2). In Morpheme Annotation Database, a compound has a maximum of eight morphemes, thus, in the corresponding Meaning Annotation Database, a compound has up to eight fields of “morpheme meaning” and “accurate meaning”.

To summarize, the database includes fields such as ID1, ID, word, word definition, morpheme 1, morpheme 1 meaning, accurate meaning 1, morpheme 2, morpheme 2 meaning, accurate meaning 2 and so on (see Figure 2-2).



A	B	C	D	E	F	G	H	I	J
ID1	ID	词	词义	语素1	语素1义	准确义1	语素2	语素2义	准确义2
40350	47563	拥护	对领袖、党	拥	47560* y ō	47560（4）	护	16902* h ù	16902（1）
40351	47564	拥挤	（1）（人	拥	47560* y ō	47560（3）	挤	18310* j ǐ	18310（1）
40352	47565	拥塞	拥挤的人马	拥	47560* y ō	47560（3）	塞	34091* s ā	34091（1）
40353	47566	拥有	领有；具有	拥	47560* y ō	47560（1）	有	48322* y ǒ	48322（1）
40354	47568	痈疽	毒疮。	痈疽	47568*;毒	47568+(1)			
40355	47570	邕剧	广西壮族自	邕	47569* y ō	47569(2)	剧	21855* 1 j	21855(1)
40356	47573	庸才	<书>指能	庸	47571* 1 y	47571(2)	才	3370* 1 c	3370(2)

Figure 2-2 The structure of Database of Morpheme Meaning Annotations for Modern Chinese Compounds

B. Establishment of Compound Database

A compound is a word consisting of two or more morphemes. In Chinese, except for retroflex suffixation [e.g., “huār” (花儿) has two Chinese characters with one morpheme], a Chinese character has a syllable, so a compound has two or more syllables. By screening compounds with two or more syllables, this study preliminarily establishes Compound Database. On the other hand, since simple words such as “pútáo (葡萄), língl óng (玲珑), xiāoyáo (逍遥), fúróng (芙蓉), qiǎokèlì (巧克力)” also have two or more syllables, they will also be in Compound Database, which will have a certain adverse effect on the authenticity and scientificity of the data. Therefore, after the initial screening of compounds, manual annotation is also necessary to do a secondary correction for the data.

Compound Database contains a total of 44,905 compounds, including some polysyllable morphemes that are not yet confirmed to be compounds. These words will be re-judged integrating word definition and morpheme meaning in the subsequent establishment of Morpheme Annotation Database and Meaning Annotation Database.

C. Establishment of Morpheme Annotation Database for Compounds

A morpheme is the smallest phonological and meaningful language unit. “Meaning” refers to lexical meaning and grammatical meaning. Zhiwei Lu (1957) proposed to use the “extension method” to identify words, namely the later generally applicable “substitution method”, which was using a known morpheme to replace another language unit that is not yet confirmed to be a morpheme. Due to restrictions on meaning selection when compiling dictionaries, it is difficult for any dictionary to exhaust all compounds. Therefore, considering some single characters that are selected into MCD 5 and MCD 7 while other compounds starting with these particular characters are not selected, or only one compound starting with these particular words is included, this research will include words and meanings in The Great Chinese Dictionary (hereinafter referred to as GCD) and The Grand Dictionary of Chinese Characters (hereinafter referred to as GDCC). For example, under the single-character entry “qì” (迄) in MCD 7, only the word “qìjīn” (迄今)

is included. If this morpheme was categorized by “substitution method”, it could be difficult to find a morpheme that can substitute “jīn” (今), which would cause challenges to judge whether “qì” (讫) is a morpheme or not. However, based on GCD, “qìjīn” (讫今) can be searched out, and another word “qǐqì” (起讫) is included as well, so “q ǐ’ (讫) can be defined as a morpheme.

Morpheme Annotation Database divides units according to characters, and computer defaults a Chinese character as a morpheme. For example, in the table below, “ānkāng” (安康) is divided into two Chinese characters, “ān” (安) and “kāng” (康), i.e., two morphemes, and they are placed in the fields of “morpheme 1” and “morpheme 2” respectively; similarly, “ānlèwō” (安乐窝) is divided into three Chinese characters and respectively placed in “morpheme 1”, “morpheme 2” and “morpheme 3”.

TABLE 2-1
EXAMPLES OF MORPHEME ANNOTATION DATABASE

ID	Word	Word definition	Morpheme 1	Morpheme 2	Morpheme 3
132	ānkāng (安康)	平安和健康 contented and in good health	ān (安)	kāng (康)	
133	Ānlā (安拉)	阿拉伯语音译词。意为真主。 A word transliterated from Arabic. Means Allah.	Ān (安)	lā (拉)	
135	ānlèwō (安乐窝)	泛指个人（构筑的）所谓安逸舒适、与世无争的生活环境。 snug retreat; cosy nest	ān (安)	lè(乐)	wō (窝)

However, since Morpheme Annotation Database divides morphemes by computer, setting a character as a unit, while characters do not necessarily correspond to words, errors of division may arise, so it is necessary to manually correct the data in later stage. As shown in the above table, “Ānlā” (安拉) is a transliterated word from Arabic, which should be one morpheme, but the computer divides it into two morphemes. Thus, in manual correction stage, “Ānlā” (安拉) is put in “morpheme 1”, whereas “lā” (拉) is deleted in “morpheme 2”. See the table below.

TABLE 2-2
MODIFIED EXAMPLES OF MORPHEME ANNOTATION DATABASE

ID	Word	Word definition	Morpheme 1	Morpheme 2	Morpheme 3
133	Ānlā (安拉)	A word transliterated from Arabic. Means Allah.	Ānlā (安拉)		

In addition, a total of 851 other words modified in the database: āfēi (阿飞), āyì (阿邑), Ēpánggōng (阿房宫), b ènd àn (笨蛋), bǎnlì (板栗), c ànl àn (灿烂), c àntou (孱头), duǒyí (朵颐), H óngm én (鸿门), Xuānwǔ (宣武), h úl ì(槲栎), hu ìy àn (会厌), hu ìhu ì(唠唠), Huáinánzǐ (淮南子), jījí (积极), l àngm àn (浪漫), Lǎozǐ (老子), Kǒngzǐ (孔子), sānmèi (三昧), sh ùndang (顺当), tǐngtuō (挺脱), w ánd àn (完蛋), suǒyǐ (所以), wèiyǔ (谓语), lěngbùdīng (冷不丁), lěngbùfáng (冷不防), xi ángsh í(翔实), Xuāntǒng (宣统), ch ích ú (踟蹰), yǔqí (与其), g àini àn (概念), zǐxū (子虚), zǒushuǐ(走水), zījiān (仔肩).

D. Establishment of Database of Morpheme Meaning Annotations for Modern Chinese Compounds

Firstly, we associate definitions of words under single-character entries in MCD 5 with morphemes in the established Morpheme Annotation Database, so that each morpheme is followed by a corresponding meaning. Generally, a Chinese character is a syllable, so we screen out words with one character in MCD 5, then we have multi-character words under single-character entries and their definitions in MCD 5, as shown in Table 2-3.

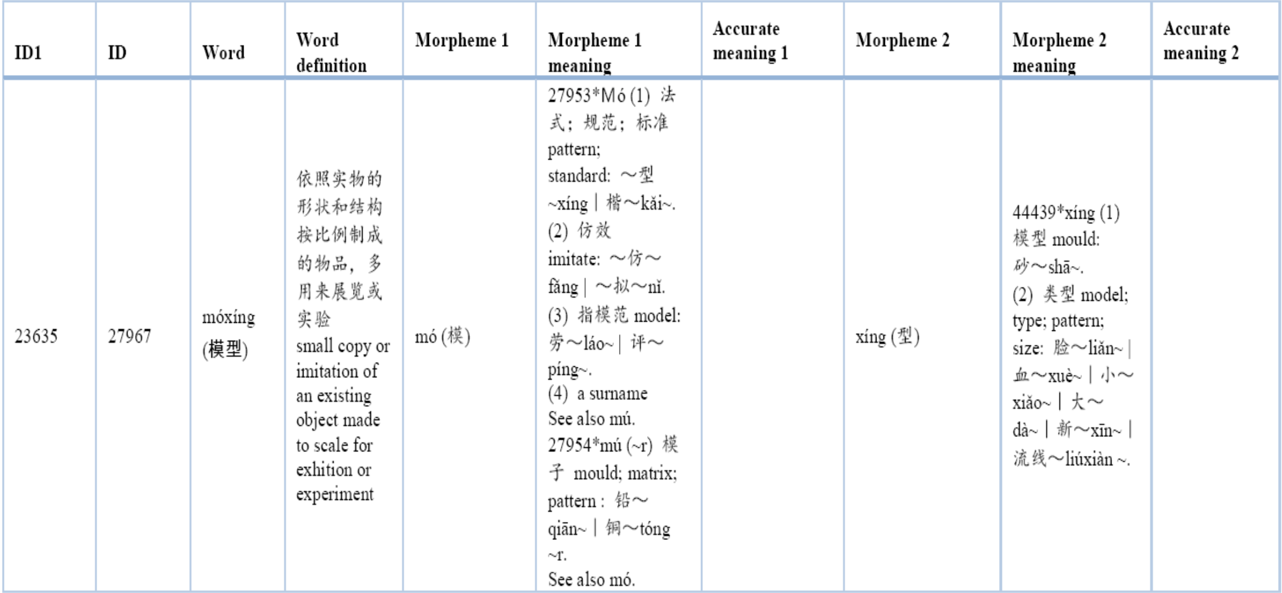
TABLE 2-3
EXAMPLES OF SINGLE-WORD ENTRIES IN *MCD*

Word	Definition
c án (蚕)	3788*C án 桑蚕、柞蚕等的统称，通常专指桑蚕。generic name for the silkworm, tussah, etc., usually referring to the silkworm
gē (歌)	13190*gē (1) 歌曲 song: 民～ m ín~｜山～儿 shān~r｜唱一个～儿 chàng yīgè~r. (2) 唱 sing: ～者~zhě |高～一曲 gāo ~ yīqǔ.
m ó (模)	27953*M ó(1) 法式；规范；标准 pattern; standard：～型~x íng｜楷～kǎi~. (2) 仿效 imitate: ～仿~fǎng｜～拟~nǐ. (3) 指模范 model:劳～ l áo~ |评～p íng~. (4) 名姓 a surname. See also m ú. 27954*M ú（～儿）模子(~r) mould; matrix; pattern：铅～qiān~｜铜～儿 t óng ~r. See also m ó.

After extracting multi-character entries in MCD 5, we associate morphemes with their meanings, following the steps below:

First, adding two fields after each morpheme in Morpheme Annotation Database, and naming them as “morpheme meaning” and “accurate meaning” respectively. Second, corresponding one to one between the extracted multi-character entries and the same morphemes in Morpheme Annotation Database. Finally, associating definitions of multi-character entries with the “morpheme meaning” field. After matching morphemes and their meanings, computer automatically generates a new ID, i.e., the “ID” column in the following table. For example, the morphemes of “móxíng” are “mó” and “xíng”. By adding fields “morpheme meaning” and “accurate meaning” after “mó” and “xíng” respectively, and definitions of “mó” and “xíng” into their corresponding field “morpheme meaning”, computer then automatically generates a new ID 27967. As shown in Table 2-4.

TABLE 2-4
EXAMPLES OF ASSOCIATIONS BETWEEN MORPHEME AND ITS MEANING



ID1	ID	Word	Word definition	Morpheme 1	Morpheme 1 meaning	Accurate meaning 1	Morpheme 2	Morpheme 2 meaning	Accurate meaning 2
23635	27967	móxíng (模型)	依照实物的形状和结构按比例制成的物品，多用来展览或实验 small copy or imitation of an existing object made to scale for exhition or experiment	mó (模)	27953*Mó (1) 法式；规范；标准 pattern; standard: ～型 ~xíng｜楷～kǎi~. (2) 仿效 imitate: ～仿～fǎng｜～拟～nǐ. (3) 指模范 model: 劳～láo~｜评～píng~. (4) a surname See also mú. 27954*mú (~r) 模子 mould; matrix; pattern：铅～qiān~｜铜～tóng ~r. See also mó.		xíng (型)	44439*xíng (1) 模型 mould: 砂～shā~. (2) 类型 model; type; pattern; size: 脸～liǎn~｜血～xuè~｜小～xiǎo~｜大～dà~｜新～xīn~｜流线～liúxiàn ~.	

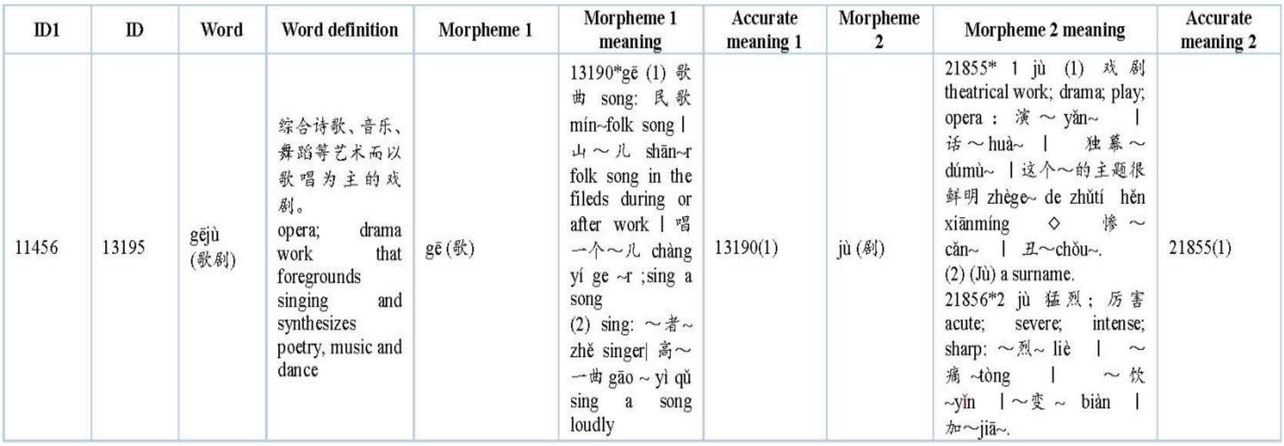
Regarding data deviations generated during the process of associating morphemes with meanings, manual annotations will be applied to ensure reliability and authenticity of the data. The steps to annotate entries are as follows:

In the first place, checking whether the compound morpheme annotations are right, and correcting them if there are mistakes. If so, we annotate meanings to ensure accuracy of the meaning annotations. Due to the drawbacks of computer annotations, there are certain errors in morpheme annotations. After data correction, two main types of morpheme annotation errors are found: first, compounds that should be annotated as multiple morphemes are annotated as single ones. For example, “gējù” (歌剧) is composed of two morphemes, but it is marked as a morpheme in Morpheme Annotation Database and is placed in the "morpheme 1” field. Second, words that should have one morpheme are annotated as multiple morphemes. For example, “Ālābó” (阿拉伯) in “Ālābóhǎi” (阿拉伯海) is a single morpheme, but it is annotated as three morphemes in Morpheme Annotation Database. In the above two types of situations, morphemes need to be corrected before their meanings are annotated. Furthermore, the method to modify morphemes is to split “gē” (歌) and “jù” (剧), putting them in “morpheme 1” and “morpheme 2” respectively, and to supplement their morpheme meanings respectively into “morpheme 1 meaning” and “morpheme 2 meaning” fields. Similarly, merging “ā” (阿), “lā” (拉) and “bó” (伯) into “morpheme 1” field, and moving the morpheme “hǎi” (海) and its definitions into “morpheme 2” and “morpheme 2 meaning” fields.

Secondly, annotating meanings for compounds with the right morpheme annotations. Choosing the exact meaning for each morpheme from “morpheme meaning” field, then filling morpheme’s ID and its meaning number into the field “accurate meaning”. Taking the term “gējù” (歌剧) (see Table 2-5) as an example, the ID of morpheme “gē” (歌) is “13190”, meaning “song” in the word “gējù” (歌剧), which corresponds to “(1) Song” in “morpheme 1 meaning”. Therefore, when manually annotating, “13190 (1)” is placed in “accurate meaning 1”. Similarly, the ID of morpheme “jù” (剧) is “21855”. In the term “gējù” (歌剧), the meaning of “jù” (剧) is “drama”, which corresponds to “(1) drama” in the field of “morpheme 2 meaning”. Therefore, “21855(1)” is annotated in “exact meaning 2” field.

Several points to note: (1) Some morphemes have only one meaning. In “morpheme meaning” field, there is only the ID of the morpheme, but no serial number annotation. Therefore, when manually annotating such morphemes, it is necessary to add “+(1)” after ID to distinguish the annotations of other morphemes in order to facilitate subsequent data analysis and processing. For example, in the word “biànlùn” (辩论), morpheme “biàn” (辩) has only one meaning, “debate”, with the ID “2577”. When annotating its meaning, it is marked as “2577+(1)”. (2) Some morphemes have multiple meanings, with one meaning corresponding to one ID, so when manually annotating the entries, special attention should be paid to the corresponding ID number to avoid reducing accuracy of the data. For example, when annotating morpheme “mó” (模) in “móxíng” (模型), “mó” (模) has two IDs, namely “27953” and “27954”. “*1” and “*2” represent the order of meanings of “mó” (模), and these two IDs correspond to the meanings of “pattern; standard”, “imitate”, “model” and “mould”. In the word “móxíng” (模型), the meaning of “mó” (模) is “pattern; standard”, which is the meaning (1) under ID27953, so “27953(1)” should be marked.

TABLE 2-5
EXAMPLES OF MEANING ANNOTATIONS



ID1	ID	Word	Word definition	Morpheme 1	Morpheme 1 meaning	Accurate meaning 1	Morpheme 2	Morpheme 2 meaning	Accurate meaning 2
11456	13195	gējù (歌剧)	综合诗歌、音乐、舞蹈等艺术而以歌唱为主的戏剧。 opera; drama work that foregrounds singing and synthesizes poetry, music and dance	gē (歌)	13190*gē (1) 歌曲 song: 民歌 mín~folk song | 山～儿 shān~r folk song in the fileds during or after work | 唱一个～儿 chàng yí ge ~r ;sing a song (2) sing: ～者~ zhě singer| 高～一曲 gāo ~ yì qǔ sing a song loudly	13190(1)	jù (剧)	21855* 1 jù (1) 戏剧 theatrical work; drama; play; opera : 演 ～ yǎn~ | 话～huà~ | 独幕～ dúmù~ | 这个～的主题很鲜明 zhège~ de zhǔtí hěn xiānmíng ◇ 惨～ cǎn~ | 丑～chǒu~. (2) (Jù) a surname. 21856*2 jù 猛烈；厉害 acute; severe; intense; sharp: ～烈~ liè | ～痛 ~tòng | ～饮 ~yǐn | ～变 ~ biàn | 加～jiā~.	21855(1)

According to incomplete statistics, the number of morphemes annotated with meanings in Meaning Annotation Database is about 105,000, including morphemes that have confirmed meanings, those that are difficult to accurately judge meanings, and those that need to supplement meanings from other dictionaries due to the lack of meanings in this database. Taking “biāo” (镖) as an example, in Meaning Annotation Database, the meaning of “biāo” (镖) is “lance; old-fashioned weapon thrown to injure or kill”, but if this meaning is used to explain “biāokè” (镖客), it is not accurate enough. So, the meaning “armed escort (of travellers or merchants’ caravans”1 is applied by referring to GCD.

III. QUANTITATIVE ANALYSIS

A. Using Frequency of Morphemes and Their Meanings

In this study, software Excel and Access are used to count the morphemes in Database of Morpheme Meaning Annotations for Compounds, arranged in order of using frequency. The statistics show that 10,688 unrepeated morphemes are in the database, and the total using frequency of all morphemes is 97,035 times, with the average frequency 9.08 times for each morpheme. 2,023 morphemes are used more than the average frequency, and 170 morphemes are used more than 100 times.

[1] *The Great Chinese Dictionary* (volume 11, page 1377)

TABLE 3-1
EXAMPLES OF MORPHEMES WITH USING FREQUENCY OVER 100 TIMES AND THEIR FREQUENCY

Morpheme	Frequency	Morpheme	Frequency	Morpheme	Frequency	Morpheme	Frequency	Morpheme	Frequency	Morpheme	Frequency
zi (子)	1111	xiǎo (小)	217	shū (书)	170	jūn (军)	138	m íng (名)	121	h ào (号)	108
er (儿)	698	fǎ (法)	216	r án (然)	168	sh í(时)	137	y òng (用)	121	gu ò(过)	107
r én (人)	386	fēn (分)	215	jiā (家)	167	p í(皮)	137	q ì(器)	120	lu ò(落)	107
shuǐ (水)	338	x íng (行)	214	yǎn (眼)	165	d ìng (定)	137	chǎng (场)	118	z h u ǎ n (转)	106
shēng (生)	338	mi àn (面)	212	m én (门)	165	y è(业)	135	x u è(血)	118	b àn (半)	106
t óu (头)	334	x ìng (性)	211	d ào (道)	163	s è(色)	135	r ì(日)	118	x ìn (信)	106
d òng (动)	320	zhōng (中)	210	tōng (通)	163	sh í(石)	134	du ì (对)	118	g àn (干)	105
xu é(学)	316	zh àn (战)	209	chǎn (产)	160	hǎi (海)	134	r è(热)	117	ch óng (虫)	105
d **ì**(地)	311	xi àn (线)	206	m ín (民)	158	qǐ (起)	134	sh í(实)	115	guǎu (管)	104
xīn (心)	300	kǒu (口)	205	lǎo (老)	157	cǎo (草)	134	z ú(族)	115	c ài (菜)	104
di àn (电)	294	fā (发)	203	lǐ(理)	157	shān (山)	133	zh èng (政)	115	ti áo (调)	104
d à(大)	292	sh ì(事)	203	l ù(路)	157	zu ò(作)	132	y ùn (运)	114	x i ā n g (相)	104
huā (花)	290	z ì(自)	199	gāo (高)	155	zh èng (正)	132	biǎo (表)	114	zī (资)	104
q ì(气)	283	w ài (外)	197	ni án (年)	150	ch éng (成)	131	pǐn (品)	114	li ào (料)	104
b ù(不)	277	zhǔ yì (主义)	191	ji ào (教)	149	jīng (经)	131	c í(词)	114	zhǒng (种)	103
gōng (工)	267	běn (本)	189	p íng (平)	149	z ì(字)	130	hu í (回)	113	m ù(木)	103
w ù(物)	263	yī (一)	189	zhǔ (主)	148	tǔ(土)	129	li àng (量)	112	b ào (报)	102
jī (机)	252	guāng (光)	184	chē (车)	146	diǎn (点)	129	jiǎo (角)	111	bǎo (保)	102
hu à(化)	244	hu ì(会)	182	y óu (油)	146	dǎ(打)	129	hu à (话)	111	m áo (毛)	102
tǐ(体)	241	b ìng (病)	180	fāng (方)	146	hu áng (黄)	127	y ì(义)	110	yǔ (语)	102
h é(合)	240	w ú(无)	179	jiāo (交)	146	z h ǎ n g (长)	126	d é (得)	110	hēi (黑)	102
gu ó(国)	240	sh àng (上)	179	chū (出)	144	l ùn (论)	125	sān (三)	110	f àng (放)	101
shǒu (手)	235	zh ì(制)	178	h óng (红)	143	yǒu (有)	124	mǎ (马)	110	bǎn (板)	101
tiān (天)	232	sh ù(数)	177	q íng (情)	142	tiě (铁)	124	yu án (原)	109	y ìng (应)	101
huǒ(火)	228	kāi (开)	177	d ù(度)	141	sh én (神)	123	m íng (明)	109		
l ì (力)	227	gōng (公)	176	zh òng (重)	140	fǎn (反)	123	n èi (内)	109		
w én (文)	227	yīn (音)	176	jīn (金)	140	t óng (同)	123	gǔ(骨)	109		
b ái (白)	224	xi à(下)	172	h òu (后)	140	y ì(意)	122	bīng (兵)	109		
fēng (风)	220	li ú(流)	170	kōng (空)	139	bi àn (变)	121	shēn (身)	108		

Although the above morphemes are used frequently in compounds, not all of them have high using frequency for all meanings. Then, this study further ranks the using frequency of all morpheme meanings in the database from high to low to recognize the most frequently used meanings in high-frequency morphemes. For convenience, this paper only shows morpheme meanings that are used more than 100 times (include 100). As shown in the table below:

TABLE 3-2
EXAMPLES OF MEANINGS' USING FREQUENCY

Morpheme	Meaning	Frequency
zi (子)	noun suffix	862
er (儿)	suffix	517
rén (人)	higher animal that can make tools and use them in labor	386
shuǐ (水)	The simplest hydroxide compound combining two hydrogen colorless, odorless and tasteless liquid oxide of hydrogen (H_2O)	318
dà(大)	(as opposed to "small") big; large; great (in volume, area, quantity, force, strength, etc.)	304
diàn (电)	physical phenomenon arising from the existence and change of electric charge, an important energy used extensively in production and daily life to provide light, heat, power, etc.	304
wù (物)	thing; matter; substance	288
dòng (动)	(as opposed to "still") change the place or position of sth..	283
xīn (心)	usually also referring to the heart; mind; feeling; intention	258
xiǎo (小)	(as opposed to "big") small; little; petty; minor;not up to the average, or not comparable in such aspects as volume, area, quantity, strength, intensity, etc.	233
bù (不)	used before verbs, adjectives and other adverbs to indicate negation	227
guó (国)	country; state; nation	225
xué (学)	subject of study; branch of learning	194
zì(自)	self; oneself; one's own	167
shǒu (手)	hand	162
huǒ(火)	fire; light and flame caused by burning	162
zhàn (战)	war; warefare; battle; fight; combat	157
bái (白)	(as opposed to "black") white; color of fresh snow or frost	157
hé(合)	(as opposed to "separate") join; combine	156
jī (机)	machine	143
bìng (病)	abnormal physiological or mental condition; disease; illness; sickness	138
jūn (军)	armed force; army; troops	130
rán (然)	adverb or adjective suffix	129
zhǔyì(主义)	-ism; systematic doctrine or theory on the objective world, society or academic issues	126
yīn (音)	sound	125
huā (花)	flower; shoot of the sporophyte of a seed plant modified for reproduction, having leaves, calyces, thalami, and pistils, some featuring splendid colours and emitting fragrance	122
yóu (油)	oil; fat; grease; petroleum; liguid fat contained in animals and plants, or mixed mineral liquids of hydrocarbon compounds	120
gōng (工)	work; labour	119
xìng (性)	property; quality; attribute of sth. resulting from a certain element contained in it	117
kǒu (口)	mouth; human or amimal's organ for taking food and uttering sounds	116
shì(事)	Matter; affair; thing; business	114
mín (民)	the people	110
huáng (黄)	yellow,like the flowers of towel gourd or sunflower	110
yǎn (眼)	eye; human or animal organ of vision or of light sensitivity	109
fēng (风)	wind; breeze; gale; air current moving approximately parallel to the ground surface, caused by uneven distribution of atmospheric pressure	109
hóng (红)	red; colour of blood or the pomegranate flower	108
guāng (光)	light; electromagnetic radiation that acts upon the retina of the eye, optic nerve, etc., making sight possible; matter that shines over an object, making it visible to the eye, such as sunlight, lamplight, and moonlight.	108

The high-frequency morphemes in Table 3-1 and Table 3-2 are simple, commonly used and close to daily life [such as shuǐ (水), diàn (电), yóu (油), guāng (光), chóng (虫), shǒu (手), huǒ (火), etc.]. By comparing Meaning Annotation Database and "accurate meaning" field, it is observed that high-frequency meanings are those in "morpheme meaning" field with serial number (1) or (2). Furthermore, there are no identical morphemes shown in the table. In other words, apart from the above morpheme meanings, no other meanings are applied more than 100 times. At the same time, other morpheme meanings with lower using frequency and more abstract and general meanings are not included in the table, indicating that among various meanings of one morpheme, only one meaning is more often applied. For example, the high-frequency meaning of "fēng" (风) is "air current moving approximately parallel to the ground surface, caused by uneven distribution of atmospheric pressure" The using frequency of it is 109 times, however, other meanings of "fēng" (风) do not exceed 100 times, indicating that not all meanings of a morpheme are frequently used, and some high-frequency meanings, such as "custom", "scene", and "attitude", are more abstract and difficult to understand. In addition, meanings with higher frequency are relatively simple and specific, such as "zì (自, self)", "kǒu (口, mouth)", etc. Otherwise, meanings are abstract and difficult to understand, such as "zì (自, from; since)", "kǒu (口, gate)", etc.

It can also be seen from Table 3-1 and Table 3-2 that morphemes sorted according to using frequency of meanings are all included in Table 3-1, but they have different sequences from Table 3-2. For example, in Table 3-1 the using frequency of morpheme “guāng” (光) is higher than that of “mín” (民), but in Table 3-2, the usage frequency of “mín” (民) is higher than that of “guāng” (光). This shows that the using frequency of a morpheme is not necessarily positively related to that of its meanings.

In addition, morphemes with high using frequency may not have meanings with higher using frequency. The reasons may be: (1) a morpheme may have many meanings, with each of them used, so their using frequency could scattered. For example, there are 212 compounds composed of morpheme “miàn” (面), and “miàn” (面) has 12 meanings with each of them applied in different compounds, so there is no certain meaning applied collectively. (2) Conditions of word-formation by morphemes are single. For example, the using frequency of morpheme “bīng” (兵) is 109 times, while that of its highest frequency meaning is only 55 times. Compounds with “bīng” (兵) include bīngqì (兵器), bīngfǎ (兵法), qíbīng (骑兵), zhǐshàngtánbīng (纸上谈兵), etc.. Although many compounds include “bīng” (兵), the use of meanings is limited to a specific condition, so the using frequency is lower than that of other common meanings.

B. *Internal Structure*

This research extracts 6,806 compounds applying morpheme meanings in Table 3-2 and sets these compounds as the research object to divide their structures. The results show: the compounds with noun-modification structure are the most, with the amount 5,119, accounting for 75.3% of the total. Followings are 729 compounds with verb-object structure, accounting for 10.7% of the total. 499 compound words have affixation structures, accounting for approximately 7.3% of the total; the total number of compounds with predicate-modification structure, subject-predicate structure, parallel-combination structure and verb-complement structure is 459, occupying around 6.7% of the total. E.g.:

Compounds with noun-modification structure: mínfǎ (民法), cháhuā (茶花), ěrjī (耳机), chūnguāng (春光), fēngshēng (风声)

Compounds with verb-object structure: tiǎozhàn (挑战), dòngxīn (动心), chùdiàn (触电), cānzhàn (参战), hébì(合璧)

Compounds with affixation structures: Ànzi (案子), diézi (碟子), chúr (雏儿), ànrán (黯然), ǒurán (偶然)

Compounds with predicate-modification structure: huǒhóng (火红), xuěbái (雪白), làhuáng (蜡黄), bùxǔ (不许), huǒhóng (火红)

Compounds with subject-predicate structure: xīnzuì (心醉), xīnxū (心虚), zìchēng (自称), shuǐshí (水蚀), mínzhǔ (民主)

Compounds with parallel-combination structure: guójiā (国家), rénmín (人民), kǒuchǐ (口齿), huācǎo (花草), shēngyīn (声音)

C. *Semantic Logic*

A large number of morphemes in compounds in modern Chinese cannot directly express meanings of words. If we only analyze morpheme meanings and internal structures of compounds, it is difficult to explain the compounds derived from metaphor, metonymy, extensions, etc. Therefore, we explore the relationship between morpheme meanings and compounds from the perspective of semantic logic of word formation.

First of all, from Database of Morpheme Meaning Annotation for Compounds, we screen out fields containing “bǐyù (比喻 fig.)” in the “meaning” field, and the selected words should be compounds containing figurative meanings. Secondly, with the same method, fields containing “zhǐ(指 refer to)” are screened out in the database, and through manual correction, compounds with figurative meaning, extended meaning, or metonymy are selected. Finally, from the compounds selected in the previous two steps, compounds containing morphemes and morpheme meanings in Table 3-2 are selected. In the end, we have 261 compounds whose morphemes cannot directly express the meaning of the word, such as kāiyóu (揩油), xīnfēi (心扉), chīxīn (吃心), chánshǒu (缠手), báiyǎn (白眼), hóng’àn (红案), kǔhǎi (苦海), yǎnzhōngdīng (眼中钉), luòshuǐgǒu (落水狗), yángchángxiǎodào (羊肠小道), èyúyǎnlèi (鳄鱼眼泪), etc.

Taking these 261 compounds as an example for a brief analysis, the semantic logic of word formation can be divided into the following categories:

The first category is: word definitions are the abstraction of entities signified by morpheme meanings. Generally, such compounds contain figurative meanings, and learners can understand the meaning of words through the similarity between the signified and the signifier. For example, “xīnfù” (心腹) originally means “heart and belly”, which are essential to human body and are used to indicate “trusted subordinate or reliable agent”; also, the term “émáo” (鹅毛) uses the light weight of goose feathers to compare to “something as light as a goose feather”; “tiěrén” (铁人) selects the hard and firm nature of iron to describe people of exceptional physical and moral strength. More such words include “chánshǒu (缠手, (of things) troublesome; hard to deal with)”, “xiǎoxiér (小鞋儿, difficulties created or unfair treatment given in secret)”, “kāiyóu (揩油, get petty gains at the expense of the government or someone else)”, “fēngshuāng (风霜, hardships experienced in life or during a journey)”, etc.

The second category is: word definitions are the abstraction of senses signified by morpheme meanings. Such morphemes refer to those that can represent sense of touch, sight, hearing, taste, etc. For example, "huǒrè" (火热) selects people's feeling of touching fire to describe temperature, atmosphere, feelings, etc., as hot as fire; "hēixīn" (黑心) literally means "black heart", and since "hēi" (黑) is commonly used in Chinese to represent issues that are "dark, secret, evil, and wicked", "hēixīn" then becomes a metaphor for "evil mind". Similar words are "xīnhán (心寒, be bitterly disappointed), jiūxīn (揪心, anxious; worried), lěngshuǐ (冷水, dampen the enthusiasm of), hóngrén (红人, favourite person of sb. in power), yǎnrè (眼热, cast covetous eyes at sth.), hóngyán (红颜, pretty woman), léimíng (雷鸣, thunderous), làshǒu (辣手, ruthless method)" and so on.

The third category is: word definitions are the abstraction of spatial meanings signified by morpheme meanings. Morphemes that represent spatial meanings are "shàng (上), xià(下), lǐ (里), wài (外), zuǒ (左), yòu (右), tou (头), gāo (高), dī (低), biān (边), shēn (深), qiǎn (浅)", etc., and the compounds formed by them include "xīndǐ (心底), kǒutóu (口头), zuóyòushǒu (左右手), hǎinèi (海内), shǒubiān (手边), méiyǎngāodī (眉眼高低)", etc. These words spatialize abstract things, with most of them used to reflect people's emotional state or value judgement. For example, people compare "zuóyòushǒu" (左右手) to right-hand man, which reflects a high evaluation of a capable assistant; "dǐ" (底) represents the bottom or base, which refers to places uneasy to be seen, such as "hǎidǐ (海底), chuángdǐ (床底), bēidǐ (杯底)". Therefore, "xīndǐ" (心底) refers to people's innermost world, containing most secret and hidden feelings and thoughts.

This research starts from the perspective of the usage frequency of morpheme meanings, the internal structure of compound words and the semantic logic relationship between word definitions and morpheme meanings, and makes a brief exploration of compound words, which has certain significance for lexical studies and natural language processing.

IV. Application Prospect

A. Establishment of a Research Platform of Modern Chinese Compound

The construction of "Morpheme Annotation Database for Compounds" starts with the morphemes of modern Chinese compounds, analyzes word formation, and statistically analyzes the using frequency of morphemes and their meanings. This can be applied as a start point for building a systematic and comprehensive platform for the study of modern Chinese compounds, in order to facilitate the study of morpheme meanings in lexicology.

B. Improvement of Accuracy of Natural Language Understanding

The database accurately annotates morphemes and their meanings, and extracts all compounds composed of morpheme meanings that are used more than 100 times. Taking this as an example, from the perspective of internal structure of compounds and logical relationship between morpheme meanings and word definitions, the rules of compound word formation have been revealed. Therefore, this research is significant for improving accuracy of machine translation and semantic analysis.

C. Development of Chinese Vocabulary Learning APP

The database classifies morphemes and their meanings. Each morpheme can be combined with different morphemes, and each morpheme (Chinese character) has a detailed explanation. As a result, APPs for Chinese vocabulary learning can be designed according to this feature to help learners quickly acquire the pronunciation, writing and meaning of Chinese vocabulary. It can also be associated with the corresponding vocabulary, adding pictures, audios, videos, etc. to improve the learner's learning efficiency.

V. Conclusions

By a close study on compounds in Modern Chinese Dictionary, this research establishes Database of Morpheme Meaning Annotations for Compounds, identifies morphemes in compounds and their meanings, and counts up the using frequency of these morphemes and their meanings in compounds. Moreover, compounds that are composed of morphemes and meanings applied over 100 times are extracted, which are studied from the aspects of inner structure, word meaning and semantic logic. The results indicate the application values of the database in establishing research platform for modern Chinese compounds, improving accuracy of natural language understanding, developing APPs for Chinese vocabulary leaning, etc. Although this research has limitations, it can provide certain references for language understanding and language generation in the field of natural language processing, and can offer data support for developing learning software on mobile terminals as well.

References

[1] Editorial Committee of The Great Chinese Dictionary. (2010). The Great Chinese Dictionary. Sichuan: Sichuan Lexicographical Publishing House Co., Ltd.11, page 1377.

[2] Fei Song. (2014). Construction and Application of Data Resource Platform for the Whole Field of International Chinese Teaching—Taking “International Chinese Teaching Database” as an Example. *Association for Modernization of Chinese Language Education*, 8.

[3] Heting Yang, Shichun Li. (2019). Study on the Distribution of Meaning Items of “Tiao（跳）” in Modern Chinese Based on Corpus. *Journal of Pingxiang University*, *01*, 80-83.

[4] Hongbo Yin. (2016). An Analysis of the Statistics of the Revision to the Part of Speech Tagging in the Contemporary Chinese Dictionary (6th Edition). *Journal of Graduate School of Chinese Academy of Social Sciences*, *03*, 108-114.

[5] Jing Wang, Lijiao Yang, Hongfei Jiang, et al. (2017). A Word Sense Annotated Corpus for Teaching Chinese as Second Language. *Journal of Chinese Information Processing*, *01*,221-229.

[6] Lilin Guo. (2018). Research on Chinese Language Characteristics——Analysis of the Teaching of Chinese Synonyms to Foreigners. *Think Tank Era*, *26*, 235-236.

[7] Rong Cheng. (2017). Some Issues on Neologism Selection and New Meaning Item Institution. *Lexicographical Studie*s, *03*, 1-9+93.

[8] Zhiwei Lu. (1957). Lexical Morphology of Chinese. Shang Hai: Zhonghua Book Company.

Yu Xu was born in Rizhao, China in 1992. She graduated from Beijing International Studies University and has obtained a master's degree in Teaching Chinese to speakers of other languages. Author of "experiencing Chinese VR audio visual oral course | intermediate 1" and "experiencing Chinese VR audio visual oral course | intermediate 2". She majored in the application of corpus in language teaching and research, Chinese vocabulary and grammar teaching, Lesson plan design and Compilation, introduction to Chinese as a second language teaching, second language acquisition, cross-cultural communication, psychology of Chinese as a foreign language.

Weiqi Chen has been teaching Chinese to speakers of other languages at Beijing International Studies University since 2002. She has written several articles and textbooks for the teaching of Chinese language, including Reading and Writing Chinese Characters. Her areas of interests include cognitive linguistics and L2 listening pedagogy.

Fei Song was born in Linyi, China in 1986. He received his PH.D. degree in linguistics from jointly trained by Central University for Nationalities and Columbia University. He is associate professor, tutor of master's degree. He mainly engaged in language intelligence, international Chinese teaching measurement research, cultural science and technology integration research. At present, he has presided over a number of scientific research projects such as the Beijing Social Science Fund project. As a backbone member, he has participated in a number of major, key and general topics of the National Social Science / national self Science Fund. He has published one monograph, two participated in writing and editing, and published more than ten papers in Chinese and foreign academic journals. He has won the first prize of Beijing University Young Teachers' teaching basic skills competition, and compiled and published the first set of VR in China Language textbooks. At present, he is the vice president of Chinese Language Institute of Beijing International Studies University.