

MARVS Revisited: Incorporating Sense Distribution and Mutual Information into Near-Synonym Analyses*

Siaw-Fong Chung and Kathleen Ahrens
National Taiwan University

In MARVS (Module-Attribute Representation of Verbal Semantics), verbs are differentiated based on eventive information, which is comprised of event modules and role modules. Huang et al. (2000) used MARVS to examine near-synonyms and suggested that it can highlight the difference between synonymous sets. This paper suggests that the operational steps underlying a MARVS analysis can be improved by analyzing the sense distribution of the near synonyms and by looking at the Mutual Information values of the collocating words. Both these steps increase the verifiability of the semantic analysis in MARVS and set the groundwork for automatic extraction of lexical meaning.

Key words: near-synonyms, MARVS, sense, Mutual Information value, 擺 *bǎi*, 放 *fàng*, 'put'

1. Introduction

MARVS (Module-Attribute Representation of Verbal Semantics) (Huang, Ahrens, Chang, Chen, Liu & Tsai 2000) is a linguistic model which is powerful in distinguishing near-synonyms using eventive information such as event modules and role modules. Therefore, two closely related words can be distinctively disambiguated based on their inherent semantic differences. The model, which was proposed in 2000, can still be improved by adding computational linguistic resources that were developed after 2000. Related work can also be found in Tsai, Huang, Chen & Ahrens (1998), and Liu (2003).

Previous models for near-synonyms were usually based on two types of analyses, namely a descriptive analysis and a quantitative analysis. Descriptive analyses usually

* We would like to thank the two anonymous reviewers for their comments on this paper. We would also like to acknowledge the NSC project grant to the second author (NSC 94-2411-H-002-038) and the Academia Sinica Investigatorship project grant "Lexicon-driven Ontology and Conceptual Structure" to Professor Chu-Ren Huang for supporting the discussion herein. We also thank Professor Chu-Ren Huang for his comments on this work.

depend on intuition with the aid of additional references such as dictionaries. Quantitative analyses, on the other hand, usually attempt to find out the differences between near-synonyms through comparing the behaviors of the synonymous pairs such as by comparing the argument types of the pairs attested.

Earlier work on synonyms tended to focus on providing descriptive information, such as that demonstrated by Collinson (1939, cited in Harris 1973:14), who attempted to list the possible differences between synonyms using nine elements: general/specific applicability, intensity, emotion, moral approbation, professionalism, written/non-written, colloquialism, local/dialect, and child talk. Today, the commonly agreed differences between synonyms are found within features such as connotations, implications, selectional restrictions, and syntactic variations (i.e. Cruse 1986, Lyons 1995, DiMarco, Hirst & Stede 1993, Edmonds 1999). Cruse, for example, considered selectional and collocational restrictions the “main effect of presupposed semantic traits of a lexical item,” which brings out the “syntagmatic companions” of words (1986:278-279). Near-synonyms, according to Cruse, are “lexical items whose senses are identical in respect of ‘central’ semantic traits, but differ... in ‘minor’ or ‘peripheral traits’” (1986:278-279). In fact, most synonyms are near-synonyms which share certain central similarities and peripheral differences. Perfect synonyms are considered rare (Taylor 2002:265).

Later work on distinguishing near-synonyms used both descriptive as well as quantitative analyses. For instance, Taylor’s (2002) analysis of ‘tall’ and ‘high’ was carried out through psycholinguistic experimentation (acceptability rating tasks) in addition to descriptive analyses of the two adjectives. Taylor claimed that ‘tall’ and ‘high’ can be differentiated using MacLaury’s (1997, 2002) Vantage theory, which distinguish near-synonyms in terms of dominant/recessive meanings. For both these adjectives, the dominant meaning emphasizes the similarity of “a fixed landmark which is the human body sanctions the application of the word to a limited range of prominently upright entities”. However, ‘tall’ has restrictions on dimensional uses whereas ‘high’ has restrictions on positional uses.

More statistical approaches to near-synonyms can be seen in the computational field. For example, a statistical analysis of near-synonyms by Church et al. (1994) use Mutual Information (MI), as well as substitutability in terms of T-scores to differentiate between the near-synonyms ‘request’ and ‘ask for’. MI values measure the degree of co-occurrences between terms, so as to determine whether a word is a collocate to another word. They found twenty-eight significant objects that collocate with both ‘request’ and ‘ask for’, among which are ‘aid’, ‘assistance’, ‘copy’, ‘dismissal’, and ‘extension’. In addition, Church et al. (1994) showed that near-synonyms can be compared in terms of their substituted words. For example, they found that ‘request’ has a higher substitutability value than ‘ask for’ when substituted by words such as ‘seek’,

‘grant’, and ‘demand’. Therefore, substitution is one way to test the similarities and differences of near-synonyms, i.e. through comparing whether or not the same words can substitute for the arguments of the two compared near-synonyms.

Also important in computational approaches to near-synonyms is the model suggested by Pustejovsky (1991). In this model, the meanings of the verbs can be generated from the nominals surrounding the verbs by examining the Qualia structures of the verbs, i.e. the structure of the nominals are co-compositional by four types of roles, shown in (1) below (Pustejovsky 1991:426-427).

- (1) a. Constitutive role (the relation between an object and its proper parts such as ‘narrative’ for a novel)
- b. Formal role (role that distinguishes the object within a larger domain such as ‘book’ or ‘disk’ for a novel)
- c. Telic role (the purpose and function of the object such as ‘read’ for a novel)
- d. Agentive role (factors involved in bring about the object such as ‘artifact’ or ‘write’ for a novel)

Based on these four roles, the author claims that the load of distinguishing verb meanings can be distributed to the nominals (or adjectives) surrounding the verbs. This model has later been used in various linguistic theories. The strength of this model comes from its prediction of the possibilities to distinguish a term from another based on their ‘part-of-relation’ (constitutive role), ‘kind-of-relation’ (formal role), ‘function relation’ (telic role), and ‘origin relation’ (agentive role). (These four relations are also stated in Croft & Cruse 2004:137.) When near-synonyms are concerned, the peripheral differences between a synonymous pair can occur at any of these four aspects. In other words, these four aspects provide alternative ways of stating the differences between a synonymous set, in addition to stating the differences in semantic features as was done in traditional semantics (i.e. [\pm female], [\pm animate], etc.) or that in the work of Collinson (1939).

The MARVS model (Huang et al. 2000) shares several assumptions with recent work on lexical semantics. For example, the first assumption of MARVS is that lexical semantic information can be used to predict grammatical behavior (cf. Dowty 1991, Levin 1993, Goldberg 1995). An additional assumption is that lexical semantics is grammatically based and mediates conceptual structures (cf. Bresnan & Kanerva 1989, Zaenen 1993, Pustejovsky 1991). Thus, the MARVS model proposes that an adequate theory of verbal semantics must be able to represent directly semantic information in a way that can be connected to grammatical structures, such as through event structure. In

addition, the lexical semantic information must have conceptual motivation (i.e. similar to the qualia structure as suggested by Pustejovsky 1991). Lastly, all lexical semantic attributes must be data-based, using information such as collocating structure, selectional constraints, or distributional patterns (Huang et al. 2000). Gathering these linguistic assumptions, MARVS turns out to be predictive when comparing words with almost synonymous meanings. This paper proposes that the distributional information of sense frequency and collocational-based information can further refine the steps needed to run a lexical semantic analysis of near-synonym verb pairs.

2. Module-Attribute Representation of Verbal Semantics

MARVS lays out eventive information in terms of event modules and role modules. (See Figure 1 below taken from Huang et al. 2000:24; arrows added.)

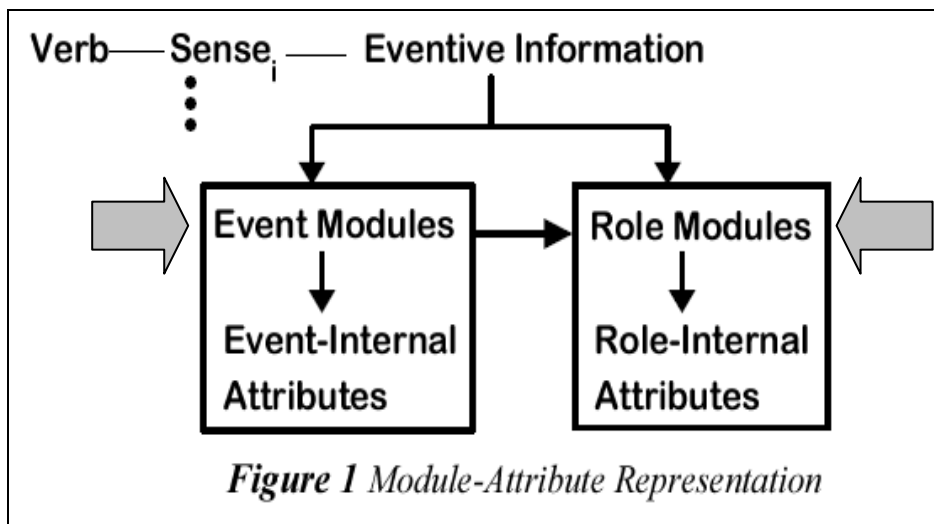


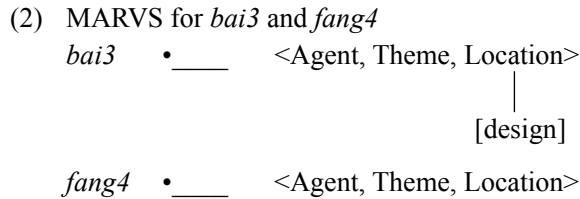
Figure 1: Eventive information in MARVS

Under each module (see arrows), there are attributes which further define the behaviors of the module. Some examples of role-internal attributes are [sentience], [volition], [affectedness], and [design]. Examples of event-internal attributes are [control] and [effect].

In Huang et al. (2000), two verbs of ‘put’ in Chinese (擺 *bǎi* and 放 *fàng*) are found to differ at the [design] of the role-internal attribute because the way of putting is different for the two verbs, with 擺 *bǎi*, but not 放 *fàng*, entailing the act of “putting

following a certain plan” as well “a resultant state”.¹

The following diagram shows the differences between 擺 *bǎi* and 放 *fàng* (Huang et al. 2000:36).



The methodology of MARVS is corpus-based, and contrasts the differentiation between the near-synonyms in terms of attributes. Therefore, it has the combination of both the descriptive approach and the corpus-based approach. This work proposes two additional steps to MARVS in order to further clarify what is needed for an accurate verbal semantic analysis. The first step includes analyzing instances from corpus so as to establish the similarities and differences between near-synonyms. The second step suggests using Mutual Information (MI) values to look for argument types that collocate with the verbs and suggests criteria for the selection of collocates based on these values, as the usual results from the calculation of MI values contain noise which has to be filtered out manually. This paper will address these issues in detail through the use of examples of two verbs of ‘put’ (擺 *bǎi* and 放 *fàng*) following Huang et al. (2000), and Ahrens, Huang & Chuang (2003).

3. Proposals for improving the MARVS analysis

Scholars such as Tognini-Bonelli (2001) distinguish between corpus-based and corpus-driven analyses. Corpus-based analysis uses corpus as resources of examples for verifying intuition. Corpus-driven analysis, on the other hand, allows the discovery of new sentence patterns for the purpose of research. MARVS is a model that is corpus-based because it, to date, has been based on selective use of sentences from a large corpus (cf. Huang et al. 2000 and Ahrens, Huang & Chuang 2003).

Ahrens, Huang & Chuang (2003), for example, suggested that the different meanings of the English ‘set’ and the Chinese 擺 *bǎi* can be represented in MARVS. However, they only took a selection of example sentences from the corpus. The original steps of

¹ Chinese terms of 擺 *bǎi* and 放 *fàng* are added in this paper. Huang et al. (2000) use *bai3* and *fang4* instead. Pinyin in this paper is generated based on the Pintone software (Teng, Cheng & Lin 2006).

MARVS are re-stated by Ahrens, Huang & Chuang (2003:470; underline and bold added).

How do we determine these collateral differences? First, we examine these near-synonym pairs by first combing a corpus for all relevant examples of the words in question. These examples are then categorized according to their syntactic function. **Third, each instance is classified into its argument-structure type.** Fourth, the aspectual type associated with each verb is determined. And fifth, the sentential type for each verb is also determined.

The underlined step above shows that the study did not collect and analyze a set of random sentences from the corpus, but only extracted relevant examples needed. However, if all the meanings of all sentences in a random set of two near-synonyms are analyzed, one can obtain information such as (a) the meaning shared by the pairs (i.e. the ‘central semantic traits’) and (b) the differences in meanings between the pairs (i.e. the ‘peripheral traits’) (cf. Cruse 1986:278-279).

Recent advances in corpora tools also allow for advances to be made in semantic analysis. For example, one can now obtain information about collocations in terms of MI values. This information can be found in the Academia Sinica Balanced Corpus of Modern Chinese (Sinica Corpus) for the analysis of the Chinese synonyms, as well as in the British National Corpus for the analysis of ‘set’ in English.² By adding this information, one can reduce manual work in generating the argument types for the synonyms (as was bolded in the quotation of methodology by Ahrens, Huang & Chuang (2003) above).

Finally, one important issue that remains in MARVS is the arbitrariness of the attributes. However, since this is also a problem for traditional semantics as well as in most feature-identifying models, this issue will not be addressed in this paper. Instead, we shall focus on adding two additional steps to the original methodology for lexical semantic analysis in MARVS. To do so, this paper re-analyzes 擺 *bǎi* and 放 *fàng* so as to demonstrate how the additional steps support the original analysis.

4. Reanalysis of 擺 *bǎi* ‘put’ and 放 *fàng* ‘put’

The revised steps for a near-synonym analysis in MARVS, with inclusion of two additional steps (the second and fourth steps) and the modification of the first step, are given in (3) below. The modified step and the two additional steps are in italic bold face.

² Sinica Corpus is available at <http://www.sinica.edu.tw/ftms-bin/kiwi1/mkiwi.sh> while the British National Corpus is available at <http://www.natcorp.ox.ac.uk/>.

(3) **Near synonym analysis in MARVS**

First, examine the near-synonym pairs by analyzing *at least the first 100 examples from the corpus*.

Second, analyze the senses either according to intuition or the meanings in WordNet so that the similarities (i.e. the pair is nearly synonymous) and differences of sense can be identified.

Third, categorize the examples according to their syntactic function.

Fourth, classify its argument-structure type based on their collocation restrictions discovered through MI values.

Fifth, determine the aspectual type associated with each verb.

Sixth, determine the sentential type for each verb.

First, we suggest that a random sample of sentences be collected from the corpus. The number of examples should be consistent for all the synonymous set attested. In the second step, we suggest that these examples are then analyzed either manually or by using a reference such as WordNet (<http://wordnet.princeton.edu/>). The purpose of this is to find out the similarities and differences in meanings between the synonymous set. This step is also important in that it proves that the items in the synonymous set are indeed synonyms, i.e. they share at least one similarity in meaning despite other differences. The fourth step suggests that collocations and MI values can be used as criteria to determine the arguments of the synonyms. As Palmer (2000) said, consistent concrete criteria have to be stated clearly for discovering sense distinction. The aim of adding our proposal is to make the MARVS model more operationalized and thus more easily applied to other verb pairs. In the next section, we shall take 擺 *bǎi* and 放 *fàng* as an example and demonstrate how these two additional steps can be conducted.³

4.1 Sense distribution analyses

First, to prove that 擺 *bǎi* and 放 *fàng* were synonymous in meaning, sentences containing these two verbs were extracted from the Sinica Corpus and analyzed. Since 擺 *bǎi* and 放 *fàng* are Chinese words, they should be searched as Chinese words in order to obtain all their senses. SinicaBow (Huang, Chang & Lee 2004) is one of the tools that provides a Chinese-English search interface for senses.⁴ For instance, the

³ Similar steps can be applied to improve the analysis of the English ‘tall’ and ‘high’ which was carried out in Taylor (2002).

⁴ SinicaBow or the Academia Sinica Bilingual Ontological WordNet is available at <http://bow.sinica.edu.tw/>.

senses of 擺 *bǎi* and 放 *fàng* are given in (4) below when searched in SinicaBow (based on WordNet 1.7.1).

- (4) Senses from WordNet 1.7.1 obtained when searched in SinicaBow
- a. 擺 *bǎi* ‘put’:
1: Arrange thoughts, ideas, temporal events, etc.
2: An apparatus consisting of an object mounted so that it swings freely under the influence of gravity
 - b. 放 *fàng* ‘put’:
1: Discharge or direct or be discharged or directed as if in a continuous stream
2: Put into a certain place: “Put your things here”
3: Locate

However, since SinicaBow is a translated database from the English WordNet, there are some senses that could have been excluded if these senses are not found in English. For example, in examples (5) and (6) below, the use of 擺 *bǎi* in (5a) is not the same as in (5b), and both are not easily represented using the meanings in (4a). Similarly, example (6a) is a meaning extension of 放 *fàng* but it is not related to the real sense of ‘locate’ or ‘put in a certain place’. If only certain examples are chosen from the corpus, additional senses of 擺 *bǎi* and 放 *fàng* may be left out. However, through the collection of random sample sentences from the corpus, this problem is necessarily addressed, as all examples (not a selected few) appearing in the sample collected have to be dealt with.

- (5) a. 擺個姿勢 *bǎi ge zīshì* ‘to pose’
b. 擺棋子 *bǎi qízǐ* ‘to lay a piece in a board game’
(6) a. 放著風箏 *fàng zhe fēngzhēng* ‘flying kite’
b. 放椅子 *fàng yǐzi* ‘to put a chair (somewhere)’

In order to collect a sample of sentences, this paper takes the first 100 sentences for each verb from the Sinica Corpus (from the total of 233 for 擺 *bǎi*, and 1,031 for 放 *fàng*).⁵ The results are shown in Table 1.⁶ The senses in Table 1 were decided upon based on the first author’s intuition, since SinicaBow has the inherent limitations mentioned above.

⁵ For the current approach, the sense analysis was carried out based on intuition. Further improvement of this method can be carried out by identifying the senses automatically. For example, Ker et al. (柯淑津等) (2007) is one of the studies that we can refer to when dealing with this issue. However, further research is still needed in this respect.

⁶ Since there are 100 sentences, the number of instances in each sense is also the percentage of each sense.

Table 1: Analysis of sense distribution for 擺 *bǎi* and 放 *fàng*

擺 <i>bǎi</i> ‘put’	%	放 <i>fàng</i> ‘put’	%
Metaphorical usages	27	Metaphorical usages	44
Arrange for display	25	Put (things)	21
Lay (baby, basin, book, dishes, garden, etc.)	15	Let go (animals, person, hand, prey, etc.)	14
Put (things)	12	Discharge (bomb, fire, firework, kite, etc.)	9
Set up	12	Keep (meat, tea leaves, things)	3
Pose	5	Play (record, music, etc.)	3
Move	4	Non-classified	3
		Add	2
		Locate (building)	1
Total	100	Total	100

From Table 1, there is an overlapped meaning of ‘put (things)’ that appears for both 擺 *bǎi* and 放 *fàng*. Examples of this sense can be seen in (7) below.

- (7) a. 錢 就 擺 在 房間 某 件 東西 裡面
qián jiù bǎi zài fángjiān mǒu jiàn dōngxi limiàn
 money just put at room some Class. thing inside
 ‘The money is put inside something (a container) in the room.’
- b. 杜象 把 這個 作品
dùxiàng bǎ zhège zuòpǐn
 Duxiang BA this art.work
 放 在 一個 木箱 裡
fàng zài yíge mùxiāng lǐ
 put at one-Class. wood-case inside
 ‘Duxiang put this piece of art inside a wooden case.’

As can be seen in (7), the use of 擺 *bǎi* and 放 *fàng* in (7) can be substituted for one another. This overlapped meaning shows that 擺 *bǎi* and 放 *fàng* are near-synonymous. Nevertheless, only through our analysis that we can see 放 *fàng* is used 21% as ‘put’ while 擺 *bǎi* has only 12% of the instances being used as ‘put’. In both 擺 *bǎi* and 放 *fàng*, the majority of their senses (27% for 擺 *bǎi* and 44% for 放 *fàng*) contain metaphorical meanings, exemplified in (8) below. For 擺 *bǎi*, its second highest meaning of ‘arrange for display’ is 25%, which is close to the percentage of its metaphorical use. Therefore, from this small sample, we can see that 放 *fàng* is used more often as metaphor than 擺 *bǎi*. However, a more large-scale analysis is still needed to validate this observation.

- (8) a. 把 全民 利益 擺 在 第一
bǎ quánmín lìyì bǎi zài dìyī
 BA all.nation profit put at first.place
 ‘To put the interests of the whole nation at priority’
- b. 霸 著 話題 不 放
bà zhe huàtí bú fàng
 dominate ZHE discussion.topic Neg. let.go
 ‘To dominate the topic of discussion (without letting go)’

The metaphorical uses found in (8) were excluded for this analysis because they will create noise in the data, if they were included as part of the other meanings.⁷

The “non-classified” use in 放 *fàng* refers to instances where 放 *fàng* is used as a noun, as in (9). There are three instances of the use in (9).

- (9) 「放 的 哲學」
fàng de zhéxué
fàng DE philosophy
 ‘The philosophy of *fàng*’

Lastly, the results in Table 1 also show that 擺 *bǎi* and 放 *fàng* are similar in one meaning, but they differ in many other meanings. These differing meanings give clues as to how one synonym differs from another. In the next section, we shall demonstrate the use of MI values to find the argument types for these two near-synonyms.

4.2 Mutual Information value

In order to find out the MI values for the arguments that collocate most frequently with each verb, the MI values for all the search results (233 for 擺 *bǎi* and 1,031 for 放 *fàng*) were calculated by the internal system of the Sinica Corpus. The window size is set from -4 to 4 (i.e. 4 words on the left or right of the key word). The MI list shown has several columns, as shown in Table 2 below.⁸

⁷ It is likely that these are the instances that were skipped over in the data collected in Huang et al. (2000).

⁸ A smaller window size, such as -1 to 1, was not used because this will exclude constructions such as BA-constructions. However, the window size of -4 to 4 might also have excluded topicalized nouns which also use 擺 *bǎi* and 放 *fàng*.

Table 2: Examples of MI values for 擺 *bǎi* ‘put’⁹

	MI	Freq (y)	Freq (x, y)	y: 詞	y: 詞類	Gloss
(a)	10.126	1	1	缸數	Na	<i>gāngshù</i> ‘number of jars’
(b)	10.126	1	1	咬鳥卦	Na	<i>yǎoniǎoguà</i> ‘fortune-telling with birds’
(c)	9.839	4	3	扭腰	VA	<i>niǔyāo</i> ‘twist one’s waist’
(d)	9.433	2	1	花椒	Na	<i>huājiāo</i> ‘a type of pepper’
(e)	9.279	7	3	炮竹	Na	<i>pàozhú</i> ‘firecrackers’
(f)	2.380	11,562	5	的	T	<i>de</i> ‘DE’

Freq (y) is the number of times the words on the rightmost column appear in the whole corpus (including texts other than 擺 *bǎi*). Freq (x, y) refers to the number of times the words y co-occur with the target word (x=擺 *bǎi*).

MI values refer to the probability of the words y collocate with x (cf. Church & Hanks 1990). For Sinica Corpus, the definition of the MI value is the calculation “between a key and the characters occurring in the specified window (i.e. the left and/or right context)” in the Sinica Corpus (Huang, Ahrens & Chen 1998:157). The MI values calculated by the Sinica Corpus is as in (10) below, “where N is the size of the corpus and *m* is the size of the selected window” (Huang et al. 1998:157):

$$\begin{aligned}
 (10) \quad I(x,y) &= \text{Log}_2 P(x,y) / P(x) \cdot P(y) \\
 &\approx \text{Log}_2 \frac{f(x,y) / m \cdot N}{f(x) / N \cdot f(y) / N} \\
 &= \text{Log}_2 f(x,y) \cdot N / m \cdot f(x) \cdot f(y)
 \end{aligned}$$

Even though MI values are indices showing whether x and y are associated, this paper suggests that one should not refer arbitrarily to MI values. This is to avoid including data that we do not need.¹⁰ For instance, in (a) and (b) of Table 2, when both x and y occurs once respectively, the probability of the two co-occurring together will be absolute and the MI value will be high.

In order to avoid this problem, this paper sets two criteria for choosing the collocated arguments for the verbs (x). These two criteria are: (a) the freq (x, y) must be higher than 3 (i.e. the x and y co-occur at least three times in the whole corpus of 擺 *bǎi*); and (b) the MI value should not be lower than 5. These threshold levels were set based on our observation of our data. These criteria, however, can be changed based on individual research, depending on how much information one needs to include. For the current

⁹ Note that these are not necessarily the top collocates. They are a sample from the results.

¹⁰ This problem has been noted by many, Kilgariff & Tugwell (2001) in particular, suggest an alternative way of measuring collocations by using saliency.

research, both 擺 *bǎi* and 放 *fàng* are searched using the same criteria and therefore, are delimited by the same conditions.

These two criteria can help avoid selecting a common term such as 的 *de* ((f) in Table 2) which occurs so often in the whole corpus that the MI value becomes very low (even though the number of times it co-occurs with 擺 *bǎi* is more than 5). Based on these criteria, the final selected arguments for 擺 *bǎi* are shown in Table 3 below.

Table 3: Collocated arguments for 擺 *bǎi* ‘put’

MI	Freq (y)	Freq (x, y)	y: 詞	y: 詞類	Gloss
9.839	4	3	扭腰	VA	<i>niǔyāo</i> ‘twist one’s waist’
9.279	7	3	炮竹	Na	<i>pàozhú</i> ‘firecrackers’
8.072	39	5	地攤	Nc	<i>dītān</i> ‘stall on the ground’
8.006	25	3	平	VC	<i>píng</i> ‘smoothen’
7.511	41	3	書架	Na	<i>shūjià</i> ‘book shelf’
7.487	56	4	桌	Nf	<i>zhuō</i> ‘table’
6.991	184	8	桌	Na	<i>zhuō</i> ‘table’
6.991	92	4	桌子	Na	<i>zhuōzi</i> ‘table’
6.507	112	3	姿勢	Na	<i>zīshì</i> ‘posture’
5.932	199	3	門口	Nc	<i>ménkǒu</i> ‘entrance’
5.881	279	4	中間	Ncd	<i>zhōngjiān</i> ‘middle’
5.680	256	3	左	Ncd	<i>zuǒ</i> ‘left’
5.627	270	3	右	Ncd	<i>yòu</i> ‘right’
5.605	644	7	起	Di	<i>qǐ</i> ‘up’
5.393	341	3	東	Ncd	<i>dōng</i> ‘east’
5.282	381	3	西	Ncd	<i>xī</i> ‘west’
5.088	1080	7	往	P	<i>wǎng</i> ‘toward’

Compared to the previous Table 2, one can see that these two criteria remove the problematic lexical items such as 缸數 *gāngshù* ‘the number of tubs’ 咬鳥卦 *yǎoniǎoguà* ‘a type of fortune-telling card picked up by a bird’. Items such as these two have high MI values with 擺 *bǎi* because the only time they appear in the corpus, they co-occur with 擺 *bǎi*. With our proposal set forth above, we excluded those that do not fit these criteria. When the same criteria applied to 放 *fàng*, the results in Table 4 are obtained.

Table 4: Collocated arguments for 放 *fàng* ‘put’

MI	Freq (y)	Freq (x, y)	y: 詞	y: 詞類	Gloss
8.080	7	4	長假	Na	<i>chángjià</i> ‘long holiday’
7.851	11	5	水燈	Na	<i>shuǐdēng</i> ‘water lantern’
7.828	9	4	倉	Na	<i>cāng</i> ‘warehouse’
7.541	9	3	成交價	Na	<i>chéngjiāojià</i> ‘transaction price’
7.486	19	6	在一塊	VH	<i>zàiyikuài</i> ‘be together’
7.423	27	8	假	Na	<i>jià</i> ‘holiday’
7.173	13	3	紅龜	Na	<i>hóngguī</i> ‘red tortoise’
6.991	26	5	四海	Nc	<i>sìhǎi</i> ‘Four Seas’
6.905	17	3	盆	Na	<i>pén</i> ‘basin’
6.815	31	5	架子	Na	<i>jiàzi</i> ‘shelf’
6.711	55	8	風箏	Na	<i>fēngzhēng</i> ‘kite’
6.560	24	3	心念	Na	<i>xīniàn</i> ‘thoughts’
6.519	25	3	武松	Nb	<i>wǔsōng</i> ‘Wusong (pronoun)’
6.480	52	6	人質	Na	<i>rénzhì</i> ‘hostage’
6.283	116	11	重心	Na	<i>zhòngxīn</i> ‘focus’
6.257	184	17	桌	Na	<i>zhuō</i> ‘table’
6.228	78	7	口袋	Na	<i>kǒudài</i> ‘pocket’
6.024	41	3	書架	Na	<i>shūjià</i> ‘book shelf’
5.905	77	5	心思	Na	<i>xīnsī</i> ‘thoughts’
5.846	49	3	炸彈	Na	<i>zhàdàn</i> ‘bomb’
5.767	53	3	枕頭	Na	<i>zhěntóu</i> ‘pillow’
5.727	92	5	桌子	Na	<i>zhuōzi</i> ‘table’
5.565	238	11	火	Na	<i>huǒ</i> ‘fire’
5.560	87	4	畝	Nf	<i>mù</i> ‘acreage’
5.489	70	3	肩膀	Na	<i>jiānbǎng</i> ‘shoulder’
5.461	96	4	化妝品	Na	<i>huàzhuāngpǐn</i> ‘cosmetics’
5.412	126	5	抓住	VC	<i>zhuāzhù</i> ‘grab’
5.376	183	7	羊	Na	<i>yáng</i> ‘goat’
5.362	106	4	客廳	Nc	<i>kètīng</i> ‘living room’
5.362	106	4	牛奶	Na	<i>niúniǎi</i> ‘milk’
5.343	135	5	心力	Na	<i>xīnlì</i> ‘mental and physical strength’
5.326	577	21	重點	Na	<i>zhòngdiǎn</i> ‘emphasis’
5.173	96	3	浴室	Nc	<i>yùshì</i> ‘bathroom’
5.113	102	3	包袱	Na	<i>bāofū</i> ‘burden’
5.034	184	5	封	Nf	<i>fēng</i> ‘seal’
5.022	484	13	下	VC	<i>xià</i> ‘below’

Comparing the lists in Tables 3 and 4, one sees that more argument types were found for 放 *fàng* when these same two criteria are used. The comparison is made more obvious by laying out the collocated arguments for the two verbs, as shown in Table 5 below.

Table 5: Collocated arguments for 擺 *bǎi* and 放 *fàng*

擺 <i>bǎi</i> ‘put’		放 <i>fàng</i> ‘put’			
扭腰 <i>niǔyāo</i> ‘twist one’s waist’	中間 <i>zhōngjiān</i> ‘middle’	長假 <i>chángjià</i> ‘long holiday’	架子 <i>jiàzi</i> ‘shelf’	心思 <i>xīnsī</i> ‘thoughts’	羊 <i>yáng</i> ‘goat’
炮竹 <i>pàozhú</i> ‘firecrackers’	左 <i>zuǒ</i> ‘left’	水燈 <i>shuǐdēng</i> ‘water lantern’	風箏 <i>fēngzhēng</i> ‘kite’	炸彈 <i>zhàdàn</i> ‘bomb’	客廳 <i>kètīng</i> ‘living room’
地攤 <i>dītān</i> ‘stall on the ground’	右 <i>yòu</i> ‘right’	倉 <i>cāng</i> ‘warehouse’	心念 <i>xīniàn</i> ‘thoughts’	枕頭 <i>zhěntóu</i> ‘pillow’	牛奶 <i>niǔnǎi</i> ‘milk’
平 <i>píng</i> ‘smooth’	起 <i>qǐ</i> ‘up’	成交價 <i>chéngjiāojià</i> ‘transaction price’	武松 <i>wǔsōng</i> ‘Wusong (pronoun)’	桌子 <i>zhuōzi</i> ‘table’	心力 <i>xīnlì</i> ‘mental and physical strength’
書架 <i>shūjià</i> ‘book shelf’	東 <i>dōng</i> ‘east’	在一塊 <i>zài yíkuài</i> ‘be together’	人質 <i>rénzhì</i> ‘hostage’	火 <i>huǒ</i> ‘fire’	重點 <i>zhòngdiǎn</i> ‘emphasis’
桌 <i>zhuō</i> ‘table’	西 <i>xī</i> ‘west’	假 <i>jià</i> ‘holiday’	重心 <i>zhòngxīn</i> ‘focus’	畝 <i>mǔ</i> ‘acreage’	浴室 <i>yùshì</i> ‘bathroom’
桌子 <i>zhuōzi</i> ‘table’	往 <i>wǎng</i> ‘toward’	紅龜 <i>hóngguī</i> ‘red tortoise’	桌 <i>zhuō</i> ‘table’	肩膀 <i>jiǎnbǎng</i> ‘shoulder’	包袱 <i>bāofú</i> ‘burden’
姿勢 <i>zīshì</i> ‘posture’		四海 <i>sìhǎi</i> ‘Four Sea’	口袋 <i>kǒudài</i> ‘pocket’	化妝品 <i>huàzhuāngpǐn</i> ‘cosmetics’	封 <i>fēng</i> ‘seal’
門口 <i>ménkǒu</i> ‘entrance’		盆 <i>pén</i> ‘basin’	書架 <i>shūjià</i> ‘book shelf’	抓住 <i>zhuāzhù</i> ‘grab’	下 <i>xià</i> ‘down’

In Table 5, the arguments in italics are those that overlap for both 擺 *bǎi* and 放 *fàng* ‘put’. We can see that only 書架 *shūjià* ‘book shelf’, 桌 *zhuō* ‘table’ and 桌子 *zhuōzi* ‘table’ are found overlapping for these two verbs and these arguments are also found under the overlapped sense of ‘put (things)’ in Table 1.

By identifying the selectional restriction through this way one can verify the following statement by Huang et al. (2000:35) that “the orientation of the placed object [of *bai3*] can be specified while only location can be specified for *fang4*.” This is seen in Table 5 above for orientations of 中間 *zhōngjiān* ‘middle’, 左 *zuǒ* ‘left’, 右 *yòu* ‘right’, etc. (shaded in Table 5), all of which are not found in the list for 放 *fàng* ‘put’. When carried out using these steps, one can then make more data-driven proposals within the MARVS model.

In addition to Sinica Corpus, there are also other resources which can be used to find information pertaining degree of collocation between words. The Chinese Sketch Engine (Kilgarriff, Huang, Rychly, Smith & Tugwell 2005) is able to provide the collocated arguments for 擺 *bǎi* and 放 *fàng* through WordSketches (such that exemplified in

Table 6).¹¹ This function of Sketch Engine provides lists of collocates according to different grammatical relations such as ‘subject’, ‘object’, ‘modifier’, etc. Table 6 below shows examples of object arguments of 擺 *bǎi* and 放 *fàng* in the Chinese Sketch Engine.

Table 6: WordSketches for the objects of 擺 *bǎi* and 放 *fàng* from the Chinese Sketch Engine

擺 <i>bǎi</i> ‘put’				放 <i>fàng</i> ‘put’			
Objects	Gloss	Freq	Saliency	Objects	Gloss	Freq	Saliency
攤 <i>tān</i>	‘stall’	257	77.72	鞭炮 <i>biānpào</i>	‘firecrackers’	363	71.49
流水席 <i>liúshuǐxí</i>	‘open-air banquet’	49	56.8	水燈 <i>shuǐdēng</i>	‘water lantern’	178	71.06
低姿態 <i>dīzītài</i>	‘low profile’	32	44.94	天燈 <i>tiāndēng</i>	‘flying lantern’	224	69.16
姿勢 <i>zīshì</i>	‘posture’	26	37.61	高利貸 <i>gāolìdài</i>	‘usury’	139	65.56
姿態 <i>zītài</i>	‘posture’	32	32.91	鴿子 <i>gēzi</i>	‘dove’	123	59.49
長龍 <i>chánglóng</i>	‘long queue’	20	30.81	風聲 <i>fēngshēng</i>	‘rumors’	51	44.95
攤販 <i>tānfān</i>	‘to set up business of a vendor’	23	28.42	厥辭 <i>juécí</i>	‘words said without serious thoughts’	12	38.69
烏龍* <i>wūlóng</i>	‘oolong tea*’	12	26.84	山雞 <i>shānjī</i>	‘pheasant’	26	38.45
攤位 <i>tānwèi</i>	‘stall’	21	26.58	炮 <i>pào</i>	‘cannon’	35	38.11
桌 <i>zhuō</i>	‘table’	15	26.45	人 <i>rén</i>	‘human’	805	35.74
宴席 <i>yànxí</i>	‘banquet’	9	25.84	長線 <i>chángxiàn</i>	‘long string’	31	35.26
派頭 <i>pàitóu</i>	‘style’	5	23.1	冲天炮 <i>chōngtiānpào</i>	‘towering cannon’	19	35.13
架子 <i>jiàzi</i>	‘shelf’	7	20.11	比率 <i>bǐlǜ</i>	‘percentage’	110	32.93
排場 <i>páichǎng</i>	‘ostentation’	5	19.85	和平鴿 <i>hépinggē</i>	‘dove as a symbol of peace’	14	29.38
擂台 <i>léitái</i>	‘a ring for contests in martial arts’	5	17.38	線狀菌 <i>xiànzuàngjūn</i>	‘string-like fungus’	6	28.16
小食攤 <i>xiǎoshítān</i>	‘small food stall’	2	16.41				

In Table 6, we can see the most salient collocates for the ‘object’ arguments of 擺 *bǎi* and 放 *fàng*.¹² Among these collocates, there are some which are also found in

¹¹ The Chinese Sketch Engine is available at <http://wordsketch.ling.sinica.edu.tw/> while the English Sketch Engine is available at <http://www.sketchengine.co.uk/>.

¹² In 擺 *bǎi* and 放 *fàng*, one collocate was removed from each list. The removed collocates are 香案 *xiāng'àn* ‘joss-tick case’ (in 擺香案 *bǎixiāng'àn* ‘the case of placing joss stick’)

Table 5 (in italics).¹³ The saliency value (fourth and eighth columns in Table 2) is more powerful than the MI value because it removes the problematic examples in Table 2 earlier with its formula (cf. Kilgarriff & Tugwell 2001). However, Sketch Engine is not sense-tagged. This means that a sense distribution analysis must be carried out based on intuition.

Nevertheless, Sketch Engine allows the analysis of senses directly from the collocates of Wordsketches, instead of reading line by line in the concordance results. This is because Sketch Engine is based on a large corpus (i.e. the Gigaword Corpus with more than one billion characters) and this makes the collocates more reliable. For example, from Table 6, we claim that the top senses for the collocates with the top saliency values (攤 *tān* for 擺 *bǎi* and 鞭炮 *biānpào* for 放 *fàng*) are ‘arrange for display’ and ‘discharge’ respectively (cf. Table 1 earlier).

In Table 6, there are also several collocates which have closer meanings to some of the words in previous Table 5 (darker shades). For instance, 擺地攤 *bǎidītān* ‘to set up a stall with goods laid on the ground’ in Table 5 is similar to 擺攤 *bǎitān* ‘to set up a stall’ in Table 6. In addition, 擺攤販 *bǎitānfàn* ‘to set up the business of a vendor’ is similar to 擺小食攤 *bǎixiǎoshítān* ‘to set up a little food vendor’. Similarly, for 放 *fàng* ‘put’, 放水燈 *fàngshuǐdēng* ‘to discharge a water lantern’ is found in both Tables 5 and 6. Moreover, 放水燈 *fàngshuǐdēng* ‘to discharge a water lantern’ is also similar to 放天燈 *fàngtiāndēng* ‘to discharge a flying lantern’.

However, a better analysis of the collocates will need to divide the long saliency list into significant and non-significant collocates so as to compare which senses are more salient than the others, as was done in Chung (2007). We thus propose this for future research.

5. Conclusion

MARVS has both features of descriptive and quantitative analyses and this paper strengthens the quantitative aspect by proposing additional evidence for Huang et al. (2000) and Ahrens, Huang & Chuang (2003)’s MARVS-based analysis of 擺 *bǎi* and 放 *fàng*. We propose two additional criteria for near-synonym analyses and suggest that

and 後稅 *hòushuì* ‘later-tax’ (in 先放後稅 *xiānfānghòushuì* ‘first release (goods) then tax (someone)’), which are both wrongly parsed. In addition, there are some segmentation issues. One of them is 擺烏龍 *bǎiwūlóng* ‘absentminded’ where 烏龍 *wūlóng* (asterisk) cannot be segmented from 擺 *bǎi* ‘put’. Otherwise, 烏龍 *wūlóng* can only be translated as ‘oolong tea’. Despite these problems, the Sketch Engine usually is able to display the most salient collocates.

¹³ These are only part of the lists. More overlapped collocates may be found later in the lists.

these criteria further allow the operationalization of the steps used to identify contrasts in near-synonyms. In addition, we propose that analysis of sense distribution and MI values can be used to state the differences between two near synonymous verbs.

The new steps proposed for MARVS combine a corpus-driven, quantitative approach with traditional semantics. This paper further integrates MI values (Church et al. 1994) into the MARVS analysis and also suggests clear criteria for the selection of MI values. Thus, this study not only provides clarification to a previously established lexical-semantic model, but also contributes methodology-wise to computational linguistic research.

References

- Ahrens, Kathleen, Chu-Ren Huang, and Yuan-hsun Chuang. 2003. Sense and meaning facets in verbal semantics: a MARVS perspective. *Language and Linguistics* 4.3: 469-484.
- Bresnan, Joan, and Jonni Kanerva. 1989. Locative inversion in Chichewa: a case study of factorization in grammar. *Linguistic Inquiry* 20.1:1-50.
- Chung, Siaw-Fong. 2007. *A Corpus-driven Approach to Source Domain Determination*. Taipei: National Taiwan University dissertation.
- Church, Kenneth W., and Patrick Hanks. 1990. Word association norms, mutual information and lexicography. *Computational Linguistics* 16.1:22-29.
- Church, Kenneth W., William Gale, Patrick Hanks, Donald Hindle, and Rosamund Moon. 1994. Lexical substitutability. *Computational Approaches to the Lexicon*, ed. by Beryl T. Sue Atkins & Antonio Zampolli, 153-177. Oxford & New York: Oxford University Press.
- Collinson, W. E. 1939. Comparative synonymics: some principles and illustrations. *Transactions of the Philosophical Society* 1939:54-77.
- Croft, William, and Alan D. Cruse. 2004. *Cognitive Linguistics*. Cambridge & New York: Cambridge University Press.
- Cruse, D. Alan. 1986. *Lexical Semantics*. Cambridge & New York: Cambridge University Press.
- DiMarco, Chrysanne, Graeme Hirst, and Mandred Stede. 1993. The semantic and stylistic differentiation of synonyms and near-synonyms. *Proceedings of the AAAI Spring Symposium on Building Lexicons for Machine Translation*, 114-121.
- Dowty, David. 1991. Thematic proto-roles and argument selection. *Language* 67.3: 547-619.

- Edmonds, Philip. 1999. *Semantic Representations of Near-Synonyms for Automatic Lexical Choice*. Toronto: University of Toronto dissertation.
- Goldberg, Adele E. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: University of Chicago Press.
- Harris, Roy. 1973. *Synonymy and Linguistic Analysis*. Toronto: University of Toronto Press.
- Huang, Chu-Ren, Ru-Yng Chang, Shiang-Bin Lee. 2004. Sinica BOW (Bilingual Ontological Wordnet): Integration of Bilingual WordNet and SUMO. Paper presented at the 4th International Conference on Language Resources and Evaluation (LREC2004). Lisbon, Portugal.
- Huang, Chu-Ren, Kathleen Ahrens, and Keh-Jiann Chen. 1998. A data-driven approach to the mental lexicon: two studies on Chinese corpus linguistics. *Bulletin of the Institute of History and Philology* 69.1:151-179.
- Huang, Chu-Ren, Kathleen Ahrens, Li-li Chang, Keh-Jiann Chen, Mei-chun Liu, and Mei-Chih Tsai. 2000. The module-attribute representation of verbal semantics: from semantics to argument structure. *Computational Linguistics and Chinese Language Processing* 5.1:19-46.
- Kilgarriff, Adam, and David Tugwell. 2001. WORD SKETCH: extraction and display of significant collocations for lexicography. *Proceedings of the ACL Workshop on COLLOCATION: Computational Extraction, Analysis and Exploitation*, 32-38. Toulouse, France.
- Kilgarriff, Adam, Chu-Ren Huang, Pavel Rychly, Simon Smith, and David Tugwell. 2005. Chinese Word Sketches. Paper presented at the ASIALEX 2005: Words in Asian Cultural Context. Singapore.
- Levin, Beth. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago: University of Chicago Press.
- Liu, Mei-chun. 2003. From collocation to event information: the case of Mandarin verbs of discussion. *Language and Linguistics* 4.3:563-585.
- Lyons, John. 1995. *Linguistic Semantics: An Introduction*. Cambridge & New York: Cambridge University Press.
- MacLaury, Robert E. 1997. Vantage theory in cognitive science: an anthropological account of categorization and similarity judgment. *Proceedings of the Interdisciplinary Workshop on Similarity and Categorization*, ed. by Michael Ramscar et al., 157-163. Edinburgh: University of Edinburgh.
- MacLaury, Robert E. 2002. Introducing vantage theory. *Language Sciences* 24.5-6: 493-536.
- Palmer, Martha. 2000. Consistent criteria for sense distinctions. *Computers and the Humanities* 34.1-2:217-222.

- Pustejovsky, James. 1991. The generative lexicon. *Computational Linguistics* 17.4: 409-441.
- Taylor, John R. 2002. Near synonyms as co-extensive categories: 'high' and 'tall' revisited. *Language Sciences* 25.3:263-284.
- Teng, Shou-hsin, Chin-Chuan Cheng, and Chin-Hsi Lin. 2006. *Pintone 2006*. [Computer Software]. Taipei: Graduate Institute of Teaching Chinese as a Second Language, National Taiwan Normal University.
- Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work*. Studies in Corpus Linguistics, 6. Amsterdam & Philadelphia: John Benjamins.
- Tsai, Mei-Chih, Chu-Ren Huang, Keh-Jiann Chen, and Kathleen Ahrens. 1998. Towards a representation of verbal semantics: an approach based on near-synonyms. *Proceedings of the 10th Conference on Computational Linguistics and Speech Processing (ROCLING-10)*, 34-48. Taipei: Association for Computational Linguistics and Chinese Language Processing.
- Zaenen, Annie. 1993. Unaccusativity in Dutch: an integrated approach. *Semantics and the Lexicon*, ed. by James Pustejovsky, 129-161. Dordrecht: Kluwer.
- 柯淑津, 黃居仁, 洪嘉馥, 劉詩音, 簡卉伶, 蘇依莉. 2007. 〈中文詞義全文標記語料庫之設計與雛形製作〉, 第十九屆自然語言與語音處理研討會海報。台北: 國立台灣大學。

[Received 29 November 2006; revised 26 September 2007; accepted 1 November 2007]

Siaw-Fong Chung
 Graduate Institute of Linguistics
 National Taiwan University
 1, Roosevelt Road, Sec. 4
 Taipei 106, Taiwan
 siawfongchung@gmail.com

Kathleen Ahrens
 Graduate Institute of Linguistics
 National Taiwan University
 1, Roosevelt Road, Sec. 4
 Taipei 106, Taiwan
 kathleenahrens@yahoo.com

MARVS理論再探： 以量化觀點比較近義詞的詞義頻率與搭配詞共現值

鍾曉芳 安可思

國立台灣大學

在 MARVS (Module-Attribute Representation of Verbal Semantics) 這個理論裡，動詞的分辨是以事件訊息為基礎，而事件訊息主要包括了事件模組和角色模組。黃居仁等 (2000) 曾以 MARVS 檢查近義詞的語意，並建議使用此理論來凸顯近義詞間的差別。本文則針對 MARVS 的理論，加強其分析結果，並加入兩個新的步驟。這兩個步驟分別是：一、量化比較和分析近義詞在語料庫的詞義，二、透過 Mutual Information 的計算比較不同近義詞的搭配詞。這兩個步驟能增加 MARVS 在語義分析的可檢驗性，也更能奠定其理論運用在自動擷取詞彙語意上的基礎。

關鍵詞：近義詞，MARVS，詞義，搭配詞共現值，擺 *bǎi*，放 *fàng*