Corpus-based Study and Identification of Mandarin Chinese Light Verb Variations

Chu-Ren Huang[1] **Jingxia Lin**[2] **Menghan Jiang**[1] **Hongzhi Xu**[1]

[1]Department of CBS, The Hong Kong Polytechnic University
[2]Nanyang Technological University

churen.huang@polyu.edu.hk, jingxialin@ntu.edu.sg,
menghan.jiang@connect.polyu.hk, hongz.xu@gmail.com

Abstract

When PRC was founded on mainland China and the KMT retreated to Taiwan in 1949, the relation between mainland China and Taiwan became a classical Cold War instance. Neither travel, visit, nor correspondences were allowed between the people until 1987, when government on both sides started to allow small number of Taiwan people with relatives in China to return to visit through a third location. Although the thawing eventually lead to frequent exchanges, direct travel links, and close commercial ties between Taiwan and mainland China today, 38 years of total isolation from each other did allow the language use to develop into different varieties, which have become a popular topic for mainly lexical studies (e.g., Xu, 1995; Zeng, 1995; Wang & Li, 1996). Grammatical difference of these two variants, however, was not well studied beyond anecdotal observation, partly because the near identity of their grammatical systems. This paper focuses on light verb variations in Mainland and Taiwan variants and finds that the light verbs of these two variants indeed show distributional tendencies. Light verbs are chosen for two reasons: first, they are semantically bleached hence more susceptible to changes and variations. Second, the classification of light verbs is a challenging topic in NLP. We hope our study will contribute to the study of light verbs in Chinese in general. The data adopted for this study was a comparable corpus extracted from Chinese Gigaword Corpus and manually annotated with contextual features that may contribute to light verb variations. A multivariate analysis was conducted to show that for each light verb there is at least one context where the two variants show differences in tendencies (usually the presence/absence of a tendency rather than contrasting tendencies) and can be differentiated. In addition, we carried out a K-Means clustering analysis for the variations and the results are consistent with the multivariate analysis, i.e. the light verbs in Mainland and Taiwan indeed have variations and the variations can be successfully differentiated.

1 Introduction: Language Variations in the Chinese Context

Commonly dichotomy of language and dialect is not easily maintained in the context of Chinese language(s). Cantonese, Min, Hakka, and Wu are traditionally referred to as dialects of Chinese but are mutually unintelligible. However, they do share a common writing system and literary and textual tradition, which allows speakers to have a shared linguistic identity. To overcome the mutual unintelligibility problem, a variant of Northern Mandarin Chinese, is designated as the common language about a hundred years ago (called 普通話 *Putonghua* 'common language' in Mainland China, and 國語 *Guoyu* 'national language' in Taiwan). Referred to as Mandarin or Mandarin Chinese, or simply Chinese nowadays, this is the one of the most commonly learned first or second languages in the world now. However, not unlike English, with the fast globalization of the Chinese language, both the term 'World Chineses' and the recognition that there are different variants of Chinese emerged. In this paper, we studied two of the most important variants of Chinese, Mainland Mandarin and Taiwan Mandarin.

Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects, pages 1–10,
Dublin, Ireland, August 23 2014.

1.1 Variations between Mainland and Taiwan Mandarin: Previous studies

The lexical differences between Mainland and Taiwan Mandarin have been the focus of research in Chinese Linguistics in the recent years. A number of studies were carried out on lexical variations between these two variants of Mandarin Chinese, including variations in the meanings of the same word or using different words to express the same meaning (e.g., Xu, 1995; Zeng, 1995; Wang,1996). Some dictionaries also list the lexical differences between Mainland Mandarin and Taiwan Mandarin (e.g., Qiu, 1990; Wei & Sheng, 2000).

By contrast, only a few of such studies were corpus driven (e.g. Hong and Huang 2008, 2013; Huang and Lin, 2013), and even few studies have been done on the grammatical variations of Mainland and Taiwan Chinese. Huang et al. (2013), the only such study based on comparable corpora so far, suggested that the subtlety of the underlining grammatical variations of these two dialectal variants at early stage of divergence may have contributed to the challenge as well as scarcity of previous studies.

1.2 Light Verbs in Light Verb Variations

The study of English light verb constructions (LVCs) (e.g., *take a look, make an offer*) has been an important topic in linguistics (Jespersen, 1965; Butt and Geuder, 2003; among others) as well as in Computational Linguistics (Tu and Dan, 2011; Nagy et al., 2013; Hwang et al., 2010; among others). Identification of LVCs is a fundamental crucial task for Natural Language Processing (NLP) applications, such as information retrieval and machine translation. For example, Tu and Dan (2011) proposed a supervised learning system to automatically identify English LVCs by training with groups of contextual or statistical features. Nagy et al. (2013) introduced a system that enables the full coverage identification of English LVCs in running context by using a machine leaning approach.

However, little work has been done to identify Chinese LVCs, especially between different variants of Chinese (cf. Hwang et al., 2010). Chinese LVCs are similar to English LVCs in the sense that the light verb itself is semantically bleached and does not contain any eventive or contentive information, so the predicative information of the construction mainly comes from the complement taken by the light verb (e.g., Zhu, 1985; Zhou, 1987; Cai, 1982). For instance, 進行 *jinxing* originally meant 'move forward/proceed', but in an LVC such as 進行討論 *jinxing taolun* proceed discuss 'to discuss', 進行 *jinxing* only contributes aspectual information whereas the core meaning of the LVC comes from the complement 討論 *taolun* 'discuss'. Chinese also differs from English in that many of the Chinese light verbs have similar usages and thus are often interchangeable, e.g., all the five light verbs 從事 *congshi*, 搞 *gao*, 加以 *jiayi*, 進行 *jinxing*, and 做 *zuo* can take 研究 *yanjiu* 'do research' as their complement and form a LVC. But Huang et al. (2013) also observed that differences in collocation constraints can sometimes be found between different variants of Mandarin Chinese. For instance, constructions like 進行投票 *jinxing tou-piao* proceed cast-ticket 'to cast votes', where the complement is in the V(erb)-O(bject) form, usually can only be found in Taiwan Mandarin. Hence, Chinese LVCs are challenging for both linguistic studies and computational applications in two aspects: (a) to identify collocation constraints of the different light verbs in order to automatically classify and predict their uses in context, and (b) to identify the collocation constraints of the same light verb in order to differentiate and predict the two Chinese variants based on the use of such light verbs. The first issue has been explored in Lin et al. (2014): by analyzing Mainland and Taiwan Mandarin data extracted from comparable corpora with statistical and machine learning approaches, the authors find the five light verbs 從事 *congshi*, 搞 *gao*, 加以 *jiayi*, 進行 *jinxing*, and 做 *zuo* can be reliably differentiated from each other in each variety. But to the best of our knowledge, there has been no previous computational study on modeling the light verb variations, or other syntactic variations of Chinese dialects or variants of the same dialect. Therefore, this paper builds on the study of Lin et al. (2014) and will adopt a comparable corpus driven approach to model light verb variations in Mainland and Taiwan Mandarin.

2 Data and annotation

Our study focuses on five light verbs, 加以 *jiayi*, 進行 *jinxing*, 從事 *congshi*, 搞 *gao* and 做 *zuo* (these words literally meant 'proceed', 'inflict', 'engage', 'do', and 'do' respectively). These five are

chosen for two reasons. First, they are the most frequently used light verbs in Mandarin Chinese (Diao, 2004); second, although the definition of Chinese light verbs is still debatable, these five are considered the most typical light verbs in most previous studies.

The data for this study was extracted from the Annotated Chinese Gigaword Corpus (Huang, 2009) maintained by LDC which contains over 1.1 billion Chinese words, consisting of 700 million characters from Taiwan Central News Agency (CNA) and 400 million characters from Mainland Xinhua News Agency (XNA). For each of the five light verbs, 400 sentences were randomly selected, half from the Mainland XNA corpus and the other half from the Taiwan CNA Corpus, which results in 2,000 sentences in total.

Previous studies (Zhu, 1985; Zhou, 1987; Cai, 1982; Huang et al., 2013; among others) have proposed several syntactic and semantic features to compare and identify the similarities and differences among light verbs. For example, while Taiwan 從事 *congshi* can take informal or semantically negative event complements such as 性交易 *xingjiaoyi* 'sexual trade', Mainland 從事 *congshi* is rarely found with such complements (Huang et al. 2013).

In our study, we selected 11 features covering both syntactic and semantic features which may help to identify light verb variations, as in Table 1. All 2,000 sentences with light verbs were manually annotated with the 11 features. The annotator is a trained expert on Chinese linguistics. All ambiguous cases were discussed with another two experts in order to reach an agreement (the features and annotation were the same with Lin et al. (2014)).

3 Modelling and Predicting Two Variants

We carried out both a multivariate analysis and machine learning algorithm to explore the possible differences existing between Mainland and Taiwan Mandarin light verbs. Our analysis shows that for each light verb, there is at least one context where the two variants of Mandarin show differences in usage tendencies and thus can be differentiated, although the differences more often lie in the presence/absence of a tendency rather than complementary distribution.

3.1 Multivariate Analysis of Light Verb Variations

As introduced in Section 1, the five or some of the five light verbs sometimes can be interchangeably used in both Mainland and Taiwan Mandarin. This indicates that the interchangeable light verbs share some features. In other words, it is unlikely that a particular feature is preferred by only one light verb and thus differentiates the verb from the others. This is also proved in Lin et al. (2014). For instance, their study finds both Mainland and Taiwan 從事 *congshi* and 搞 *gao* significantly prefer nominal complements (POS.N). Therefore, to better explore the light verb differences in the two variants, we adopt a multivariate analysis for this study.

The multivariate analysis we used is polytomous logistic regression (Arppe 2008, cf. Han et al. 2013, Bresnan et al. 2007), and the tool we used is the Polytomous() function in the Polytoumous package in R (Arppe 2008). The polytomous logistic regression is an extension of standard logistic regression; it calculates the odds of the occurrence of a particular light verb when a particular feature is present, with all other features being equal (Arppe, 2008). In addition, it also allows for simultaneous estimation of the occurrence probability of all the five light verbs.

Before we discuss the light verb variations based on multivariate analysis, we will show that the polytomous multivariate model adopted is reliable for our study. Table 2 presents the probability estimates of Mainland and Taiwan light verbs calculated by the model. The results indicate that the overall performance of the model is good: the most frequently predicted light verb (in each column) corresponds to the light verb that actually occurs in the data (in each row) (see the numbers in bold).

In addition, the recall, precision, and F-measure of the estimates given in Table 3 show that each light verb in each variant can be successfully identified with a F-score better than chance (0.2), while the performance varies from light verb to light verb, which is thus consistent with the results in Lin et al. (2014). The only exception is 搞 *gao* in Mainland Mandarin, but the low F-score of 搞 *gao* (0.14) is consistent with the linguistic observation that this verb is rarely used as a light verb in Mainland Mandarin. More detailed information of the factors that can distinguish the five light verbs in each

variant can also be found in Table 4. In the following of this section, we focus on the variations of each light verb in Mainland and Taiwan Mandarin.

Feature ID	**Explanation**	**Values (example)**
1. OTHERLV	Whether a light verb co-occurs with another light verbs	Yes (開始進行討論 ***kaishi*** *jinxing taolun* Start proceed discuss 'start to discuss') No (進行討論 *jinxing taolun* proceed discuss 'to discuss')
2. ASP	Whether a light verb is affixed with an aspectual marker (e.g., perfective 了 *le*, durative 著 *zhe*, experential 過 *guo*)	ASP.le (進行了戰鬥 *jinxing-**le** zhandou* 'fighted') ASP.zhe (進行著戰鬥 *jinxing-**zhe** zhandou* 'is fighting') ASP.guo (進行過戰鬥 *jinxing-**guo** zhandou* 'fighted') ASP.none (進行戰鬥 *jinxing zhandou* 'fight')
3. EVECOMP	Event complement of a light verb is in subject position	Yes (比賽在學校進行 ***bisai*** *zai xuexiao jinxing* game at school proceed 'The game was held at the school') No (在學校進行比賽 *zai xuexiao jinxing **bisai*** at school proceed game 'the game was held at the school')
4. POS	The part-of-speech of the complement taken by a light verb	Noun (進行戰爭 *jinxing **zhanzheng*** proceed fight 'to fight') Verb (進行戰鬥 *jinxing **zhandou*** proceed fight 'to fight')
5. ARGSTR	The argument structure of the complement of a light verb, i.e. the number of arguments (subject and/or objects) that can be taken by the complement	One (進行戰鬥 *jinxing **zhandou*** proceed fight 'to fight') Two (進行批評 *jinxing **piping*** proceed criticize 'to criticize') Zero (進行戰爭 *jinxing **zhanzheng*** proceed fight 'to fight')
6. VOCOMP	Whether the complement of a light verb is in the V(erb)-O(bject) form	Yes (進行投票 *jinxing **tou-piao*** proceed cast-ticket 'to vote') No (進行戰鬥 *jinxing **zhan-dou*** proceed fight-fight 'to fight')
7. DUREVT	Whether the event denoted by the complement of a light verb is durative	Yes (進行戰鬥 *jinxing **zhandou*** proceed fight-fight 'to fight') No (加以拒絕 *jiayi **jujue*** inflict reject 'to reject')
8. FOREVT	Whether the event denoted by the complement of a light verb is formal or official	Yes (進行國事訪問 *jinxing **guoshi fangwen*** proceed state visit 'to pay a state visit') No (做小買賣 *zuo **xiao maimai*** do small business 'run a small business')
9. PSYEVT	Whether the event denoted by the complement of a light verb is mental or psychological activity	Yes (加以反省 *jiayi **fanxing*** inflict retrospect 'to retrospect') No (加以調查 *jiayi **diaocha*** inflict investigate 'to investigate')
10. INTEREVT	Whether the event denoted by the complement of a light verb involves interaction among participants	Yes (進行討論 *jinxing **taolun*** proceed discuss 'to discuss') No (加以批評 *jiayi **piping*** inflict criticize 'to criticize')
11. ACCOMPEVT	Whether the event denoted by the complement of a light verb is an accomplishment	Yes (進行解決 *jinxing **jiejue*** proceed solve 'to solve') No (進行戰鬥 *jinxing **zhandou*** proceed fight-fight 'to fight')

Table 1: Features used to differentiate five Chinese light verbs.

Predicted / Observed	*congshi*		*gao*		*jiayi*		*jinxing*		*zuo*	
	ML	TW	ML	TW	ML	TW	ML	TW	ML	TW
congshi	**131**	**64**	1	87	62	39	1	10	5	0
gao	69	8	**16**	**139**	86	36	16	16	13	1
jiayi	1	0	1	0	**192**	**190**	6	6	0	4
jinxing	31	18	9	34	47	80	**62**	**67**	51	1
zuo	50	24	5	16	44	114	4	14	**97**	**32**

Table 2: Probability estimates of Mainland (ML) and Taiwan (TW) light verbs.

	Recall		Precision		F-measure	
	ML	TW	ML	TW	ML	TW
congshi	0.66	0.32	0.46	0.56	0.54	0.41
gao	0.08	0.70	0.5	0.5	0.14	0.58
jiayi	0.96	0.95	0.45	0.41	0.61	0.58
jinxing	0.31	0.34	0.70	0.59	0.43	0.43
zuo	0.49	0.16	0.58	0.84	0.53	0.27

Table 3: Recall, precision, and F-measure of the polytomous multivariate estimates.

	congshi		*gao*		*jiayi*		*jinxing*		*zuo*	
	ML	TW	ML	TW	ML	TW	ML	TW	ML	TW
(Intercept)	(1/Inf)	(1/Inf)	*0.02271*	(1/Inf)	(1/Inf)	(1/Inf)	(1/Inf)	(1/Inf)	(1/Inf)	(1/Inf)
ACCOMPEVTyes	(1/Inf)	(0.3419)	*0.09863*	(1/Inf)	***56.25***	***11.33***	*0.1849*	(0.1607)	(1/Inf)	*0.2272*
ARGSTRtwo	*0.2652*	*0.1283*	***2.895***	(0.7613)	***76.47***	(Inf)	(1.481)	(0.7062)	*0.2177*	(1.217)
ARGSTRzero	(1.097)	(0.6219)	***3.584***	***7.228***	(1/Inf)	(4.396)	(1.179)	*0.5393*	*0.245*	*0.2068*
ASPle	(0.7487)	(1/Inf)	(0.1767)	(1/Inf)	(0.8257)	(0.3027)	(0.9196)	(Inf)	(1.853)	***32.98***
ASPno	(Inf)	(0.9273)	(1.499)	(0.6967)	(Inf)	(Inf)	(0.2307)	(Inf)	(0.2389)	(0.2385)
ASPzhe	(1.603)		(1/Inf)		(0.4571)		(Inf)		(1/Inf)	
DUREVTyes	(Inf)	(Inf)	(2.958)	(Inf)	(1/Inf)	(1/Inf)	(Inf)	(0.9575)	(Inf)	(Inf)
EVECOMPyes	(1/Inf)	(1/Inf)	(1.726)	(0.8491)	(1/Inf)	(1/Inf)	***3.975***	***8.113***	(1.772)	(0.5019)
FOREVTyes	(2.744)	*0.0867*	(1.227)	(Inf)	(Inf)	(Inf)	(0.7457)	(1.437)	*0.2679*	(1.467)
INTEREVTyes	*0.03255*	*0.1896*	(0.5281)	(1/Inf)	(0.5432)	(0.951)	***18.67***	***10.47***	*0.08902*	(0.398)
PSYEVTyes	(1/Inf)	(1/Inf)	(1/Inf)	(1/Inf)	***19.87***	(1.395)	(1/Inf)	(1/Inf)	(0.9619)	(3.323)
VOCOMPyes	(0.1346)	*0.18*	(3.043)	(2.35)	***23.54***	(Inf)	(1.086)	***3.161***	(0.5344)	(0.5956)

Table 4: Multivariate analysis of light verb variations in Mainland and Taiwan Mandarin.

Table 4 summarizes the results estimated by the Polytomous multivaraite analysis. The numbers in the table are the odds for the features in favor of or against the occurrence of each light verb: odds larger than 1 indicate that the chance of the occurrence of a light verb is significantly increased by the feature, e.g., the chance of Mainland 加以 *jiayi* occurring is significantly increased by ARGSTRtwo (76.47: 1), followed by ACCOMPEVTyes (56.25: 1), VOCOMPyes (23.54: 1), PSYEVTyes (19.87: 1); odds smaller than 1 indicate that the chance of the occurrence of a light verb is significantly descreased by the feature, e.g., the chance of Mainland 進行 *jinxing* occurring is significantly decreased by ACCOMPEVTyse (0.1849: 1); in addition, "inf" and "1/inf" refer to odds larger than 10,000 and smaller than 1/10,000 respectively, and non-significant odds (p-value < 0.05) are given in parentheses, regardless of the odds value.

Table 4 finds that Mainland and Taiwan Mandarin indeed show some variations in each light verb. Furthermore, the variations of each light verb mainly lie in non-complementary distributional patterns. That is, as highlighted in dark grey colour in Table 4, the odds differences are more often between non-significance (odds in parentheses) and significance (odds larger or smaller than 1), rather than between significant preference (odds larger than 1) and significant dis-preference (odds smaller than 1). In other words, the difference of a light verb in the two variants is more **comparative**, rather than

contrastive. This explains why the variations are not easily found by traditional linguistic studies. The following summarizes the key variations of each light verb.

從事 *congshi*

從事 *congshi* in both Mainland and Taiwan Mandarin has no feature significantly in its favor and it is significantly disfavored by ARGSTRtwo (taking two-argument complements, e.g., 研究 *yanjiu* 'to research') and INTEREVTyes (taking complements denoting interactive activities, e.g., 商量 *shangliang* 'to discuss'). However, Taiwan 從事 *congshi* is differentiated from Mainland 從事 *congshi* in that the former is also disfavored by FOREVTyes (taking complements denoting formal events, e.g., 研究 *yanjiu* 'to research') and VOCOMPyes (taking complements in the form of V(erb)-O(bject), e.g., 投票 *toupiao* 'cast a vote'), whereas the latter is not. The finding that Taiwan 從事 *congshi* is less likely to take formal event as its complement is consistent with that in Huang et al. (2013).

搞 *gao*

Both Mainland and Taiwan 搞 *gao* are significantly favored by *ARGSTRzero* (taking zero-argument complements, i.e. noun complement in this study). However, compared with Taiwan Mandarin, Mainland 搞 *gao* is more likely to take two-argument complements (ARGSTRtwo), but less likely to take complements denoting accomplishment events (ACCOMPEVTyes, e.g., 解決 *jiejue* 'to solve'), and it is also disfavored by the aggregate of default variable values (i.e. the intercept, 0.02: 1).

加以 *jiayi*

Both Mainland and Taiwan 加以 *jiayi* are favored by the feature ACCOMPEVTyes (accomplishment complement such as 解決 *jiejue* 'to solve'), but the chance of occurrence of Mainland 加以 *jiayi* increases with the presence of two-argument complements (ARGSTRtwo), complements in VO form (VOCOMPyes), and complements denoting mental or phychological activities (PSYEVTyes, e.g., 反省 *fanxing* 'to introspect').

進行 *jinxing*

Both Mainland and Taiwan 進行 *jinxing* have INTEREVTyes (taking complements denoting interactive activities) and EVECOMPyes (allowing event complements in subject position, e.g., 會議進行順利 *huiyi jinxing shunli* meeting procced smoothly 'The meeting proceeded smoothly') in their favor. However, 進行 *jinxing* in Mainland Mandarin is less likely to take accomplishment complements (ACCOMPEVTyes); whereas 進行 *jinxing* in Taiwan Mandarin is more disfavored by *ARGSTRzero*, but more likely to take complements in VO form, which is also consistent with the findings in Huang et al. (2013).

做 *zuo*

The occurrence of 做 *zuo* in Mainland Mandarin is decreased by factors such as ARGSTRtwo, FOREVTyes, and INTEREVTyes, whereas the occurrence of 做 *zuo* in Taiwan Mandarin is decreased by ACCOMPTEVTyes, but significantly increased by ASP*le*. It is obvious to linguists that 做 *zuo* in both Mainland and Taiwan Mandarin are frequently found with the perfective marker 了 *le*, but our analysis reveals that the affixation 了 *le* to Taiwan **做** *zuo* is much more frequent than that in Mainland.

3.2 Clustering Analysis of Light Verb Variations

We adopted a vector space model (VSM) to represent the use of light verbs. The features in Table 1 could be expanded to 17 binary features. For example, ASP could be expanded into four binary features: ASP.le, ASP.zhe, ASP.guo, ASP.none. Each instance of a light verb in the corpus was represented by a vector with 17 dimensions. Each dimension stores the value of one of the 17 binary features determined by the context where the light verb is used.

Cluster ID		0	1	2	3	4	5	6	7	8	9
congshi	TW	39	43	1	84	2	21	4	4	1	1
	ML	62	48	0	83	1	4	1	1	0	0
gao	TW	38	141	0	0	9	10	2	0	4	0
	ML	88	64	3	8	11	5	10	4	6	4
jiayi	TW	152	0	6	28	11	2	0	4	0	0
	ML	117	3	6	62	18	2	5	14	1	1
jinxing	TW	26	79	7	2	38	30	0	3	15	1
	ML	23	80	16	0	55	22	5	2	1	0
zuo	TW	20	3	0	2	23	130	20	2	1	6
	ML	23	44	3	16	38	45	20	11	8	3

Table 5: The distribution of data origin by the clustering result.

Then we adopt a clustering algorithm K-Means to identify the variations of light verbs in Taiwan and Mainland Mandarin. The assumption is that the instances of a light verb will form different clusters in the hyperspace according to the distances among them. Each cluster reflects a special use of a light verb. For example, there could be one cluster, where all the instances take non-accomplishment event argument, e.g., 加以分析/研究/評論 *jiayi fenxi/ yanjiu/ pinglun* inflict analyze/ research/ comment 'to analyze/ research/ comment', etc.

In this sense, if there are light verb variations between Mainland and Taiwan Mandarin, the light verbs will be distributed to two clusters, one with data mainly from Mainland Mandarin, whereas the other mainly from Taiwan Mandarin. Meanwhile, if a cluster contains much more data from one variant than the other, it indicates the usage of a light verb is mainly restricted to the variant with more data; or if a cluster contains data of similar amount from both Mainland and Taiwan Mandarin, it indicates that the two variants share common usages regarding the light verbs. Therefore, for each light verb, all 400 examples from both Mainland and Taiwan Mandarin are mixed together for the analysis.

As the K-Means algorithm requires an input of the number *N* of the clusters, the selection of *N* is then an issue we need to consider. Remembering that the clusters reflect the use of a light verb rather than data origin, the selection of N should be based on the consideration of how many different uses a light verb may have. As there are 17 expanded binary features, the whole space of the values of the vectors is 2^{17} = 128K. However, the number of different uses for a light verb should not be too large. There is no problem if *N* is set slightly larger than the real number of different uses of a light verb. For example, if there are 5 different uses for a light verb and we set N=6, then we can imagine that there may be two clusters that reflect the same use of the light verb. On the contrary, if *N* is set too small, all different uses will be mixed together. Then, the clustering result may not be able to show any interesting result we expected. In our experiments, we set *N*=10 for all the five light verbs. Especially, we use the WEKA (Hall et al., 2009) implementation of the simple K-Means for our experiments. The result is shown in Table 5. The key variations of each light verb are summarized as follows.

從事 *congshi*

Cluster 5 shows that Mainland 從事 *congshi* prefers to take complements denoting formal or official events in Mainland Mandarin. However, Taiwan 從事 *congshi* does not show such preference as it can take both formal and informal events. Clusters 6 and 9 show that Taiwan 從事 *congshi* can also take complements in VO form, e.g., 進行開票 *jinxing kaipiao* proceed ballot counting 'to proceed with ballot counting', but this is not preferred by Mainland 從事 *congshi*.

搞 *gao*

Clusters 6 and 7 together show that the argument of Mainland 搞 *gao* can occur in the subject position in addition to the complement position, but such word order is rarely found in Taiwan data. Cluster 3 shows a possibility for Mainland 搞 *gao* to take arguments denoting events involving interactions of participants (e.g., 討論 *taolun* 'to discuss'). In addition, Cluster 9 shows the possibility

that Mainland 搞 *gao* can take complements describing informal events, while the complements to Taiwan Mainland 搞 *gao* are more often formal events (especially political activities).

加以 *jiayi*

Cluster 7 suggests Mainland 加以 *jiayi* show a preference over complements denoting mental or psychological events. However, although Clusters 1 and 6 show some difference between Mainland and Taiwan 加以 *jiayi*, our closer examination of the original data found that such differences actually do not reflect any variant-specific uses.

進行 *jinxing*

Cluster 6 suggests that Mainland 進行 *jinxing* show a preference over the aspectual marker 了 *-le*, but such preference is not seen in Taiwan 進行 *jinxing*. Cluster 8 shows a preference by Taiwan 進行 *jinxing* that it could take VO compound (e.g., 投票 *toupiao* cast-ticket 'to vote') as complements, while this rarely happens in Mainland.

做 *zuo*

Clusters 1 and 3 show that in Mainland Mandarin, it is common for 做 *zuo* to take the aspectual marker 了 *-le*, but such use of 做 *zuo* in Taiwan is not as common as in Mainland.

To sum up, the results from the machine learning method are consistent with that from the multivariate statistical analysis in Section 3.1. Bringing together, we find that while the light verbs in Mainland and Taiwan Mandarin show similarities (as the speakers of these two regions can communicate without difficulty), there are indeed also variations in the two variants.

4 Concluding Remarks

Our study is the one of the first comparable corpus driven computational modeling studies on newly emergent language variants. The automatic identification of Mainland and Taiwan syntactic variations has very significant linguistic and computational implications. Linguistically, we showed that our comparable corpus driven statistical approach can identify comparative differences which are challenging for human analysis. The fact that newly emergent variants differ from each other comparatively rather than contrastively may also have important linguistics implications. In addition, by successfully differentiating these two variants based on their uses of light verbs, the result also suggests that variations among such newly emergent variants may arise from categories that are semantically highly bleached and tend to be/or have been grammaticalized. Computationally, the ability of machine learning approaches to differentiate Mainland and Taiwan variants of Mandarin Chinese potentially contributes to overcoming the challenge of automatic identification of subtle language/dialect variations among other light verbs, other lexical categories, as well as other languages/dialects.

Acknowledgements

The work is supported by a General Research Fund (GRF) sponsored by the Research Grants Council (Project no. 543512) and NTU Grant no. M4081117.100.500000.

References

Arppe, Antti. 2008. Univariate, bivariate and multivariate methods in corpus-based lexicography - a study of synonymy. *Publications of the Department of General Linguistics*, University of Helsinki, volume 44.

Butt, Miriam and Wilhelm, Geuder. 2003. On the (semi) lexical status of light verbs. *Semi-lexical Categories*, Pages 323-370.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina, and R. Harald Baayen 2007. Predicting the dative alternation. In: Cognitive Foundations of Interpretation. Boume, G., I. Kraemer, and J. Zwarts. Amsterdam: Royal Netherlands Academy of Science, pp. 69-94.

Cai, Wenlan. (1982). Issues on the complement of *jinxing* (“进行”帶賓問題). *Chinese Language Learning (漢語學習)* (3), 7-11.

Diao, Yanbin. 2004. *現代漢語虛義動詞研究 (Research on Delexical Verb in Modern Chinese)*. Dalian: Liaoning Normal University Press.

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann and Ian H. Witten. 2009. The WEKA data mining software: An update. *SIGKDD Explorations*, 11(1):10-18.

Han, Weifeng, Antti Arppe, and John Newman. 2013. Topic marking in a Shanghainese corpus: from observation to prediction. *Corpus Linguistics and Linguistic Theory* (preprint).

Hong, Jia-fei, and Chu-Ren Huang. 2013. 以中文十億詞語料庫為基礎之兩岸詞彙對比研究 (Cross-strait lexical differences: A comparative study based on Chinese Gigaword Corpus). *Computational Linguistics and Chinese Language Processing*. 18(2):19-34.

Hong, Jia-fei, and Chu-Ren Huang. 2008. 語料庫為本的兩岸對應詞彙發掘. (A corpus-based approach to the discovery of cross-strait lexical contrasts). *Language and Linguistics*. 9 (2):221-238.

Huang, Chu-Ren. 2009. *Tagged Chinese Gigaword Version 2.0*. Philadelphia: Lexical Data Consortium, University of Pennsylvania. ISBN 1-58563-516-2

Huang, Chu-Ren and Jingxia Lin. 2013. The ordering of Mandarin Chinese light verbs. In *Proceedings of the 13th Chinese Lexical Semantics Workshop*. D. Ji and G. Xiao (Eds.): CLSW 2012, LNAI 7717, pages 728-735. Heidelberg: Springer.

Huang, Chu-Ren, Jingxia Lin, and Huarui Zhang. 2013. World Chineses based on comparable corpus: The case of grammatical variations of jinxing. *《澳门语言文化研究》*, pages 397-414.

Hwang, Jena D., Archna Bhatia, Clare Bonial, Aous Mansouri, Ashwini Vaidya, Nianwen Xue, Martha Palmer. 2010. PropBank annotation of multilingual light verb constructions. *Proceedings of the Fourth Linguistic Annotation Workshop, ACL 2010*, 82–90. Jespersen, Otto. 1965. *A Modern English Grammar on Historical Principles. Part VI, Morphology*. London: George Allen and Unwin Ltd.

Lin, Jingxia, Hongzhi Xu, Menghan Jiang and Chu-Ren Huang. 2014. Annotation and classification of light verbs and light verb variations in Mandarin Chinese. *COLING Workshop on Lexical and Grammatical Resources for Language Processing*. Dublin, August 24.

Nagy, István, Veronika Vincze, and Richárd Farkas. 2013. Full-coverage identification of English light verb constructions. *In Proceedings of the International Joint Conference on Natural Language Processing*, pages 329-337.

Qiu, Zhipu, 1990. 大陸和台灣差別詞典 *(Dictionary of Mainland and Taiwan Mandarin)*. Nanjing University press.

Tu, Yuancheng and Dan Roth. 2011. Learning English light verb constructions: Contextual or statistical. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*. Association for Computational Linguistics.

Wang Tiekun and Li Xingjian, 1996. 兩岸詞彙比較研究管見 (Research on lexical differences between Mainland and Taiwan Mandarin), *World Chinese* (《*華文世界*》), volume 81.

Wei Li and Sheng Yuqi, 2000. 大陸及港澳台常用詞對比詞典. (Comparative Dictionary of Lexical use in Mainland, Hong Kong, Macau and Taiwan), Beijing Industry University Press.

Xu Danhui, 1995. 兩岸詞語差異之比較 (Lexical difference between Mainland and Taiwan Chinese). *1st symposium on Cross-Strait Lexical and Character differences* (第一屆兩岸漢語語彙文字學術研討會論文集).

Zeng Rongfen, 1995. 兩岸語言詞彙整理之我見 (Opinion on cross-Strait language differences)*1st* symposium *on Cross-Strait Lexical and Character differences* (第一屆兩岸漢語語彙文字學術研討會論文集).

Zhou, Gang. 1987. 形式動詞的次分類 (Subdivision of dummy verbs). *Chinese Language Learning (漢語學習)*, volume 1, pages 11-14.

Zhu, Dexi. (1985). 現代書面漢語里的虛化動詞和名動詞 (Dummy verbs and NV in Modern Chinese). *Journal of Peking University (Humanities and Social Sciences)* (*北京大學學報(哲學社會科學版)*, volume 5, pages 1-6.