Comparable Corpus Based Approach to Description and Identification of Varieties of Chinese

Chu-Ren Huang[1] & *Menghan* Jiang[1] & *Jingxia Lin* [2]
[1] The Hong Kong Polytechnic University;
[2] Nanyang Technological University

- Mandarin Chinese is one of the most commonly learned first or second languages in the world now.

- **Mainland China and Taiwan**: official language
- **Singapore** : one of the four official languages.
- **Hong Kong**: the "biliterate and trilingual" (兩文三語) policy

Past Work on Varieties of Chinese

- Lexical variations and grammatical variations are two important aspects of variation studies
- A number of studies have been done on lexical variations
  - case studies (e.g., Sheng 1996)
  - corpora (e.g., Hong&Huang,2013)
  - statistical (e.g., Hong&Huang,2013)
  - comparative dictionaries (e.g., Li 2010, T'sou and You 2010)

- However, studies on grammatical variation is not comparable to those of lexical variations
  - most existing studies are only listing or brief discussions of individual grammatical constructions
  - e.g., Chen (1986), Lu (2002), Chew (2007), on Singapore Chinese, Shi et al. (2006) on Hong Kong Chinese, and Diao (2000), Tseng (2003), Huang et al. (2014), and Lin et al. (2014) on Taiwan Chinese.

Challenges to Study of Grammatical Variations among Varieties

- Same language, mutually intelligible
  - Introspection not applicable;
  - observation hardly effective for lack of 'red-flagged' ungrammatical usages;
  - Convergence greatly outnumber divergence
  - Yet, people know there are differences!
- We (Huang et al. 2013) hypthesize that
  - Grammatical variations among varieties are frequency or preference based tendencies;
  - Are best discovered by study of comparable corpus

Issues

- Corpus Linguistics/NLP
  - How to extract grammatical variations
  - How to apply knowledge of grammatical variations to classify different language varieties
- Linguistic/Theoretical
  - What kind of variations are allowed among varieties of the same language
  - Can the emergence of grammatical variations inform us about possible language changes?

Our studies

A comparable corpus-based approach

- Two case studies comparing Mainland and Taiwan Chinese
  - **Light Verb Construction**
  - **VO Compounds**

  +These two constructions are chosen exactly because they have 'light-weight' (i.e. less rigid) grammatical constraints hence we hypothesize that this where variations may arise.
- Comparable corpora from the Tagged Chinese Gigaword corpus 2.0 (Huang 2009, over 1.1 billion characters)
  - Taiwan Central News Agency (700 m. characters/500 m. words)
  - Mainland Xinhua News Agency (400 m. characters/300 m. words)

Case Study 1: **L**ight **V**erb **C**onstruction

Similar to English light verbs: ***take*** *rest,* ***give*** *advice,* ***give*** *description*

Semantically Bleached

- Containing no eventive (predicate-argument) information
  - The predicative content mainly comes from its taken complement

    *進行**討論** jin4xing2 tao3lun4 'have a discussion'*

- Do not have strong selectional Restrictions
  - They can take a wide range of objects, including deverbal nouns, eventive nouns, and even coerce eventive meaning from concrete nouns.
  - Different light verbs are sometimes interchangeable with the same nominal object

Underspecified Selecitonal Restriction of Chinese Light Verbs

- 從事*cong2shi4*,搞*gao3*, 加以*jia1yi3*, 進行*jin4xing2*, 做*zuo4* are among the most frequently used (also most typical) light verbs in Modern Chinese

- The use of these five light verbs are sometimes interchangeable

- 從事/搞/加以/進行/ 做研究
  cong2shi4/gao3/jia1yi3/jin4xing2/zuo4 yan2jiu1
  "to do research"

Underspecified Selecitonal Restriction of Chinese Light Verbs II

- Collocation constraints are sometimes found with these light verbs,

- **進行**/*加以/*從事/**搞**/*做賽事,

  jin4jing2/*jia1yi3/*cong2shi4/***gao3***/*zuo4 bi3sai4*

  "play a game"

- *進行/**加以**/*從事/*搞/*做考慮

  jin4jing2/jia1yi3***/*cong2shi4/*gao3/*zuo4 kao3lv4*

  "give consideration"

Observation on Variations of Light Verb Usages in Chinese Varieties

- Even with the very limited collocation constraints, variations still exist: Taiwan light verbs tend to take more types of NPs and even VPs as its complements

- 進行感恩之旅/君子之爭
  Jin4xing2 gan3en1zhi1lv3/ju1zi3zhi1zheng1
  "to proceed with a 'thanksgiving trip'/'gentlemen's dispute'"

- 進行抹黑/開票
  Jin4xing2 mo3hei1/kai1piao4
  "to proceed with 'mud-slinging'/'ballot counting' "
  -------(Huang et al. 2013)

Methodology

- A comparable-corpus-driven statistical approach
- 加以*jia1yi3*, 進行*jin4xing2*, 從事*cong2shi4*, 搞*gao3*, 做*zuo4* in Mainland Mandarin and Taiwan Mandarin
- Statistical methods and tools
  - Univariate analysis + multivariate analysis
  - Polytomous package in R (Arppe 2008)

Data

- Chinese Gigaword corpus

- Random sample: 200 sentences for each of the five light verbs in Mainland and Taiwan corpora
  - 1,000 in total for Mainland Chinese
  - 1,000 in total for Taiwan Chinese

- 12 linguistic factors collected from past studies of Mainland/Taiwan Variations are Tested

			Value levels
Co-occur with other light verbs "OTHERLV"	**開始 進行 比賽 *kai1shi3/jin4xing2/bi3sai4***	**"start the game"**	**Yes, no**
Take aspectual marker: 著，了，過 "ASP"	**昨天進行了比賽 zuo2tian1/jin4xing2/le0/ bi3sai4**	**"played the game yesterday"**	**No, le, zhe, guo**
Event complement is at subject position "EVECOMP"	**比賽在學校進行 bi3sai4/zai4/xue2xiao4/ jin4xing2**	**"play the game at school"**	**Yes, no**

POS **“POS”**	進行比賽（N） ***jin4xing2/bi3sai4*** 進行戰鬥（V） ***jin4xing2/*** ***zhan4dou4***	**“play the game”** **“fight the battle”**	**N, V**
Argument structure **“ARGSTR”**	進行調查（two） ***jin4xing2/*** ***diao4cha2***	**“carry on investigation”**	**One, two, zero**
VO compound as argument **“VOCOMP”**	進行投 票 ***jin4xing2/*** ***tou2piao4***	**“carry on voting”**	**Yes, no**

Spontaneous/ controllable event "SPONTEVT"	進行投票 ***jin4xing2/tou2piao4***	**"carry on voting"**	**Yes, no**
durative event "DUREVT"	進行比賽 ***jin4xing2/bi3sai4***	**"play a game"**	**Yes, no**
formal event "FOREVT"	進行訪問 ***jin4xing2/fang3wen4***	**"pay an official visit"**	**Yes, no**
psychological activity "PSYEVT"	加以考慮 ***jia1yi3/kao3lv4***	**"give consideration"**	**Yes, no**
event involving interaction of agent and patient "INTEREVT"	進行溝通 ***jia1yi3/gou1tong1 inflict/communicate***	**"do communication"**	**Yes, no**
accomplishment complement "ACCOMPEVT"	進行修正 ***jin4jing2/xiu1zheng4 proceed/correct***	**"make corrections/ amendments"**	**Yes, no**

Univariate analysis of Chinese light verbs

- Chi-squared tests for the significance of the co-occurrence of the factor with individual light verbs

- Chisq.posthoc() function in the Polytomous package automatically transforms the results (Standardized pearson residuals e_{ij} (Agresti 2002)) into signs
  - "+": $e_{ij} > 2$, statistically significant overuse of the light verb with the factor
  - "-": $e_{ij} < -2$, statistically significant underuse of the light verb with the factor
  - "0": e_{ij} [-2,2], lack of statistical significance

Comparison of Mainland and Taiwan light verbs -univariate analysis

			congshi		gao		jiayi		jinxing		zuo	
	factor	N	ML	TW	ML	TW	ML	TW	ML	TW	ML	TW
1	POS.N	585	+	+	+	+	-	-	0	-	0	-
2	POS.V	1415	-	-	-	-	+	+	0	+	0	+
3	ARGSTR.one	376	0	+	-	-	-	-	0	+	+	0
4	ARGSTR.two	1039	-	-	0	-	+	+	0	-	-	+
5	ARGSTR.zero	585	+	+	+	+	-	-	0	-	0	-
6	VOCOMP.no	1939	0	0	0	0	0	+	0	-	0	0
7	VOCOMP.yes	61	0	0	0	0	0	-	0	+	0	0
8	EVECOMP.no	1919	+	+	-	0	+	+	-	-	-	0
9	EVECOMP.yes	81	-	-	+	0	-	-	+	+	+	0
10	ASP.guo	9	0	0	0	0	0	0	0	0	0	0
11	ASP.le	155	-	-	-	-	-	-	+	-	+	+
12	ASP.no	1835	+	+	+	+	+	+	-	+	-	-
13	ASP.zhe	1	0		0		0		+		0	
14	SPONTEVT.no	1		0		+		0		0		0
15	SPONTEVT.yes	999		0		-		0		0		0
16	DUREVT.no	35	-	0	0	0	+	+	-	0	-	0
17	DUREVT.yes	1965	+	0	0	0	-	-	+	0	+	0
18	FOREVT.no	66	0	+	0	-	-	-	0	0	+	0
19	FOREVT.yes	1934	0	-	0	+	+	+	0	0	-	0
20	PSYEVT.no	1981	0	0	0	0	-	0	0	0	0	-
21	PSYEVT.yes	19	0	0	0	0	+	0	0	0	0	+
22	INTEREVT.no	1870	+	+	0	+	+	0	-	-	+	0
23	INTEREVT.yes	130	-	-	0	-	-	0	+	+	-	0
24	ACCOMPEVT.no	1904	+	+	+	+	-	-	+	+	+	0
25	ACCOMPEVT.yes	96	-	-	-	-	+	+	-	-	-	0

Key results:

ML and TW 做*zuo4* show opposite usage tendency of the feature ARGSTR.two

ML and TW 進行*jin4xing2* show opposite usage tendencies of the features ASP.le and ASP.no

But the difference is more between a significant and non-significant feature, rather than between a significant positive vs. a significant negative feature

Polytomous Logistic Regression

- 加以/進行/從事/搞/做 研究.
- *Jia1yi3/jin4xing2/cong2shi4/gao3/zuo4 yan2jiu1*
- "to do research"

- Five light verbs as the possible outcome
  - Estimate the probability of presence of each of the potential light verb

- Polytomous logistic regression
  - An extension of standard logistic regression
  - allows for simultaneous estimation of the probability of multiple outcomes (light verbs in the current study)

Comparison of Mainland and Taiwan light verbs in multivariate polytomous regression

	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*	(0.7615)	*76.47*	(Inf)	(1.481)	(0.7062)	*0.2177*	(1.217)
ARGSTRzero	(1.097)	(0.6251)	3.584	7.177	(1/Inf)	(4.382)	(1.179)	*0.5393*	0.245	0.2075
ASPle	(0.7487)	(1/Inf)	(0.1767)	(1/Inf)	(0.8257)	(0.3027)	(0.9196)	(Inf)	(1.853)	32.98
ASPno	(Inf)	(0.9291)	(1.499)	(0.6946)	(Inf)	(Inf)	(0.2307)	(Inf)	(0.2389)	(0.2386)
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)	(0.8534)	(1/Inf)	(1/Inf)	3.975	8.115	(1.772)	(0.5016)
FOREVTyes	(2.744)	*0.08674*	(1.227)	(Inf)	(Inf)	(Inf)	(0.7457)	(1.441)	*0.2679*	(1.467)
INTEREVTyes	0.03255	0.1896	(0.5281)	(1/Inf)	(0.5432)	(0.951)	18.67	10.46	*0.08902*	(0.3979)
PSYEVTyes	(1/Inf)	(1/Inf)	(1/Inf)	(1/Inf)	*19.87*	(1.395)	(1/Inf)	(1/Inf)	(0.9619)	(3.323)
SPONTEVTyes		(Inf)		(1/Inf)		(1/Inf)		(Inf)		(Inf)
VOCOMPyes	(0.1346)	*0.18*	(3.043)	(2.351)	*23.54*	(Inf)	(1.086)	*3.16*	(0.5344)	(0.5956)

The difference between two Chinese varieties for the same light verbs, however, is between statistically significant vs. non-significant pairs.

10-fold corss validiation result of ID3 algorithem on both corpora

	Precision		Recall		F-Measure	
	TW	ML	TW	ML	TW	ML
jingxing	0.442	0.593	0.311	0.423	0.365	0.494
gao	0.681	0.449	0.557	0.347	0.612	0.391
zuo	0.61	0.57	0.537	0.562	0.571	0.566
jiayi	0.634	0.72	0.946	0.9	0.759	0.8
congshi	0.528	0.583	0.579	0.724	0.552	0.646
Average	0.58	0.586	0.588	0.599	0.574	0.585

Confusion matrix of the classification with ID3 algorithm on both corpora

	jingxing		**Gao**		**zuo**		**jiayi**		**congshi**	
	TW	ML	TW	ML	TW	ML	TW	ML	TW	ML
Jingxing	**61**	**83**	15	27	36	40	38	11	46	35
Gao	20	16	**113**	**70**	13	23	24	39	33	54
Zuo	24	25	8	28	**108**	**118**	39	25	22	14
Jiayi	5	11	0	6	5	6	**192**	**206**	1	0
Congshi	28	5	30	25	15	20	10	5	**114**	**144**

Classifying five light verbs by automatic clustering

	Mainland					Taiwan				
	0	1	2	3	4	0	1	2	3	4
jinxing	2	32	110	23	37	30	10	77	20	64
gao	2	33	116	41	11	120	23	30	0	31
zuo	0	36	80	14	81	19	4	47	5	132
jiayi	68	0	161	0	0	0	0	1	6	196
congshi	0	67	66	21	46	90	20	68	0	22

- the simple K-Means clustering algorithm (number of examples indicated in each cluster)
- Light verb 加以 *jiayi* behaves quite differently from the other four light verbs in both Mainland and Taiwan corpora
  - Closer examination found a tendency that all other light verbs mostly take activity complements but fewer accomplishment complements in both Taiwan and Mainland corpora.
- Cluter 0 is mainly formed by instances of 加以 *jiayi* from Mainland corpus
  - closer examination of the examples found that the cluster mainly includes sentences where jiayi takes complements denoting accomplishment events, e.g. gaizheng ‘to correct’ and jiejue ‘to solve’. However, Taiwan jiayi mainly takes complements denoting activity events, and thus almost all instances of Taiwan jiayi are mixed with those of the other light verbs

Classifying five light verbs by automatic clustering

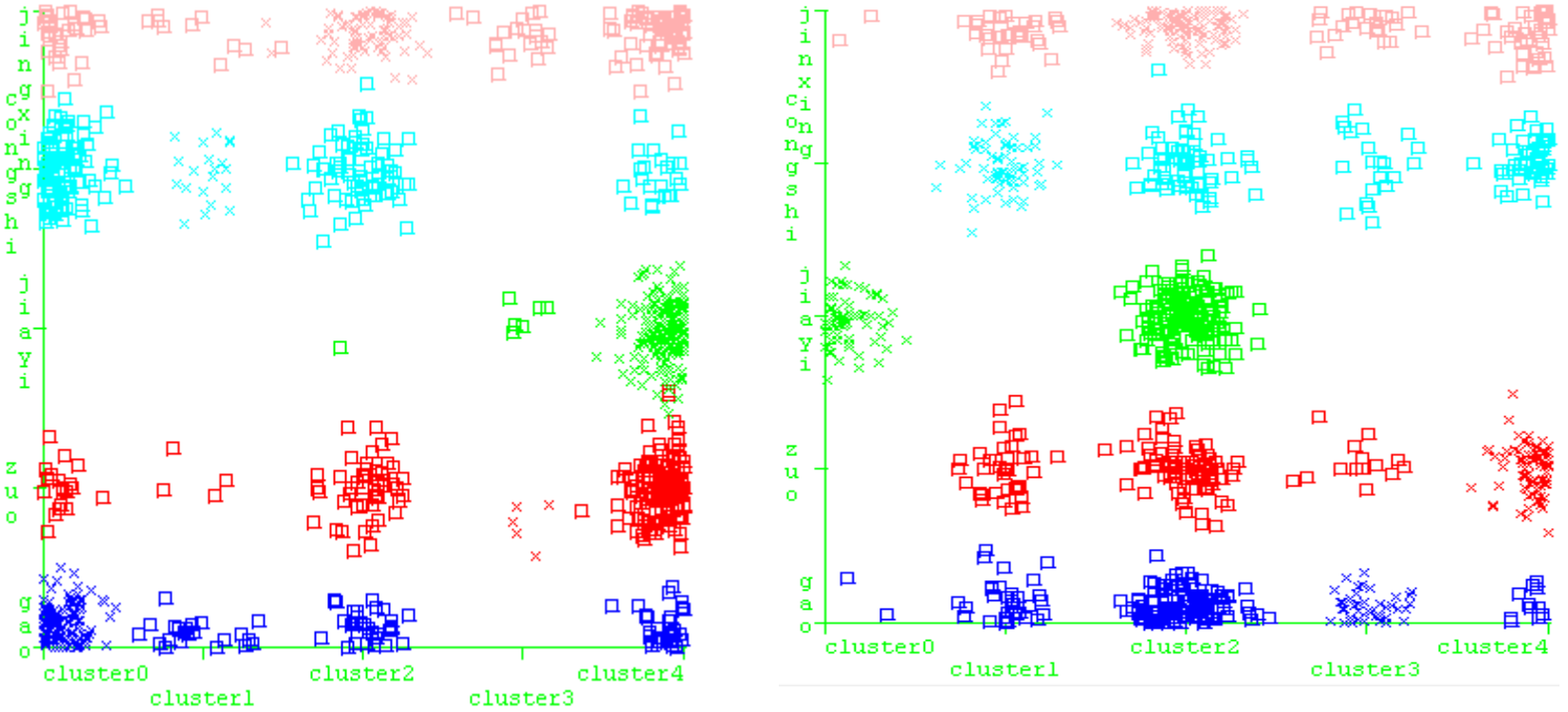


Figure 1. Clustering result on Taiwan (Left) and Mainland (Right) corpora. X-axis is the clusters, Y-axis is the light verbs.

Light verb 加以 *jiayi* behaves quite differently from the other four light verbs in both Mainland and Taiwan corpora,
Meanwhile, we can also find a cluster which is mainly formed by instances of 加以 *jiayi* from Mainland corpus (i.e. cluster 0).

Summary of Case Study 1

- In this case study,
  - We identified the subtle variations between the light verbs in Taiwan and Mainland Chinese (cf. Huang et al. 2013).
  - both statistical methods and machine learning technologies were used to cross-verify the result.
  - The methodology used in this case study can be adopted for future studies on other light verbs as well as other lexical categories.

Case Study 2: VO Compounds

- Lexicalized VO compounds is a challenge in both theoretical and computational accounts as they
  - Can occur in discontinuous positions (V and O occur separately, called ionization by Y.R. Chao 1968)
  - As a lexical verb, they cannot take direct object in general. For instance, 开刀 ‘to perform operation on, to operate on…’ is composed of open+knife, has a transitive meaning (and argument structure), but its ‘object’ can only be introduced in non-canonical positions, such as prepositional objects, possessive subject, or topics.
- However, a few VO compound verbs can be used transitively
  - 关心他人 guan1xin1 (close+heart)/ta1ren1 ‘show concern for him’

Variation in VO compounds

- Mandarin Chinese disyllabic VO compounds have been observed to have different degrees of transitivity
  - ???观光意大利 guan1guang1/yi4da4li4 'travel in Italy'
  - *出丑别人 chu1chou3/bie2ren2 embarrass someone??

- The degree of transitivity of VO compound can also differ in different varieties of Chinese
  - 幫忙 bang1mang2 'give a hand':
    幫忙 他 bang1mang2/ta1 'give him a hand'
    √ Taiwan Mandarin and Singapore Chinese
    ???Mainland Mandarin

Methodology

- A comparable corpus-based approach
- Comparison between Mainland and Taiwan
- 234 VO compounds
  - Collected based on corpus as well as previous studies
- Chinese Gigaword

(1). Differ in the type of the taken object?

Syntactic type

Thematic role

Syntactic complexity

Semantic polarity

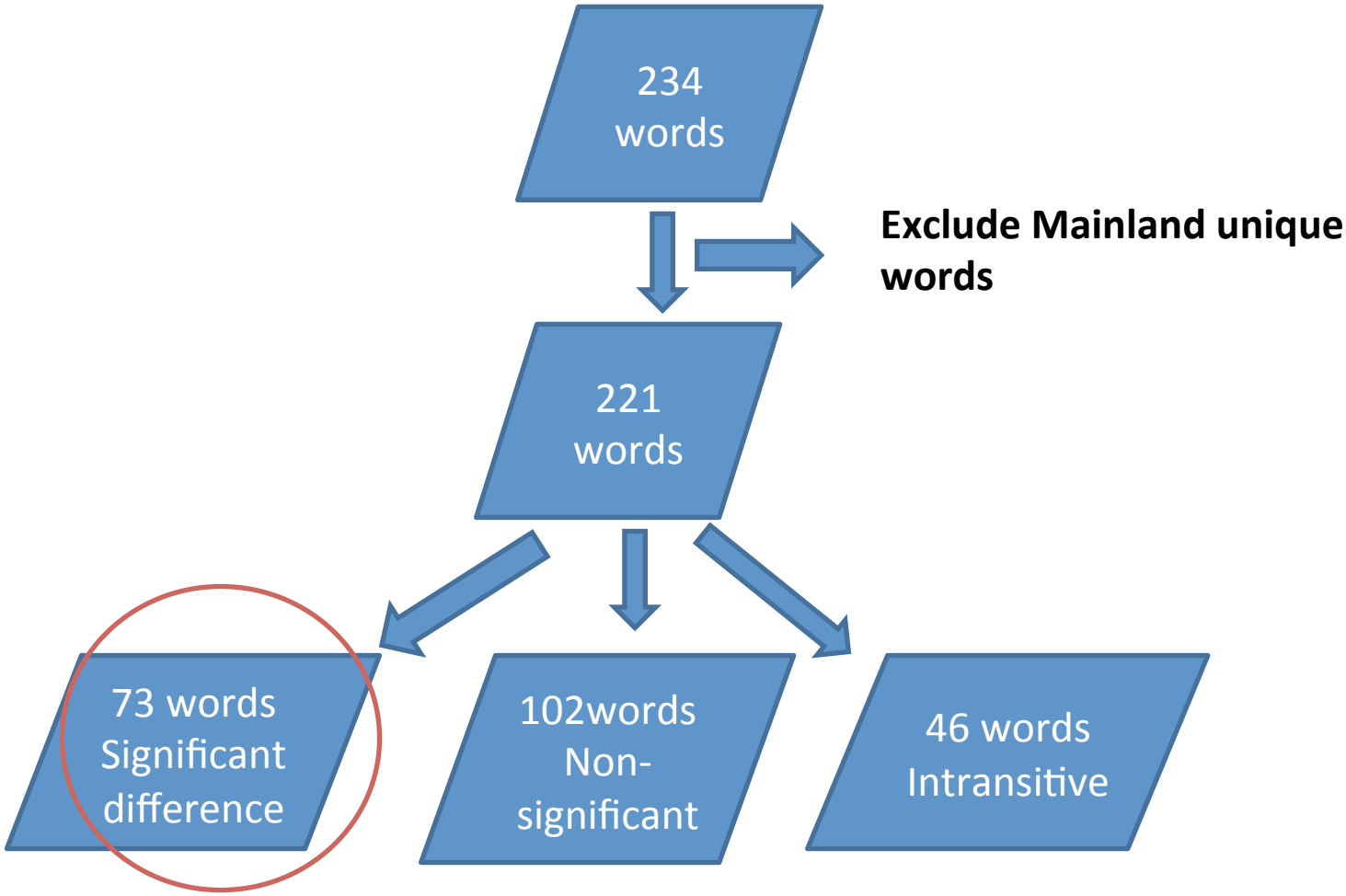
Pronominality

(2). Differ in transitivity?

Relative frequency of transitive usages

(2). Transitivity difference

- Relative Frequency = transitive tokens/all the tokens
- Z-test : significance of the difference between two independent proportions
- P-value<0.05



The result of Z-test

Statistically significant, but……

For example,

	TW	ML	Ratio	P value
借道 Jie4dao4	271/311	60/87	1.26	<0.05
涉足 Sh4zu2 ‘be included in’	1157/1269	1203/1481	1.12	<0.05

Z-test and P value failed to obtain linguistically significant results.

We propose to use instead likelihood ratio of transitivity:

Likelihood Ration R = RelFreg A/RelFreq B

Filter by likelihood of ration of transitivity:

Ratio	Total	TW higher	ML higher
Ratio>3	33	27	6
Ratio>5	23	21	2
Ratio>10	9	9	0
Absolute difference	7	7	0

In general, in terms of their transitivity frequency, **Taiwan VO compounds are more likely to be used in a transitive way.**

Prominent significant difference (Ratio >=10)

9 words TW higher

	TW	ML	Ratio
媲美 pi4mei3 ‘rival’	727/1021	28/1030	26.19
中意 zhong1yi4 ‘like’	192/540	8/1337	59.42
把關 ba3guan1 ‘guarantee’	182/743	11/1547	34.45
過境 guo4jing4 ‘transit’	341/1000	33/1000	10.33
獻計 xian4ji4 ‘offer advice’	6/84	2/1000	35.71
移民 yi2min2 ‘immigrate’	455/2000	1/1000	227.5
接壤 jie1rang3 ‘neighbor on’	34/922	1/2269	83.67
撤軍 che4jun1 ‘pullback’	23/1000	1/1000	23
聯手 lian2shou3 ‘join hands’	10/1000	1/1000	10

- Absolute difference:

	TW	ML	Examples in TW
撤兵 che4bing1 'pullback'	1/197	0/46	撤兵西岸地區
垂愛 chui2ai4	5/37	0/2	老天特別垂愛鐘嶽岱
領航 Ling3hang2 'pilot'	76/810	0/169	有能力領航國家發展
觀光 Guan1guang1 'visit'	4/1000	0/5224	觀光義大利
轉行 Zhuan3hang2 'switch job'	18/392	0/167	轉行影視界
失望 Shi1wang4 'disappoint'	3/1000	0/1000	我很失望他未早日全心全力處理重要問題
過目 Guo4mu4 'look over'	22/317	0/65	過目所有的展品幻燈片

Summary of Case Study 2

- The differences between the two variants often lie in the presence/absence of a **tendency**.
- In general, Taiwan VO compounds are more likely to be used **in context with higher transitivity**
- **Likelihood Ratio** identified high transitivity uses of a small number of VO compounds (esp. from Taiwan Mandarin)

Conclusion

- Comparative corpus based empirical model can automatically classify variations among near synonyms (i.e. light verbs) and different varieties of the same language.
- Variations similar to grammaticalization (e.g. increase of transitivity of VO compounds) are shown to have lexical diffusion like features by likelihood ratio comparison

- 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.
- Huang, Chu-Ren, Lin, Jingxia, Hongzhi Xu, and Menghan. 2014. Corpus-based Study and Identification of Mandarin Chinese Light Verb Variations. The Workshop on Applying NLP Tools to Similar Languages, Variations, and Dialects (VarDial), at the the 25th International Conference on Computational Linguistics. Dublin, Ireland.
- Jiang, Menghan, Jingxia Lin, and Chu-Ren Huang. *A comparable corpus-based study of VO compound variations between Mainland and Taiwan Mandarin. T*he 23rd annual conference of the International Association of Chinese Linguistics (IACL). 26-28 August. Hanyang University, Seoul, Korea.
- Lin, Jingxia, Hongzhi Xu, Menghan Jiang and Chu-Ren Huang. 2014. Annotation and Classification of Light Verbs and Light Verb Variations in Mandarin Chinese. The Workshop on Lexical and Grammatical Resources for Language Processing (LG-LP 2014), the the 25th International Conference on Computational Linguistics. Dublin, Ireland.
- Xiong Jiajuan and Chu-Ren Huang. 2015. Dongzuo'action' as a nominalizer in Taiwan Mandarin : a corpus driven study. To be presented at XXVIIIes Journées de Linguistique d'Asie Orientale. 28th Paris Meeting on East Asian Linguistics July 2. CLRAO-EHESS,

Thank you!

(1). Preference difference

Features	Variables
1. Syntactic type	NP, Deverbal noun, VP, Clause
2. Thematic role	patient, theme, goal, benefactive, source, instrument, locative
3. Syntactic complexity	length of object
4. Semantic polarity	Positive, Neutral, Negative
5. Pronominality	personal, demonstrative, interrogative pronoun

(1). Preference difference:

1. Variation in syntactic type

1.1 Taiwan tends to take event-denoting objects (deverbal noun, VP and clause) while Mainland prefer to take common NP as the object.

1.1 Taking VP and deverbal noun
e.g., 插足 cha1zu2 'interfere'

	Taiwan	Mainland
VP: 35/185	插足經營貿易等產業 Cha1zu2/jing1ying2/mao4yi4/ deng3/chan3ye4 Interfere/manage/commerce/and so on/industry 'interfere the management of industries such as commerce'	NA
Deverbal noun: 7/185	插足電視營運 Cha1zu2/dian4shi4/ying2yun2 interfere/cable TV/business 'interfere the business of the Cable TV'	NA
Common NP	插足造船業 Cha1zu2/zao4chuan2ye4 Interfere/shipbuilding industry 'Interfere the shipbuilding industry'	插足国际市场 Cha1zu2/guo2ji4/shi4chang3 Interfere/international/market 'interfere the international market'

1. Variation in syntactic type

1.2 Taking clause as the object

- Taiwan VO compounds have the tendency of taking clause as the object while Mainland compounds do not.

Taiwan Example:

e.g., 曝光 bao4guang1 'expose'

~她以前也曾是護產科護士

'to expose that she used to be a nurse'

- **2. variation in thematic role：**

Taiwan VO compounds tend to take more types of objects. e.g.,

	TW			ML
投資	Goal： 19/31	Quantity： 12/31		Only quantity: 20/20
	中鋼有意參與投資台翔航太工業公司 ‘Zhong Gang wants to participate the investment of TaiXiang Hang Tai industry company’	中鋼公司投资約六百七十億元 ‘Zhong Gang Company invested about 670000000’		总部又每年投资上亿元 ‘The haedquarter invests more than 100000000 every year’

3. Variation in syntactic complexity of the taken objects:

Taiwan VO compounds tend to take more complicated objects.

e.g.,

	TW	ML
插足	~這個世界著名潛力極大的旅遊市场 ~zhe4ge4/shi4jie4zhu4ming2/ qian2li4ji2da4de0/lv2you2/shi4chang3 ~this/world well known/great potential/tourism market 'this world well known tourism market with great potential '	~中国市场 ~zhong1guo2/shi4chang3 Chinese market

4. Variation in semantic polarity：

e.g.,

从事：

- TW：從事性交易/非法勾當 **(can be negative)**

  cong2shi4/xing4jiao1yi4 or fei1fa3gou4dang4

  ‘do sex trade/illegal activity’

- ML：从事国事访问

  cong2shi4/guo2shi4/fang3wen4

  ‘make state visit’

5. Variation in pronominality：

e.g.,
TW examples：

沖到山谷幫忙他
Chong1dao4/shan1gu3/bang1mang2/ta1
‘rush to the valley and give him a hand’

不滿意他
Bu4man3yi4/ta1
‘not satisfied with him’

我們應該感恩他們
Wo3men0/ying1gai1/gan1en1/ta1men0
‘We should be grateful for them

Selected references

- Arppe, A. (2008) Univariate, bivariate and multivariate methods in corpus-based lexicography – a study of synonymy. Publications of the Department of General Linguistics, University of Helsinki, No. 44. URN: http://urn.fi/URN:ISBN: 978-952-10-5175-3.
- Arppe, A. (2009) Linguistic choices vs. probabilities – how much and what can linguistic theory explain? In: Featherston, S. & S. Winkler (eds.) The Fruits of Empirical Linguistics. Volume 1: Process. Berlin: de Gruyter, pp. 1–24.
- Arppe, A. (in prep.) Solutions for fixed and mixed effects modeling of polytomous outcome settings.
- Han, Weifeng, Arppe, Antti & Newman, John (2013). Topic marking in a Shanghainese corpus: from observation to prediction. *Corpus Linguistics and Linguistic Theory* (preprint).
- Butt, M., & Geuder, W. (2001). On the (semi) lexical status of light verbs. *Semi-lexical Categories,* 323-370.
- Cattell, R. (1984). *Composite Predicates in English*. Syntax and Semantics Volume 17. Sydney: Academic Press Australia.
- 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.

Selected References

- 蔡文澜. (1982). 进行带宾问题. *汉语学习*. (3), 7-11.
- 陈建民.(2000).内地与香港的词语比较，*语文建设*. 第4期。
- 刁晏斌. (2004) *现代汉语虚义动词研究*， 辽宁师范大学出版社。
- 刁晏斌. (1998). 也谈 “动宾式动词+ 宾语” 形式. *语文建设*, (6), 39-41.
- 陆俭明. (1995) 关于新加坡华语规范化问题，《新加坡联合早报》1995年6月16日
- 陆俭明. (1996) 新加坡华语语法的特点，*新加坡南大中华语言文化学报*.
- 陆俭明. (2002) 新加坡华语句法特点及其规范问题, *新马华人传统与现代的对话*，南洋理工大学。