A Lexico-Semantic Analysis of Chinese Locality Phrases

- A Topic Clustering Approach

August F.Y. Chao
Department of M.I.S, National Chengchi University Taipei, Taiwan
fychao.tw@gmail.com

Siaw-Fong Chung
Department of English National Chengchi University Taipei, Taiwan
sfchung@nccu.edu.tw

Abstract

In this paper, we present a novel approach using LDA (Latent Dirichlet Analysis, Blei, David, Andrew, Michael, and Jordan, 2003) to analyze synonym groups appearing in fixed frames containing Chinese locative phrases, such as [*zái* noun phrase (*yǐ*/*zhī*) *shàng*/*xià*/etc. *biān*/*miàn*/etc.], and to understand noun meanings related to the syntactic forms of locative phrases. We mapped the different noun phrases to their collocating synonym groups before we generated similarity comparison among different combinations. We collected locative phrases using 11 monosyllabic locative words and 5 locative compound-formation patterns from Sketch Engine, and we aligned these compounds with Chinese Synonym Forest (Mei, Zhu, & Gao 1983) before clustering. A Hive Plot (Krzywinski, Birol, Jones, and Marra, 2012) visualizer was constructed in order to help understand the relationship of locative nouns and their synonym groups. The results showed not only the semantic meaning within a locative phrase, but also the corresponding semantic meanings among locative phrases.

1 Introduction

Locative phrases express the locative and directional information in relation to a certain object or entity. In Chinese, locative phrases have the following structure (Li and Thompson, 1989, pp 390):

zái noun phrase ~ (*locative particle*)
'at'

Within this structure, the locative particles can be monosyllabic or disyllabic compounding with prefixes, like *yǐ* and *zhī*, as well as suffixes, like biān, miàn, and tóu. (See Table 1).

Many scholars have done research on locative phrases to understand the language context through frame reference (Hsu and Tai, 2001; Liang and Wang, 2010) and image schema (Liang and Wang, 2010; Wang and Hsieh, 2011), but not on cross-comparing the different locative words with their suffix/prefix combinations.

Table 1 Combinations of Chinese Locative Nouns

	Suffix			Prefix	
	~邊 *biān*	~面 *miàn*	~頭 *tóu*	以 *yǐ* ~	之 *zhī* ~
上 *shàng*	上邊	上面	上頭	以上	之上
下 *xià*	下邊	下面	下頭	以下	之下
前 *qián*	前邊	前面	前頭	以前	之前
後 *hòu*	後邊	後面	後頭	以後	之後
左 *zuǒ*	左邊	左面	N/A	N/A	N/A
右 *yòu*	右邊	右面	N/A	N/A	N/A
裡 *lǐ*	裡邊	裡面	裡頭	N/A	N/A
外 *wài*	外邊	外面	外頭	以外	之外
東 *dōng*	東邊	東面	東頭	以東	之東
西 *xī*	西邊	西面	西頭	以西	之西
南 *nán*	南邊	南面	南頭	以南	之南
北 *běi*	北邊	北面	北頭	以北	之北
內 *nèi*	N/A	N/A	N/A	以內	之內
中 *zhōng*	N/A	N/A	N/A	N/A	之中

In this study, we collected data from the Chinese Giga-word corpus[1] (Ma, and Huang, 2006) in Sketch Engine to retrieve the combinations of

[1] Giga-word corpus contains 2466840 news articles in Taiwan's CNA and Mainland China's XIN.

Chinese locative nouns in Table 1, and we categorized each compound into synonym groups according to the categories provided by the Chinese Synonym Forest (Mei, et. al. 1983) disregarding the part-of-speech information [2]. Then we adapted the LDA (Latent Dirichlet Analysis, Blei, et. al. 2003) methods to cluster each synonym group to extract meaningful groups of combinations existing in our data set. Instead of a network view, we used Hive Plot (Krzywinski, et. al. 2012) to visualize the comparison result of each locative noun combination. The graphical decomposition of concept categories in locative phrases, hopefully, would benefit the analysis of Chinese locative nouns.

2 Methodology

2.1 Latent Dirichlet Allocation

The LDA model involves drawing samples from Dirichlet distributions and from multinomial distributions. This method is widely used in biomedical studies and can profile genes (Flaherty, Giaever, Kumm, Jordan, and Arkin, 2005) by considering DNA sequences are simple 4-letter combination (A, T, G, and C). The formally probabilistic generative process is defined (Blei and Lafferty, 2009) as:

1. For each topic *k*, draw a distribution over words $\phi_d \sim Dir(\alpha)$.
2. For each document *d*,
   a) Draw a vector of topic proportions $\theta_d \sim Dir(\beta)$.
   b) For each word *i*,
      i. Draw a topic assignment $z_{d,i} \sim Mult(\phi_d), z_{d,i} \in \{1,..., K\}$
      ii. Draw a word $w_{d,i} \sim Mult(\phi_{z_{d,i}}), w_{d,i} \in \{1,..., V\}$

where *K* is a specified number of topics, *V* is the number of words in vocabulary; *Dir(α)* is a *K*-dimensional Dirichlet; *Dir(β)* is a *V*-dimensional Dirichlet; and $z_{d,i}$ is the *i*-th word in the *d*-th document.

[2] The locative suffixes and prefixes are also interfered by the concept combination in locative phrases. Because lack of part-of-speech information in Chinese synonym forest, we can't not create explicit formation for locative phrases.

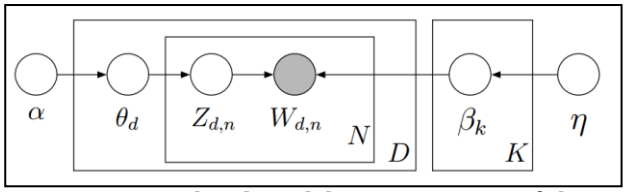


Figure 1 A graphical model representation of the latent Dirichlet allocation (LDA). (Nodes denote random variables; edges denote dependence between random variables. Shaded nodes denote observed random variables; unshaded nodes denote hidden random variables. The rectangular boxes are "plate notation," which denote replication.)

In large corpus experiments, LDA topic model can explain why some parts of the data are similar by observing different sets of various words' probabilities in topics, such as arts, budgets, children and education word groups (Blei, et. al., 2003).

2.2 Chinese Synonym Forest

The Chinese Synonym Forest (or Chilin 同義詞詞林, Mei *et al.*, 1983) is a collection of 5300 Chinese synonyms. In this synonym forest, synonyms were categorized into 3 levels hierarchical groups. The top level of this hierarchy is the upper concept labeled from "A" to "L" including *human*, *object*, *time*/ *space*, *abstract entities*, *characteristic*, *movements*, *psychological*, *phenomenon-condition*, *activities*, *relationship*, *auxiliaries*, and *honorifics*, (see Appendix I). Within each top level, there are several middle and specific synonym groups, and each one has its own group code representing the hierarchical information and synonym relationship symbols: "=" means a semantic equal group, "#" means semantic unequal but in the same group, and "@" is a self-enclosed and independent group. The extended version of the Chinese Synonym Forest by the HIT IR Lab expended the original 3 level hierarchies to 5, deleted rarely usage words, and included modern words from news corpus. Table 2 (next page) shows several examples from the extended version (hereafter Chilin).

In Table 2, we can see that each synonym group has a unique code: the initial capital letter represents the top level concept, the last symbol represents the semantic relationship within a synonym group, and the other letters or numbers in between represent the position of a word in a synonym hierarchy.

Table 2 Samples of Extended Chinese Synonym Forest. (The synonym meanings are translated by the authors.)

Synonym Groups	Synonym Meanings
Aa01B03# 良民 順民	obedient civilians
Aa01C05@ 眾學生	students
Bp20B03= 招子 **幌子** 市招	signboard *
Bg02B07# 超聲波 低聲波 聲波	sonic wave
Bg03A01@ 火	fire
Dd15A09= **幌子** 招牌 牌子 旗號 金字招牌	brand *

This synonym list has two major problems while applying it in computation algorithm: first, because of the lack of clearly definition of each synonym group, we can only conjectured the meaning; second, because Chinese compounds have many senses, a word can be found in many synonym groups, such that *幌子 huǎngzi* (asterisked in Table 2) originally means 'signboard of hotel' and it also commonly means 'brand' (a metaphor when referring to performing an activity under the guise of the name). Despise the problems presenting above, Chilin is the state-of-art collection of synonym wordlist.

2.3 Statistics of Collected Data

We used 11 directional words: *上 shàng*, *下 xià*, *前 qián*, *後 hòu*, *裡/裏 lǐ* *外 wài*, *東 dōng*, *西 xī*, *南 nán*, *北 běi*, and 5 prefixes/suffixes: *～邊 biān*, *～面 miàn*, *～頭 tóu*, *以 yǐ～*, *之 zhī～* to collect data from the Chinese Giga-word corpus, and the results of locative nouns, disregarding the presents of *zái*, can be found in Table 3 below. The reason of excluding *左 zuǒ*,*右 yòu*, *內 nèi*, *中 zhōng* is because these words cannot be found in all 5 prefixes/suffixes.

Because the Chinese Giga-word corpus is a news corpus, we found that not all the combinations can be found. The usage of *~tóu* is significantly lower than other formations in every locative word and statistics shows that the usages of *shàng*, *xià*, *qián*, *hòu*, *lǐ*, *wài* as prefixes of *biān*, as well as *dōng*, *xī*, *nán*, *běi* as suffixes of *yǐ* cannot be found in news corpus. Even *dōng*, *xī*, *nán*, *běi* as media addressing directional information are barely found using *biān* as suffix.

Table 3 Statistics of Collected Data

	~邊 *biān*	~面 *miàn*	~頭 *tóu*	以 *yǐ* ~	之 *zh*~
上 *shàng*	11	788	51	1557	15559
下 *xià*	6	169	3	8273	7547
前 *qián*	3	1085	154	31618	12596
後 *hòu*	9	1028	215	22051	3751
裡 *lǐ*	9	1086	97	0	0
外 *wài*	28	1254	154	4370	1918
東 *dōng*	139	33	0	0	424
西 *xī*	147	66	3	0	874
南 *nán*	118	20	0	0	390
北 *běi*	199	78	0	0	731

2.4 Clustering using LDA and Hive Plot

Pustejovsky (1991:437) points out that "much of the lexical ambiguity of verbs and prepositions is eliminated because the semantic load is spread more evenly throughout the lexicon to the other lexical categories." Here, we differentiate different uses of locative phrases through observing their groups of nouns in a fixed frame. In order to find the meanings corresponding to the locative nouns appearing in the fixed frame [*zái* noun phrase (*yǐ*/*zhī*) *shàng*/*xià*/etc. *biān*/*miàn*/etc.] in all combinations in Table 3 above, we used LDA to cluster in the nouns appearing in each combination by first mapping each noun to its synonym group in Chilin. Before we used Chilin, we needed to translate the original synonym list which is in simplified Chinese into traditional encoding. In order to avoid any translation problems, we uses Simplified/Traditional Chinese conversion table[3] with maximum matching phrases for the conversion. In our study, in order to retrieve the patterns in Table 3, we used *zái* as a keyword to locate any fixed locative phrases. Thus, the pattern we are looking for is [*zái* noun phrase (*yǐ*/*zhī*) *shàng*/*xià*/etc. *biān*/*miàn*/etc.]. To locate this pattern, we searched for occurrences of *zái* within the left window size of 3 from all locative compounds. When mapping each compound onto the synonym group codes, some compounds may be located in more than one synonym group. We enlisted all synonym group codes before doing LDA process, because LDA topic model considering each vocabulary entry (here is

[3] Simplified / Traditional Chinese conversion tables can be retrieved on Wikipedia source code web site: http://svn.wikimedia.org/svnroot/mediawiki/trunk/phase3/includes/ZhConversion.php

synonym code) to be multinomial distribution to each specific topic (see in 2.1). While clustering each mapped locative phrase containing several translated synonym group codes into topics, each topic (or cluster) also presents a multinomial distribution. While clustering, we set the minimum data set to 5 to filter out *xià tóu* and *xī tóu*, and commanded LDA to cluster each locative phrase in to 5 topics with parameters chunk-size at 10% of dataset during 20 passes. The selected results as follows:

Table 4 Selected Results of LDA model at 5 clusters

Cluster for :北邊
#1 0.166*3-Ka + 0.122*1-Kc + 0.120*2-Kc + 0.082*1-Bn + …
#2 0.250*1-Cb + 0.144*1-/Nca + 0.139*1-Kd + 0.073*2-Cb + …
#3 0.150*2-Kb + 0.141*1-Kb + 0.097*3-Kb + 0.089*3-Di + …
#4 0.152*1-/Nb + 0.102*2-Ka + 0.069*2-Gb + 0.064*1-/Nab + …
#5 0.321*1-Di + 0.134*1-/Nc + 0.130*2-Di + 0.098*2-Jd + …
Cluster for :下面
#1 0.167*3-Ka + 0.116*2-Ka + 0.082*1-Di + 0.079*1-Hi + …
#2 0.176*1-Kd + 0.141*1-Bo + 0.141*2-Dn + 0.060*2-Kd + …
#3 0.175*1-Kc + 0.114*3-Kd + 0.107*1-Bp + 0.078*2-Kb + …
#4 0.149*2-Kc + 0.146*1-Kb + 0.089*1-/Na + 0.057*2-Ka + …
#5 0.155*3-Jb + 0.138*2-Bn + 0.105*1-Bn + 0.089*3-Kc + …

LDA creates 5 clusters according to the percentage of the existing synonym groups. Because all information, including parts of speech, were inputted into LDA, if a part of speech is considered more prominent by LDA (as in 0.152*1-/**Nb** above where the initial is marked with a "/"), the part of speech will be listed as one of the most important features of a particular cluster. The elements without the "/" represent the Chilin synonym group codes.

However, the results in Table 4 are hard to us to recognize the patterns of correlation between synonym groups and parts of speech in locative nouns. Therefore we adapted the Hive Plot (Krzywinski, *et al*, 2012) method to construct a network viewer for comparison by using each coefficient weight larger than 0.05 (Figure 2). Furthermore, we also incorporated the synonym-group-occurring frequency into the Hive Plot diagram if this occurring frequency ratio is higher than 0.005 within each pattern. As the occurring frequency represents the most prominent pattern in each locative phrase, the LDA topic coefficients are able to show the significant differences among topics with each locative prefix/suffix formation.

In our study, we used GENSIM (Řehůřek and Sojka, 2010) library to perform the LDA clustering of the essential synonym groups appearing in different formations of locative nouns, and we analyzed the coefficient within each generating groups.

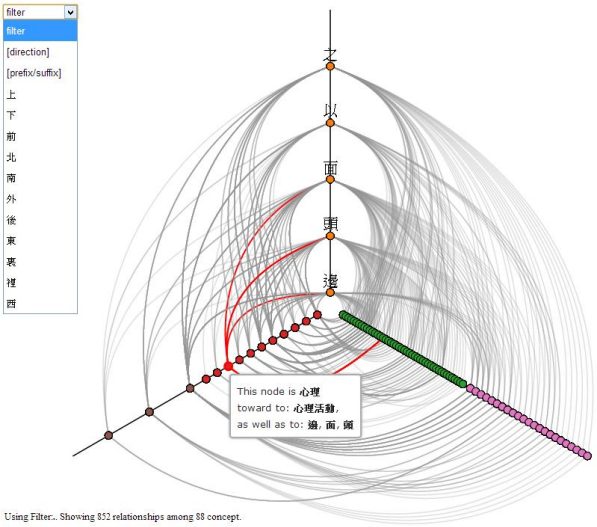


Figure 2 The Hive Plot viewer for Locative Clusters. Red Single links from *psychological activities* to *biān, tóu*, and *miàn.*

In Figure 2, we used the Hive Plot to integrate information of LDA clusters and Chilin hierarchy within an interactive diagram for 5 patterns of locative noun formations. The upper left drop-down bar shows the select function, and the bottom text provides the basic statistics of the current diagram. There are 3 axes in each diagram: the upward axis is the pattern we used in LDA (the top two nodes are prefixes and the last three are suffixes); the right-downward axis is the mapped (sub)synonym groups (green nodes)[4] or P.O.S tag (pink nodes) in the LDA topics or high occurring frequency nodes; and the left-downward axis is the upper-level synonym groups (red nodes) and the initial letter of P.O.S tags (the coarse categories of 'N', 'V', etc.) (brown nodes) corresponding to the nodes locating in right-downward axis.

Each node in the upward axis represents a single LDA clustering processing, which processes all locative nouns matching the node label (e.g., all patterns containing *biān*) , and the links (grey) represent the aggregated coefficient (weighting) of each clustered topic. While users put the mouse on any node, the over-layer will show the information about the mouse-over node including the meaning of the node and the target nodes it links to. In Figure 2, currently it shows the *psychology 心理* has three links toward to *頭tóu, 邊biān*, and *面miàn*, as well as one link toward

[4] Sub-synonym groups refer to the lower subordinate synonym groups under each synonym group in Chilin.

psychological activities 心理活動. The linkage represents a general idea of common usage (occurring frequency) which LDA clusters regard as significant difference among topics for this 3 patterns of locative noun formations.

3 Results

3.1 Prefixes/Suffixes of Locative Nouns

In order to understand the locative phrases, it is necessary to provide some of the meanings of the prefixes/suffixes in this section. Noun phrases are appearing in the fixed frame [*zái*~ (*yǐ*/*zhī*) *shàng*/*xià*/etc. *biān*/*miàn*/etc.] may occur with 5 prefixes/suffixes: *～邊 biān*, *～面 miàn*, *～頭 tóu*, *以 yǐ～*, *之 zhī～*. Each prefix/suffix word has its specific meaning, according to http://www.zdic.net/. *Biān* means "edge, margin, side, border", *miàn* means "face; surface; plane; side, dimension", *tóu* means "head; top; chief, first; boss", *yǐ* means "by means of; thereby, therefore; consider as; in order to", and *zhī* means "marks preceding phrase as modifier of following phrase; it, him/her, them; go to". When these prefixes/suffixes form compounds with locative words such as *zái*, they can be used to describe the locative information of *time/space* and *abstract entity* (Figure 3), as well as *object*, and *characteristics* (Figure 4), for example.

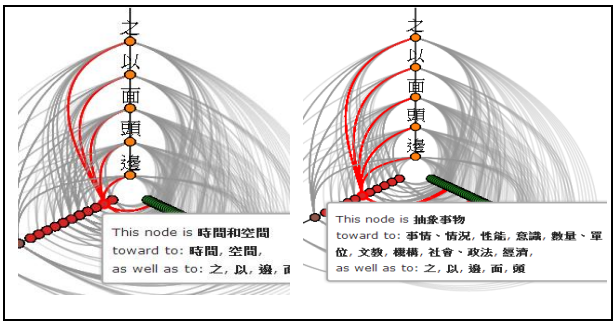


Figure 3 Prefixes/suffixes Connected to the Synonym Groups of *time/space* and *abstract entity*.

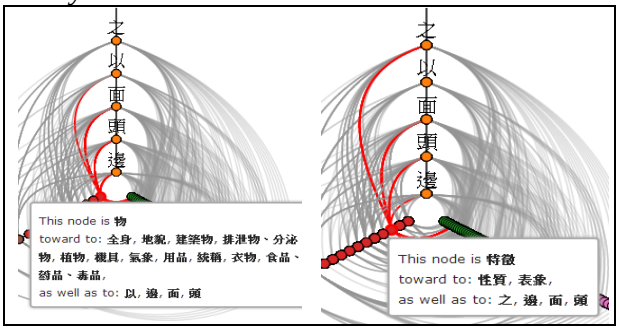


Figure 4 Connections of *object* and *characteristics* toward to prefix/suffix axis.

In Figure 3, we can see *time/space* and *abstract entity* are all connect to every prefixes and suffixes, and "toward to" illustrate the links from the node which your mouse pointing on to sub-categories, and the top-ward dimension is the nodes of prefixes and suffixes which can be compounded with locative nouns (*上 shàng*, *下 xià*, *前 qián*, *後 hòu*, *裡/裏 lǐ* *外 wài*, *東 dōng*, *西 xī*, *南 nán*, *北 běi*), such as "*上面*"(*shàng miàn*), and "*以下*"(*yǐ xià*).

When expressing location information of the synonym group *object* (in Figure 4), the data showed that nouns compounding with all prefix patterns (*biān*, *miàn*, *tóu*) and the suffix pattern of *yǐ* are significant, because these physical features like "edge", "surface", and "top" can only be found in the physical existence of objects. And the connection of *zhī* is not presenting in Figure 4, because of less frequent and not significant in LDA clusters. Similar to *object* the synonym group called *characteristics* is an upper group of *shape*, *appearance*, *color/taste* for physical objects, and *nature*, *moral*, *circumstances*, and others for abstract events. Therefore, all prefix patterns (*biān*, *miàn*, *tóu*) can be used while expressing the locative information related to physical objects. As to the prefix formation of *zhī*, it usually modifies the features of abstract events. More examples like *human* and *movement* are presenting in Figure 5.

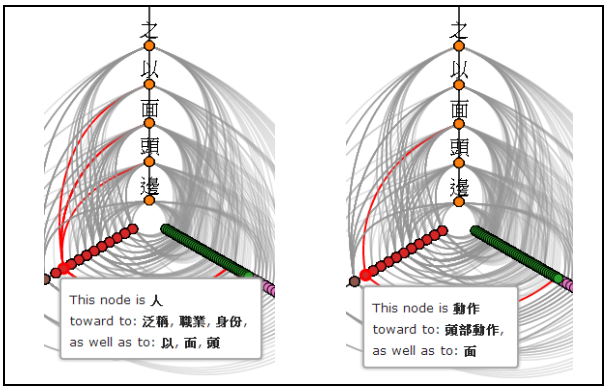


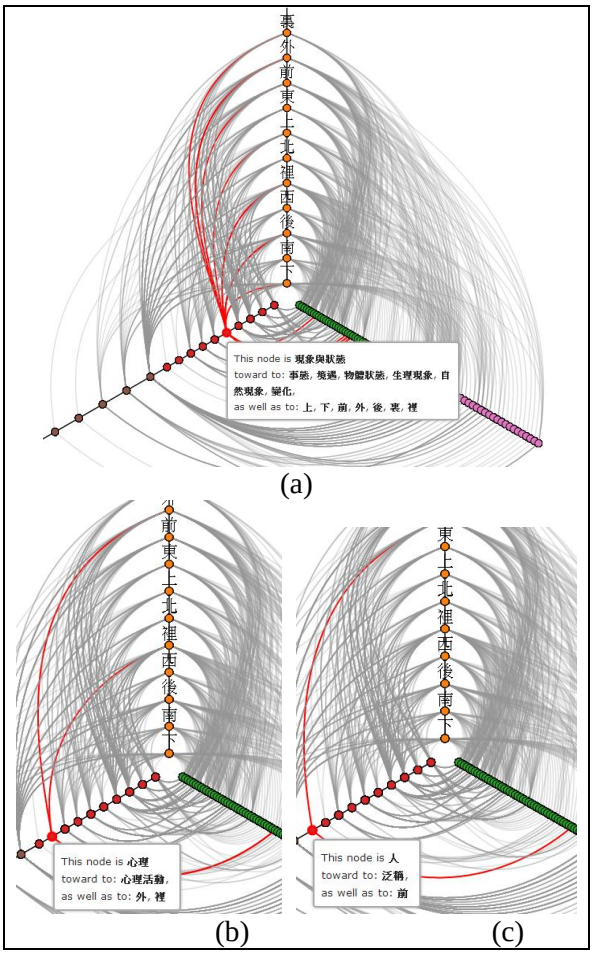
Figure 5 Connections of *human* and *movements*

When referring to directional information about *human*, linkage to sub-synonym (green nodes) groups such as *occupation* or *social positions* (*labors*, *land lord*, *loyal families*, and etc.) were seen. Data also show that only *yǐ*, *tóu*, and *miàn* are found in corpora. Considering the meaning of *tóu* and *miàn*, it is clear to find "face" and "head" senses of *human*. Another interesting finding is the linkage of *activities* to *head movements*

(*kissing*, *blinking*, *listening*, *biting* … and etc.) and to *miàn*.

3.2 Directional Words

In this section, we analyzed 11 directional words (excluding *zuǒ*, *yòu*, *nèi*, *zhōng* but including two forms of *裡/裏 lǐ*; cf. Table 1), and we also used the same setting in 3.1 (using LDA model to create 5-topic clusters and combining most frequent synonym groups to plot Hive diagram). After mapping all noun phrases onto Chilin synonym groups, we used Hive Plot for interactive investigation and we enlisted the selected results in Figure 6.



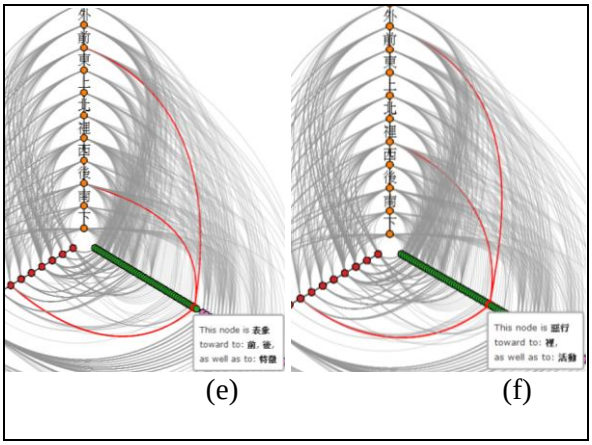


Figure 6 Results of comparing directional words. (a) focuses on *phenomenon-condition*; (b) focuses on *psychology*; (c) focuses on *human*; (e) focuses on *appearance*; (f) focuses on *sin activities*

In Figure 6, we show differences when comparing directional words in the LDA results,. We selected some interesting patterns for discussion, as follows: in Figure 6(a), *phenomenon-condition* (including sub-synonym groups) is not connected to *dōng*, *xi*, *nán*, *běi*, because it is awkward to address any direction of a *phenomenon* or *conditio*; in (b), *psychology* nouns exclusively use *裡/裏 lǐ* and *外 wài*, such as "*在 發現 裡面*" (discover something inside..) and "*在 支持 之外*" (besides supporting); in (c), *human* nouns only show the usage of *前 qián* "*在 民眾 之前*"(before citizens). In Figure 6(e), if we observed the nodes on the right-downward axis, we can find even more interesting usages of directional nouns. For example, only *前 qián* and *後 hòu*, such as "*在 明朗 之前*" (before event clear) and "*在 明朗 之後*"(after event clear), can be addressing *appearance* (sub-synonym groups are *less*, *fertile*, *bare*, *dense*, *sparse*, and etc.), as well as the only usage of *sin activities* and *lǐ*, such as "*在 治罪 條例 裡面*" (in offences ordinance), can be found in news corpora. Unfortunately, we cannot enlist all 1,509 relationships among 104 concepts in Hive Plot of all directional words. The above are just some interesting observations.

3.3 Combination of Directional Words and Patterns

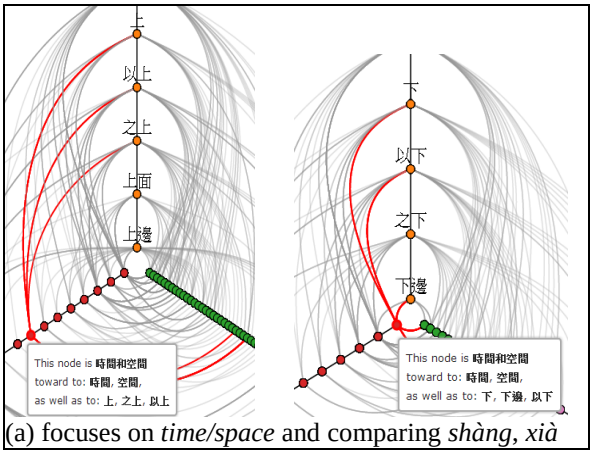
It is possible to dig in each formation of directional words and prefixes/suffixes combination. We used the same method to create

Hive Plot for each directional word and prefix/suffix formation pattern. We combined all statistical results of LDA and the occurring frequency of synonym groups using the same directional words. Similarly, in this paper, we just selected some findings for discussion.

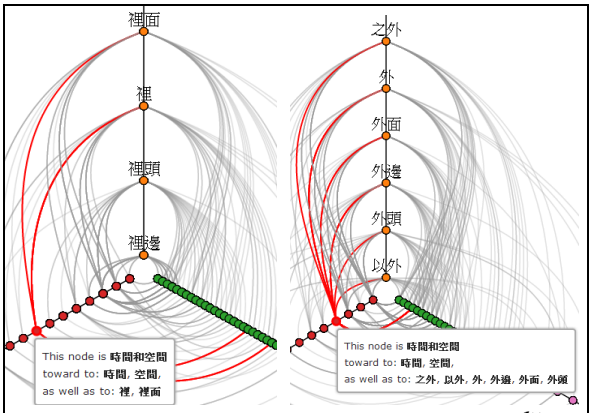
(A) *Time/space*

In Chilin synonym groups, *time/space* is the upper-level groups of *time* (its sub-synonym groups including: *A.D.*, *B.C.*, *end of year*, *four seasons*…and etc.) and *space* (its sub-synonym groups including: *position, direction, neighborhood, surrounding*, etc.). Our data showed very interesting results while comparing each opposite direction in pairs (Figure 8).

In Figure 7(a), we can see *shàng* is used in the suffix patterns (*zhī* and *yǐ)*, and *xià* is used with the prefix *biān* and suffix *yǐ*. If we consider the semantic senses of *biān* – "edge, margin, side, border", example like "*在 朝往 下邊*" (To the following), it seems that *xià biān* shows a distance closer to an observed point. On the contrary, when addressing *shàng*, data showed that most uses ignored the distance with regarding to the observation point. In addition, in (b), we can see *lǐ* and *wài* are totally different. When expressing *time/space* in *wài*, just like all other directional nouns (*qián*, *hòu*, *dōng*, *xī*, *nán*, *běi*), they are connected to every prefix/suffix pattern. As to *lǐ*, no matter its sub-synonym groups are *time* or *space*, we can find only one linkage to *miàn* which means "face; surface; plane; side, dimension", such as "*在 時間 裡面*" (during that time).



(a) focuses on *time/space* and comparing *shàng*, *xià*

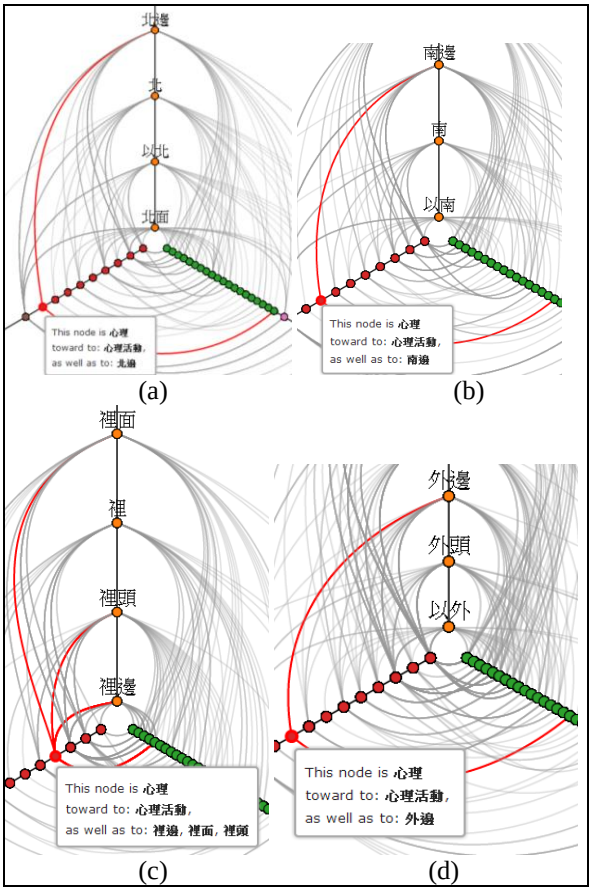


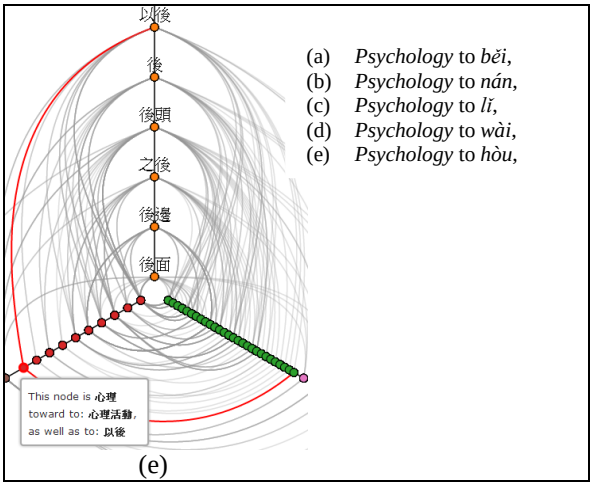
(b) focuses on *time/space* and comparing *lǐ and wài*

Figure 7 comparing *time/space* in pairs of directional words with opposite meanings

(B) *Psychology*

In Chilin, *psychology* has only two sub-synonym groups, *psychological activities and psychological status.* However, we can only found *psychological activities* is connected in collected corpora. In Figure 2, we found that nouns in *psychology* synonym groups are usually used with *biān*, *tóu*, and *miàn*, and we could only find 5 linked graphs in every locative noun and pattern combination (Figure 8).





(a) *Psychology* to *běi*,
(b) *Psychology* to *nán*,
(c) *Psychology* to *lǐ*,
(d) *Psychology* to *wài*,
(e) *Psychology* to *hòu*,

(e)

Figure 8 Connections of synonym group *psychology* in different locative nouns

First, the synonym group *psychology* is linked to *běi* (a) and *nán* (b), and the linkages are relatively lower than other linkage weighting in the same graph, such as "*在 支持 中國 北邊*" (supporting northern China) and "*在 規劃 濁水溪 以南*" (planning south of Zhuoshui River). When expressing locative information, *psychology* nouns use only suffix *biān* in our corpora. More significant results can be found in locative nouns *lǐ*, *wài*, and *hòu*. In Figure 8(c), when using locative noun *lǐ* to address directional information of synonym group *psychology*, we can only find evidences support using suffixes of *biān*, *tóu*, and *miàn*, but not with prefixes. The LDA result of no linkage to pattern node *lǐ* is different from all other graphs. In Figure 8(d), *psychology* nouns are relatively close to the observation of *object* because both use *wài biān* instead of using suffixes *tóu* and *miàn*. In Figure 8(e), locative information of *psychology* appears as *abstract entities* by using *yǐ hòu*.

(C) *Example of Generating Locative Structure*
Although the complexity of analyzing Chinese locative nouns which accompany with 5 different suffixes and prefixes, it is possible to generate locative structure for a locative nouns. We take *裡 lǐ* as an example. In Figure 7(b), while addressing concept regarding to *time/space*, the frequency of using suffix combination, *miàn*, is significant in diagram, and the usage of *miàn* only can be found in compounds in *space* category, if we look into right-downward axis (sub-synonym groups).(Figure 9)

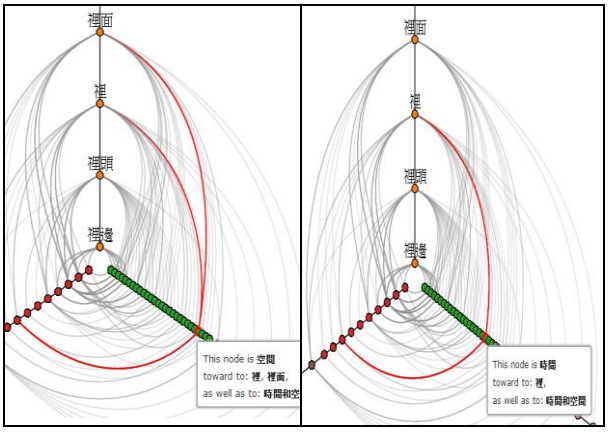


Figure 9 Connections of sub-synonym groups, *time* and *space*, to different locative nouns

The translations of using *lǐ miàn* and *lǐ* are different, because the translated senses depend on the concept before locative nouns (here, one is *time* period, and the other is *space*). For example, "*夏日*" is compounds addressing summer days and be collected in *times* category, therefore "*在 夏日 裡*" is translated into "in/during summer days". As to *space*, locative nouns like "*在 城市 裡面*" can be translated into "in/inside city", other suffixes and prefixes, such as *～邊 biān*,*～頭 tóu*, *以 yǐ～*, *之 zhī～*, are rarely found in corpus.

4 Conclusions

Locative phrases are formatted compounds which contain directional nouns and referring scope at the same time. The combinations of locative phrases are difficult for us to analyzing the formation and to establish formal rules for representation and composition for locative nouns. Our study tries to re-categorize all nouns appearing in a certain fixed frame. The semantic meaning of the nouns can be seen in our study by observing their concepts. Instead of using human judgments, we propose a novel method by using LDA model and its clustered topics parameters, as well as integrating the statistical frequency and Chinese Synonym Forest hierarchical information to inspect the differences between locative nouns and prefix/suffix formation through Hive Plot interface. In this study, we discover several findings regarding locative nouns and syntactic locative phrases using synonyms nouns. Our study is limited by the news genre of Giga-word corpus in Sketch Engine. It is possible to use different machine learning mechanisms, and to adapt interactive visual investigating method to help us understand

more relationships beyond statistical data. As Pustesjovsky (1995:26) points out that "the ways in which words carry multiple meanings can vary", by observing the nouns in a fixed frame, we can see how different, and some closely-related, locative phrases vary in their concepts.

Acknowledgements

This research is supported by National Science Council grant 101-2410-H-004-176-MY2 directed by Siaw-Fong Chung.

References

J. J. Mei, Y. M. Zhu, Y. Q. Gao, and H. X. Yin. 1983. Tongyici Cilin (Chinese Synonym Forest).

J. Pustejovsky, 1991. The generative lexicon. Computational linguistics, 17.4 (1991):409-441.

J. Pustejovsky, 1995. The Generative Lexicon: A Theory of Computational Lexical Semantics. Cambridge, MA: The MIT Press.

D. M. Blei,A. Y. Ng, and M. I. Jordan. 2003. Latent dirichlet allocation. The Journal of Machine Learning research, 3:993-1022.

C. N. Li and S. A. Thompson. 1989. Mandarin Chinese: A functional reference grammar. Univ of California Press.

Yuqi Sheng. Modern Chinese online course (現代漢語網絡課程), http://www.yyxx.sdu.edu.cn/chinese/, visited on 2013/06/01.

Ya-chen Hsu and James H.I. Tai. 2001. An analysis of the Chinese spatial term shang in three reference frames. Unpublished dissertation.

Kan-Yuan Liang and Song-Mu Wang. 2010. Study on the Image Schemata of Modern Chinese Cian: Based on Frames of Reference. Unpublished dissertation.

Chao-Mei Wang and Ching-Yu Hsieh. 2011. The Cognitive Analysis of Mandarin Locative Expressions from the Perspective of Emotion: With Reference to Shang and Xia. Unpublished dissertation.

W. Y. Ma and C. R. Huang. 2006. Uniform and effective tagging of a heterogeneous giga-word corpus. In 5th International Conference on Language Resources and Evaluation (LREC2006):24-28.

M. Krzywinski, I. Birol, S. J. Jones, and M. A. Marra. 2012. Hive plots—rational approach to visualizing networks. Briefings in Bioinformatics, 13(5):627-644.

D. Blei and J. Lafferty. 2009. Topic Models. In A. Srivastava and M. Sahami, editors, Text Mining: Classification, Clustering, and Applications. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series.

R. Řehůřek and P. Sojka. 2010. Software framework for topic modelling with large corpora. In Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks:46-50.

Flaherty, P., Giaever, G., Kumm, J., Jordan, M. I., and Arkin, A. P. 2005. A latent variable model for chemogenomic profiling. Bioinformatics, 21(15):3286-3293.

Gao, Z.M., Extraction and Integration of Chinese Lexical Semantic Information, Proceedings of the 19th Conference on Computational Linguistics and Speech Processing, ROCLING 2007, Taipei.

Appendix I: Chinese Synonym Forest (Chilin 同義詞林)

Chinese synonym group code and translated senses by the authors.

Codes	Group Name	Translated sense	Codes	Group Name	Translated sense	Codes	Group Name	Translated sense
A	人	*People*	C	時間和空間	*Time/Space*	H	活動	*Activities*
Aa	泛稱	*General term*	Ca	時間	*Time*	Ha	政治活動	*Political activities*
Ab	男女老少	*Men and Women*	Cb	空間	*Space*	Hb	軍事活動	*Military activies*
Ac	體態	*Posture*	D	抽象事物	*Abstract*	Hc	行政管理	*Administration activities*
Ad	籍屬	*Nationality*	Da	事情、情況	*Things/situation*	Hd	生產	*Production activities*
Ae	職業	*Profession*	Db	事理	*Affair*	He	經濟活動	*Economic activities*
Af	身份	*Identity*	Dc	外貌	*Appearance*	Hf	交通運輸	*Transportation*
Ag	狀況	*Status*	Dd	性能	*Performance*	Hg	教衛科研	*Education/Research activities*
Ah	親人、眷屬	*Relatives/dependents*	De	性格、才能	*Character/Talent*	Hh	文體活動	*Sports*
Ai	輩次	*Seniority*	Df	意識	*Awareness*	Hi	社交	*Social activities*
Aj	關係	*Relationship*	Dg	比喻物	*Metaphor*	Hj	生活	*Life*
Ak	品性	*Moral*	Dh	臆想物	*Imaginary*	Hk	宗教生活	*Religious*
Al	才識	*Ability*	Di	社會、政法	*Social, political and legal*	Hl	迷信活動	*Superstitious*
Am	信仰	*Faith*	Dj	經濟	*Economy*	Hm	公安、司法	*Public security, justice*
An	丑類	*Bad title*	Dk	文教	*Culture*	Hn	惡行	*Sin activities*
B	物	*Object*	Dl	疾病	*Disease*	I	現象與狀態	*Phenomenon-condition*
Ba	統稱	*General term*	Dm	機構	*Agency*	Ia	自然現象	*Natural phenomenon*
Bb	擬狀物	*Proposed substance*	Dn	數量、單位	*Quantity/Unit*	Ib	生理現象	*Physiological condition*
Bc	物體的部分	*Part-of*	E	特徵	*Feature*	Ic	表情	*Expression*
Bd	天體	*Astronomical*	Ea	外形	*Shape*	Id	物體狀態	*Object condition*
Be	地貌	*Landforms*	Eb	表像	*Table*	Ie	事態	*Situation*
Bf	氣象	*Meteorological*	Ec	顏色、味道	*Color/Taste*	If	境遇	*Circumstance*
Bg	自然物	*Natural*	Ed	性質	*Nature*	Ig	始末	*Begin and end*
Bh	植物	*Plant*	Ee	德才	*Moral*	Ih	變化	*Changes*
Bi	動物	*Animal*	Ef	境況	*Situation*	J	關聯	*Relevance*
Bj	微生物	*Microorganism*	F	動作	*Movement*	Ja	聯繫	*Contact*
Bk	全身	*Whole*	Fa	上肢動作	*Upperr limb movements*	Jb	異同	*Differences*
Bl	排泄物、分泌物	*Excretions/secretions*	Fb	下肢動作	*Lower limb movements*	Jc	配合	*Coordinate*
Bm	材料	*Material*	Fc	頭部動作	*Head movements*	Jd	存在	*Exist*
Bn	建築物	*Building*	Fd	全身動作	*Full body movements*	Je	影響	*Affect*
Bo	機具	*Machines*	G	心理活動	*Psychology*	K	助語	*auxiliaries*
Bp	用品	*Articles*	Ga	心理狀態	*Psychology status*	Ka	疏狀	*Sparse*
Bq	衣物	*Clothing*	Gb	心理活動	*Psychology activities*	Kb	仲介	*Agency*
Br	食品、藥品、毒品	*Food/medicines/drugs*	Gc	能願	*Wishes*	Kc	聯接	*Link*
						Kd	輔助	*Aid*
						Ke	呼歎	*Call*
						Kf	擬聲	*Onomatopoeia*
						L	敬語	*Honorifics*