Word Space Modeling for Measuring Semantic Specificity in Chinese

Ching-Fen Pan
Department of English
National Taiwan Normal University
debbychingxp@hotmail.com

Shu-Kai Hsieh
Department of English
National Taiwan Normal University
shukai@gmail.com

Abstract

The aim of this study is to use the word-space model to measure the semantic loads of single verbs, profile verbal lexicon acquisition, and explore the semantic information on Chinese resultative verb compounds (RVCs). A distributional model based on Academia Sinica Balanced Corpus (ASBC) with Latent Semantic Analysis (LSA) is built to investigate the semantic space variation depending on the semantic loads/specificity. The between group comparison of age-related changes in verb style is then conducted to suggest the influence of semantic space on verbal acquisition. Finally, it demonstrates how meaning exploring on RVCs is done with semantic space.

1 Introduction

The issue of 'word space' has been gaining attention in the field of distributional semantics, cognitive and computational linguistics. Various methods have been proposed to approximate words' meanings from linguistic distance. One of the most popular models in distributional semantics is Latent Semantic Analysis (LSA) with dimension-reduction technique, Singular Value Decomposition (SVD)(Landauer and Dumais, 1997; Karlgren and Sahlgren, 2001; Sahlgren, 2002; Widdows et al., 2002). The backbone of LSA is the co-occurrence distributional model in which words are conceived as points scattered in a texts-built n-dimensional space(Lenci, 2008). Rather than trying to predict the best performing model from a set of models, this study highlights the extent to which word space or semantic space measured from a vector-based model can access the verbal semantics and has influence on verbal acquisition.

This paper is organized as follows: Section 2 profiles the variation of semantic space affected by the semantic loads of single verbs. Section 3 discusses the correlation between the developing change in verbal lexicon and word space from the experimental data collected by M3[1] project. It will reveal how semantic space facilitates early child verbal learning. Section 4 demonstrates how to assess the meaning of Chinese resultative verb compounds (RVCs) from semantic space. The results of this work are finally concluded in Section 5.

2 The Variation of Semantic Space Between Two Verb Types (G/S) in LSA

The goal of this section is to examine the semantic variation between two verb types, generic versus specific verbs. It first creates a taxonomy for the classification of various verb groups (generic verbs versus specific verbs) based on the semantic distance with Latent Semantic Analysis (LSA) and Cluster Analysis.

2.1 Distributional Model Based on Sinica Corpus

The distributional model built in this survey is based on the Chinese texts collected in Academia

[1] Model and Measurement of Meaning: A Cross-lingual and Multi-disciplinary Approach of French and Mandarin Verbs based on Distance in Paradigmatic Graphs. Project website: http://140.112.147.149:81/m3/

Sinica Balanced Corpus (ASBC)[2]. It includes 190 files containing about 96000 word types[3]. The original matrix (M) is further decomposed into the product of three matrices (TSD^T). These matrices are then reduced into *k* dimensions. In the following reconstruction process based on *k* dimensions, it multiplies out the truncated matrices $T_kS_kD'_k$ and then gets a M_k matrix (the approximation of X)(Landauer et al., 1998; Sahlgren, 2005; Widdows and Ferraro, 2008). The following shows an example of finding the nearest neighbors of the word *da* (打 / to hit) via two methods (see Table 1). For the convenience of visualization and cluster analysis, Euclidean distance is applied in the following study.

	qu 'go'	*na* 'take'	*zhao* 'find'
Cosine	0.928	0.926	0.920
Distance	0.377	0.382	0.397

Table 1: Associating words of *da* 'hit'.

2.2 Semantic Clustering

The primary objective of cluster analysis is to examine the formation of a taxonomy: whether G verbs and S verbs form two groups separately. The clusters also help us grasp the semantic space among verbs as well as the potential semantic relation of them. Based on the distance matrix of lexical items generated in the last section, this part applied cluster analysis on the selected 150 verbs/observations[4]. For the convenience of comparison, each verb is coded with its type and a serial number like *zuo* (做/ to do) is G1 and *si* (撕/ to tear) is S27[5].

Once the similarity measure is done, the next procedure is to combine similar verbs into groups. The clustering procedure starts with each verb/observation in its own cluster, and combines two clusters together step by step until all the verbs are in a single cluster[6] The cluster dendrogram is plotted is Figure 1, in which clusters are formed from the bottom to the top.

Figure 1 demonstrates that the highest split separates these verbs into two big groups: the left branch group and right branch group drawn in different squares. The constituents of the two branches are listed in Table 2. It is clear that most of the constituent parts of the left group are G verbs whereas S verbs count as majority in the right group. If the left group is considered as a group formed with G verbs and right group with S verbs, the hit ratio[7] of G verbs (74.6%) is much higher than that of S verbs (57.1%). The clustering algorithm that we applied shows some structure, but there is no accurate separation of these two verb types. A detailed investigation of the relationship between the verb type and the distance is discussed in the next section.

	left group	right group
Generic verbs	59 (64.1%)	18 (33.3%)
Specific verbs	20 (21.7%)	24 (44.5%)
Undetermined verbs	13 (14.1%)	12 (22.2%)
Hit ratio	74.6%	57.1%

Table 2: Distribution of G/S verbs in two big clusters.

[2] ASBC website:
http://dbo.sinica.edu.tw/ftmsbin/kiwi1/mkiwi.sh

[3] The hapax legomena (words occur only once in the whole data) are not included in the matrix. The total word types including hapax amount to 220000 or so. To avoid time and computer consuming, we excluded those hapax from the co-occurrence matrix.

[4] These 150 verbs are single verbs selected from the experimental data. In the previous study of classification, these verbs are divided into two types (G:generic versus S:specific). There are 78 G verbs and 45 S verbs, along with 27 U(undetermined) verbs. It is noticeable that U verbs do not count as one type of verbs. They are floating verbs between G and S. We keep their identity as U and examine their potential characteristics in a binary cluster analysis.

[5] In fact, only 146 of 150 verbs are being classified because four words are missed in Sinica Corpus. To avoid confusion, we still call them 150 verbs in cluster analysis.

[6] Agglomerative method is implemented in the process in which single points are agglomerated into larger groups. This is termed a hierarchical cluster procedure that explores the co-relational structure of these single verbs. In complete linkage, all objects in a cluster are linked to each other with the longest distance. The use of the longest distance in complete linkage makes the least similar pair of objects group together. In other words, the maximum distance of the group results from the linkage of objects with minimum similarity.

[7] The hit ratio is calculated as follows:
hit ratio of G in the left group: $59/(59 + 20) = 74.6\%$
hit ratio of S in the right group: $24/(18 + 24) = 57.1\%$
It is noticeable that U verbs are temporarily ignored here.

2.3 Distance Variation in Small-G/S-clusters

Following the line of argumentation, this section demonstrates how distance varies within small-G-clusters and small-S-clusters. In order to examine the distance difference, small-G-cluster (or small-S-cluster) is defined as a cluster formed with the nearest twenty words of the G verb (or S verb) target.[8] In the example of one G verb *yong* (用/use) coded as G5, the closest twenty words are almost G verbs and the only one S verb is the farthest word *xie* (寫/write) (see Figure 2). The distance examination of the small cluster is applied to all of the 150 verbs studied in this survey. Table 3 has illustrated the comparison of verb types and the distance in the small cluster. As expected, the semantic distance is significantly affected by the verb type of the target word in the small cluster. The distances among words in most of the small-G-clusters range between 0.4 and 0.8. In contrast, over eighty percent small-S-clusters obtain a distance from 0.8 to 1.2. As for those U verbs which can not be decided as generic or specific in the manual tagging because of the lacking of agreement, they have distance between 0.6 and 1. Their distance shows an overlap with part of G verbs and part of S verbs. It confirms that U verbs are in a fuzzy zone between G verbs and S verbs.

In summary, G verbs are words with more senses and they appear more frequently in various context. Based on their high frequency distribution, G verbs construct a solid relation with each other in small-G-clusters. In contrast, S verbs are words with restricted meanings and they have relatively limited distributional patterns. Due to their low variety of patterns, S verbs are not easy to have tight relations with other words. It shows that words with generic meaning have high distribution variety and the distances among them are much shorter. The lack of polysemous feature makes the specific verbs be short of various distributional patterns and lose the opportunities to form close semantic relation with others. The semantic space among G verbs is short enough to form a solid cluster whereas S verbs are relatively remote from each other in semantic space. The distance of each verb cluster can help assess the verb category as generic (G) or specific (S). Approximately 75% of generic verbs form small clusters with distance lower than 0.8 while more than 80% of specific verbs acquire a

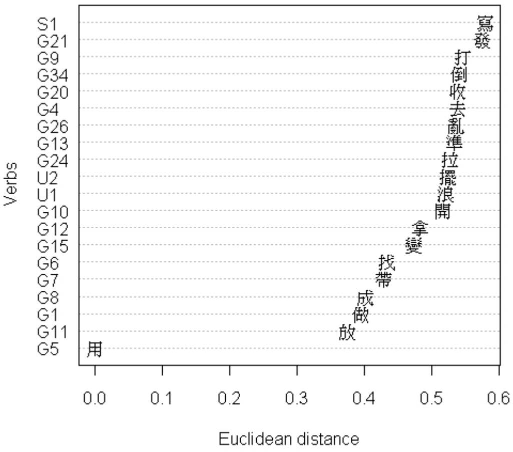


Figure 2: The small-G-cluster of *yong* (用/use).

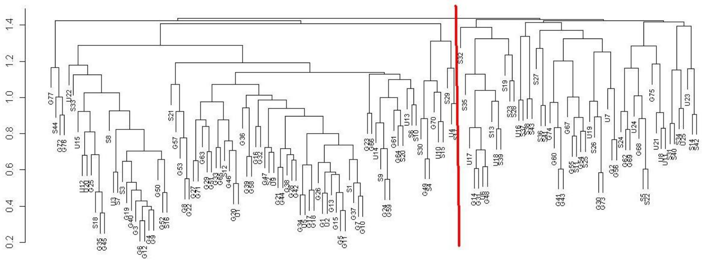


Figure 1: Agglomerative hierarchical cluster analysis of 150 verbs.

[8] In order to test the representative power of small-clusters with 20 words, we have examined the clusters with 25 and 30 words as well. In all of the cases, the curves in 20-word cluster don't change significantly when the sample size is set to 25 or 30. The small-G/S-clusters with the sample size (N=20) is justified as representative.

distance greater than 0.8 . As to the verbs of indeterminacy, they are averagely scattered in a fuzzy zone between G and S verbs. Over 70% U verbs are centering the distance 0.8, which suggests that words near distance 0.8 are likely to be undetermined verbs. This analysis has proved that semantic space varies in accordance with verb's meaning specificity. The distributions in context represent not only the linguistic behaviors but the semantic contents of lexical items.

3 The Influence of Specificity on Acquisition

This section assesses the influence of semantic space on the acquisition of the verbal lexicon. With the examination of Specific verb (S verb) progress, this study proposes that Generic verbs (G verbs) are acquired earlier than S verbs due to the closer semantic space. It also testifies whether the S verb development is a developing trend parallel with the acquisition of conventional verbs(Chen et al., 2008; Hsieh et al., 2009)[9] from the experimental data collected by M3 project. Based on the developing trend of conventional lexical items, the following parts analyze the relation of meaning specificity and the acquisition of lexical items.

3.1 Decreasing in Lexical Variation

The section is concerned with lexical variation among participants within the same age group. It measures type-token ratios of each group and profiles the lexical variation[10] in verbal acquisition. Data analyzed in this part include five groups of respondents' usages of verbs to four different films, each of which pictures one event. Respondents are assigned into five groups according to their age: 3-year-old, 5-year-old, 7-year-old, and 9-year-old groups have 20 respondents separately while 60 respondents are in the Adult group composed of people in their twenties. In respondents' answers, only one single verb is extracted from each respondent in this study. The number of verbs in each group is equal to the amount of participants. The first analysis begins with the lexical variation or lexical flexibility in these five groups. It is done with the ratio of lexical variation: the amount of word type is divided by the amount of word token, as shown in Table 4. The greater number of the ratio means the lexical variation is more abundant and the smaller ratio means a low diversity of word types. The ratio of lexical variation in these four films all show a decreasing trend from 3-year-old groups to adult groups. The quantity of different verbs is higher in children group (3y, 5y,7y, 9y) than that in adult group. That is, children appear more creative in event description tasks while adults are confined in the conventional usage. With the decreasing trend of lexical variety, the next step is to propose an increasing trend of specific verb

[9] They rearranged the five groups of participants into three units and then investigated the learning trend by Replacing Rate (Frequency of $V2_{freq}$ / Frequency of $V1_{freq}$). By defining adults' usages as the conventional one called V1, children's second highest frequency verb is counted as V2. Along with the increase of age, the number of V2 drops slowly whereas the amount of V1 increases gradually.

[10] Lexical diversity or sometimes called lexical variation is used to mean a combination of lexical variation and lexical sophistication. It is also referred to an indication of a combination of vocabulary size and the ability to use it effectively(Malvern et al., 2004). However, lexical variation or lexical diversity doesn't mean lexical richness in this study. In other kinds of experiment like writing tests, adults should perform better than children in lexical diversity. But the experimental data applied in this study is action-naming task. The trend of lexical variation may perform in an opposite way.

Distance	0.4-0.6	0.6-0.8	0.8-1.0	1.0-1.2
Small-G-cluster	24 (31.2%)	32 (41.6%)	17 (22.0%)	4 (5.2%)
		Total:72.8%		Total:27.2%
Small-S-cluster	0 (0)	6 (13.6%)	19 (43.2%)	19 (43.2%)
		Total:13.6%		Total:86.4%
Small-U-cluster	1 (4%)	8 (32%)	11 (44%)	5 (20%)

Table 3: Comparison of verb types (G/S) and semantic distance within small cluster.

usage when the age raises. It will show that the change is from various generic verbs to one or two specific verbs rather than various specific verbs.

Films	carrot-peel	paper-crumple	plank-saw	glass-break
3y	0.35	0.55	0.2	0.33
5y	0.25	0.47	0.2	0.2
7y	0.3	0.2	0.25	0.1
9y	0.21	0.105	0.157	0.157
Adult	0.016	0.083	0.066	0.066

Table 4: The ratio of lexical variation ($ratio =$ word type/word token).

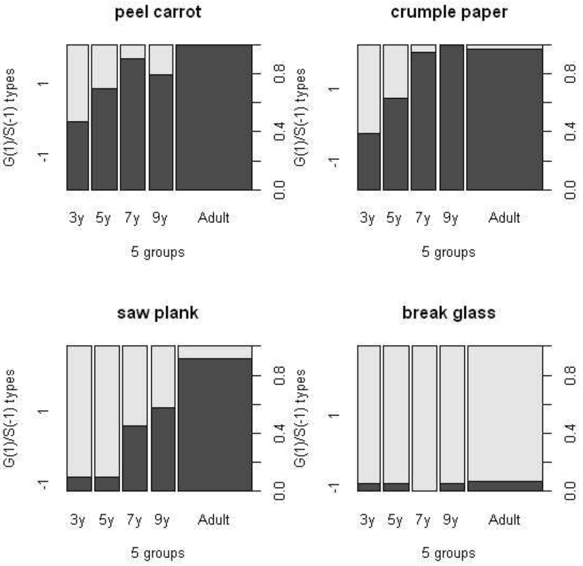


Figure 3: The proportion of S (-1) verbs to G (1) verbs from 5 groups of respondents to four events.

3.2 Increasing in Specific Verbs

With regard to the aim of the investigation, the findings reported above provide evidence of the changing trend of lexical variety in action-naming tasks. The next step is to discover the developing trend of verb type (G/S) usage. According to the annotation result of verb category, each verb in the data is now transferred into either generic (label as G or 1) or specific (S or -1) and the proportions of S verbs is plotted as Figure 3.

3.2.1 The Non-proportionality of S Verb among Age Groups

A closer investigation is then implemented for non-proportionalities by chi-squared test(Baayen, 2008). Although the proportion of S verb changes more or less in different groups, it is still need to confirm that whether S verbs are more frequently used by adults than children. The hypothesis is formulated as follows:

H_0: The proportions of the two verb types (G verb vs. S verb) do NOT vary in five age groups.

With Pearson's chi-square test for four sets of data. It is reported that the small p-values (9.779e-07, 1.324e-09, and 1.191e-13) in the first three sets of data (carrot-peel (f_6), paper-crumple (f_2), and plank-saw (f_16)) suggest a non-proportionality of S verb in different age groups. However, the p-value (0.8467) obtained in the last data set (glass-break (f_3)) is too large to suggest a significant variation of S verb proportion in different age groups. It proves that the proportions of S verb change with the participant's age in the three event-naming tasks but that doesn't happen in the glass-break (f_3) event. Except for the data in glass-break (f_3) event, the null hypothesis doesn't hold in the analysis.

3.2.2 The Relationship between S Verb and Age

In order to test the correlation of S verb proportion and age variation, four groups (3y, 5y, 7y, 9y) are merged into one group called Child versus Adult group. The data are now represented by two by two contingency tables with one categorical dependent variable (verb types) and one categorical independent variable (age). Here summarizes the hypothesis:

H_0: The frequency of the two verb types (G verb vs. S verb, the dependent variable) do NOT vary depending on participants' age (Child vs. Adult, the independent variable).

The result has shown that the small p-values (2.803e-05, 0.001225, 1.754e-12) verify the significant difference of S verb in Child group and Adult group with regard to the three data sets in carrot-peel (f_6), paper-crumple (f_2), and plank-saw (f_16). Along with the correlation examination, the effect size is revealed with correlation coefficient from 0 (no correlation) to 1 (perfect correlation)(Gries, 2009). According to the Phi value in this table, only the data in plank-saw (f_16) has a correlation coefficient (0.612) greater than 0.5. That is, the correlation between S verb usage and age group is considered as significantly correlated in the one data set (plank-saw (f_16)). As for the other two data sets (carrot-peel (f_6) with phi:0.379, paper-crumple (f_2) with phi: 0.297), the correlation is not particularly strong but it is still highly significant. Over half of the data sets exhibit a significant non-proportionality of S verb usage in different age groups but the correlation of S verb and participants' age requires.

In relation to the aim of this study, it has shown that meaning specificity functions as a factor in the development of verbal lexicon. The results of the analysis also show a significant variety of S verb between children and adults. It is plausible to suppose that verbs with specific meaning are acquired later than those with generic meanings. This developing trend suggests that a closer semantic space among G verbs facilitates the acquisition of verb meanings whereas a distant space among S verbs causes difficulties in meaning acquiring. Once those verbs with specific meanings are picked up, most of them will become the so-called conventional verbs. When the conventional use to an action is a specific verb, the progress of S verb usage is more obvious. The usage of verbs with specificity meaning is a developing trend of language acquisition.

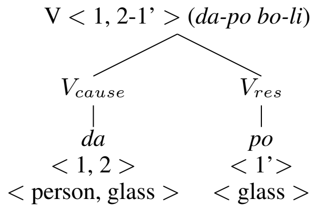
4 Meaning Exploring on Chinese Resultative Verb Compounds (RVCs)

In the verb-event co-occurrence matrix, verbs elicited from the same event are considered to be verbs have the same object in a verb-object co-occurrence matrix. With the distributional model, it then shows how meaning specificity affects the linguistic behavior and semantic content of Chinese resultative verb compounds (RVCs). Those RVCs with similar distributional patterns will present a high semantic relation. This semantic relation could result from the meaning of the first verbal morpheme (V_{caus}) or the second one (V_{res}). It is further proposed that the verb type (generic or specific) of V_{caus} would affect the whole meaning content of V-V compounds.

4.1 The RVC Structure in the Data

A Chinese resultative verb compound (RVC) consists two main elements: the first element (V_{caus}) expresses a causing event or a state while the second element (V_{res}) denotes a resulting event or the aspectual properties of the object. According to the Aspectual Interface Hypothesis(Tenny, 1989), the property of an internal argument can measure out the event. In the Chinese example, *da-po bo-li* (打破玻璃 / hit-break glass), the state of the object *bo-li* (玻璃 / glass) is changed into smashed and this change points out an end point of the event. The resultative *po* (破 / broken) is an delimiting expression which refers to the property of the object. In addition to defining the second element of an RVC as a delimiting expression, other surveys label it as V_{res} which requires the saturation of arguments. Four possible V-V compound argument structures are proposed in Li's (1990) works. In the following studies, most of RVCs require an argument structure like (1). The first verbal morpheme (V_{caus}) has a theta-grid <1, 2> and the second morpheme (V_{res}) has <1'>. V_{caus} requires an external argument (a person) and an internal argument (a glass). The internal argument (a glass) is identified with the argument of V_{res}. Since the internal argument of V_{caus} has to be identified with the argument of V_{res}, it raises the issue that which one functions more prominent in choosing the object of a V-V compound. From the study of RVCs' distributional pattern, it examines which one (V_{caus} or V_{res}) is more salient and also dominates the argument selection of a V-V compound.

(1)



4.2 Semantic Assessment

The semantic links among words are built by measuring the linguistic distances among them. In order to examine the semantic information of RVCs, a sub-sample with thirty-six verbs is selected to do cluster tasks. The semantic relationships of word in the sub-sample is visualized as a clustering tree, as shown in Figure 4. The figure shows that an RVC with a G verb as its V_{caus} ($GV_{caus} - V_{res}$) build a close relation with other RVCs which have the same V_{res} with it. Take the most extreme G verb *da* (打/hit)as an example, *da-lan* (打爛/hit-ruin) is closer to *pai-lan* (拍爛/hit with palm and ruin) than *da* (打/hit). On the other hand, an RVC with an S verb as its V_{caus} ($SV_{caus} - V_{res}$), are grouped with those having the same V_{caus}. The RVC, *ju-kai* (鋸開/saw-open), with a S verb *ju* (鋸/saw) as its head, forms a cluster with *ju* (鋸/saw) and *ju-duan* (鋸斷/saw-crack).

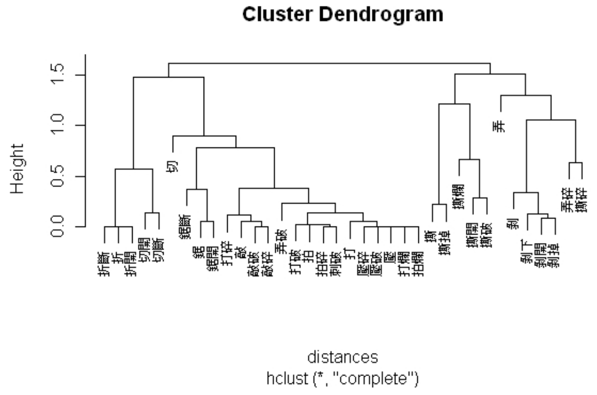


Figure 4: Semantic clustering of selected verbs.

With regard to the semantic relation of RVCs shown in the cluster plot, the next step is to justify the proportion of RVCs with the structure $GV_{caus} - V_{res}$ in which V_{res} selects a G verb as its V_{cause}. As Table 5 shows, the proportion of $GV_{caus} - V_{res}$ and $SV_{caus} - V_{res}$ is 50% respectively. That is, half of the selected seven V_{res} pick up a G verbs as its head while the other half words go with S verbs. Those V_{res} preferring a G head to a S head are *sui*, *po*, *lan*, *duan*; those preferring a S verb to a G verb head are *kai*, *diao*, *xia*. According to the semantic content these resultative verbs, *kai*, *diao*, *xia* describes the direction of the action and the motion of objects and they are defined as 'path' V_{res} in Ma and Lu's (1997) work. As for *sui*, *po*, *lan*, *duan* called as 'result' V_{res}, they mainly express the result of the object affected by the action. The outcome reported here suggests that 'result' V_{res} is apt to have a G verb as its head verb whereas 'path' V_{res} tends to pick up a S head verb. The proposal in literatures that V_{res} tends to choose a G head verb is justified as valid when the V_{res} expresses the meaning of 'result' rather than 'path.'

	GV_{caus}	**S**V_{caus}
'result' V_{res}		
sui (碎/smash)	da, nong, pai, ya, qiao	si
po (破/break)	da, nong, ya, qiao	si, ci
lan (爛/ruin)	da, pai	si
duan (斷/crack)	qie	
Proportion	47%	15%
'path' V_{res}		
kai (開/open)	qie	zhe, ju, si, bo
diao (掉/fall)		zhe, ju, si, bo
xia (下/down)		bo
Proportion	3%	35%

Table 5: $GV_{caus} - V_{res}$ versus $SV_{caus} - V_{res}$.

In summary, words with small distance resulting from their similar distributional patterns can be interpreted to be semantically similar in a semantic cluster. The result of semantic clustering has suggested that the meaning of RVCs depend on either the V_{caus} or the V_{res}. The meaning of $GV_{caus} - V_{res}$ is more determined by V_{res} because GV_{caus} is more polysemous and the V_{res} becomes a prominent role to dominate the meaning of $GV_{caus} - V_{res}$. In contrast, $SV_{caus} - V_{res}$ focuses on the part of SV_{caus} since

SV_{caus} expresses its meaning specific enough. In addition, the property of V_{res} also affects the category of its head verb. When V_{res} like *sui* belong to the 'result' V_{res}, it tend to choose a G verb as its V_{caus}. On the other hand, the 'path' V_{res} like *xia*, its head verb is apt to be a S verb. It is suggested that 'path' V_{res} is more likely to have a G verb than 'path' V_{res}. As the empirical study illustrates the semantic information on Chinese RVCs are affected by the semantic space of words.

5 Conclusion

In this paper, we argue the following points: firstly, the distributional model shows that the semantic space differ clearly in accordance with the specificity of verbs. The G verbs form tight relations with each other and become a larger cluster whereas the semantic space among S verbs is too distant to become a solid group. Secondly, semantic space has influence on the acquiring of words' meanings. Generic verbs are earlier and easier acquired due to the closer semantic space among words. The developing trend of specific verb lexicon parallel with conventional usage suggests a language acquisition phenomenon. Finally, the G/S verbs play an influential role in Chinese resultative compounds. The resultative verb becomes more prominent when the first verb is with a generic meaning. The 'result' V_{res} is apt to have a G verb as its head verb whereas 'path' V_{res} tends to pick up a S head verb. We believe that results of our analysis will shed light on semantic assessment and make predictions for lexical acquisition.

References

Baayen, R. H. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics using R.* Cambridge University Press.

Chen, P., M.-A. Parente, K. Duvignau, L. Tonietto, and B. Gaume. 2008. Semantic approximations in the early verbal lexicon acquisition of chinese: Flexibility against error. *The 7th Workshop on Chinese Lexical Semantics.*

Gries, Stefan Thomas. 2009. *Quantitative Corpus Linguistics with R: A Practical Intriduction.* Routledge.

Hsieh, Shu-Kai, Chun-Han Chang, Ivy Kuo, Hintat Cheung, Chu-Ren Huang, and Bruno Gaume. 2009. Bridging the gap between graph modeling and developmental psycholinguistics: An experiment on measuring lexical proximity in chinese semantic space. Presented at The 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC 23). Hong Kong: City University of Hong Kong., December 3-5.

Karlgren, J. and M. Sahlgren. 2001. From words to understanding. In Uesaka, Y., Kanerva P. and H. Asoh, editors, *Foundations of Real-World Intelligence*, pages 294–308.

Landauer, T. K. and S. T. Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240.

Landauer, T. K., P. W. Foltz, and D. Laham. 1998. Introduction to latent semantic analysis. *Discourse Processes*, 25:259–284.

Lenci, A. 2008. Distributional semantics in linguistic and cognitive research. *From context to meaning: Distributional models of the lexicon in linguistics and cognitive science, special issue of the Italian Journal of Linguistics*, 20/1:1–31.

Li, Yafei. 1990. On v-v compounds in chinese. *Natural Language and Linguistic Theory*, 8:177–207.

Ma, Zhen and Jian-Ming Lu. 1997. Xingrongci zuo jieguobuyu qingkuang kaocha yi (形容詞作結果補語情況考察(一)). *Hanyuxuexi (漢語學習)*, 1:3–7.

Malvern, David D., Brian J. Richards, Ngono Chipere, and Pilar Duran. 2004. *Lexical diversity and language development : quantification and assessment.* New York : Palgrave Macmillan.

Sahlgren, M. 2002. Random indexing of linguistic units for vector-based semantic analysis. *ERCIM News*, 50.

Sahlgren, Magnus. 2005. An introduction to random indexing. In *Proceedings of the Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering (TKE).* Copenhagen, Denmark.

Tenny, Carol. 1989. The aspectual interface hypothesis. In *Proceedings of NELS 18.* University of Massachusetts at Amherst.

Widdows, Dominic and Kathleen Ferraro. 2008. Semantic vectors: a scalable open source package and online technology management application. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard Joseph Mariani Jan Odjik Stelios Piperidis Daniel Tapias, editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Widdows, Dominic, Scott Cederberg, and Beate Dorow. 2002. Visualisation techniques for analysing meaning. In *Fifth International Conference on Text, Speech and Dialogue (TSD 5)*, pages 107–115. Brno, Czech Republic.