Exploring Chinese Verbal Lexicon Developmental Trend with Semantic Space

Ching-Fen Pan[a]

Department of English, National Taiwan Normal University,
No. 162, Sec. 1, HePing East Road, Taipei, Taiwan R.O.C. 106
debbychingxp@hotmail.com

Abstract. This study assesses the influence of semantic space on the acquisition of verbal lexicon. The studied one hundred and fifty action verbs extracted from the experimental data in M3 project are classified into clusters in terms of meaning specificity. The semantic space variation of different clusters is examined in the distributional model based on Academia Sinica Balanced Corpus (ASBC) with Latent Semantic Analysis (LSA). With semantic distance measured in the distributional model, this survey captures the homogeneity of verbs within the cluster and reveals the heterogeneity between the clusters. We compare the semantic space variation to the age-related changes in verb style, and explore the potential influence of word space on verbal lexicon acquisition.

Keywords: Mandarin verbs, semantic space, latent semantic analysis

1 Introduction and Motivation

The investigations of the semantic flexibility (Duvignau et al., 2005) in verbal lexicon between children and adults in the M3[1] project motivate and facilitate this study. The purpose of M3 project is to show the importance of semantic verbal approximations in both French and Chinese mandarin early lexicon acquisition. According to Duvignau's (2005) work, the nature of flexibility can be shown in the use of words and their representation of meanings. This is strongly supported by the evidence of metaphorical utterances, especially nominal metaphorical expressions. In additional to concerning the nominal lexicon, Lordat research laboratory at the University of Toulouse, in France, has shed light on the metaphorical utterances based on verbal lexicon. It has been pointed out that children will try to make an analogy to a previous event and apply the learned verb to the current event because of the lack of a conventional verb to describe the current event. Such usage should be rather considered in terms of semantic approximations, an attempt to utter with the conventional verb due to the semantic cognitive flexibility. From this kind of utterances expressed by children, one can test the semantic flexibility of each verb and semantic proximity among various verbs.

The studied 150 verbs[2] in this paper are single verbs selected from the experimental data of M3. In the experimental procedure, all of the subjects were asked to describe orally each of the action films. In each film, there appearing a woman who picks up an object and performs specific actions. The scenery remains the same but object and action change each time. Verbal utterances

[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/

[2] In the previous study of classification, these 150 verbs are manually tagged as G or S (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.

in the data collect from M3 project will be scrutinized according to the two criteria—specificity and conventionality. This paper sticks to the issue of specificity. The survey begins with modeling the semantic space of physical activity verbs, restricted to single verbs. The between group comparison of age-related changes in verb style is then conducted to suggest the influence of semantic space on verbal acquisition.

This paper is organized into the follow sections: Section 1 introduces the background and motivation of the study. Section 2 presents the distributional model based on Academia Sinica Balanced Corpus. It reveals how meaning specificity (semantic loading) affects the semantic distance of individual verbs and verb clusters. Section 3 provides evidence of the changing trend of lexical variety in action-naming tasks and discovers the developing trend of verb type (G/S) usage. The relationship between word space and age-related changes in verb style is then revealed. The results of this work are finally concluded in Section 4.

2 Semantic Space Modeling

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. The common used technique for measuring out semantic distance in a distributional model is LSA. The notion of distributional hypothesis is that the semantic similarity of two lexical items is derivable from the similarity of their distributional patterns (Lenci, 2008). With semantic distance as similarity measure, cluster analysis groups similar verbs together. This is to capture the homogeneity of verbs within the cluster and reveal the heterogeneity between the clusters (Hair et al., 1998). The following parts first introduce the construction of the distributional model via the statistical package R and then explain how 'meaning specificity' (semantic loading) affects the semantic distance of individual verbs and verb clusters.

The distributional model built in this survey is based on the texts in Academia Sinica Balanced Corpus (ASBC)[3]. The mathematical tool applied is Latent Semantic Analysis (LSA) with a well-known linear algebra, Singular Value Decomposition (SVD) (Landauer and Dumais, 1997; Karlgren and Sahlgren, 2001; Sahlgren, 2006; Widdows and Ferraro, 2008). Measuring the word/row vectors in a geometric space is to approximate the semantic space between words. The shorter the distance is, the closer the meaning could be. The following shows an example of finding the nearest neighbors of the word *da* (打 / to hit) via two methods (see Table 1).

Table 1: Associating words of *da* (打 / hit).

	qu (去/go)	*na* (拿/take)	*zhao* (找/find)	*chi* (吃/eat)
Cosine	0.9287147	0.9269788	0.9209483	0.9130709
Distance	0.3775852	0.3821550	0.3976221	0.4169630

Following the line of argumentation, this section then 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 Generic (G) verb (or Specific (S) verb) target[4]. In the example of one G verb *yong* (用/use) coded as G5, the

[3] ASBC website: http://dbo.sinica.edu.tw/ftmsbin/kiwi1/mkiwi.sh. It includes 190 files containing about 96000 word types. 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] 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 most 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.

closest twenty words are almost G verbs and the only one S verb is the farthest word *xie* (寫/write) (see Figure 1). In addition, the distance among these G verbs are between 0.4 and 0.6. In contrast, the nearest words of the S verb *qia* (掐/pinch) include S verbs and G verbs with distance over 0.8 (see Figure 2). The distance examination of the small cluster is applied to all of the 150 verbs studied in this survey. Table 2 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.

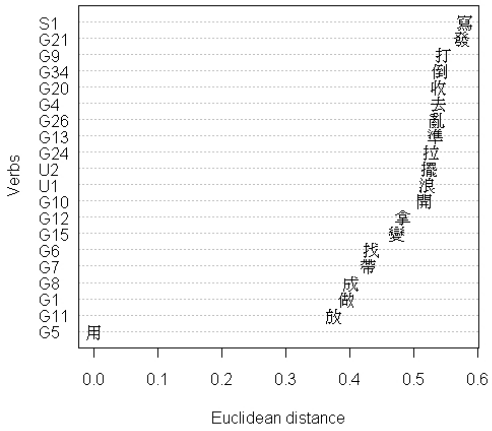

Figure 1: The small-G-cluster of *yong* (用/use).

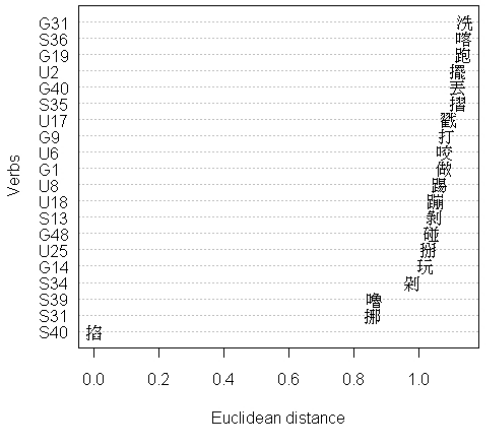

Figure 2: The small-S-cluster of *qia* (掐/pinch).

Table 2: Comparison of verb types (G/S) and semantic distance within small cluster.

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%)

The semantic relation depends on the words' meanings so that words with more meanings are apt to build connections with other words. That is, 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.

In sum, the distance among different verb types has shown a great variation. 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 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. It further confirms that the distributional semantics is semantics at all. The distributions in context represent not only the linguistic behaviors but the semantic contents of lexical items.

3 The Influence of Semantic Space on Verbal Acquisition

As noted above, the semantic space of verbs varies along with the meaning specificity of the word. Words with low semantic specificity form a closer semantic space while high specificity causes distant space. With the examination of Specific verb (S verb) progress, it is proposed 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's (2008) paper stated that children describe events with non-conventional lexical items initially. They can only learn the typical usage when they grow up. Instead of picking up a verbal lexicon most adults use, children appear to be more creative in action naming task. Other surveys of the data have shown that there is a developing trend children learn adult conventional verb in naming these four action events: carrot-peel, paper-crumple, plank-saw, and glass-break in Hsieh's (2009) paper.[5] Based on the developing trend of conventional lexical items, the following parts analyze the relation of meaning specificity and the acquisition of lexical items.

[5] 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.

3.1 Lexical Variation Decreasing

The concern here is with lexical variation among participants within the same age group. It measures type-token ratios of each group and profiles the lexical variation[6] 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. Table 3 gives a general picture of the data structure. 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 variety of verb types among participants in each group is transferred into plot in Figure 3. It is clear that 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 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.

Table 3: A general picture of single verbs elicited in five groups of respondents to four films.

Films	carrot_peel	paper_crumple	plank_saw	glass_break
3y	qie, xiao, nong, na...	nong, si, chai...	qie, nong, bo...	qiao, da, nong...
5y	xiao, qie, bo...	rou, bao, nie...	qie, ju...	qiao, da, chui...
7y	xiao, bo, ge...	rou, nie, zhe...	qie, ge...	qiao, da, chui...
9y	xiao, qie...	rou, nie, bao...	qie, ju...	qiao, da...
Adult	xiao	rou, nie...	ju, qie...	qiao, da...

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

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

3.2 Specific Verb Increasing

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 develop-

[6] 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.

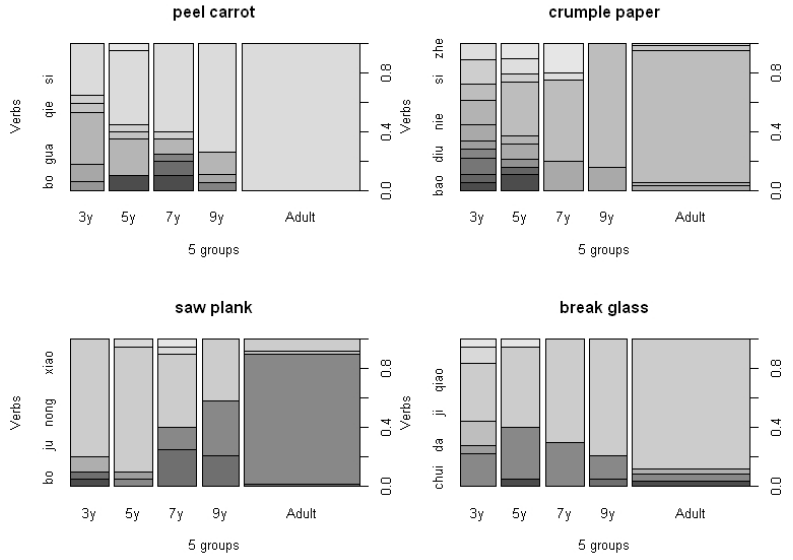

Figure 3: The lexical variety of verbs in four films.

ing trend of verb type (G/S) usage. According to the algorithm of measuring semantic specificity of verbs, each verb in the data is now transferred into either generic (label as G or 1) or specific (S or -1), as shown in Figure 4. In examine the change of S verb proportion, Table 5 lists the exact number of proportions of S verbs. The proportions are calculated by dividing the number of S verbs by the sum of S and G verbs ($prop = S/(S+G)$) in each group. At first sight, there seems to be an increasing trend of S verb usage in the three events (carrot-peel, paper-crumple and plank-saw). In these three events, the most frequently used verbs in adult group are *xiao* (削/peel with knife), *rou* (揉/crumple) and *ju* (鋸/saw) and all of them are classified as specific verbs in the previous studies. However, the highest frequency verb *qiao* (敲/knock) in the glass-break event is coded as generic and the proportion of S verb in each group in this film are relatively lower than that of other three films.

Table 5: The proportion ($prop = S/(S+G)$) of S verb in each group in four different events.

Films	carrot-peel	paper-crumple	plank-saw	glass-break
3y	0.47	0.38	0.10	0.055
5y	0.70	0.63	0.10	0.050
7y	0.90	0.95	0.45	0.000
9y	0.78	1.00	0.57	0.052
Adult	1.00	0.96	0.91	0.066

3.2.1 The Non-proportionality of Specific 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 hypotheses are formulated as follows:

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

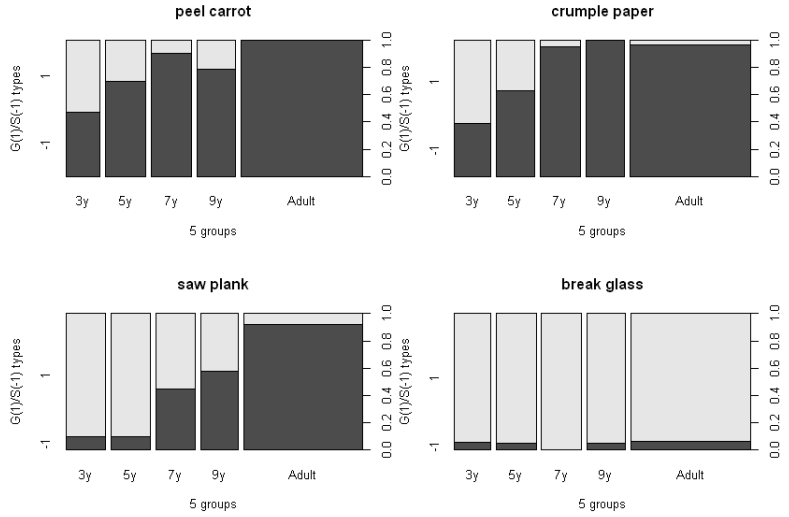

Figure 4: The proportion of S (-1) verbs to G (1) verbs from 5 groups of respondents to four events.

H_1: The proportions of the two verb types (G verb vs. S verb) do vary in five age groups.

With Pearson's chi-square test for four sets of data, the results are shown in Table 6. 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.

Table 6: The proportionality test of S verb in five age groups for four data sets.

carrot-peel (f_6): x-squared = 33.4243, df = 4, p-value = 9.779e-07
paper-crumple (f_2): x-squared = 47.2945, df = 4, p-value = 1.324e-09
plank-saw (f_16): x-squared = 66.5874, df = 4, p-value = 1.191e-13
glass-break (f_3): x-squared = 1.3858, df = 4, p-value = 0.8467

3.2.2 The Relationship between Specific 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 summarize the hypotheses:

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).
H_1: The frequency of the two verb types (G verb vs. S verb, the dependent variable) do vary depending on participants' age (Child vs. Adult, the independent variable).

As the results shown in Table 7, 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.

Table 7: The correlation test of verb types (S vs. G) and age groups (Child vs. Adult) for four data sets.

carrot-peel (f_6): x-squared = 17.5473, df = 1, p-value = 2.803e-05, Phi:0.379
paper-crumple (f_2): x-squared = 10.4528, df = 1, p-value = 0.001225, Phi:0.297
plank-saw (f_16): x-squared = 49.7413, df = 1, p-value = 1.754e-12, Phi:0.612
glass-break (f_3): x-squared = 0.1154, df = 1, p-value = 0.7341, Phi:0.062

4 Conclusion

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 experimental data focused on four different events demonstrate the change of language variety and verb type choices. The lexical variation among participants decreases but the amount of specific verbs increases when the age of participants increases. In compared with children, adults are more consistent with verb choices in the same action-naming task. Those verbs used by adults are almost verbs with a specific meaning in the study. In contrast, verbs produced by children are more inconsistent and most of the verbs are with generic meanings. 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.

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*.

Duvignau, K., M. Fossard, B. Gaume, and M.-A Pimenta. 2005. From early lexical acquisitions to the 'disacquisition' of verbal lexicon: Verbal metaphor as semantic approximation. In *Proceedings of the 2nd Conference on the metaphor in language and thought*. Unversidade Federal Fluminense.

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

Hair, J.F. Jr., R.E. Anderson, R.L. Tatham, and W.C. Black. 1998. *Multivariate Data Analysis*. Englewood Cliffs, NJ : Prentice-Hal, 5th edition.

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.

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.

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. 2006. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Ph.d. dissertation, Department of Linguistics, Stockholm University.

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).