Mutual information and corpus-based approaches to English reduplication

Shih-ping Wang
Ming Chuan University and University of Nottingham
spwang@mcu.edu.tw

The aim of this project is to use corpus-based approaches to investigate reduplicative fixed expressions in English, e.g., *sooner or later*, *first and foremost*, *part and parcel*, etc. The probabilistic relations between two adjacent words were examined. A reduplication corpus has been constructed and the frequency of each token was calculated based on its occurrence in the British National Corpus (BNC). Then a word list with 116 items was proposed for related research in terms of SARA software, the built-in tool of BNC. Frequency is often considered as the main factor to decide the word order in a conjoined phrase (Fenk-Oczlon, 1989, 2001), but the frequency-based argument is shaky and not always reliable. Mutual information (MI) was therefore employed to calculate the probability of collocation and assess collocational significance. MI can be used to decide what to look for in a concordance (Church and Hanks, 1990). The higher the mutual information, the more genuine the association between two words.

Keywords: BNC, corpus, reduplication, frequency, mutual information, and z-score

1. INTRODUCTION

The aim of this project, grant-funded by the National Science Council (NSC) is to use corpus-based approaches to investigate reduplicative fixed expressions in English, e.g., *sooner or later*, *first and foremost*, *part and parcel*, etc. Reduplication is important in language studies. Its word order at the phrasal level is explored in the present study.

A reduplication corpus (1,700 items) had been accumulated. The frequency of each token was calculated based on its occurrence in the British National Corpus (BNC). A reduplicative wordlist with 232 items was established based on the self-constructed corpus. Then a questionnaire with 116 items, extracted from the wordlist, was proposed for related searches using the built-in SARA software.

Frequency is important in creating a word list. It is also considered as the main factor to decide which item goes first in a conjoined phrase (Fenk-Oczlon, 1989, 2001). According to Fenk-Oczlon's rule, high frequency comes before low frequency in binomials. However, frequency is not the only criterion to decide the word order in fixed expressions. The probabilistic relations between two adjacent words should be examined as well. Mutual information (MI) and z-score are therefore employed to assess collocational significance (Church and Hanks, 1990). Both MI and z-score may provide useful insights into direction of collocability. Therefore the primary question of current researches is to explore fixed reduplications and collocational significance. Three aspects, *frequency*, *MI* and *z-score*, are discussed to evaluate which method is more appropriate to decide the word order in binomials and how they shed new light on word order and collocational analysis in terms of corpus-based approaches.

2. LITERATURE REVIEW

2.1 Multiword units and reduplicative fixed expressions

In the research of word usage, linguists have recently turned their attention to multiword units (MWUs), which are strings of words acting as a unitary lexical item with a single meaning (Carter, 1998; Moon, 1998). MWU includes compound words, phrasal verbs, fixed phrases, idioms and proverbs (Schmitt, 2000, pp. 99-100). Freezes or fixed expressions consist of irreversible conjoined phrases and fixed reduplicatives (Pinker and Birdsong, 1979). Reduplicative MWUs can be compounds, fixed expressions and other types, e.g., *first and foremost* (*248*)[1], *deaf and dumb* (*276*), *bits and bobs* (*49*) and the like.

Reduplicated word-formation varies in English; examples consist of ablaut and rhyming terms: 'i-æ' (riprap), 'i-o' (ping-pong), *super-duper* and *hocus-pocus*. Terminology to describe the phenomenon also shows a discrepancy and includes: *fixed expressions*, *freezes*, *binomials* and *frozen locutions* (Pinker and Birdsong, 1979; McCarthy, 1990; Landsberg, 1995; Moon, 1998). Binomials or trinomials are usually irreversible combinations with other conjunctions whose order may be different from language to language, e.g., *sooner or later (503)* [2] (McCarthy, 1998, 130-131).

Basically fixed expressions can be divided into the concise formal types as shown in Table 1 (Carter, 1998; Moon, 1998):

Table 1 Formal types of fixed expressions

Types of freezes	*Examples*
irreversible conjoined phrases	wear and tear hook, line and sinker first and foremost
fixed reduplicatives	
· vowel alternations (ablaut)	pitter-patter ping-pong
· rhyming terms	super-duper razzle-dazzle hocus-pocus

2.2 Freezes, word order and frequency-based arguments

Fenk-Oczlon (1989, 2001) argued that the frequency-based approaches can be simply used to solve some old questions. A new rule was proposed for the decision of word order in freezes: *high frequency before low frequency*, which implies 'more frequent tokens come before less frequent ones.' For example, the frequency of 'plus' (*7,767*) in the fixed expression, 'plus or minus' is higher than that of the second one, 'minus' (*1,776*); the frequency of occurrences for their collocation is 66. According to Fenk-Oczlon, the new rule achieves the highest accuracy with 84% correct predictions in his corpus. Frequency is considered

[1] The Arabic numerals mean the frequency of each token in the British National Corpus.

[2] In Mandarin Chinese the word order is opposite, i.e., *chi-tsao* ('late-early') → 'sooner or later'.

as the main factor to decide which item goes first in a conjoined phrase (Landsberg, 1995). However, *is frequency the only criterion to decide the word order in fixed expressions*? It is often maintained that the probabilistic relations between two adjacent words should be considered as well when dealing with the fixed expressions (Moon, 1998; Schmitt, 2000).

2.3 Mutual information, z-score and frequency in the British National Corpus

There are three major kinds of scores frequently proposed to assess the collocational significance of each co-occurrence: i.e., mutual information (MI) score, t-score and z-score (McEnery and Wilson, 1996; Hunston, 2002). MI-score is probably the best known among them. T-score and z-score are most similar in terms of how they are calculated. MI and z-score are the two formulae mostly used to calculate the relationship of significant collocations. MI-score and z-score can be calculated using the SARA software, a built-in tool in the BNC, by which the z-score is generally recommended. Therefore, only both *MI* and *Z-scores* together with *frequency* will be disussed in the current research.

MI is a measure of the strength of collocation, provides a summary of what company words keep and thus is used for assessing collocational significance (Aston and Burnard, 1998). The higher the MI score, the more genuine the association between two words (Oakes, 1998), which can be calculated in terms of the following formula (Church and Hanks, 1990; Stubbs, 1995):

$$\boxed{I = \log2\ ((f(n, c) \times N) / (f(n) \times f(c))}$$

• I = MI = mutual information; n = node; c = collocate
• f(n, c) is the collocation frequency
• f(n) is the frequency of node word (the query focus)
• f(c) is the frequency of the collocate
• N is the number of words in the corpus (corpus size):
  - If I (n; c) > 3, then the pairs tend to be significant or 'interesting'[3].
  - If I (n; c) ~ 0, then the pairs are less interesting.
  - If I (n; c) < 0, then x and y are in complementary distribution.

The z-score is the number of standard deviations from the mean frequency. It is used to measure how likely it is that the focus/node and collocate are related. The higher the z-score is for an item related to the node word, the greater is its degree of collocability with that word (McEnery and Wilson, 1996).

[3] An MI score greater than (or equal to) 3 may indicate a significant collocational link (Church and Hanks, 1990: 24; Hunston, 2002: p. 71).

2.4 Problems and Research Questions

Frequency is commonly assumed to influence the word order in fixed expressions. However, is frequency the main factor to decide which item goes first in a conjoined phrase? In addition, the current topic of reduplication and binomial freezes is often neglected because it presents problems for those theorists. Therefore the research questions of the present study will focus on the following topics:

- Reduplication at the lexical level;
- Frequency explored to see whether it is the major factor to influence word order in freezes;
- The frequency and probability of collocation for reduplicative freezes, calculated in terms of MI (Church and Hanks, 1990; Moon, 1998);
- MI and z-score employed to calculate the probability of collocation and assess collocational significance.

Frequency, *MI* and *z-score* are explored to see how they shed new light on word order and collocational analysis in terms of an integrated methodology.

3 METHOD & PROCEDURE

3.1 Data and Analytical Tools

The data for the present study include the author's own reduplication corpus and the BNC. The author's data have been gathered in the field, from research papers, dictionaries, websites, newspapers, advertisements, slogans, etc. since 1986. The ongoing collection (about 1,700 tokens) has undergone two stages for this research. The criteria for the establishment of the reduplication corpus are mainly based on the *pattern* (Wang, 2002a):

- The form of each token should be *reduplicated* in various types (onset, rhyme, etc.): e.g., ***first** and **f**oremo**st*** (247), ***th**i**s** and **th**a**t*** (202), and ***t**own and **g**own* (8).
- The pattern may be reduplicative binomial or trinomial expressions:
  - ***Full copy**, X_1 {conj., prep} X_1, e.g., so and so, and all in all;*
  - ***Partial reduplication**, X{conj., prep, art}Y, e.g., wine and dine (10), tit for tat (92), trick or treat (15), and bric-a-brac (52) ;*
  - ***Triplet and others**: X_1 X_2 X_3, e.g., tic tac toe (3) and Milly Molly Mandy (2).*

Only the second pattern, partial reduplication, is explored for this current study. The following are the main foci to be investigated in order to demonstrate how reduplication is used pervasively in day-to-day discourse:

- Corpus: using the BNC and Constructing the author's own corpus;
- Calculating the frequency for each selected token in the BNC;
- New power extracting data to be used in the questionnaire;
- Using Google to surf general websites to download examples of reduplication;
- Instruments: using SPSS and SARA for statistical analysis;
- Using SARA searches for the frequency of each token, MI and z-score.

3.2 Procedure and Method

The initial step is to collect data, analyze it and then create the compiled reduplication corpus (1,700 items). The first 232 items were extracted as wordlist 1, and then 116 items were selected as a questionnaire (wordlist 2). All tokens are chosen according to the following principles:

- *All types are mainly based on the patterns, [X] + and/or + [Y], including MWU, fixed expressions and idioms.*
- *All tokens identified in the corpus as reduplications undergo SARA searches for their frequency of occurrence, MI-score and z-score.*
- *ANOVA and Sheffé tests were used to investigate which approach underwent significant difference in terms of frequency, MI-score and z-score.*

Figure 1 summarizes the basic procedure for the current studies:

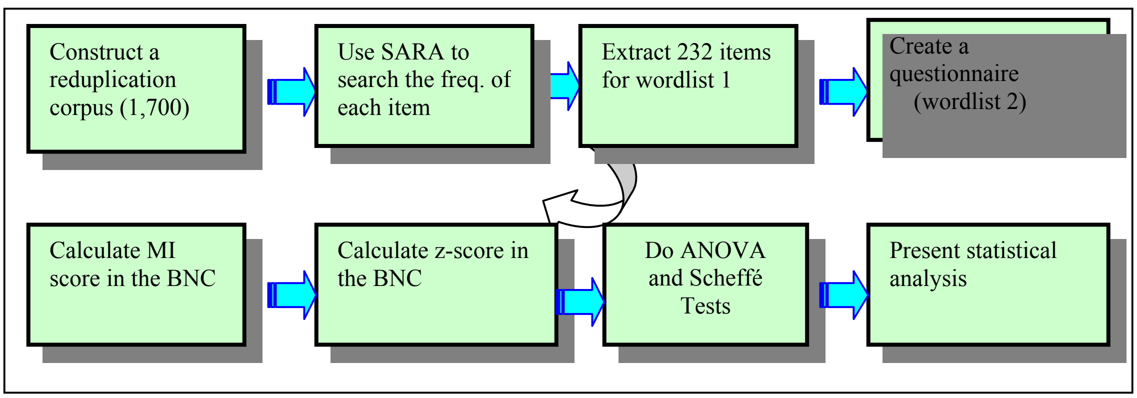

Figure 1 Flowchart for the data-processing

4. RESULTS and SAMPLE ANALYSIS

4.1 Frequency counts and frequency-based method

4.1.1 Procedure 1: a word list with the frequency

Fenk-Oczlon (1989) proposes frequency-based arguments to investigate freezes or fixed expressions. Therefore, the frequency for each token in the BNC was calculated first. The abridged ranking order in Table 2 is based on the results of frequency searches, i.e. the collocational F(x,y). For example, the following F(x) represents 'either' (27152), and F(y), 'or' (367981). The frequency of their collocation, F(x, y) 'either…or…' is 22111. The MI-score of 22.30 and the z-score of .1178 are calculated respectively.

Table 2 Abridged results for frequency-based grouping with MI and Z score

No.	grouping	F(x)>F(y)	F(x)	F(y)	F(x, y)	MI	Z score	Tokens
1	High		27152	367981	22111	22.30	.1178	either X or Y
2	High	+	1931791	197273	1152	30.50	.4021	in and out (X in and X out)
3	High		1710	38424	503	30.40	.6051	sooner or later
4	High	+	36609	6086	458	32.70	1.4491	once or twice
5	High	+	2629	724	276	34.00	3.2693	deaf and dumb
6	High	+	120825	604	247	33.90	3.3861	first and foremost
7	High		454441	1119521	202	23.00	.0686	this and that
8	High		5328	12249	156	28.70	.5973	upper and lower
9	High	+	197273	192000	152	24.50	.1436	out and about
10	High	+	51557	717	134	32.10	2.2891	part and parcel
11	High	+	8335	4715	112	29.10	.8123	positive or negative
12	High	+	24401	9171	83	27.30	.5036	mother and daughter
13	High		78	111	69	32.50	4.1748	comings and goings
14	High	+	7767	1776	66	29.00	1.0207	plus or minus
15	High		1187	2048	67	28.80	.9577	fame and fortune
16	High	+	74666	30303	53	24.30	.1978	last but not least
17	High		1490	4324	54	27.20	.5917	odds and ends
18	High		3424	9978	48	25.60	.3671	rough and ready
19	High	+	3282	92	49	31.80	3.8643	bits and bobs
20	High	+	23864	10059	46	25.50	.358	black and blue
21	High		5124	19891	44	24.40	.2489	birth and death
22	High		2663	9361	43	25.40	.3588	mix and match
N=116 ;		n=76	Correct rate for Fenk-rule (n/N) = 65.5%					

4.1.2 Procedure 2: "High frequency before low frequency"

Fenk-Oczlon's rule (1989; hereby *Fenk-rule*), "high frequency before low frequency," indicates that the frequency of F(x) is higher than that of F(y), i.e., $F(x) > F(y)$. There are only 76 items in the wordlist (total = 116) whose F(x) is bigger than F (y). This means that the correct rate of *Fenk-rule* is 65.5% in our test, not so accurate as 84% correct predictions in Fenk-Oczlon's statement. For example, the frequency of the first part (1710) of 'sooner or later' is lower than that of the second one (38424). It is evident that Fenk rule fails to come up consistently with the correct prediction.

4.1.3 Procedure 3: Frequency-based grouping

Using the ANOVA test, there is no significant difference for frequency-based groups and Z-score groups. Only MI between groups shows a significant difference (*p<.01; see Table 3). Based on the ranking of collocational frequency of occurrences, all the items were divided into three groups, higher (1), middle (2) and lower (3) groups.

Table 3 ANOVA for frequency-based grouping

ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
FXY.0	Between Groups	22662708	2	11331354	2.069	.133
	Within Groups	4.60E+08	84	5475709.9		
	Total	4.83E+08	86			
MI.0	Between Groups	600.342	2	300.171	16.437	.000
	Within Groups	1533.961	84	18.261		
	Total	2134.303	86			
ZSCORE.0	Between Groups	7251390.2	2	3625695.1	2.585	.081
	Within Groups	1.18E+08	84	1402822.6		
	Total	1.25E+08	86			

The Scheffé test further indicates that only MI between groups exhibits significant difference (Table 4). It is evident that something else is needed to make up the shortcomings of frequency-based argument.

Table 4 Scheffé test for frequency-based grouping

Multiple Comparisons

Scheffe

Dependent Variable	(I) GROUP	(J) GROUP	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
FXY.0	1.00	2.00	1169.0814	613.382	.169	-359.4983	2697.6611
		3.00	1184.0000	705.544	.250	-574.2500	2942.2500
	2.00	1.00	-1169.0814	613.382	.169	-2697.6611	359.4983
		3.00	14.9186	613.382	1.000	-1513.6611	1543.4983
	3.00	1.00	-1184.0000	705.544	.250	-2942.2500	574.2500
		2.00	-14.9186	613.382	1.000	-1543.4983	1513.6611
MI.0	1.00	2.00	2.8205*	1.120	.047	2.902E-02	5.6120
		3.00	7.2955*	1.288	.000	4.0845	10.5064
	2.00	1.00	-2.8205*	1.120	.047	-5.6120	-2.902E-02
		3.00	4.4749*	1.120	.001	1.6835	7.2664
	3.00	1.00	-7.2955*	1.288	.000	-10.5064	-4.0845
		2.00	-4.4749*	1.120	.001	-7.2664	-1.6835
ZSCORE.0	1.00	2.00	70.9095	310.465	.974	-702.7845	844.6035
		3.00	707.7545	357.112	.147	-182.1876	1597.6967
	2.00	1.00	-70.9095	310.465	.974	-844.6035	702.7845
		3.00	636.8450	310.465	.128	-136.8490	1410.5390
	3.00	1.00	-707.7545	357.112	.147	-1597.6967	182.1876
			-636.8450	310.465	.128	-1410.5390	136.8490

he mean difference is significant at the .05 level.

4.2.1 Procedure 4: Mutual Information (MI-based grouping)

The MI-based grouping is based on the results of calculating the association of mutual information between two items (e.g., *binomials*) in terms of using the Sara tool. It can be divided into three new groups, i.e., higher group (the first 23%), lower group (the last 23%), and middle group (the rest). Again, the ANOVA test shows there is no significant difference for frequency-based ranking between groups. On the other hand, both MI and Z- score groups indicate significant differences between groups.

Table 5 ANOVA for MI-based grouping

ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
MI	Between Groups	1585.789	2	792.895	121.425	.000
	Within Groups	548.514	84	6.530		
	Total	2134.303	86			
ZSCOREMI	Between Groups	61326443	2	30663221	40.396	.000
	Within Groups	63762044	84	759071.958		
	Total	1.25E+08	86			
FXY.MI	Between Groups	4805388.3	2	2402694.2	.422	.657
	Within Groups	4.78E+08	84	5688297.0		
	Total	4.83E+08	86			

The Scheffé test pinpoints that MI-based grouping illustrates the significant differences only for the relations between MI and Z-score groups except for one comparison.

Table 6 Scheffé test for MI-based grouping

Multiple Comparisons

Scheffe

Dependent Variable	(I) GROUP	(J) GROUP	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
MI	1.00	2.00	6.5869*	.670	.000	4.9176	8.2561
		3.00	11.9773*	.770	.000	10.0572	13.8973
	2.00	1.00	-6.5869*	.670	.000	-8.2561	-4.9176
		3.00	5.3904*	.670	.000	3.7211	7.0596
	3.00	1.00	-11.9773*	.770	.000	-13.8973	-10.0572
		2.00	-5.3904*	.670	.000	-7.0596	-3.7211
ZSCOREMI	1.00	2.00	1716.5628*	228.377	.000	1147.4356	2285.6900
		3.00	2194.8227*	262.691	.000	1540.1837	2849.4617
	2.00	1.00	-1716.5628*	228.377	.000	-2285.6900	-1147.4356
		3.00	478.2599	228.377	.118	-90.8673	1047.3871
	3.00	1.00	-2194.8227*	262.691	.000	-2849.4617	-1540.1837
		2.00	-478.2599	228.377	.118	-1047.3871	90.8673
FXY.MI	1.00	2.00	-379.7368	625.176	.832	-1937.7064	1178.2329
		3.00	154.5909	719.109	.977	-1637.4650	1946.6468
	2.00	1.00	379.7368	625.176	.832	-1178.2329	1937.7064
		3.00	534.3277	625.176	.695	-1023.6420	2092.2974
	3.00	1.00	-154.5909	719.109	.977	-1946.6468	1637.4650
		2.00	-534.3277	625.176	.695	-2092.2974	1023.6420

*. The mean difference is significant at the .05 level.

4.2.2 Procedure 5: Z Score-based grouping

Likewise, given the z-score-based grouping, the frequency-based ranking between groups displays insignificant difference. The Scheffé test verifies the observations that various groups based on MI grouping exhibit significant differences.

4.3 Summary

Given the ANOVA tests and multiple comparisons, the key grouping methods can be summarized in terms of Table 7:

Table 7 Nine Different grouping types and the results of ANOVA tests

Grouping \ *Ranking*	*Frequency F (x, y)*	*MI*	*Z Score*
Frequency-based	□	□□	□
MI-based	□	□□	□□
Z Score-based	□	□□	□□

**p<.01

It is apparent that no matter what kind of grouping method is used, there is no significant difference for frequency ranking groups. However, MI ranking always shows significant difference between groups. If frequency-based grouping is adopted, there is no significant difference for z-score ranking. Multiple comparisons such as Scheffé confirm the above summary. Therefore frequency is not the main factor to decide the word order and the collocations in frozen expressions. It is evident that MI is more important to assess the probabilistic collocation between two adjacent words.

5. CONCLUDING REMARKS

The aim of this research, using corpus-based approaches, has been to explore fixed expressions in terms of frozen types, frequency, correct rate, collocation, and MI score. A corpus with a ranked wordlist was constructed. The frequency and probability of collocation for reduplication were calculated in terms of MI and z-score. Frequency was explored, but it is not the major factor to influence the word order in freezes. The MI-based method is instead proposed to confirm observations of word order and to revise Fenk-Oczlon's arguments (1989). In addition, ANOVA and post hoc tests, Scheffé, were employed to evaluate which grouping method is appropriate to come up with robust statistics for collocational significance. The key points for this research are summarized along the following lines:

- Given corpus-based approaches, Fenk's rule, high frequency before low frequency, is not always reliable. Among the 116 items in the present corpus, only 76 items meet this rule. The correct rate is 65.5%.
- MI-based approaches play a useful role in aligning, reinforcing or making up the shortcomings of frequency-based arguments.
- MI provides a quick guide to decide what to look for in the collocation pairs. It is used to calculate the probabilistic collocation of two observed items, f(x) and f(y). If the first item, f(x), is larger than f(y), the MI score is higher.

Statistics such as percentage coverage, and frequency of occurrences in a corpus are required to reinforce and constitute relevant arguments and research approaches. MI and z-scores are useful reference points while choosing fixed expressions to discuss. Further studies integrating probabilistic methods are definitely needed.

ACKNOWLEDGEMENT
The author is grateful to Michael McCarthy, John Ohala, Stuart Davis, Ming-wen Wu for their very useful comments. The research is partially supported by a grant, NSC 91-2411-H-130-008, from the National Science Council.

REFERENCES

Aston, G. & Burnard, L. 1998 *The BNC handbook*. Edinburgh: Edinburgh University Press.
Biber, D., Conrad, S.&, Reppen, R. 1998 *Corpus linguistics: investigating structure and use.* Cambridge: Cambridge University Press.
Birdsong, D. 1995 Iconicity, markedness, and processing constraints in frozen locutions. In M. Landsberg (Ed). Syntactic iconicity and linguistic freezes, pp. 31-45.
Carter, R. 1998 Vocabulary: applied linguistic perspectives. London: Routledge.
Church, K., &, Hanks, P. 1990 Word association norms, mutual information, and lexicography. Computational Linguistics, 16(1), 22-29.
Cook, G. 2000 *Language play, language learning*. Oxford: Oxford University Press.
Cooper, W. & J. Ross 1975 World order. *Chicago Linguistic Society,* 11, 63-111.
Fenk-Oczlon, G. 1989 Word frequency and word order in freezes. *Linguistics,* 27, 517-556.
Fenk-Oczlon, G. 2001 Familiarity, information flow, and linguistic form. In J. Bybee & P. Hopper (Eds) *Frequency and the Emergence of Linguistic Structure*, pp. 431-448. Amsterdam: John Benjamins.
Hoey, M. 1991 *Patterns of lexis in text*. Oxford: Oxford University Press.
Hunston, S. 2002 Corpora in applied linguistics. Cambridge: Cambridge University Press.
Landsberg (Ed). 1995 *Syntactic iconicity and linguistic freezes: the human dimension*. New York: Mouton de Gruyter.
Landsberg, M. 1995 Semantic constraints on phonologically independent freezes. In Landsberg (Ed). *Syntactic iconicity and linguistic freezes: the human dimension*, pp. 65-78.
McCarthy, M. 1990 *Vocabulary*. Oxford: Oxford University Press.

McCarthy, M. 2001 Good listenership made plain: British and American non-minimal response tokens in everyday conversation. MS, University of Nottingham.
McEnery, T. & Wilson, A. 1996 *Corpus linguistics*. Edinburgh: Edinburgh University Press.
Moon, R. 1998 *Fixed Expressions and Idioms in English*. Oxford: Oxford University Press.
Nation, P. 1982 Beginning to learn foreign vocabulary: a review of the research. RELC Journal, 13 (1), 14-36.
Oakes, M. 1998 *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press.
Pinker S. & Birdsong, D. 1979 Speaker's sensitivity to rules of frozen word order. *Journal of Verbal Learning and Verbal Behavior*, 18, 497-508.
Schmitt, N. 2000 *Vocabulary in language teaching*. Cambridge: Cambridge University Press.
Schmitt, N. & McCarthy, M. 1997 *Vocabulary*. Cambridge: Cambridge University Press.
Sinclair, J. 1991 *Corpus, concordance, collocation*. Oxford: Oxford University Press.
Stubb, M. 1995 Collocations and semantic profiles: on the cause of the trouble with quantitative studies. Functions of Language, 2 (1), 23-55.
Tao, H. & McCarthy, M. 2001 Understanding non-restrictive which-clauses in spoken English, which is not an easy thing. Language Sciences, 23, 651-677.
Wang, S.P. 2001a Integrating corpus-based and vocabulary learning approaches into a linguistics project. Paper presented in 46th International Linguistics Association, New York: New York University.
Wang, S.P. 2001b Reduplication and repetition in applied linguistics. The Proceedings of 2001 International Conference on the Application of English Teaching, 200~222. Taipei: Crane Publishing Co.
Wang, S.P. 2002a Corpus-based approaches and discourse analysis to reduplication and repetition, paper submitted to journal.