	Report on a free continuous word association test (part 2)
	MUNBY, Ian
	, 135: 55-74
	2008-03-00

Report on a free continuous word association test (part 2)

Ian MUNBY

Key words: cue (word), prompt (word), reaction, and stimulus (word) are used co-terminously

INTRODUCTION

This second experiment was a follow-up of the 2006 replication of the multiple word association probe by Kruse, Pankhurst, and Sharwood-Smith (1987) reported in Munby (2007). The 1987 study compared the associations produced by a group of 15 Dutch third-year university students of English with a group of 7 native speakers of English in a test which used a specially designed software to collect up to 12 responses for each of a set of 10 stimulus words.

The aim was twofold: first, to attempt to find a link between the non-native subject WAT scores (Word Association Test) and two proficiency measures (a cloze test and a grammar monitoring test), and, second, to measure differences, if any, between native and non-native subject performance on the WAT. The scoring system is detailed in Munby (2007). With the Dutch subjects in 1987, correlations between proficiency scores and WAT scores were weak. Further, in all three WAT measures, no significant difference between the two groups was reported. The experiment was hugely influential since it seemed to prove that the free continuous WAT was inadequate as a proficiency measure. In the 2006 replication, native subjects outscored non-native subjects in the WAT but correlations between non-native WAT scores and proficiency measures were also weak.

This 2007 study featured the following alterations:

i) a new set of cue words was introduced, including two from the 2006 experiment that showed some promise as good prompts.

ii) a test of vocabulary size (yes/no test, Meara, 1992) was introduced

iii) there was no re-test of the Word Association test for non-native subjects.

Experiment One (2006), a replication of Kruse, Pankhurst, and Sharwood-Smith (1987), revealed a number of flaws in the 1987 test design which may have undermined its conclusions.   The choice of stimulus words was one problem with item analysis revealing that some cue words were considerably more effective than others in discriminating learners of different levels.   The assumption is that there may be some important selection criteria that cue words should satisfy before meriting inclusion in the test.   These criteria were not considered in the 1987 test although, coincidentally, two of the prompt words produced response profiles that correlated more strongly with subject proficiency in the 2006 study.

AIMS

(i) To compile a criteria-based selection of stimulus words for the WAT from the Kent-Rosanoff lists (1910) which allow the non-native subjects the best possible chance of producing native-like responses.   It was necessary to select items from these lists because the normative data-the Minnesota norms lists-was restricted to 1, 000 responses to each of the 100 items on the original Kent-Rosanoff lists

(ii) To elicit responses from learners with these cues and compare them to their performance on two proficiency measures (a grammar monitoring test and a cloze test) and an additional test of vocabulary size to discover whether the new cue words result in closer correlations between learner proficiency and multiple response WAT performance.

(iii) To confirm that responses to the following two prompt words from the 2006 study-*sickness* and *anger* produce significant correlations with the proficiency measures.

(iv) To confirm preliminary findings of the 2006 test which suggest that, although results depend to a great extent on the stimuli chosen, other perceived design flaws in the original experiment may have to be dealt with in order to establish a link between association norms-based scoring lists and non-native subject proficiency in future experiments.   In other words, we wish to provide further important evidence that there were several serious problems with the original 1987 study.   Most of these flaws concern the scoring system.   For example, the number of responses a subject can enter within the 30-second time limit is not accurately assessed because no more than 12 responses can be entered.   Many of the native subjects, and even some non-native subjects, entered the full compliment within half or three quarters of the time allowed.   The stereotypy measures are also unsatisfactory since the normative data (the Minnesota lists, Jenkins, 1952) is drawn from a collection of single or

primary responses to the one hundred Kent-Rosanoff stimuli from 1, 000 subjects. Although the number of responses on the lists is large, it seems likely that these lists fail to tap more distant, or remote associations in the native speaker lexicon. For example, the response *cough* to the stimulus *sickness* does not appear on the norms lists but often appears amongst sets of native and non-native responses in both the 2006 and 2007 studies, but never as a primary response. A norms list drawn from the same testing instrument would certainly be more suitable for measuring both native and non-native associative behaviour.

An additional problem with the scoring system concerns the weighted stereotypy measure employed by Kruse, Pankhurst, and Sharwood-Smith (1987). With this system, a maximum of 144 points is awarded to a subject who enters a primary response that matches the primary response on the norms lists, such as *sleep* for the cue *dream*. Since primary responses are inevitably high frequency words, this system lacks fundamental power to discriminate learners of different levels and is therefore unlikely to unearth a link between learner proficiency and WAT performance.

CRITERIA FOR SELECTION OF CUES FROM THE KENT-ROSANOFF LIST

A further problem concerns the selection of the cue words themselves. Analysis of correlations between cloze scores and WAT scores for each cue word in the 2006 reveals remarkable variation in cue word effectiveness. With reference to Table 1 below, correlations between cloze score test scores and WAT scores for the most successful cue word *sickness* are almost the same (number of responses 0.48, non-weighted stereotypy, 0.44) as the total scores for these measures for all nine cue words put together (0.45 and 0.49 respectively). In contrast, two cue words, *mutton* and *short*, produced correlations with the cloze test that were not significant with any of the three WAT measures.

Table 1. Correlations between non-native subject cloze scores (Test 1, 2006) and WAT performance per stimulus word.

Stimulus	No. responses	Non-weighted	Weighted
high	0.32*	0.33*	−0.09ns
sickness	0.48**	0.44**	0.39**
short	0.23ns	0.23ns	−0.02ns
fruit	0.47**	0.23*	0.04ns
mutton	0.24ns	0.03ns	0.15ns
priest	0.36**	0.39**	0.30*
eating	0.33*	0.29*	−0.06ns
comfort	0.35*	0.19ns	0.09ns
anger	0.48**	0.40**	0.35**

1-sided p-value *$p = <0.05$ **$p = <0.01$

The cue word selection process was therefore the most important part of this 2007 follow-up study.

HOW DO FINDINGS FROM 2006 STUDY INFORM US ABOUT CUE WORD SELECTION?

On the evidence of Table 1 above, it was decided to reuse two cues from the 1987 and 2006 studies, *sickness* and *anger*, because they both produced significant correlations with the cloze test.   With all the cue words, with the exception of *mutton*, the number of responses measure and the non-weighted stereotypy measure correlate more strongly with the cloze test scores than the weighted stereotypy measure.   However, with *mutton*, correlations in all measures are not significant.   The remaining six cue words correlated less well with proficiency. Possible reasons for this are cue word frequency, the influence of superordinates, and the phenomenon of excessively predictable responses.   Regarding word frequency, several non-native subjects had not entered any responses for the following two cue words: *priest* and *mutton*, the suggestion being that they were unknown words.   *Priest* falls in the BNC 3-4K range while *mutton* is in the 6-7K range.   For this reason they seem unsuitable as prompt words and the assumption is that low frequency words of this kind should be avoided in future.   Regarding the influence of superordinates, correlations for *fruit* seemed promising in the number of response measure but not in the stereotypy measures, probably because low level subjects had performed well by simply listing names of fruit, such as *apple, pear*, and *orange*.   The remaining cues may have been relatively ineffective because they invited excessively predictable responses.   For example, the cue *eating* failed to discriminate low-level subjects from high perhaps because easy points were won on stereotypy measures for predictable responses such as *drinking, food, lunch, and dinner*.   Similarly, the two adjec-

tives *high* and *short* also enabled low level subjects to enter sets of responses which did not differ significantly from those produced by higher level subjects due in part to the availability of simple associations such as the polar opposites of these adjectives (*low, tall, and long*). Since they are high frequency adjectives, highly frequent collocations were also to be found on the norms lists, such as *high-school* and *short-hair*. Finally, *comfort* was interesting in its failure to produce promising correlations because, like *anger* and *sickness*, it is also an abstract noun. Nevertheless, closer inspection shows the same forces were at work with easy points won for *comfort* associations such as *chair, bed*, and *sofa*.

In sum, the above analysis of the prompt words used in the 2006 study suggests that cues which (a) are unknown to some subjects taking the test, (b) are superordinates that invite simple, readily accessible sets, or (c) provoke excessively predictable reactions should be avoided in future studies.

WHAT LIGHT DOES THE LITERATURE SHED ON CUE SELECTION PRINCIPLES?

Little has been reported in the literature on the subject of what constitutes good and bad, or effective and ineffective, prompt words in free continuous word association tests. However, three issues are identifiable and these are (i) the tendency of some cues to provoke dominant primary responses, (ii) the tendency of some cues to provoke idiosyncratic responses, and (iii) the influence of grammatical form class.

First, regarding the tendency of some cues to provoke dominant primary responses, Meara (1983) had warned of the dangers of using high frequency words as prompts, particularly adjectives which produce their polar opposites such as *black-white, soft-hard*. By extension, nouns marked for sex which produce their complimentary terms, such as *boy-girl*, and *king-queen* were branded unsuitable. Meara (1983) also claims that about 60% of cues on the Kent-Rosanoff list are of this type.

Wolter (2001) also commented on problems caused by using cues with dominant primary responses, such as *black* which produced *white* in 751 out of 1, 000 cases on the Minnesota norms lists whereas the secondary response was *dark* (54 respondents) and *cat* (26) as tertiary. Wolter (2002) developed this idea further with the formulation of a 15-60 rule for selecting cues from the Edinburgh Associative Thesaurus, or EAT (1973). Prompt words whose

primary response generated more than 15% of the total number of responses. '60' refers to the requirement that the total proportion of non-idiosyncratic responses (or responses entered by at least two respondents) make up at least 60% of the total. However, since up to 12 responses are being elicited, the negative influence of the dominant primary response may not be such a significant factor, even though 144 points can be won on the weighted stereotypy scale for responses such as *low* for the cue *high*. It should also be borne in mind that dominant primary responses restrict the number of remaining responses available for norming purposes on lists compiled exclusively from primary responses such as the Minnesota Norms Lists and the EAT.

The prompt words the screening would thus produce is limited to the items available on the original Kent-Rosanoff lists since, at present, norming data is not available for any other cue words. Although some similarity in response trends can be identified across published norms lists, such as EAT and the Postman and Keppel lists (1970), there are important differences, even with primary responses to some cues.

Second, there is some support in the literature (Wolter, 2002) for the notion that cues which elicit a large number of idiosyncratic responses should be avoided in multiple response word association tests. An idiosyncratic response is a response only provided by one respondent in a group of informants. In this way, if there is evidence in a norms lists that a cue word produces a large number of idiosyncratic responses, it may not be effective in measuring approach to native speaker associative behaviour in non-native subjects. Initially, it was decided that prompt words generating more than 500, or 50% shared primary responses out of 1000 on the Minnesota norms lists should not normally be considered. The rule of 60, or the prerequisite of the norms list having at least 60% non-idiosyncratic responses was ignored because *sickness*, a successful prompt from the previous study, elicited only 30 non-idiosyncratic responses out of a total of 76 different responses on the norms lists, or 40% and *anger* had 60 non-idiosyncratic responses out of 168 or 35%.

Third, form class of prompt word is another issue considered by Wolter (2002) and his decision to use verbs only may have been a factor contributing to the inconclusive results of the study. This is not especially relevant to the selection process for this 2007 study since all verbs on the Kent-Rosanoff list also function as nouns including gerunds such *eating* and *working*. Indeed, as Namei (2004) points out, there are only 71 nouns and 29 adjectives on the

Kent-Rosanoff lists. The resulting effect on the selection process for this study is that it is hard to find a good balance of nouns and adjectives as cue words, if variety of form class is important. However, Miller and Fellbaum (1991) make a strong case for the noun, or the nominal concept, as the basis for the organization of the lexical memory. In this sense, given that adjectives on the Kent-Rosanoff list are low in both cue word quality and number, it is tempting to dismiss them altogether. Although 2 adjectives were finally chosen, the prospect of focusing on nouns may prove a promising research angle since one aspect of assessment of cue word effectiveness could be narrowed down to analysis of the properties, or distinguishing features of particular nouns. However, at this stage, there is not enough evidence what these features or properties may be.

THE CUE SELECTION PROCESS

Ineffective cues from the 2006 study were not included. Next, a list of questions was drafted to match the criteria and check the cues for potential problems. If the answer is "yes" to one or more of the following, the cue may qualify as a poor stimulus word, or a stimulus that is unlikely to elicit native-like responses from more proficient subjects.

(a) Is the stimulus likely to be confused with a similar sounding word in L1 (Japanese) ?
Since *trouble* had the potential to be confused with *travel*, it was eliminated from the list.

(b) Is the stimulus likely to be unknown to many of the subjects taking the tests (eg. *mutton, priest*) ? 1K words are preferable but others from 2K and above which are likely to be known by all non-native subjects, such as *spider*, may be included. Judgments concerning whether or not non-native subjects were likely to be familiar with a word or not were based on personal experience of teaching students of similar level to the lowest level subjects for many years. The following 16 prompts were discounted on this basis:
2K words: *citizen, command, cottage, justice, religion, rough.*
3K words: *bible, bitter, needle, soldier, thief, whistle.*
4K words: *stem, swift*
5K words: *sour, stove*

(c) Is the stimulus likely to produce a "dominant primary" response, such as an adjective or other word that produces its polar opposite (eg. *high-low*) or a noun which is marked for sex which tends to produce the opposite sex in response (eg. *king-queen*) ?

61

Although this may not appear too important since 12 responses are being elicited, if the primary response on the Minnesota norms list is high-in the range of 500 or above out of 1000 responses-it limits the number or variety of the remaining responses on the list. This in turn limits opportunities for participants to score points on the non-weighted stereotypy measure. It also influences results on the weighted stereotypy measure, where lower level subjects can easily score the maximum 144 points on the scale by entering *apple* as primary response to *fruit*, for example. The following 39 prompts were discounted on this basis: *table, dark, deep, soft, black, hand, chair, sweet, woman, cold, slow, white, sleep, carpet, girl, hard, eagle, lamp, bread, boy, light, bath, hungry, long, whiskey, square, butter, loud, bed, heavy, tobacco, scissors, quiet, salt, king, blossom, beautiful, hammer and smooth.*

(d) Is the stimulus likely to generate superordinates involving simple sets such as *fruit*? The following 11 prompts were not included on this basis: *music* (genre of music), *foot, head, stomach* (parts of the body), *red, blue, green, yellow* (colours) *lion, sheep* (animals) and *cabbage* (vegetables). It was expected that *music* would generate types of music such as classical, pop, rock, jazz, rap, punk, some of which did not exist at the time of norms list compilation. The main problem was the risk that subjects of different levels would all score equally well by producing simple sets of responses.

(e) Is the stimulus likely to elicit proper nouns, such as *river-Mississippi, City-Minneapolis, ocean-Pacific*? The following 4 prompts were not included on this basis: *mountain, river, ocean, city.*

(f) Is the stimulus likely to generate too many simple, obvious responses? If the norms lists include too many simple, obvious responses there will be a risk that subjects of different levels of proficiency will perform equally well or score highly in a similar fashion.
The following 7 cues were discounted for this reason. Responses adjudged to be too obvious from among the top 12 on the norms lists appear in brackets.
House [home, door, garage, roof, windows, room]
Working [hard, sleeping, playing, man, resting, job, tired]
Child [baby, mother, adult, boy, small, young, kid, little]
Baby [boy, child, cry, mother, girl, small]
Moon [stars, sun, night, light]
Street [avenue, road, cars, lights, city, walk, house, corner]

Cheese [mouse, bread, eat, food, milk, yellow, cake, butter]

Finally, *health* and *doctor* were eliminated because they were too similar to the cue *sickness* to be reused from the previous study. *Butterfly* was also eliminated because it was too similar to *spider*. This screening process eliminated 82 items from the list and left the following 8 cue words for 8 remaining places in the set: afraid, earth, dream, joy, memory, spider, window, wish [8].

METHOD

Subjects

Control group: 24 native speakers of English. The breakdown by nationality was USA: 10, Canada, 9, New Zealand 2, United Kingdom 2, and Australia 1 and by gender: male 21, female 3. With only two exceptions, all native subjects were, or had been, Japan-based teachers of English.

Experimental group: 86 Japanese English majors at Hokkai Gakuen University, Japan, from six different classes including 49 first years, 15 second years, and 22 third and fourth years. These students ranged in level from early to high intermediate but the majority was in the early intermediate range.

Test materials

With the exception of the vocabulary size test, all testing materials remain unchanged from the 2006 replication test. The materials were:

(i) The word association test (software IM06a) including two practice items (*cheese* and *lion*) as in the 2006 experiment. The following prompt words appeared in this order:

afraid, earth, dream, joy, memory, sickness, spider, window, anger, wish. Subjects did not receive any training in the different response types available, but were simply told to type in any words that the stimuli made them think of. Subjects were also encouraged to type in as many words as they could think of, up to 12, within the time allowed of thirty seconds, to refrain from dictionary use, to type in single, English words, to avoid proper nouns if possible, and to notice that the timer would deactivate while they were typing, allowing them 30 seconds pure thinking time. They were also informed that there were no right or wrong answers. After entering their personal data, subjects click on *next word* and the cue word appears on the screen. Subjects type in their responses and press the *return* key each time

until either the time allowed has expired or the maximum twelve responses has been entered, whereupon the stimulus word disappears. Subjects click on *next word* to repeat the above process until they reach the end of the set of 10 prompt words.

(ii) Two proficiency measures used in the 2006 study: a 50 word cloze test and a 50-item grammar error recognition test. Subjects were given 30 minutes to complete each test

(iii) One additional testing measure not included in 1987 & 2006 test: A yes/no vocabulary test (Meara, 1992). In this test, subjects are presented with a list of 300 words from five frequency bands (drawn from and grouped into five frequency bands: IK to 5K) and asked to check the box next to each item if they knew the meaning of the word, but not to check it if they weren' t sure. Each frequency band contains 60 words, including 40 real words and 20 non-words. Points are deducted if non-words are checked. There was no time limit although subjects typically finished the test in about 10 minutes.

Procedure

The order of the tests was as follows:

Session 1

Non-native subjects from six intact classes took the WAT in a CALL lab followed by the vocabulary size test. 45 minutes.

Session 2 (1 week later).

Non-native subjects took the cloze test followed by the grammar monitoring test. 60 minutes.

Native subjects only took the WAT.

There was no re-test of the WAT. The 1987 & 2006 tests show that non-native subjects perform slightly better, on average, on Test 2 than Test 1. However, Table 2 (Munby, 2007) show that test/re-test correlations are strong enough in each of the three measures to show that results are not random and probably the result of a practice, or test-retest effect.

Scoring, treatment and processing methods

The responses were measured using three scoring systems:(i) number of responses entered, (ii) non-weighted stereotypy, and (iii) weighted stereotypy. The non-weighted stereotypy measure is a straight count of the number of responses that also appear on the

Minnesota norms lists (Jenkins, 1952). Each subject's total score was obtained by summing the scores obtained on responses to all nine stimulus words The weighted stereotypy measure was based on specific scores for stereotypic responses according to whether they were native speaker primary, secondary, or subsequent responses. This takes into account the order in which the responses occur in the norms lists producing a weighted stereotypy score. This scoring system is unchanged from the 2006 study.

Six kinds of problematic responses were also treated in the same way as in the 2006 study. First, misspelled responses, such as *goast (ghost) in response to the cue *afraid*, were corrected if the word was identifiable. Second, plural or singular equivalents of items on the norming list such as *oceans* (*ocean* appears on the norms list as a response for *earth*) were accepted. Third, where multi-word unit responses were entered, such as *elementary school* for the cue *memory*, the response is accepted as scoring if one of the words in the unit, *school* in this case, appears on the norms lists. Fourth, if the same response to a cue word is entered more than once, it is discounted from the scoring. Fifth, responses that were L2 words or otherwise did not exist in English, such as *sirsology* or *vomitory*, were discounted. Finally, if the cue word is entered as a response it is also discounted.

Results from one of the ten cues (*spider*) was dropped because it correlated least successfully with scores from the cloze test which produced the strongest correlations of the three proficiency measures. This reflects the conditions of the original 1987 experiment, and the 2006 replication, where responses from one of the ten cue word were dropped because of a mistake. The WAT was scored by hand, although Tex-Lex Compare (Tom Cobb, http: // www.lextutor.ca/text_lex_compare/) was used to identify which responses also appeared on the norms lists for the non-weighted stereotypy count.

RESULTS

Table 2. (Munby, 2006 and 2007 studies) **Mean scores, standard deviations, and theoretical maximum for all scoring methods of the word association test.**

	Non-native speakers	Native speakers	
	Mean (SD)	Mean (SD)	Maximum
A (2006)	61.8 (25.3)	94.6 (12.7)	108
A (2007)	55.3 (21.0)	93.6 (13.2)	108
B (2006)	32.2 (10.2)	46.3 (9.86)	108
B (2007)	27.9 (10.0)	48.8 (12.1)	108
C (2006)	1,378 (346)	2,141 (496)	5,971
C (2007)	1,015 (307)	1,934 (527)	5,948

A=number of responses.   B=weighted stereotypy.   C=non-weighted stereotypy.

With reference to Table 2 above, the data supports the 2006 findings that in a free continuous word association test (based on Kruse, Pankhurst, and Sharwood-Smith, 1987), a group of native speakers will achieve higher mean scores than a group of non-native speakers in all three measures: number of responses, non-weighted stereotypy, and weighted stereotypy.   Although it has to be said that 8 of the 24 native speakers in the 2007 test had also taken the test in the 2006 study, the suggestion is that the performance of native speakers with similar backgrounds in a similar age group is remarkably standard and that this does not appear to change very much no matter what the prompt words are.   For example, the mean number of responses in 2006 was 94.6 while in 2007 this was 93.6, with only a small difference apparent in the stereotypy measures.

Table 3. **Munby 2007 test.**
Correlations between the association scores and the proficiency measures.
2006 scores are included in brackets

	Cloze test	Grammar test	Yes/No test
A	r=0.25* (0.45*)	r=0.16ns(0.34*)	r=0.15n. s
B	r=0.41**(0.49*)	r=0.29**(0.30*)	r=0.24*
C	r=0.43**(0.32*)	r=0.35**(0.16n. s.)	r=0.19*

*p<.05   **p<.01   1-sided p-value

A=number of responses.   B=weighted stereotypy.   C=non-weighted stereotypy.

With reference to Table 3 above, 2006 & 2007, although correlations between the proficiency measures and WAT scores were expected to be stronger than in 1987 & 2006 due to (supposed) improved cue selection, they turned out to be weaker with Tests A (number of

responses) and Test B (non-weighted stereotypy) with both cloze and grammar monitoring tests. In the light of earlier comments on the system of weighted stereotypy, the finding that seems most out of line with expectations was that the strongest correlations of all were found between the cloze scores and the weighted stereotypy measures (r=0.43). While there was no data in the original Kruse, Pankhurst, and Sharwood-Smith probe (1987) and 2006 replication to suggest that weighted stereotypy is an effective measuring tool, for the first time there is evidence that it could be. Correlations between grammar monitoring scores and weighted stereotypy scores were also stronger (r=0.35) than with the non-weighted stereotypy measure (r=0.29) when the reverse was to be expected judging from the pattern set in both the 2006 study and in the original 1987 study. A further unexpected finding was that in the 2006 study, the strongest, "flagship" correlations were to be found between the cloze test scores and the non-weighted stereotypy measure (r=0.49). In the 2007 study, the figure was lower at r= 0.41 when it was expected to be higher if the cue words were genuinely superior in quality.

Nevertheless, correlations between the cloze and all three measures of the WAT (number of responses, non-weighted stereotypy, and weighted stereotypy) are higher than for the grammar monitoring test, and the same was true in the 2006 study. Also, although all correlations were weak, they are at least significant with the exception of correlations between the number of responses and the grammar test and the yes/no vocabulary test.

Another finding that did not match up to expectations concerns the yes/no vocabulary test (Meara, 1992). While it was predicted that correlations with the WAT would be stronger than with both the cloze and grammar monitoring, they turned out to be the weakest of the three measures. It appears that the decision to use a yes/no test which was limited to 40 words (and 20 non-words) for each of the first five 1, 000 level frequency bands (1K-5K) resulted in some "ceiling effect" where the true extent of the vocabulary knowledge of the higher level subjects was not picked up. For example, 21 out of 86 subjects scored more than 375 out of a maximum of 500 points on this test, or 75% knowledge of the words tested. The mean score was also too high at 327, with a standard deviation of 56.6. Seven students scored above 400, with the highest score being 471. Clearly, a vocabulary size test including additional items from the 5K-10K levels would have been more likely to produce stronger correlations with WAT scores.

With reference to the mean WAT scores per stimulus (Table 4), non-native speaker

performance does not even approach native speaker performance with any of the nine cue words in any of the three WAT measures. As mentioned earlier, this was not the case with the 1987 and 2006 experiments. This could be viewed as a sign of cue selection success.

Table 4. (Munby, 2007 test) scores for each stimulus word.
Note 1: 2006 scores for two of the prompt words *anger* and *sickness* (marked with an asterisk in the left-hand column) are included in the row below the 2007 scores.
Note 2: 2006 non-native subject scores are for Test 1 (not Test 2, the re-test).

Stimulus	Non-native subjects			Native subjects		
	A	B	C	A	B	C
1. AFRAID	4.93	2.28	69.9	10.29	4.88	227.17
2. EARTH	7.93	4.94	106	11.42	7.92	207.62
3. DREAM	6.06	2.29	139.47	9.96	4.5	275.88
4. JOY	6.65	2.31	75.81	10.08	4.58	184.0
5. MEMORY	6.29	2.78	102.27	10.25	4.29	167.08
6. SICKNESS	6.81	4.36	141.40	11	6.25	231.54
6. SICKNESS*	8.36	5.11	155.36	10.84	6.16	247.28
7. WINDOW	6.78	4.03	125.34	11.13	6.21	216.46
8. ANGER	4.55	1.86	49.65	9.92	4.58	179.08
8. ANGER*	6.18	2.84	77.16	10.24	4.44	161.20
9. WISH	5.38	3.03	205.63	10.04	5.54	249.92

A = number of responses.   B = weighted stereotypy.   C = non-weighted stereotypy.

Correlations between non-native WAT performance per stimulus word and cloze scores show that some cues are more effective than others, as in the 2006 study (see Table 5). Although generally correlations are weaker with all cue words than in the 2006 study, five of the eight new cues: *afraid, earth, dream, memory* and *window* appear more effective than the two best-performing cues *sickness* and *anger* from the 2006 study which were re-used in this 2007 study. It is also worth noting from comparison of the 2006 and 2007 studies that *sickness* elicits slightly better results than *anger*. *Joy, spider*, and *anger* appear to be the weakest with no significant correlations to be found with any of the three WAT measures. This is most unusual since *anger* had been the second most effective cue in the 2006 study. There seems to be no reasonable way to explain this finding.

Table 5. Correlations between WAT performance per stimulus word and non-native subject cloze scores (2007). 2006 scores for the prompt words *anger* and *sickness* are included in brackets.

Munby Stimulus	Number of responses	Stereotypy measures Non-weighted stereotypy	Weighted stereotypy
1. AFRAID	0.22*	0.20*	0.27**
2. EARTH	0.19*	0.30**	0.30**
3. DREAM	0.33**	0.36**	0.29**
4. JOY	0.14ns	0.17ns	0.05ns
5. MEMORY	0.19*	0.35**	0.03ns
6. SICKNESS	0.19*(0.48**)	0.24*(0.44**)	0.19*(0.39**)
7. SPIDER	0.17ns	0.10ns	0.05ns.
8. WINDOW	0.24*	0.26*	0.09ns
9. ANGER	0.17ns (0.48**)	0.12ns (0.40**)	0.05ns (0.35**)
10. WISH	0.13ns	0.25*	0.30**

1-sided p-value   *p= <0.05   **p= <0.01

Discussion

Despite evidence of recurrent patterns, the new set of cue words produced some findings that ran against expectations. The following is an attempt to account for these anomalies, identify the recurrent patterns, and finally to review what is known or remains unknown about multiple response, or "free", WAT with particular attention paid to choice of prompt word.

As detailed earlier in the results section, correlations between WAT performance and proficiency measures were weaker than in the previous study (Munby, 2007), despite expectations to the contrary. This could in part be attributable to the quality of the non-native subjects. Since it was a quantitative study, it had been hoped that a larger quantity of subjects (86 in the 2007 study compared with only 45 in 2006) would lead to stronger correlations between WAT scores and proficiency scores. The reverse turned out to be the case.

There were four key differences between the experimental group of non-native subjects in the 2006 study and the 2007 study group. The first applies to participant level of English. Although some second, third, and fourth year subjects in the 2007 study had also taken part in the 2006 study, the general level of the subjects was lower than in the 2006 study. For example, the mean cloze score was 18.5 (2007) but 21.4 (2006), while the mean grammar monitoring score was 20.5 (2007) but 24 (2006). Second, 27 of the 45 subjects in the 2006 study had volunteered to take part in the study whereas all 86 students who completed all four tests

in the 2007 study had simply been asked to take the tests in class time possibly leading to a difference in motivation and attitude. Third, with three of six participating classes being first year students, the distribution of proficiency test scores was not as balanced in the 2007 study as it had been in 2006 with the majority of cloze scores falling in the lower band this year. Also, WAT scores used for correlations in Table 3 in the 2006 study were average scores from the test and re-test leading to some slight inflation in participant scores, an advantage not enjoyed by the 2007 group. A recalculation of the WAT Test 1 scores with the two proficiency measures in Table 3.2 below bears this out, but only to a very small extent.

Table 6. Correlations between the association scores and the proficiency measures.
The mean WAT scores for both Test 1 and the re-test (test 2) in 2006 are included in the column entitled "2006 mean", while correlations with WAT and test 1 only appear in the next column "2006 test 1". 2007 correlations appear in the "2007" column. In each case, with the exception of number of response correlations with the cloze test, Test 1 correlations are weaker than combined Test 1 and re-test scores.

	Cloze test			Grammar monitoring test		
	2006 mean	2006 Test 1	2007	2006 mean	2006 Test 1	2007
A	$r=0.45^*$	$r=0.45^*$	$r=0.25^*$	$r=0.34^*$	$r=0.33^*$	$r=0.16$ns
B	$r=0.49^*$	$r=0.48^*$	$r=0.41^*$	$r=0.30^*$	$r=0.27^*$	$r=0.29^*$
C	$r=0.32^*$	$r=0.29^*$	$r=0.43^*$	$r=0.16$(n. s.)	$r=0.12$(n. s.)	$r=0.35^*$
$^*p<.05$						
A=number of responses. B=weighted stereotypy. C=non-weighted stereotypy.						

Nevertheless, the lower mean proficiency level of the subjects is perhaps reflected in the data shown in Table 2 where mean scores for all three WAT measures: number of responses, non-weighted stereotypy, and weighted stereotypy are lower in the 2007 study, which could be interpreted as an encouraging sign of a definite underlying pattern in WAT performance. In other words, the free continuous word association test, while not showing much promise as an accurate measure of proficiency, as argued by Kruse, Pankhurst, and Sharwood-Smith in their original probe (1987), does have the power to reflect subject level, albeit quite weakly. For example, the mean non-weighted stereotypy score was 32.2 in WAT Test 1 (2006) but fell to 27.9 in 2007, all very much in line with the lower mean proficiency scores in 2007.

As mentioned earlier, it seems reasonably clear that the cloze test correlates more strongly with the WAT scores than the grammar-monitoring test. Also, the non-weighted stereotypy measure correlates more strongly with cloze scores than the number of responses

measure. The major break in the pattern implied by 1987 and 2006 results is that the weighted stereotypy measure is more effective than the other two measures: number of responses and non-weighted stereotypy. This could be due to improved cue selection where the absence of stimulus words eliciting dominant primary responses or simple, readily accessible sets prevented lower level subjects from scoring as many easy points on the weighted stereotypy measure as they did in the 2006 test. With reference to Table 1, the 2007 non-native subjects' mean score on the number of response measure stood at 89% of the 2006 group. For the non-weighted stereotypy, this was 87% but the mean weighted stereotypy score was only 74% of the mean achieved by the 2006 group.

The criteria for cue selection employed in this 2007 study seems to have been more effective in other ways too. In word association test 1, 2006, the problem of having cues which were unfamiliar to many subjects taking the test, was much in evidence. 14 subjects out of 45 failed to enter any responses for *priest* and 8 provided no reactions to *mutton*. In the 2007 study, only two subjects out of 86 failed to supply any responses to the cue *anger*, and one participant failed to supply any responses to the prompt word *afraid*. In the 2006 study there were also numerous cases of clang responses, such as **priestmentalization*, and responses to *mutton* that suggested that some of the subjects did not know that it was a kind of meat. This is, however, a problem that cannot be eradicated completely. In this study, there were two examples of subjects entering the following responses for *anger* that appeared to be related to age: *generation, young,* and *people* in one case and *old* in the other. The following responses to the cue *sickness*: *happy, love, priceless, smile, good, wonderful, OK* also imply misunderstanding or non-familiarity with the prompt word unless of course the subject intends to supply a list of positive reactions or "negative concept defying" associations. However, it seems more likely that the above associations are for the preceding cue *memory*, and that this subject had continued oblivious to the change in cue word.

While the first five cue selection criteria, (a) to (e), seem to be satisfactory as far as one can tell, some uncertainty surrounds criteria (f) "Is the stimulus likely to generate too many simple, obvious responses?" To a certain extent, most responses to IK words may seem obvious.

It is certainly difficult to explain why some cues seem to be more effective than others, and even more of a challenge to account for the phenomenon of some cues, such as *anger*,

71

proving reasonably effective in one study (2006) with one group of non-native subjects, but not with another (2007). With reference to Table 3, *dream* appears to be the best cue, although this may not prove to be true with different subjects, norms lists, proficiency measures and scoring systems. The possibility remains that the best 8 of the 90 remaining cues from the Kent-Rosanoff list (1910) were not chosen, but there appears to be no way of knowing which ones.

It' s also possible that increasing the number of cue words in the test from 10 to 15 would have helped to isolate better cues, but the reasons for it would likely still remain unclear. There again, extending the test in this way might alter the test conditions with subject stamina or concentration affected towards the end of the test. It is also possible that WAT performance is affected by a presentation order effect, wherein subjects tend to produce more responses to words presented first or last. While *afraid*, the first cue in the 2007 study, does produce a lower mean number of responses 4.93, anger, the ninth cue, produces 4.55. If stamina were a problem, the average number of responses for the 10th. cue *wish* would be lower, but it was not. Certainly there is no apparent bell shape effect in the data in Table 4. Nevertheless, it future tests, candidates should probably be allowed more than 2 practice words.

Window, which usually functions as a concrete noun, a mere piece of glass, is an interesting cue because it appears to be moderately effective, certainly better than the abstract nouns *joy* and *anger*. In this sense, any part of a building, such as *door*, may meet with similar success. On the other hand, *earth* was possibly the second best cue, but it is not clear why. Certainly it evoked the largest mean number of responses (7.93) of all 10 cues. It is tempting to think that this is because it is somehow elemental, or represents a radical concept that cannot be further reduced thus allowing it to trigger a wealth of associations. It certainly has a wide coverage, or displays polysemous qualities, allowing subjects to enter a variety of associations related to planet, world, and soil. If this was the case, then cue words such as *god, air, art, money*, or *fire*, along with *sickness*, might be the right kind of cue for a future experiment.

The first of two final points regarding cue selection criteria concerns cues that tend to elicit a large number of idiosyncratic responses. The issue was discussed in the literature section, but the implications were largely ignored in the cue selection process, largely on

account of the success of *anger* in the 2006 study that also tended to produce a large number of idiosyncratic responses.   However, the evidence in Table 4 suggests that *joy* was a poor choice of cue word.   This is perhaps because it produces the lowest proportion of scoring responses on the non-weighted stereotypy measure (35%) in relation to the mean number of responses entered.   Indeed it was quite common for subjects to enter sets of responses such as: *sports movie TV reading traveling driving* which did not match any responses on the norms lists.

Finally, there are some signs that intra-set criteria need to be considered more carefully. Many subjects, both native and non-native, re-entered the same responses for similar cues such as *fear* for *afraid, spider*, and *anger*.   There also appeared to be some overlap with responses to *joy, memory*, and *dream*.

It now seems increasingly likely that increases in proficiency in learners of English cannot be effectively measured simply in terms of the number of native-like associations produced in a free continuous WAT.   This is simply because a lower level learner can score quite well on this kind of test without having to produce any low frequency responses at all, beyond the two thousand word band.   However, there is some evidence to suggest that, with increases in proficiency, subjects produce more responses, perhaps indicative of either a larger productive vocabulary knowledge or an improved ability to access lexical knowledge. The improved correlations with the weighted stereotypy measure could also be interpreted as a sign that higher-level learners do begin some form of approach to native-like associative behaviour in terms of the organization of the lexicon.   Alternatively, learners may be moving towards the norms of highly proficient Japanese speakers of English, rather than native speakers of English, or towards both at different times.   We need two new norms lists to find this out, and compilation is under way.

References

Jenkins, J. J. (1970). The 1952 Minnesota word association norms. In L. Postman & Keppel (Eds.), *Norms of word association*. New York: Academic Press.

Kent G. H and A. J Rosanoff, (1910) A Study of Association in Insanity. *American Journal of Insanity* 6737-96 and 317-390.

Kiss, G. R., Armstrong, C., Milroy, R., & Piper, J. (Eds.). (1973). *An Associative Thesaurus of English and its Computer Analysis*. Edinburgh: University Press.

Kruse, H, J. Pankhurst, and M. Sharwood-Smith. (1987). A Multiple Word Association Probe. *Studies*

in Second Language Acquisition 9, 2, pp 141-154.

Meara, P. (1992) Yes/no test. Retrieved online from: http: //www.lextutor.ca/tests/yes_no_eng/test_ 1/

Meara, P. (1982). Word association in a foreign language: A report on the Birkbeck Vocabulary Project. *Nottingham Linguistic Circular, 11* (2), 29-37.

Miller, G. and C. Fellbaum. (1991). Semantic networks of English. Cognition 41: 197-229.

Munby, I. D. (2007) Report on a free continuous word association test. *Gakuen Ronshu* no. 132. The Journal of Hokkai-Gakuen University.

Namei, S. (2004). Bilingual lexical development: a Persian-Swedish word association study. International Journal of Applied Linguistics. Vol. 14. No. 3. pp. 363-388.

Wessa, P. (2007). Free Statistics Software, Office for Research Development and Education, version 1. 1. 21-r4, URL http: //www.wessa.net/

Wolter, B. (2001). Comparing the L1 and L2 mental lexicon: A depth of individual word knowledge model. *Studies in Second Language Acquisition, 23,* 41-69.

Wolter, B. (2002). Assessing proficiency through word associations: is there still hope? *System,* 30 (2002), 315-329.