Chapter 7. Knowledge of word associations

Tess Fitzpatrick and Ian Munby

1. Introduction

Meaning	form and meaning	R	What meaning does this word form signal?
		P	What word form can be used to express this meaning?
	concepts and referents	R	What is included in the concept?
		P	What items can the concept refer to?
	associations	**R**	**What others words does this word make us think of?**
		P	**What other words could we use instead of this one?**

Nation lists knowledge of associations as the third of three aspects of meaning which are "involved in knowing a word" (Nation 2001: 27). Evidence of this knowledge, he explains, lies in the answer to the question "What other words does this make us think of?". This apparently straightforward question forms the foundation of word association research, and the rubric for many word association tasks. Importantly, it does not ask what other words mean the same, or sound the same, or can be found in the same place, although the responses to the question might include words connected in all those ways and many others besides. Word association research is compatible with lexical models which use the metaphor of "network" or "web" to describe the organisation of the mental lexicon (Aitchison 2003: ch8, Wilks 2009). The associative links elicited in word association tasks are assumed to represent the strongest and most salient links in individual lexical and semantic networks (Albrechtsen et al 2008: 32) and therefore allow us to identify similarities and variations in these networks between individuals. In Fodor's metaphor of "the mental lexicon [as] a sort of connected graph, with lexical items at the nodes with paths from each item to the other" (1983: 80), word association analyses focus on the "paths" chosen.

2. Developing word association networks

Work by Riegel and Zivian (1972), Politzer (1978), Read (1993), Söderman (1993), Sökmen (1993), Schmitt (1998) and others is indicative of a clear belief among second language researchers that word association patterns can inform us in some way about L2 word knowledge, and about the way in which the mental lexicon operates. However, there is some debate about how these patterns should best be interpreted. Meara (1996a) has described vocabulary knowledge as consisting of three dimensions: size (or breadth), depth, and accessibility (or structure), and word association data have been used at various times to illustrate all three. Politzer (1978), for example, finds that the number of responses given to a cue word increases as an individual's proficiency increases, and so uses his data to glean information about vocabulary size. Word association tasks have also been called upon to shed light on the depth of an individual's vocabulary knowledge (see especially Read 1993 and 1998). Wolter (2001) discusses his word association study findings in terms of both breadth and depth of knowledge, but goes on to suggest that they indicate a difference in structure, too, between the L1 and L2 lexicons. This is a complex notion, though, and Wolter hypothesises that the way in which the lexicon is structured is in fact a function of the quality of word knowledge.

The implication running through the research outlined above is that word association behaviour can tell us about such aspects of the lexicon as size, depth and organisation.

The extensive use of association tasks in investigations of the L1 in childhood (Entwisle et al 1964, Ervin 1961) show that they can also be used to identify developmental changes in the lexicon. Perhaps it is a logical extension, then, to use the same tools to investigate the developing L2 language system, and in particular to draw inferences about proficiency levels from association behaviour.

With a few exceptions (e.g. Fitzpatrick 2009, Riegel and Zivian 1972, Wilks, Meara and Wolter 2005, Wolter 2006), word association tasks have been used in a very specific way in second language acquisition research; to investigate the proficiency of learners. This approach has grown out of the use of associations in first language acquisition research which, in turn, was developed from earlier psychology research and practice. A century ago word association tasks were used as a tool for psychiatric diagnosis, with research focussing very clearly and centrally on the way in which words are connected in prompt-response pairings, and, specifically, on the idea that some of these pairings could be considered normal, or frequent, or predictable (e.g. Kent and Rosanoff 1910). From this developed observations about the ways in which association patterns evolve in early L1 development (see K. Nelson 1977 for an overview), and the ways in which adult L1 users seem to have preferences for certain association types as illustrated by word association lists such as those in Postman and Keppel (1970). This body of research established standards of predictable word association behaviour, and led to the acceptance of certain word association behaviour "norms". These were, for example, that young L1 users tend to prefer syntagmatic responses, adult L1 users tend to prefer paradigmatic responses, and that for many English stimulus words, adult L1 users will tend to give the same responses (e.g. *black>white*, *bread>butter*).

Given these established patterns, it seemed logical for second language researchers to look for ways of evaluating L2 proficiency by comparing response behaviour with that of the L1 user. In other words, if an L2 learner responds to the word *black* with *white*, we might consider him to be more "native speaker like" than the learner who responds *yellow*. Measures of L2 proficiency did not only examine the response items produced; the type of association made was also a focus for many studies. This focus was based on the hypothesis that L2 learners would mirror the observed L1 development pattern, with responses shifting from predominantly syntagmatic to predominantly paradigmatic as proficiency increased. If, as the previous chapter suggests, this knowledge is still developing among learners up to and even beyond the age of 10, then there may be good reason for expecting to see such a pattern of change in L2 knowledge and performance. Politzer (1978) was probably the first explicitly to test that hypothesis, and it is surely no coincidence that his paper appeared at a time when second language acquisition theory was heavily influenced by models of first language acquisition and development (Gass and Selinker 2008: 30). As the hypothesis predicted, Politzer's subjects produced a higher proportion of paradigmatic responses in their L1 than in their L2, and he reports significant but weak correlations between the number of L2 paradigmatic responses and various measures of L2 proficiency. However, this finding was not consistent with previous research (e.g. Davis and Wertheimer 1967), nor with many subsequent studies. Meara (1983), Söderman (1993), Fitzpatrick (2006) and Nissen and Henriksen (2006) have found that L2 users do not necessarily move systematically from syntagmatic responses to paradigmatic responses in the way that L1 users seem to. Other studies (e.g. Sökmen 1993) found that not only response *types*, but also response *items*, did not become

more native-like as proficiency increased, so that even the responses of quite proficient learners were less predictable than those of native speakers.

In many ways this is a surprising finding. Riegel and Zivian (1972) and Read, in his Word Associates test (e.g. 1993), suggest that collocation is an important determinant of response, and a tendency to define also influences association choice. We might expect that both these influences would result in increasingly native-like responses, as proficiency progresses; learners will become more aware of common collocations, and will become more and more likely to know the synonymous items needed for definition-type responses. However, the development of the L2 lexicon is susceptible to other influences too, which might cause it to deviate from L1 patterns of development. Sökmen (1993) and Politzer (1978), for example, consider association behaviour to be closely linked to classroom practice, with Politzer specifically suggesting that the use of drilling techniques in class will increase the tendency towards syntagmatic responses. A further confounding influence on association behaviour is that of cultural input. Kruse, Pankhurst and Sharwood Smith (1987) emphasize this as problematic, and give the example of *apple > gravity* to illustrate the culture-specific nature of some responses. This suggests that observed differences in the response patterns of native and non- native speakers might have as much to do with cultural awareness as with proficiency level. In other words, the response patterns of the most proficient non-native speakers might still differ significantly from those of native speakers.

Frustratingly, it seems that association behaviour can be influenced by all and none of the above. Collocation may well be a strong factor in determining responses, but not to the degree that corpus collocation lists can accurately predict native or non-native speaker responses. Responses often take the form of definitions, but in some cases, where a definition is almost certainly available to the task participant, it is not given. Wolter's mixed findings from his attempt to use word association responses to measure proficiency "like those of past studies, do not support the notion that word associations in a foreign language are clearly linked to proficiency" (2002: 326). He adds, though, that "the results do not seem to suggest that there is no relation at all ... I still believe that a word association/proficiency measure can be developed…".

Multiple attempts have been made, then, to measure proficiency by analysing association responses both by type (e.g. paradigmatic, syntagmatic, clang) and by item (identifying how stereotypical, or how native-like, responses are). Although none of these studies has conclusively identified a clear connection between proficiency and association behaviour, most conclude, like Wolter, that, if appropriately developed and designed, they have the potential to reveal important information about the developing lexicon. Schmitt's paper, in which he proposes an improved procedure for handling word association data, concludes "The use of word associations holds a great deal of promise in the areas of L2 vocabulary research and measurement. This promise has been rather limited by somewhat unsophisticated methodology" (1998: 400). The main methodological components of a word association study are the choice of cue words, the mode of presentation and response (spoken or written, one response or multiple, etc.), and the way in which responses are analysed. The last of these is perhaps the most important in terms of experimental design, and the two main techniques for handling response data are described below.

The L2 word association studies which have their roots in the first language research of the 1960s focus on the types of association made, rather than on the specific items provided as responses. The properties of the associations can be categorized in a number of ways, but the most common categories, especially in earlier studies, are paradigmatic, syntagmatic and clang. In defining these categories we are in fact taken full circle back to Nation's aspects of word knowledge, which he broadly divides into form-based knowledge, meaning-based knowledge and use-based knowledge (2001: 27). Clang responses are form-based in that they are words with phonological similarities to the stimulus word, paradigmatic responses are meaning-based as they are from the same word class and with related meanings, and syntagmatic responses are use-based because they are commonly found alongside the stimulus word in a text. In some studies (e.g. Fitzpatrick 2006 and 2009) the link between classification methods and Nation's word knowledge framework is even more explicit, with categories and subcategories matching exactly those in his framework, such as *collocation* and *word parts*, in addition to *form* and *meaning.*

Typically, studies compare the patterns and changes in these response types for different user groups (e.g. native or non-native speakers) and proficiency levels. Examples of this sort of study include Politzer (1978), Söderman (1993), Nissen and Henriksen (2006), Sökmen (1993), Albrechtsen et al (2008) and Fitzpatrick (2006). Within this strand of study there is a degree of variation in terms of category definitions and parameters. Politzer, Söderman, and Nissen and Henriksen, for example, use the standard three-way (paradigmatic, syntagmatic, clang) classification in their studies. Others, though, have criticised this system as being difficult to use and imprecise in nature. Meara, for example, notes that "I have always found that this distinction is very difficult to work in practice, especially when you cannot refer back to the testee for elucidation" (1983: 30), and Wolter is similarly concerned that "there are always some responses that may quite reasonably (and accurately) be classified in more than one category" (2001: 52). Maréchal addressed this problem by including a category (P/S) "to cover those cases where it is difficult to decide whether a response is paradigmatically or syntagmatically related to the stimulus" (Singleton 1999: 234 citing Maréchal 1995). Other researchers have attempted to devise more transparent and user-friendly classification systems. Sökmen (1993), for example, categorises responses as *collocation, contrast, coordinate, synonym, classification (supra/subordinate), affective, word form,* or *nonsense*. Fitzpatrick (2006, 2007, 2009) models her system on the three-way meaning-based (≈paradigmatic), position-based (≈syntagmatic) and form-based (≈clang) categories, but adds the following subcategories: *defining synonym, specific synonym, lexical set, conceptual association, forward collocation, backward collocation, change of affix, similar in form only*. Choice of classification system is also dependent on the information the researcher wishes to elicit about the mental lexicon. Albrechtsen et al (2008), for example, include a frequency dimension in their somewhat sophisticated categorisation system, listing the following response types: *repetition/translation, form-related, chaining, high frequent non-canonical but semantically related, high frequent canonical, low frequent canonical,* and *low frequent non-canonical but semantically related* (2008:48). The *canonical* responses refer to a further dimension of response analysis, the stereotypy of response. Albrechtsen et al's study is unusual and innovative in that it combines a response-type analysis with a response-item analysis, the second main technique for analysis of association responses.

A second group of word association studies, then, focuses on the fact that certain lexical items have particularly strong connections in the lexicon, and that in many cases language users will share these strong links. Examples of such links in English would be *bread>butter, man>woman, black>white*. These studies typically use lists of native speaker response norms (e.g. the Postman and Keppel lists (1970), the Edinburgh Associative Thesaurus (Kiss et al 1973), the Florida State University norms (Nelson et al 1998)) to determine how "native-like" is the association behaviour of learners. Studies which have compared this "stereotypy" of responses with general measures of proficiency include Randall (1980), Schmitt (1998) and Wolter (2002), all of whom report findings which are inconclusive in themselves, but which, they claim, indicate the potential usefulness of this kind of test. Kruse et al, however, finding only a weak correlation between response stereotypy and proficiency scores, conclude that "word association tests do not show much promise for the specific role created for them in L2 research" (1987: 153). This paper had a rather negative effect on contemporary researchers working in this area, as Meara describes (2009: xii). Meara also observes, though, that with hindsight certain methodological features of the Kruse et al study are revealed as problematic (perhaps this is an example of the sort of "unsophisticated methodology" we have noted Schmitt referring to). In the remainder of this chapter, then, we review Kruse, Pankhurst and Sharwood Smith, and report an original study which is based on theirs, but which attempts to address aspects of their methodology which may have adversely influenced their findings.

Adopting a methodology initially developed by Randall (1980), Kruse et al (1987) investigate the viability of using a multiple response word association test to measure L2 learner proficiency. Their subjects were 15 third year English majors at a Dutch university (Dutch L1) and a control group comprising 7 native speakers of English. For the purposes of the study, a computer program was designed to display and collect a maximum of 12 responses to 10 stimuli: *man, high, sickness, short, fruit, mutton, priest, eating, comfort, and anger* (though data for *man* was erratic and therefore excluded from analysis). These stimulus words were chosen at random, one each from 10 categories of stimuli of different strengths devised by den Dulk (1985) according to the Postman and Keppel norms list (1970), and were intended to improve on the types of stimulus used by Randall. No restrictions were put on response type and informants were instructed to type in all the single English word responses they could think of, up to a maximum of 12 for each cue. The task was administered using a computer programme which allowed participants thirty seconds to type their answers for each cue (excluding actual typing time).

The word association responses were scored in three different ways:

1) Number of responses. This is a straight count of the total number of responses entered for the 9 cue words.
2) Non-weighted stereotypy score. This is a straight count of the total number of responses that match responses listed on the Postman and Keppel norms list (1970).
3) Weighted stereotypy. This is an order-related scoring system from 12 to 1. If a subject provides the response *low* as her first (or primary) response for the stimulus *high*, she scores 144 (12 x 12) because *low* is listed as a primary response on the norms list. If it is her secondary response then the score for

this response would be 132 (11 x 12). If she provides *school* as her fifth response, her score would be 88 (8 x 11) since *school* is listed second on the norms list.

The validity of this test was assessed through a correlation analysis with two language proficiency tests: a cloze test and a grammar error monitoring test. The cloze was a 50-gap test where every sixth or seventh word had been deleted. To assess reliability, the non-native subjects completed the word association test on two separate occasions about two weeks apart, but the control group took the test only once. For the non-weighted stereotypy measure, native speakers' scored higher than either of the non-native group scores, with 25.7 (compared to 23.4 and 22.9). For the other two measures, though, the non-native test time 2 mean score exceeded that of the control group, with 76.8 (tt1) and 82.8 (tt2) for the non-natives and 79.9 for the control group in the number of responses measure, and 1475 (tt1) and 1542 (tt2) for the non-natives and 1509 for the controls in weighted stereotypy (Kruse et al. 1987: 150).

Test-retest correlations were significant, but not particularly high (.76, .66 and .55 respectively for test measures A, B and C), indicating only a moderate consistency of performance across the two test sessions. In order to create a single set of scores for each non-native speaker subject, the two test session scores were combined. Correlations between these scores and proficiency measures were then calculated, with all three test measures correlating significantly ($p<.05$) with the cloze test (Number of responses $r=.441$; Non-weighted stereotypy $r=.547$; Weighted stereotypy $r=.535$), but only the Number of responses measure correlating significantly with the grammar test (Kruse et al. 1987: 151).

The authors describe their results as "disappointing" for four reasons. First, they see no clear difference between native and non-native performance on the test. Second, correlations with the proficiency measures were low. Third, since the highest of those correlations was between the simple "number of responses" measure and the grammar test, there would appear to be no need to measure responses for stereotypy, or quality of response in terms of native-speaker likeness. Finally, the test-retest demonstrated that test performance was not particularly consistent. They conclude by suggesting that factors other than language proficiency, perhaps the effects of cultural background knowledge and intelligence, affect association responses, and that therefore association tasks cannot be used to measure proficiency in a straightforward way.

As hinted at by Meara (2009: xii), it is possible that the conclusions of this study were premature and undermined by a methodology which was flawed in a number of ways. First, not only was the subject group small for a study based on quantitative analyses and aiming for results which could be generalised to a larger population, but also the non-native speaker subjects had studied English through the Dutch education system, and had completed 3 years as English majors at tertiary level. They can therefore be assumed to be highly proficient. This fact somewhat tempers the authors' conclusion that there is no useful difference between native and non-native performance on the word association test. Table 7.1 shows that the native speakers do in fact outperform non-native speakers in all three measures in test 1 (and also if the means of the two test times are used). However, they were not asked to take the test a second time, making it impossible to determine whether the higher test 2 scores for non native

speakers were due to increased proficiency or a test practice effect. The discriminatory power of the word association test is not, therefore, as fully explored as it would have been had the subjects represented more diverse, and less advanced, proficiency levels, and had all subjects taken the test twice.

Second, the format of the task was multiple response, with subjects instructed to provide up to 12 responses to each cue. Responses produced in this format often reveal evidence of "chaining", where subsequent responses are prompted by previous ones, rather than by the cue word. The authors then score these multiple responses against normative data (from Postman and Keppel 1970) drawn from a collection of single (i.e. primary) responses to 100 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. It is precisely responses of this kind that subjects are more likely to provide when confronted by a multiple response testing format. However, the validity of Kruse et al's test depends on the assumption that the lexical retrieval behaviours involved in producing single word responses are identical to those involved in producing multiple responses. The apparently principled weighted stereotypy scoring system is therefore the product of an "immediacy" score from the individual (represented by primary, secondary, tertiary etc. responses) and a "popularity" or "degree of commonality" score from the norms list (representing the percentage of people who produce that item as a primary response). The construct represented by these scores is in fact, then, somewhat opaque, and results in the awarding of a maximum 144 points for supplying a primary response which matches the most frequent response on the norms lists, while only giving one point for a low stereotypy twelfth response.

The use of a weighted stereotypy scale (12-1) also belies the actual distribution of responses on a norms list. For example, a subject who in response to *high*, supplies *low, school*, and *mountain* as first, second and third responses will score 144, 121, and 100 points respectively for each response on the weighted stereotypy scale. This does not reflect the response distribution of these items on the lists (675, 49, and 32). A further problem related to the use of these norms lists is that the norms lists were not contemporary to the study; they were published 17 years before the Kruse et al. study, and indeed were compiled some years before that.

A third issue with this study which requires further exploration is that results may be highly dependent on the stimuli chosen. Meara (1983) points out, for example, that (i) high frequency words generally elicit very similar responses in both L1 and L2, (ii) words such as *high* invariably produce their polar opposites such as *low*, and (iii) high frequency stimuli produce high frequency and rather obvious responses that are unlikely to discriminate between learners of different levels with any sensitivity. Five of the stimuli used by Kruse et al are highly frequent (in the 1st 1000 of the BNC) and at least two have polar opposites, making them, according to Meara's analysis, susceptible to particular association behaviours.

Finally, the authors' interpretation of the low correlations between the proficiency measures and the word association test scores as "disappointing" is perhaps misplaced. The three tests inevitably measure different aspects of language knowledge and use, and strong correlations should therefore not be expected. The cloze test is likely to measure more elements of linguistic competence than the

grammar monitoring (Fotos 1991), including lexical knowledge, and the finding of positive and significant correlations between it and the word association test measures could equally be interpreted as an argument for the validity of the latter.

In conclusion, the results of the study apparently question the usefulness of word association tasks in L2 research, and indeed seemed to dampen enthusiasm for L2 word association research for several years. However, the reservations listed above certainly give us cause to question the authors' interpretation of their findings and their conclusions and, importantly, there are still some useful characteristics of word association tasks revealed here, which may relate to levels of proficiency. In the final section of this chapter we will describe a study which attempts to exploit the strengths of this sort of word association test, while addressing some of its shortcomings, in order to reassess Kruse et al's rather negative conclusions.

3. Current Research Work

The study described here, then, aims to design and test an improved version of the measure used by Kruse et al. by devising an alternative list of prompt words, and using a different norms list for scoring. The degree to which the new version of the test can be used as an indicator of proficiency is assessed using three proficiency measures: a cloze test, a TOEIC test (listening and reading) and a single-word L1>L2 translation test.

3.1 The development of cue words

As noted above, the nature of the cue words used by Kruse et al mean that they provoke particular kinds of association behaviour (e.g. opposites such as *high>low*). One of the aims of our study was to see whether cue words which are selected in a more principled and informed way might help to differentiate more clearly between learners of different proficiency levels. Although there is some tendency for frequently occurring cue words to prompt frequent responses (Meara 1983), we decided that it was important to select cues likely to be known by learners (as opposed to *mutton*, for example, from the original Kruse et al cue list). Cue words were therefore all taken from the 0-1000 band of the BNC (British National Corpus) lists in order to maximise the likelihood that all learners, including low level ones, were able to produce associations to all cues. Each of these 1000 words was screened to determine whether it met the following criteria:

- The word is not likely to produce a "dominant primary" response; specifically, it does not have a polar opposite (e.g. *hot>cold*) and is not the first of a binominal pair (e.g. *food>drink*; *king>queen*)
- The word is not likely to generate hyponyms or superordinates (in the way that, for example, *fruit* might prompt *apple*, or vice versa)
- The word is not a proper noun (some words on the 0-1000 BNC list are proper nouns such as *Germany* and *America*).
- The word is not likely to elicit proper nouns (in the way that, for example, *river* might prompt *Mississippi*, or *ocean>Pacific*).
- The word does not have a phonological equivalent in the L1 (Japanese in this case), or the potential to cause confusion because of the existence of a similar sounding loan word.

125 of the first 1000 BNC words met all the above criteria. In order to minimize the likelihood of similar responses being given for different cues, and of cue words being echoed in responses to other cues, we then discarded cue words with the same popular response, or with a popular response overlapping another cue on the list. To do this we used norms from the Edinburgh Associative Thesaurus (Kiss et al 1973). A popular response was defined as one which accounts for 6% or more of the total responses. For example, *body* stimulates the response *soul* on 10% of occasions, which means that the cue *heart*, producing *soul* on 7% of occasions, cannot be used as a cue alongside *body*. This selection process resulted in the following list of 50 cue words:

AIR	CHOICE	GAS	MEAN	SCIENCE
BEAR	CHURCH	HAPPEN	MOVE	SET
BECOME	CLASS	HEART	NATURE	SHARE
BLOW	CROSS	HOSPITAL	PACK	SORRY
BREAK	CUT	KEEP	PART	SPELL
BOAT	DRAW	KILL	POINT	STAGE
CALL	DRESS	KIND	POLICE	SURPRISE
CASE	FAIR	LEAD	POWER	TIE
CATCH	FIT	LINE	READY	WORLD
CHANCE	FREE	MARRY	RULE	USE

Our task was now to identify the 10 words from this list (to match the number of cues used by Kruse et al) which had the greatest potential to discriminate between learners of different proficiency levels. To do this, we ran a preliminary word association test study with 82 participants (L1 Japanese). Their responses to the 50 cues were scored for stereotypy against a native speaker norms list, and the results for each cue word were compared with their scores on a TOEIC test. The 10 words with the strongest correlations with the TOEIC scores and therefore selected for use in the main study were *air, break, choice, church, heart, keep, lead, pack, police, sorry*.

Our purpose was now to compare the sensitivity (in terms of proficiency discrimination) of these cues with that of the original list from Kruse et al. In order to do this we administered a word association test, alternating cues from the two sets so that any order of presentation effect was minimised. Responses to the 2 cue sets were then separated out again for scoring and analysis.

3.2 Participants

The participants in the main study were 71 Japanese learners of English at tertiary level and included both first and second year students ranging in level from elementary to intermediate. They were presented with the word association test, and instructed to enter up to 12 English responses to each of the 20 stimulus words. They were requested to provide only single word responses, to avoid using dictionaries, and to try to avoid proper nouns or chained responses (where the response is prompted by the previous response rather than by the cue). Participant scores from three additional tests were used as proxies for proficiency level. The first of these was a 50-gap cloze test, similar to that used by Kruse et al. The second was an L1>L2 translation test

adapted from Webb (2008), with 120 single word target items, selected from three frequency bands. The third test comprised two parts of a TOEIC examination (listening and reading) which the students took as part of their university course requirement. These three tests were completed within a week of participants taking the word association test.

3.3 Results

Responses to the Kruse et al cues and to the new set of cues were scored separately. Responses to the original set of cues were scored using the Postman and Keppel norms lists (1970), as in the original 1987 study. As detailed above, though, use of these norms is potentially problematic both because they were compiled from a single-response task, and because they are by now 40 years old. A new norms list was therefore compiled for the new set of cues, by asking 114 native speakers of English to provide five responses to each cue. The word association test was processed and scored in the same way as in the Kruse et al study, except that the "weighted stereotypy score" was excluded due to the problems associated with its calculation, which we discussed above. So, for each cue and each subject, a "number of responses" score was obtained by summing the number of responses given, and a "stereotypy" score reflected the number of responses that matched words on the respective norms lists. The resulting scores were then compared with those from the three proficiency measures, and the correlations between these can be seen in Table 7.3.

Table 7.3**.** Correlations between the word association test scores and the proficiency measures, for the Kruse et al cues and norms, and the Fitzpatrick and Munby cues and norms

Cues and norms	Kruse et al.	Fitzpatrick and Munby	Kruse et al.	Fitzpatrick and Munby
Measures	A	A	B	B
Cloze	.425*	.310*	.520*	.662*
Webb	.533*	.394*	.606*	.676*
TOEIC	.459*	.371*	.534*	.700*

*$p < 0.01$

A= number of response measure, B= stereotypy measure

3.4 Conclusions

Three aspects of these results are worthy of note. Firstly, comparing the correlations between the two word association task measures using the Kruse et al. cues, and the cloze test with the equivalent correlations in Kruse et al.'s original study, we see that they are remarkably similar: .425 and .520 here, and .441 and .547 in the original study. This indicates that the relationship between performances on a cloze test and the word association test relate to each other in a broadly consistent way.

Secondly, all three proficiency measures correlate significantly with all word association test measures. The correlations with the stereotypy scores are consistently stronger than those with the number of response measure. This contradicts Kruse et

al's rather tentative finding that "the [number of] response test would be the best overall predictor of proficiency" (1987: 150) and indicates that the quality of responses, as measured by norms lists (stereotypy), reveals more about a learner's L2 competence than the quantity of responses they produce. The implication here is that, with gains in proficiency, learners of English tend to move towards patterns of native speaker like organization in associative performance. The fact that there is also a positive correlation between the number of responses produced within a time limit and the proficiency measures suggests that learners become more fluent in their response behaviour with gains in L2 ability. This could be because learners at higher levels of proficiency are generally able to demonstrate more fluent, or efficient, accessibility to L2 vocabulary in their lexicons than their lower level peers. These tentative conclusions are consistent with claims that learning an L2 involves the gradual building of lexical networks that approach those of native speakers in terms of structure, dynamics and accessibility.

The main aim of this study, though, was to develop a version of the test presented by Kruse et al. which could better differentiate between learners at different levels of proficiency. The correlations in the two stereotypy columns on the right of Table 7.3 indicate that by using a specifically selected set of cues, and a specifically compiled norms list, the test can indeed be improved to better reflect proficiency, whether the latter be measured in terms of vocabulary knowledge, listening/reading skills, or through a cloze test.

4. Practical implications and suggestions for further research

Taken together, the Kruse et al study and the adaptation of it we present here illustrate well both the promise and the pitfalls of L2 word association research. We opened this chapter with the premise that knowledge of associations is a component of word knowledge, and the significant correlations we find between stereotypy of response and proficiency level support this and encourage us to echo Schmitt's optimism about the "promise" of such studies (1998:400). However, the differences between test results using the same cues and norms lists as Kruse et al, and results using cues and norms lists compiled in a more principled and considered way warn us of the potential pitfalls of such research. Schmitt goes on to say that the "promise has been rather limited by somewhat unsophisticated methodology" (1998:400). This perhaps understates precisely how sophisticated, careful and theory-driven the methodology for word association studies must be; cue words, task type and the compilation and application of norms lists can, as we have seen, all have a powerful influence on scores and findings. Nevertheless, by tweaking the design of the study by Kruse et al which had rejected word associations as unpromising to L2 research, we have, we hope, demonstrated that this strand of research is worthy of further investigation.

By understanding the nature of associational links we can identify their role in lexical processing and lexical retrieval. In this chapter we have investigated these links in the context of second language proficiency. This kind of study, though, helps us to hone methods and theoretical frameworks which can be applied to other conditions in which lexical retrieval is an issue, such as dementia, aphasia and even healthy ageing. Meara has commented that dimensions of word knowledge are "not properties attached to individual lexical items: rather they are properties of the lexicon considered as a whole" (1996: 37); this is perhaps most true of the dimension

"associative knowledge", and future research would benefit from using it to formulate a holistic representation of individual lexicons.

5. Questions for discussion

1) Why might language learners develop networks of L2 word associations which are a) different from other language learners; and b) different from native speakers?

2) Is it beneficial to target the building of word association networks in teaching activities? How might this be done?

3) To what extent can word association responses be used as objective measures of L2 proficiency?

4) How much word association information do you think can be usefully transferred from one language to another?

5) Is word association information in a foreign language something you think could be usefully taught in class? If you wanted to do this, how would you set about it?

Acknowledgement:
The authors are grateful to Paul Meara for designing software specifically to gather data for the word association study presented in this chapter.