PERGAMON

System 26 (1998) 389–401

SYSTEM

Quantifying word association responses: what is native-like?

N. Schmitt*

Department of English Studies, University of Nottingham, Nottingham NG7 2RD, UK

Abstract

Word associations are beginning to be used in the areas of L2 vocabulary research and measurement, but traditional methodology has limited their potential. Three problems in particular have been identified as weaknesses. First, the difference between common and uncommon associations has not been captured by previous methods. Second, traditional methodology has accepted single associations as sufficient evidence of association "knowledge". Third, there has previously been no principled way to determine a threshold where association performance becomes native-like. The new procedure proposed in this paper addresses these problems and results in a four-level descriptive system of association behavior which includes a principled way of determining whether L2 word associations are native-like. This new methodology provides an enhanced way of incorporating word associations into future investigations of vocabulary learning and assessment.

1. Introduction

Eliciting word associations is an established method of probing the mental state and abilities of subjects. Research flowered in the early part of this century with the compilation of a number of word association norming lists (Kent and Rosanoff, 1910; Woodrow and Lowell, 1916; O'Conner, 1928; Schellenberg, 1930). Word associations had a second coming in the 1960s and 1970s, mostly in the field of psychology, where association tests were commonly used to assess the cognitive development and social attitudes and assimilation of L1 children (Szalay et al., 1970; Johnston, 1974; Rierdan, 1980). About the same time, early research into L2 associative behavior began (Riegel et al., 1967; Riegel and Zivian, 1972). More recently,

*Corresponding author. Tel.: +44-115-929-6823; fax: +44-115-951-5924; e-mail: norbert.schmitt@nottingham.ac.uk

0346-251X/98/$19.00
PII: S0346-251X(98)00019-0

researchers interested in nonnative English competence have adopted the procedure in their investigations, in an effort to determine how well English words are known. The method involves finding whether nonnative's associations are similar to those of native-speakers (e.g. Meara, 1980, 1983; Kruse et al., 1987; Read, 1993, 1998). Although using word associations in L2 research and assessment is relatively new, it holds great promise, since much richer information can often be gained from association responses when compared to conventional item types. For instance, Schmitt (1995, p. 114) gives an example of a Japanese subject's responses to *commit*:

commit— together meeting people

Where a traditional vocabulary item would merely show that the meaning of *commit* was not known, the association responses strongly suggest that the source of the problem is a confusion between *commit* and *committee*.

The elicitation of word associations is a relatively simple procedure, which is one of its attractions. Traditionally, subjects are given a stimulus word and asked to produce the first response which comes to mind. For example, the stimulus *massive* typically elicits responses like *huge*, *big*, *large*, and *mountain*. These responses are then matched against a list of norm responses, which have been collected from a large number of target respondents. When assessing nonnative speakers' responses, the norming responses are typically collected from native speakers. If a nonnative speaker's response appears on the norm list of associations, it is considered native-like. The attribute on which L2 responses are judged must logically be native-likeness, because it does not make sense to judge them in terms of *correct* vs *incorrect* or *is an association* vs *is not an association* (any association given, however unusual, is by definition an association, unless the subject is lying).

This traditional procedure is subject to a number of drawbacks, however. Perhaps the most important is that it does not take into consideration the differences in norm responses. Let us examine the responses which 99 British university students gave for *dark*.

light	41	body	1	negro	1
night	16	close	1	quiet	1
fear	4	corner	1	scare	1
black	3	dark	1	see	1
bright	3	darkness	1	shadow	1
room	3	fresh	1	sky	1
ages	2	frightening	1	sleep	1
alley	2	gloomy	1	slow	1
brown	2	god	1	sun	1
bench	1	ground	1	winter	1
blue	1	man	1		

(Edinburgh Associates Thesaurus, Internet resource http://www.cis.rl.ac.uk/proj/psych/eat.html; background information available in Kiss et al., 1973.)

Light is clearly the most frequently given response, certainly much more frequent than *winter* for example, and it would seem uncontroversial to assert that *light* is also a more central, core, or native-like association. The problem lies in the fact that in matching subject association responses with associations on the norming list, no consideration for these differences is usually given. What is needed is a method of weighting the various norm list associations in order to give L2 subjects more credit for producing typical or frequent associations than for producing associations given by one or a few norming respondents.

The second weakness is that traditional methodology relies on a single response to determine a subject's association state for any stimulus word. Although asking for "the first word which comes to mind" is implicitly assumed to tap into the strongest mental connections between words in the mind, subjects will not necessarily give the most "typical" response initially. They may well give an idiosyncratic response first and a very typical one second. Asking for multiple responses gives the subject additional chances to supply these more typical associations, and thus may well be a fairer measure. In addition, even if a subject is able to give a typical response on a single-response task, this is still only one unit of information. On the other hand, providing multiple typical responses would supply a more convincing illustration that the stimulus word is incorporated into a subject's lexicon in a way similar to a native speaker. Thus it can be argued that requiring multiple responses better captures the richness of a subject's association network.

A third weakness is the inability to determine in any principled manner whether an L2 subject's performance on an association task is native-like or not. The situation is fairly straightforward if the subject produces a response like *light* or *night*, but can we be so confident in proclaiming their response native-like if it is *bench* or *fresh*? If a native respondent had an off day and produced a highly unusual association which then appeared on the norm list, would it be sensible to consider an L2 subject's response native-like because it matched the unusual norm association? There is no principled way of deciding whether a norming respondent's associations are actually reasonable or not, resulting in an assumption that all of the norm associations are "natural". Furthermore, the traditional method of matching one subject response to a norm list results in simplistic *native-like* vs *not native-like* decision-making, which has no provision for partial degrees of nativeness.

The degree to which a subject's association response(s) are native-like is bound to fall on a continuum, the same as most other language knowledge or ability. In order to better describe the degree of nativeness, a revised procedure is required which addresses these issues. The rest of this paper will propose and discuss such a procedure, which was developed by the author as one tool to study longitudinal L2 vocabulary acquisition (Schmitt, 1998*a*).

2. Compiling a norming list

The first step in developing a new procedure for quantifying association responses was to compile a new norming list. Several association norms currently

exist (Edinburgh Associates Thesaurus, Internet resource; Postman and Keppel, 1970; Russell and Jenkins, 1954), but they were not felt to be suitable for this study. They are either quite old, based on children's responses, based on single response elicitations, or a combination of these. Also, the above norms fail (with the exception of The Edinburgh Associates Thesaurus) on the principle that norming respondents should be as similar as possible to the subjects to be measured. The subjects the author would eventually measure by the new procedure were nonnative speakers of English studying in British universities, so native-speaking British university students were chosen as the norming group, since they were as similar to the L2 subjects as possible, with the exception of mother tongue.

Once the norming respondents were decided upon, the elicitation instrument had to be developed. A total of 17 stimulus words were selected, a number which piloting showed was at or nearing the maximum respondents were willing to take the time to answer. Eleven came from the above-mentioned longitudinal vocabulary acquisition study, where the main criterion was a degree of polysemy (*abandon, brood, circulate, convert, dedicate, illuminate, launch, plot, spur, suspend,* and *trace*). The other six (*massive, peak, rare, subtle, surging,* and *trend*) were taken from the author's study into TOEFL (TOEFL Practice Tests, 1995) vocabulary items (Schmitt, 1998*b*). Heeding Meara (1983), no high-frequency stimulus words were used, which was another reason for not using existing association norms, because most revolve around high-frequency Kent–Rosanoff stimuli (Kent and Rosanoff, 1910). In order to more fully explore association performance, it was decided to ask for three responses per stimulus word, following Schmitt and Meara (1997). The 17 stimulus words were placed on an instrument with three blanks attached to each word, in the following manner:

abandon _____ _____ _____

The norming group consisted of 27 1st-year students studying Modern English Language at the University of Nottingham, 36 1st- and 2nd-year business students and 28 French majors at Nottingham Trent University, and 9 University of Nottingham students randomly approached on the campus and asked to complete the association task. Thus for each stimulus word, a total of 100 respondents gave associations. The norming respondents were given the instrument with the instructions "Write the first three words you think of when you see each prompt word on the three lines provided." The French majors completed the instrument in class, while all other respondents were allowed to take it home to answer.

The elicitation process usually resulted in 300 responses (100 respondents × 3 responses), but occasionally a response was illegible, which in the worst case (*spur*) brought the total down to 297. These norming responses were first tallied on lists with the criterion that any response with a different orthographic form counted as a separate association (*religion, religions*). Then the lists were condensed by combining words at Level 1 of the Bauer and Nation (1993) morphological hierarchy. This included any base word and its inflections (e.g. *control* + *controls, controlled,* and *controlling*). This was done because these words should all have the same underlying

meaning and therefore be the same association. For example, despite the morphological difference, *religion* and *religions* seem to be the same association for the stimulus word *convert*. However, the situation is not so clear with derivations (*set/setter*, *mood/moody*, *move/movement*). Different members of a word family do not always have the same associations (*massive—huge, attack* √; *mass—huge, attack* ?, *massively—huge, attack* ?), so derivations were counted as separate associations on the lists. This seems to be relatively standard procedure, as even early studies lemmatized association responses (Kent and Rosanoff, 1910). The result at this point was a list of association responses for each stimulus word with a tally of how often each response was given.

3. A procedure for weighting association responses

The three most frequent responses were identified and their frequency of response added together. For example, for the stimulus word *abandon*, the top three responses were *leave* (85), *desert* (28), and *alone* (16). So the "best" performance possible would be to produce these 3 most frequently given associations, which would yield a maximum score of 129 points. But of course few norming respondents gave all three top responses. Therefore, each respondent's score was divided by 129 to gain an *association proportion* figure. Let us take the actual results of a respondent to illustrate the procedure. Her three responses were *neglect* (7), *leave* (85), and *rid* (1). Summing the associations ($7+85+1=93$) and dividing the result by the maximum possible score ($93 \div 129 = 0.721$) leaves us with an association proportion of about 0.72. To further illustrate, let us take the responses from an L2 subject who was later tested using this procedure: *surrender* (0), *hopeless* (0), and *forget* (7) for an association proportion of 0.05 ($7 \div 129$). Instead of simply declaring his first response non-native-like, this procedure has allowed him to demonstrate some nativeness, however minimal. The upshot is that using this procedure, it is possible to derive a numerical score which takes into account the typicality of the association responses.

4. Establishing guidelines for the association proportions

Although the association proportion may capture the typicality of association responses, it is useless without some benchmarks as to what magnitude of association proportion can be considered relatively strong or weak. To provide this guidance, we must look at the behavior of the native-speaking respondents themselves to discover the magnitude of association proportions they achieve. The association proportion for each norming respondent was calculated for each stimulus word. All of these were averaged and the mean association proportion derived, again for each stimulus word. The summary statistics are illustrated in Table 1.

The average mean proportion for all 17 words for all respondents was 0.52. This average figure is not particularly informative however, as the proportions vary quite widely depending on the stimulus word. For stimulus words like *massive* and

Table 1
Association proportion scores for native-speaking respondents

Stimulus words	Maximum raw score	Mean proportion	STD proportion	No. of different associations
Abandon	129	0.68	0.25	92
Brood	75	0.45	0.23	93
Circulate	71	0.43	0.19	101
Convert	139	0.67	0.24	95
Dedicate	63	0.36	0.21	144
Illuminate	112	0.67	0.28	89
Launch	75	0.45	0.23	98
Massive	187	0.71	0.24	47
Plot	75	0.44	0.26	102
Peak	131	0.58	0.21	56
Rare	76	0.47	0.25	88
Spur	95	0.41	0.25	109
Subtle	56	0.36	0.23	129
Surging	62	0.41	0.21	109
Suspend	104	0.59	0.31	102
Trace	99	0.47	0.23	93
Trend	114	0.67	0.28	104
Average	97.82	0.52		97.05

convert, there was a rather high maximum possible score, indicating that there was a high degree of agreement of the respondents' association responses (300 would indicate complete agreement). This also resulted in a rather high proportion score. Other stimulus words, like *dedicate* and *subtle*, elicited a wide range of associations, with the most frequent ones being given by a relatively small number of the respondent group. This resulted in lower figures for the maximum possible score and the mean association proportion score. These results make it fairly clear that different stimulus words elicit different group association behavior. It is thus difficult to formulate any blanket association proportion criterion which would work for any stimulus word. This suggests it is probably necessary to use norming data collected for each individual stimulus word to evaluate subject responses.

The mean association proportion gives us something on which to base our interpretation of responses given by nonnative speakers, since we now have some idea of native respondents' behavior. The continuum from 0 to 1.00 for possible association proportions can be broken most reasonably into three obvious levels, at least initially. First, if no responses are given which match those on the norming list (0 score), that would indicate the subject has demonstrated no native-like associations for that stimulus word. Second, if several very common responses are given, then the word association performance for a particular stimulus word can be considered equivalent to that of an average native-speaker. L2 subjects who achieve an association proportion equivalent to or higher than the native norming group mean are clearly at this level. A third level in which the associations are partially, but not typically, native-like exists between the first two.

As is usual with most clines, the extremes of the association proportion continuum are easy to define, but this leaves the really interesting question of whether we can set a threshold criterion for when associations become native-like. The threshold must exist somewhere between 0.00 and the mean association proportion, but before we can place it more precisely, we must define more clearly what we will accept as native-like. This definition must take two things into consideration. First, there is a great deal of agreement among responses given by native speakers. Second, and conversely, native speakers also typically give a number of idiosyncratic responses. However, in a three-response task, the distribution of the idiosyncratic responses is quite interesting. The idiosyncratic responses were tallied for each of the 1700 cases (17 words × 100 respondents), and it was found that only 28 (1.65%) included three idiosyncratic responses to a stimulus word, 187 (11.0%) had two, 573 (33.71%) included one, and 912 (53.65%) had none. From these figures it can be seen that giving three idiosyncratic responses is not at all typical of this norming group, and giving two is not all that common either. On the other hand, it was quite usual for native speakers to give one unique response, and producing no unique response was the most common behavior of all.

These results suggest the following approach concerning the definition of native-likeness. While it is true that native speakers occasionally give three unique responses to a prompt word, this is an unusual situation. Thus it would be unrealistic to use this as a minimal threshold of native-likeness, especially considering the level of commonality among most native responses. It is more reasonable to take the native *group* behavior as the criterion instead of any individual native speaker, some of whom may not be typical of the norming group overall. So while it is impossible to state that giving three idiosyncratic responses is not native-like, it is possible to say that this behavior is not typical of the norming group. This suggests that the best way to develop a threshold of native-likeness is to describe what is *typical* of a native norm group as a whole, but with the caveat that a very limited number of the native speakers will themselves be defined as atypical.

If we decide that individual native respondents who gave three idiosyncratic responses (and perhaps others with very low association proportions) are not really typical of native speaker performance, then we need some way of determining a native-like threshold which lies above their scores. One could manually examine all responses to find which respondents gave only three idiosyncratic ones and then set the threshold just above their score. This is a principled method, but it does not address the problem of respondents who achieved only a slightly higher association proportion, for example, two idiosyncratic responses and another with a value of 2. A respondent with such a low score would not really be typical of the group behavior either. Determining the threshold by intuitively deciding which sets of association responses are typical and which are not is obviously too subjective, as well as being very time-intensive. The best and most principled method of setting the threshold seems to be the use of descriptive statistics. One could take the mean association proportion and subtract one standard deviation to derive a figure which would disregard approximately the bottom one-sixth of the respondent performances. While this method worked satisfactorily for some words, it usually cut too

many performances off which were clearly still in the mainstream of group performance. On the other hand, subtracting two standard deviations set the threshold too low. After some trial and error, it was found that subtracting 1.5 standard deviations succeeded in eliminating the responses which seemed atypical (three unique responses and others which added up to a very low association proportion), while not discarding too many responses which seem more in line with group behavior.

To illustrate this, let us take the responses for the stimulus words *abandon, dedicate*, and *rare*, which represent the words with the highest, lowest, and central mean association proportions. They have mean proportions of 0.684, 0.356, 0.467 and standard deviations of 0.252, 0.211, 0.248, respectively. The respondents with the lowest association proportions for each word are as follows:

abandon **(Idiosyncratic responses in bold)**

1a.	**house**	**junkyard**	**bus**	0.02
2a.	hope	**game**	**car**	0.04
3a.	child	**refuge**	**homeless**	0.05
4a.	**kitten**	child	**family**	0.05
5a.	loose	**trip**	**do**	0.05
6a.	wild	gay	hope	0.07
7a.	neglect	**ignore**	**redundant**	0.07
8a.	ship	hope	desolate	0.08
9a.	lost	**Moses**	child	0.09
10a.	lonely	alone	**frightened**	0.17
11a.	lost	left	alone	0.19
12a.	lost	island	alone	0.19
13a.	**jettison**	reject	desert	0.25
14a.	loose	desert	**sacrifice**	0.26
15a.	desert	isolate	alone	0.36
16a.	leave	**wreck**	**destroy**	0.67

dedicate

1b.	**follower**	**poet**	**athlete**	0.05
2b.	**resolve**	discipline	**follow**	0.06
3b.	**motivation**	**selfless**	sacrifice	0.06
4b.	**message**	hardworking	**reward**	0.06
5b.	involve	**value**	**constant**	0.06
6b.	body	**mind**	**soul**	0.06
7b.	**hard worker**	study	**student**	0.06
8b.	tribute	**football**	**Bryan Robson**	0.06
9b.	**no change**	**stuck with**	heart	0.06
10b.	**work hard**	challenge	**success**	0.06
11b.	determined	**win**	**succeed**	0.06
12b.	assign	**allocate**	**sign**	0.08

13b.	baptise	**christen**	**religion**	0.08
14b.	passion	**supported**	determined	0.08
15b.	sacrifice	**affections**	loyal	0.10
16b.	**trust**	monument	poem	0.11
17b.	commitment	**D.J.**	**request**	0.13

rare

1c.	**none**	**kill**	**melt**	0.04
2c.	**unpopular**	distinctive	**inspiring**	0.05
3c.	one	never	**whole**	0.08
4c.	meat	**lonely**	uncooked	0.11
5c.	**banana**	rabbit	different	0.11
6c.	**one off**	strange	meat	0.11
7c.	species	**occurrence**	**good men**	0.13
8c.	old	expensive	**original**	0.13
9c.	old	never	expensive	0.16
10c.	animal	**ivory**	**diamond**	0.16
11c.	exotic	**luxury**	precious	0.17
12c.	animal	disease	infrequent	0.18
13c.	meat	breed	expensive	0.18
14c.	bloody	extinct	limited	0.20

Abandon is representative of words which have a high level of agreement among the subjects' responses. It has an exceptionally frequent primary (*leave*) which was given by 85 out of the 100 respondents. Thus the respondents are essentially split into two groups, those who gave *leave* and those who did not. Since such a high percentage of native-speaking respondents produced *leave*, the criteria for native-likeness for the stimulus *abandon* would ideally require this response. Using the formula of *mean association proportion−1.5 STDs* results in a cut-point of 0.306. Reaching this score doesn't necessitate giving *leave*, with the production of the secondary and tertiary responses *desert* or *alone* being sufficient for that (13a). Any subject giving *leave* will clearly be above the threshold however. There are some subjects who gave what appear to be quite reasonable associations (i.e. 6a and 8a), but who would not make the native threshold using the "1.5 STD" procedure. We would have to set the threshold at about 0.06 to include these and still limit the more idiosyncratic performances. Using less strict criteria, such as subtracting 2.5 STDs to achieve a lower threshold would accomplish this, as would simply shaving off the bottom 5 scores to eliminate them. But as we will see, these methods will not work with stimulus words with lower communality.

At the other end of the spectrum, *dedicate* is a stimulus word with a great deal of diversity of response. There are a large number of idiosyncratic responses, and consequently numerous very low association proportions. Looking at Examples 1b–17b, it becomes clear that it is quite common to give two idiosyncratic responses to this low communality prompt word. There is no large "jump" in association

proportion as there was for *abandon*, just a gradual increase. Thus the only respondent whose performance is arguably not typical is 1b, who gives three idiosyncratic responses. The "1.5 procedure" provides a cut-point figure of 0.040, which does not exclude the 1b responses. However, if we consider what a nonnative respondent must achieve in order to reach this threshold, it still seems to work reasonably well. The maximum raw score for *dedicate* is 63, which means that a nonnative matching only one or two idiosyncratic responses on the norming list would not reach the native-like threshold ($1 \div 63 = 0.016$; $2 \div 63 = 0.032$). But if the nonnative matches three unique responses ($3 \div 63 = 0.048$), their performance is considered native-like. This threshold may seem rather low, but given the degree of diversity and idiosyncrasy of the natives, setting it any higher would not accurately reflect native behavior. Coming back to the point made in the previous paragraph, it is clear that the possible alternative solutions offered for *abandon* earlier will not work here.

The example of a word in the middle of the association proportion range is *rare*. In this case the suggested procedure works quite well, excluding the one subject with only idiosyncratic responses, as well as two other low scores.

The end result is that subtracting 1.5 STDs from the mean association proportion provides a threshold of native-likeness that seems to perform well in the mid-part of the association proportion spectrum, and that also gives reasonable cut-points for words at the high- and low-communality extremes. This is in spite of the fact that association proportion distributions are usually oddly skewed. The procedure may classify more native respondents as atypical than might be hoped for, but it does succeed in providing a threshold high enough that one can be confident in the native-likeness of those who achieve it. Analyzing the procedure over all 17 stimulus words shows that the procedure is not ideal in every case, but suggests that it does provide a workable solution to the problem of developing a weighting standard which performs reasonably well for stimulus words with widely varying association behavior.

One problem in setting thresholds and cut-points is that respondents with figures just above the threshold are seldom so dissimilar from respondents with figures just short of the threshold, making it difficult to categorize the two as different. This is sometimes the case with this procedure, but by the sixth respondent, the clearance is normally at least +0.05, and usually much more. Table 2 shows the association proportions of the six respondents above, but nearest, the threshold.

5. A four-level description of native-likeness

Adding the native-like threshold to the previous three levels, we can now define a principled four-level description of how native-like L2 subjects' responses are.

5.1. Level 0

- Association proportion = 0.
- Produced no native-like associations.

Table 2
Lowest five native association proportions which cleared native-like threshold level

Stimulus words	Threshold level	Respondents nearest, yet above, threshold level					
		1	2	3	4	5	6
Abandon	0.306	0.36	0.67	0.67	0.67	0.67	0.67
Brood	0.094	0.11	0.13	0.13	0.15	0.15	0.16
Circulate	0.148	0.15	0.15	0.15	0.17	0.17	0.20
Convert	0.307	0.63	0.63	0.63	0.63	0.63	0.64
Dedicate	0.040	0.05	0.06	0.06	0.06	0.06	0.06
Illuminate	0.253	0.33	0.36	0.71	0.71	0.71	0.71
Launch	0.107	0.13	0.15	0.16	0.16	0.16	0.16
Massive	0.343	0.37	0.42	0.42	0.42	0.42	0.43
Plot	0.059	0.07	0.08	0.09	0.09	0.09	0.09
Peak	0.266	0.28	0.28	0.32	0.42	0.42	0.44
Rare	0.096	0.11	0.11	0.12	0.13	0.13	0.16
Spur	0.033	0.05	0.05	0.06	0.06	0.07	0.07
Subtle	0.021	0.05	0.07	0.07	0.07	0.07	0.09
Surging	0.094	0.10	0.10	0.13	0.13	0.13	0.15
Suspend	0.129	0.13	0.15	0.15	0.15	0.15	0.16
Trace	0.121	0.13	0.14	0.22	0.22	0.22	0.23
Trend	0.250	0.26	0.72	0.72	0.72	0.72	0.72

5.2. *Level 1*

- Association proportion = > 0 and < threshold proportion.
- Produced one or more associations which appear on the norming list, but not ones which are typical. Thus the association responses overall are not yet typical of the native norming group.

5.3. *Level 2*

- Association proportion = ≥threshold proportion and < mean association proportion.
- Native-like productive association performance.

5.4. *Level 3*

- Association proportion≥mean association proportion.
- Native-like productive associations similar to those of the top portion of the native norming group.

Note that both Levels 2 and 3 are labeled as native-like. Considering that a large portion of the native respondents fall under the mean association proportion and into Level 2 (and a few even into Level 1), it is probably unwise to argue that Level 3 performance is any more native-like than Level 2 performance. However, a Level

3 performance necessarily includes more of the most commonly given responses than Level 2, so we can be even more *confident* in labeling a Level 3 performance as native-like.

6. Conclusion

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. The descriptive procedure proposed in this paper has several advantages over previous methods of determining the nativeness of association responses. First, it quantifies the association responses in a way which results in a tangible figure being produced. Second, it takes into account the differences in typicality of association response. Third, the description of association performance is based on more than a single unit of information. Finally, the procedure provides a principled way of determining whether any association performance is native-like or not, with group typicality as the criterion.

This procedure has already proved of use in research into L2 vocabulary acquisition (Schmitt, 1998*a*) and an analysis of TOEFL vocabulary items (Schmitt, 1998*b*). Moreover, there is no reason why word association procedures like this, with further study and refinement, could not be used as an alternative way to test vocabulary. The proposed methodology will hopefully prove useful in future enquiries into both vocabulary learning and measurement.

Acknowledgements

The initial idea for this procedure was first generated in a brainstorming session with Paul Meara. He also offered insightful comments, along with John Read, on an earlier version of this paper. Comments from members of the Language Testing Research Group at the University of Lancaster, particularly Charles Alderson and Caroline Clapham, helped to sharpen my thinking on associations prior to embarking on this study. Thanks to Sheridan Graham, Barry Harrison, Hillary Hillier, Michael McCarthy, and Anoma Siriwardena for facilitating the data collection.

References

Bauer, L., Nation, I.S.P., 1993. Word families. International Journal of Lexicography 6, 1–27.

Johnston, M.H., 1974. Word associations of schizophrenic children. Psychological Reports 35, 663–674.

Kent, G.H., Rosanoff, A.J., 1910. A study of association in insanity. American Journal of Insanity 67, 37–96, 317–390.

Kiss, G.R., Armstrong, C., Milroy, R., Piper, J., 1973. An associative thesaurus of English and its computer analysis. In: Aitkin, A.J., Bailey, R.W., Hamilton-Smith, N. (Eds.), The Computer and Literary Studies. Edinburgh University Press, Edinburgh, pp. 153–165.

Kruse, H., Pankhurst, J., Sharwood Smith, M., 1987. A multiple word association probe in second language acquisition research. Studies in Second Language Acquisition 9, 141–154.

Meara, P., 1980. Vocabulary acquisition: a neglected aspect of language learning. Language Teaching and Linguistics Abstracts 13, 221–246.

Meara, P., 1983. Word associations in a foreign language: a report on the Birbeck Vocabulary Project. Nottingham Linguistic Circular 11, 29–38.

O'Conner, J., 1928. *Born That Way*. Williams and Wilkins, Baltimore.

Postman, L., Keppel, G., 1970. *Norms of Word Association*. Academic Press, New York.

Read, J., 1993. The development of a new measure of L2 vocabulary knowledge. Language Testing 10, 355–371.

Read, J., 1998. Validating a test to measure depth of vocabulary knowledge. In A. Kunnan (ed.) *Validation in Language Assessment* pp. 41–60. Mahwah, NJ: Lawrence Erlbaum.

Riegel, K.F., Ramsey, R.M., Riegel, R.M., 1967. A comparison of the first and second languages of American and Spanish students. Journal of Verbal Learning and Verbal Behavior 6, 536–544.

Riegel, K.F., Zivian, I.W.M., 1972. A study of inter- and intralingual associations in English and German. Language Learning 22, 51–63.

Rierdan, J., 1980. Word associations of socially isolated adolescents. Journal of Abnormal Psychology 89, 98–100.

Russell, W.A., Jenkins, J.J., 1954. The complete Minnesota Norms for responses to 100 words from the Kent–Rosanoff word association test. University of Minnesota, Department of Psychology, Tech. Rep. No. 11.

Schellenberg, P.E., 1930. A group free-association test for college students. Unpublished doctoral dissertation, University of Minnesota. Published in an abridged form in Tinker, M.A., Baker, K.H., 1938. *Introduction to Methods in Experimental Psychology*. Appleton-Century, New Yvior 6, pp. 213–214.

Schmitt, N., 1995. Verbal Suffix and Word Association Knowledge of Japanese Students. Unpublished MPhil Thesis. University of Wales, Swansea, UK.

Schmitt, N., 1998a. Tracking the incremental acquisition of second language vocabulary: A longitudinal study. Language Learning 48(2), 281–317.

Schmitt, N., 1998b. The relationship between TOEFL vocabulary items and meaning, association, collocation, and word class knowledge. Language Testing (in press).

Schmitt, N., Meara, P., 1997. Researching vocabulary using a word knowledge framework: word associations and verbal suffix knowledge. Studies in Second Language Acquisition 19, 17–35.

Szalay, L.B., Windle, C., Lysne, D.A., 1970. Attitude measurement by free verbal associations. The Journal of Social Psychology 82, 43–55.

TOEFL Practice Tests, 1995. Educational Testing Service, Princeton, NJ.

Woodrow, H., Lowell, F., 1916. Children's association frequency tables. Psychology Monographs 22(5), No. 97.