

A Comparative Study on Long-Term Vocabulary Recall: Word Lists and a Vocabulary Trainer

การศึกษาเชิงเปรียบเทียบเกี่ยวกับการฟื้นความจำคำศัพท์
ในระยะยาว ระหว่างการใช้รายการคำศัพท์และตัวฝึกเรียนคำศัพท์

Thomas Krey*

ictthomas@mahidol.ac.th

Abstract

This study was conducted in order to examine recall rates on previously learned vocabulary. The 2 methods investigated were word lists for exams and a vocabulary trainer using flashcards. In order to gain better insights in the effectiveness of these methods participants were tested for both treatments in a repeated measure design over a period of 6 months. The main focus was on recall rates for previous recall failures. Results show significantly higher recall scores for the vocabulary trainer-method at every stage. The gap in recall scores between both methods was most noteworthy in the field of never accurately recalled items. Percentages of omission errors were also remarkably higher on word lists, though by a smaller margin. However, the gap in interference errors was less significant. The implementation of special mistake review sessions after the previous trial, an indispensable element of the vocabulary trainer-method, appears to have influenced the differences. The conscious raising effect of self-production of cards for test items seem to have contributed to the outcome as well. Based on the findings individual adaptations for vocabulary learning strategies could be advisable.

Keywords : long-term memory, retrieval, spacing effect, vocabulary trainer, word lists

บทคัดย่อ

การศึกษานี้กระทำขึ้นเพื่อตรวจสอบถึงอัตราการฟื้นความจำต่อคำศัพท์ที่ได้เรียนไปแล้ว การศึกษานี้ได้ใช้วิธีการสองอย่าง คือการใช้รายการคำศัพท์เพื่อการสอบ และบัตรคำศัพท์จากตัวช่วยฝึกเรียนคำศัพท์ เพื่อเข้าใจถึงประสิทธิภาพของวิธีการทั้งสองนี้ ผู้เข้าร่วมศึกษาได้รับการทดสอบโดยใช้วิธีการศึกษาทั้งสองอย่าง โดยกระทำชำๆ กันหลายครั้งในตลอดหกเดือนที่ผ่านมา ชุดเน้นสำคัญอยู่ที่อัตราการ

* Lecturer At Mahidol University International College, Salaya Campus

พื้นความจำคำศัพท์สำหรับความลืมเหลวในการจำครั้งก่อนหน้านี้ ผลการศึกษาแสดงให้เห็นอย่างนี้ นัยยะสำคัญว่าในทุกขั้นตอนของการใช้ตัวฝึกเรียนคำศัพท์นั้นมีอัตราการพื้นความจำที่สูงกว่า ช่วงระหว่างคะแนนในการรือพื้นความจำระหว่างสองวิธีมีนัยยะสำคัญมากในเรื่องการไม่เคยจำคำศัพท์ได้ถูกต้องเลย อัตราส่วนเป็นปีอร์เซ็นต์จากข้อผิดพลาดในการจำคำศัพท์ไม่ได้เลย จากการใช้รายการคำศัพท์เพียงแค่ครั้งเดียว มีความสูงกว่าอั้งมีนัยยะสำคัญ แม้ว่าข้อแตกต่างจะมีเพียงเล็กน้อย อั้งໄร์กีตานช่องว่างจากความผิดในเรื่องการจำสับสน ไม่ค่อยมีนัยยะสำคัญเท่าใดนัก การได้ใช้เวลาในการทบทวนข้อผิดพลาดหลังจากการทดลองในครั้งที่แล้ว ถือเป็นหลักการสำคัญของตัวช่วยฝึกเรียนคำศัพท์และเป็นสิ่งที่ทำให้ผลลัพธ์ของห้องสองวิธีมีความแตกต่างกัน การเข้าใจในสิ่งที่ศึกษาอย่างถ่องแท้ก็ยังเกิดขึ้นจากการที่ผู้เรียนได้ทำการสร้างบัตรคำศัพท์ขึ้นมาใช้ด้วยตนเองอีกด้วย ดังนั้นจากการศึกษา กลยุทธ์และวิธีการศึกษาคำศัพทน่าจะได้รับการแนะนำต่อการใช้ในแต่ละบุคคล

คำสำคัญ: ความจำระยะยาว, การสืบค้น, ช่วงห่างของเวลาในการเรียนรู้แต่ละครั้ง, ตัวช่วยฝึกเรียนคำศัพท์, รายการคำศัพท์

1. Introduction

“If you want to forget something, put it on a list.” (Stevick, 1976, cited in Lewis, 1993). This quote reflects a widespread negative attitude among researchers against wordlist learning. A deeper analysis of research articles reveals that critical points of view towards those lists proceed from two different perspectives.

One group of researchers claims the superiority of a ‘*natural approach*’ and ‘*language acquisition*’ and opposes explicit instruction of vocabulary in general (see Judd, 1978; Krashen, 1982, 1989). Direct, intentional vocabulary learning is considered superficial, and translation in general is seen as ‘a crutch’ (Rivers and Temperley, 1978), because many L2 – L1 ‘*paired-associates*’ are not considered truly equivalent. It is also claimed that translation does not address

properly the issue of the polysemous feature of many lexical items. This position towards explicit language learning became very influential in the 70s and 80s of the last century. Consequently, learning by word lists was considered inefficient and became less prominent in curriculum development and language book design.

A second group of researchers does not oppose intentional, direct vocabulary learning as one learning method among others. ‘*Learning by context*’ is seen as rather slow, as it does not address the necessity for beginners to quickly develop a large basic vocabulary (see Meara, 1980, 1995; Laufer, 1991; Nation, 2001; Waring, 2002).

Consequently, the effectiveness of bilingual word lists has been compared in several studies with methods using L2 contexts. Nation (1982) reviewed research

on the usage of ‘paired associates’. His main conclusion is that foreign language learners can acquire many new lexical items by this method in quite a short time. Similar conclusions are drawn by other researchers (e.g., Carter and McCarthy, 1988; McCarthy, 1990; Clipperton, 1994; Folse, 2004 and Waring, 2004). In several studies, higher scores were reported for de-contextualized word lists than context of different levels of elaboration (e.g., Dempster, 1987; Laufer and Osimo, 1991; Prince, 1995; Laufer and Shmueli, 1997b; Groot, 2000; Herman, 2003; Folse, 2006).

The referenced results and theoretical conclusions clearly contradict one-sided positions against direct vocabulary learning in general and list learning in particular. However, a closer look into the results shows the limitations of advantages for list learning to short range effects, most impressive in immediate post tests (see Crow and Quigley, 1985; Groot, 2000). Despite the referenced benefits, word list usage is not only criticized by advocates of a ‘natural approach’, but as well by many researchers who do propose direct vocabulary learning in general. Critical aspects in the research literature were found as follows:

1. Word lists can tempt learners to just read word pairs rather than recalling them from memory by covering the target side (see Landauer and Bjork, 1978).

2. Lists have a fixed word order,

in which each item has to be learnt. Some items are persistently unknown or forgotten, while others are known without effort. Thus, list learning leads both to overexposure and underexposure (see Nation, 2001).

3. Learners can often remember the answers to the following words on a list, even before they encounter them again, because they can remember the position of items. Particularly, items at the beginning and the end of a list are learned better than items in the middle due to the ‘*primacy and recency effect*’ (Baddeley, 1990).

4. Word lists are often presented in alphabetical order. This structure can lead to ‘*serial order effects*’ (Waring, 2004) and to interference with ‘*synforms*’ (Laufer, 1997a), i.e. morphologically similar lexical items.

5. Some word lists are structured in lexical sets. Study results show that this structure can lead to inhibition of vocabulary learning due to semantic interference with competing items. (Tinkham, 1997).

6. Word lists can lead to two-tier memory structures. “Sometimes learners remain forever unable to recall some items and those items that can be recalled are always recalled” (Mizuno, 1999).

7. The usage of word lists is often unit and exam related. Once the unit or exam is passed, many items are discarded from memory. Though cramming might be partially helpful for short-term goals, the pace

of forgetting is fast (Thalheimer, 2006).

8. Rigid word lists neither address interpersonal differences among learners nor do they tap learners' motivation to participate in vocabulary collection.

Regarding alternative strategies, many of the referenced researchers favor flashcard usage over word lists as an initial step in the incremental process of learning vocabulary (e.g., Schmitt and McCarthy, 1997; Hedge, 2001; Nation, 2001; Thornbury, 2002; Waring, 2004). The main advantages are seen as follows:

1. Improved facilitation of retrieval (Nation, 2001).

2. A learner cannot remember an item by its position, because packs of cards are easily shuffled; this effect also enhances retrieval efforts (Waring, 2004).

3. Flash cards facilitate the control of repetition procedures as it is easier to separate known items from incorrectly recalled or unknown items than on a rigid word list (see Nation, 2001).

However, in order to gain the referenced benefits of flash cards one has to use them in a productive way, because no learning method is good in itself. Suggestions in the research literature were found as follows:

1. Produce flashcards based on a thoughtful selection of appropriate items (not just a random collection).

2. Self production is considered a

core principle of this method (e.g., Mondria and Mondria-de Vries, 1994; Schmitt and Schmitt, 1995; Waring, 2004). The outcome of several studies indicates deeper processing strategies by writing than by reading-only approaches (e.g., Thomas and Dieter, 1987).

3. Each item is recommended to be learnt in '*chunks of a language*' (Ellis, 1997) and not as an isolated word, presenting the target item with typical collocations or as less fixed lexical units is seen to enable the possible benefits of the '*lexical approach*' (Lewis, 1993).

4. A comprehensive '*relearning schedule*' (Pimsleur, 1967), based on findings in learning theory has to be developed. For example, the '*expanded rehearsal strategy*' by Bjork (1978) proposes staggered longer intervals for accurately recalled items and shorter intervals for unknown or forgotten items. A similar approach, called the '*Low-First Method*' was developed for CALL by Mizuno (1999). It comprises three principles:

First: a flexible rehearsal system based on individual item retrieval rates. Thus, forgotten or partially known items are reviewed more frequently than known items.

Second: the omission of items successfully retrieved two sessions in a row. Third: the transition to a new learning session when only few items are left

unrecalled in the previous session. Experimental studies, conducted by Mizuno (1999, 2002) show superior recall rates for the combined three principles over any approach using only one or none of them.

The literature in favor of direct vocabulary learning shows strong support for the referenced ways of using flashcards. However, many advocates of this method remain silent on what learners should do in the process of compiling hundreds of word cards. How should learners store those cards and separate known items from unknown ones? How should they address repeated recall failures and the dynamic relationship between retention and forgetting and organize distinguished intervals for correctly and incorrectly recalled items? In order to overcome these problems some researchers suggest using dynamic 'vocabulary trainers' (e.g., Mondria and Mondria-de Vries, 1994; Cobb, 1998; Waring, 2004).

This learning method was originally devised some 35 years ago by the *psychologist* Leitner (1972) and since then it has grown quite popular in western countries, both in a card file box format and increasingly in software applications. The card box version is separated into several compartments of staggered widths. It is not only a container for flashcards, but also a dynamic review system in which correctly recalled items are promoted to the next compartment, while incorrectly recalled

items have to be demoted to (or kept in) the first compartment. Items in the first compartment are reviewed the most frequently, items in the last compartment are reviewed the least frequently.

In addition to the referenced benefits of flashcard usage, the advantages of vocabulary trainers are seen as follows:

1. They help to keep scores of cards in order by storing them in a systematic way.
2. They facilitate both the organization of re-encounters and retrieval of items. This combination of distributed rehearsal sessions and retrieval practice is seen as a very effective tool for memorizing vocabulary (Bjork, 1978). It is also seen as reducing interference induced errors and "error fossilization" (Selinker, 1972).
3. They are considered a time saving learning tool, as only persistently forgotten or unknown items are reviewed frequently. Thus, overexposure to known items is avoided, while conscious raising strategies for unknown or forgotten items can be applied (Mizuno, 1999).
4. These card file boxes are designed to give clear feedback about one's learning progress. A learner can visually estimate their progress in centimeters.
5. The increasingly wider compartments are arranged in order to conduct review sessions in staggered intervals. By this means the widely recognized benefits of the

'spacing effect' (Dempster, 1988; Thalheimer, 2006) are used.

6. File card boxes and their electronic equivalents are seen as more flexible than rigid word lists. Items can be easily shuffled in order to avoid the *'serial effect'* (Waring, 2004). Items that are consistently recalled accurately can be omitted and new lexical items can continuously be added to the storage system.

7. Learners can use the method in a more individualistic way. They can choose the number of items learnt at a time, the compartment of the card file box they wish to review, the spacing of intervals for review sessions etc. This individual approach to learning is also seen as enhancing motivation (e.g., Mondria and Mondria-de Vries, 1994).

However, both in research theory and educational practice the state of research concerning both methods (flash cards and word lists) seems inconclusive. On the one hand, the usage of flash cards is often recommended, but on the other hand, support by experimental, comparative studies for their advantages could not be found. While scores of vocabulary trainers are offered by sales companies as a *'miracle cure against forgetting'* (vtrain; Supermemo; Memorylifter etc.), the benefits of their practical usage remain unclear. Additionally, while word lists are shunned by many researchers, they are still popular with language learners

and remain integrated in several language books and courses.

Little is also known about the long-term effects of different vocabulary learning strategies, as most studies are confined to acquisition of new items over rather short periods of several days or weeks. An exception are the long-term studies of Bahrick et al. (1984, 1987, 1993). Moreover, the literature research did not reveal a single study comparing the effectiveness of the two methods under scrutiny. This might be understandable, as the effects of direct, intentional vocabulary learning are often examined in comparison to indirect, incidental learning and the studies do not focus on the particular technique used. However, the lack of research seems surprising, as many advocates of flash card usage emphasize its benefits in opposition to rigid word lists without producing any experimental evidence. Thus, it was decided to shed more light on the claimed advantages of vocabulary trainers, the suggested disadvantages of word lists, as well as their respective long-term effects and to start experimental research beyond administering questionnaires and in-class observations. Based on arguments of previous research the following hypothesis is proposed, concerning the usage of the two investigated methods of vocabulary learning: Long-term vocabulary recall scores will be higher with the vocabulary trainer method than the word list method.

2. Method

2.1 Participants

Thirty - five students of two intact Elementary German II classes at Mahidol University International College in Thailand took part in the study. Both sections were taught by different instructors. At the start participants had learnt 44 hours of German in the previous Elementary German I level. There were 25 female and 10 male participants, aged 17 - 22. Most of them were first year students. The course was part of their General Education requirements, and only a small percentage is going to take German as a minor program in their second year. Thirty - four of them were Thai nationalities, and one was Burmese, thus, German is a third language for them after English, which is the basic classroom language at the college. Participants became familiar with both examined methods - word lists and vocabulary trainers - during Elementary German level I. Results of a pre-survey revealed that some participants had a long history of using word lists, and others hardly reviewed vocabulary by any method. Flash cards were not used before participants enrolled in the Elementary German I class at the college. Class time for both sections was 10.00 - 12.00 twice a week. The second part of the study took place during Elementary German III, which 28 participants attended. The 4 remaining participants were called in

individually for review trials on the same days. During the six months of the study, 3 participants left the university. Thus, the data of 32 participants were collected. The decision to use intact classes was meant to facilitate the integration of participants with different language proficiency levels, particular weak performers, who might be reluctant to deliberately take part in an out-of-classroom experiment.

2.2 Materials

Sixty vocabulary items from Elementary German I, items not recycled in Elementary levels II and III, were collected from units 1 - 3 of the textbook and workbook in use. Those items were encountered for the last time between two - four months before the start of the study. The low language proficiency of the participants safely excludes the possibility, that they used the German language in the mean time and thus, were accidentally exposed to items of the study trials. From each of the three units more or less equal numbers of items were chosen, prominently nouns, (because they occur most often topic related and are rarely recycled), but also verbs, adjectives, adverbs, prepositions and multi-word units, in order to confront participants with parts of speech of different difficulty. The length of items varied from one to four syllables. Neither international words nor English cognates nor derivates were listed. Both high - and low - frequency items from unit 1 - 3 were

included. In the focus of interest were previously learned items which could not be recalled or were incorrectly or imperfectly recalled during the study period. On the contrary, items which could be accurately recalled from the start were considered less important, because the recall efforts cannot be clearly linked to one of the examined methods.

3. Design

The study was conducted in a with-in participant, repeated measures design, thus, each participant was examined for both treatments. It was a classroom related, longitudinal case study, comprising both quantitative and qualitative approaches. The independent variable was the method of reviewing: the word list approach and the vocabulary trainer method. Counterbalancing of test trial schedules was used in order to address carry-over effects. However, a strict 'AB BA' - structure was not applicable due to method related rehearsal intervals. As variables are impossible to control over a period of 6 months, the study could not be strictly experimental. However, possible factors affecting internal and external validity like maturation, carry-over effects and chance factors were interfering with both methods and did not favor one of them. In order to strengthen the external reliability of the study a replication will be conducted.

3.1 Procedures

3.1.1 Preliminary Remarks

1. All test trials were integrated into regular classroom sessions. No detailed information about the aims of the study were given, however, the participants were well aware that the study trials were not regular classroom activities. The aim of this procedure was to avoid overly artificial attention to the consecutive trials. All pretests and delayed posttests were conducted unannounced (though - due to high frequency - not unexpected). Participants used code names throughout the study, in order to avoid compromising the data.

2. As measurable comparisons between the two methods were an important aim of the study, a basic adjustment was necessary. Both flashcards and word lists are often used as a read-only method. Recall failures or inaccurate recalls are rarely addressed by note taking. However, in order to facilitate collection of quantitative data, participants had to put down all definitions of target items on prepared blanked lists.

3.1.2 Pretest Design

At the start, two pretests were administered providing participants with a list of 60 lexical items in German, which they had to translate. Cued recall was used, as it is considered to facilitate retrieving information from memory and not only recognizing previously stored informa-

tion. The chosen translation direction was L3 German – L2 English. The main reason was the aspect of word knowledge which the study was focused on: word meaning. (Neither spelling, parts of speech, morphology, collocations, register etc. were focused on). Moreover, at this early stage of language acquisition, the development of a receptive vocabulary was considered more important than a productive vocabulary. A last reason was the quite high English proficiency of the participants. Thus, they were not supposed to have major difficulties in quickly putting down correctly spelled English definitions. The opposite would be true for writing German target items. Consequently, it would sometimes be difficult to distinguish gross spelling errors from recall failures if the L2 - L3 direction had been used. The lists for pretests were designed as follows:

1. Instructions on how to fill the blanked lists;
2. Four warm up buffers (conducted together);
3. The list of 60 items in the following format:
 - a) Cue words on the left side;
 - b) Cue sentences in German (with the cue item highlighted in order to enhance noticing) in the middle. The formulation of cue sentences followed widely the '*encoding specificity principle*' (Tulving, 1983), according to which a cue should be from the same context with which the item was

originally encountered. Cue sentences are also included in order to put emphasis on the 'lexical approach' (Lewis, 1993);

- c) blanks for English definitions on the right side.

In order to reduce the impacts of chance factors on a randomly chosen day, the list was administered twice within 48 hours. Additionally, the items on the second list were shuffled, in order to counterbalance the 'serial order effect' and the possible influence of fatigue and tediousness. At this early stage of language acquisition, it is hard to avoid listing items from the same semantic sets, because only a few basic topics had been covered at the point. However, semantic set items on the pretest lists were positioned as far as possible from each other. Participants were also encouraged:

- to additionally use the cue sentences in case they could not instantly recall the item;
- to write down any association with the item to show that they had a rough idea of its meaning instead of no understanding at all. This procedure refers to the incremental process of language acquisition;
- to write the translation in their mother tongue Thai, if they could not recall the English equivalent, but were confident to know its meaning.

After conducting the pretests, the translations were marked by a modified 'Vocabulary Knowledge Scale', (Wesche and

Paribakht, 1996). As language acquisition is 'rarely an all-or-nothing proposition' (Folse, 2006), partial gains in word knowledge were included in the scores:

- Fully accurate recalled items: 3 points
- Partially accurate recalled items 2 points (e.g. correct meaning, but wrong part of speech):
- Correct semantic field, 1 point antonym etc. :
- Recall failures or misplaced 0 point definitions:

Spelling mistakes were not counted (unless the item was misspelled beyond recognition), even if the mistake led to another, but semantically not related lexis (e.g. *widowed*- the target item, and *window*- the spelling error).

3.1.3 Posttest Design

After having examined the participants twice without '*knowledge of correct response*' correct response, the 60-item list was divided in two lists of 30 items, providing each participant with two more or less equally difficult lists of known, partially known and forgotten items, distinguished by equally distributed parts of speech, semantic set items and most importantly by participants' personal item difficulty. This is determined by considering the participants' performance on the two pretests. For example: A participant managed 20 correct recalls, but committed 20 partial errors and 20 complete

errors on the combined 60-item pretests. Then 10 correctly recalled items, 10 partial and 10 complete errors would be equally distributed to each of the two 30-item lists for this participant. In order to further diminish the possible occurrence of unequal distribution of items, the participants were then divided in two nearly equal groups of 17 and 18. One group received their personal first 30 items as a word list and the second part as flash cards, related to the referenced vocabulary trainer. The other group was administered their personal first part as flash cards, the second part as a word list. During the subsequent six months the participants were treated several times with both methods.

The following conditions were administered in a similar manner:

- The design of cue sentences;
- Retrieval of items by cued recall;
- self-pacing in order to avoid speed-accuracy-trade-offs;
- distributed review trials;
- '*Knowledge of correct response*' through presenting the correct definitions after a trial;
- no permission of note taking for recall errors.

However, in order to reflect the distinguishing underlying concepts there were also key differences:

1. Flashcards were self-produced by participants, because note taking is consi-

dered a deeper processing strategy than reading-only (see Thomas and Dieter, 1987). Conversely, word lists were provided as pre printed versions, as participants are most familiar with this means of presentation.

2. For both methods a specific procedure was designed in order to enhance awareness of recall success and failure. Word list items had to be ticked off by a translation list matching the personal item list for recall success and failure for each participant.

Flash cards had to be moved according to the referenced procedure. In order to avoid flipping cards accidentally (or even intentionally) during review trials each target side was covered by a second attached blank flashcard. Regarding both methods the participants were asked to pay special attention to recall failures, but no particular technique was recommended.

3. First 'knowledge of correct response' for flash card items was given by writing and designing the cards 90 minutes before a first posttest. Only previously unknown or partially known items were included in trial 1. This specific design follows the referenced 'Low-First Method' (Mizuno, 1999), as 'low' compartments of the vocabulary trainer for unknown items will be reviewed more often than 'high' compartments for known items to counter effects of the 'forgetting curve' (Ebbinghaus, 1885). In case of further recall failures, the cue sentence had to be added on the card.

First 'knowledge of correct response' for word lists was provided by reading the items 90 minutes before the first posttest, including all 30 items. The correct response was given for all 30 items, though the research literature is not conclusive regarding its effects. In case of recall success the effects seem to be low, however, in cases of recall failures the chances of recall are often improved in the next review session (e.g., Bahrick and Hall, 2005; Pashler et al., 2005). In the later posttests the correct response was presented after the specific study trial.

4. Flash card items were shuffled in every consecutive trial, while word lists were always presented in the same sequence in order to examine possible impacts of the 'serial order effect'.

5. Word list items were tested 4 times during the first 3 months shortly before quizzes and exams (in order to simulate their real usage under classroom conditions), and once in the latter part of the study (approximately five months after the first re-encounter.) Flash card items were tested according to the referenced procedure, which stretches beyond external, exam related time lags: shorter time lags and more frequent review trials for unknown items, longer time intervals for known items, staggered approximately by the exponential factor of 3 (3, 9, 27, 81).

6. Following the principles of the 'Low-First Method', accurately recalled

flash card items were omitted in subsequent trials. Conversely, word lists were always presented as a whole, not distinguished by recall failure and success. However, in order to examine long-term recall scores in a last

trial after 6 months all flash card items had to be presented again, with the erroneous items first, and the known items in a separate review trial.

Table 1: Review Trial Plan for Word Lists and a Vocabulary Trainer

<i>Word list trials (WL)</i>	<i>Posttests (PT)</i>	<i>Vocabulary trainer trials (VT)</i>
	<i>Week 1* pretest all 60 items</i>	
	<i>Week 1 pretest all 60 items (shuffled)</i>	<i>Week 2 PT 1/1 VT (based on pretest errors)</i>
		<i>Week 2 PT 1/2 VT (based on 1/1 VT errors)</i>
<i>Week 3 PT 1 all 30 WL items</i>		
<i>Week 3 PT 2 all 30 WL items</i>	<i>Week 5 PT 1/3 VT (known items from pretests)</i>	
<i>Week 7 PT 3 all 30 WL items</i>		<i>Week 9 PT 2/1 VT (based on 1/1 - 1/3 VT errors)</i>
		<i>Week 9 PT 2/2 VT (based on 2/1 VT errors)</i>
		<i>Week 12 PT 2/3 VT (known items from 1/1 - 1/3)</i>
<i>Week 13 PT 4 all 30 WL items</i>		<i>Week 16 PT 3/1 VT (based on 2/1 - 2/3 errors)</i>
		<i>Week 16 PT 3/2 VT (based on 3/1 errors)</i>
<i>Week 21 PT 5 all 30 WL items</i>		<i>Week 22 post test 3/3 VT (all 30 items)</i>

* the exact dates will be assigned during the study period

4. Results

4.1 Statistical Analysis

The data regarding the effects of both treatments were analyzed using Wilcoxon Signed-Rank tests. This non-parametric test was chosen in order to analyze the differences between sets of two related

samples, as a within participant - and repeated measures design was used. Furthermore, the data were not normally distributed (positively skewed) and not randomly sampled (intact language classes), so a parametric *t*-test analysis could not be conducted and the different scores could only be ranked

in size. The tests were performed with either the vocabulary trainer method or the word list method as within-participant factors. In order to support the directional research hypothesis, with the Wilcoxon Signed-Rank tests the sum of the positive ranks [scores on the vocabulary trainer (VT) method] had to be greater than the sum of the negative ranks [scores on the word list (WL) method]. $P < 0.1$ was chosen as the appropriate alpha level for all tests. The effect size of the pair wise comparisons was calculated with Cohen's d by defining the VT as the experimental method and the WLs as the control method. As the direction of the effect was predicted in the research hypothesis, one tailed-measurements had to be applied. However, the comparisons between recall scores could be only an approximation. Due to referenced review procedures, neither the number of trials nor the number of items learned per trial were equal. All VT-trials were separated by known and unknown items, and specific 'Low- First'-trials (LF) were added. WLs were throughout given as a whole. The data of 32 participants were collected.

4.1.1 Pretest 1 and 2 (Without 'Knowledge of Correct Response' [KCR])

Pretest 1 and 2 yielded a mean score of 32.2 out of 60 items. Though the mean score for both pretests was very similar (+ 0.8 items), there was a trade-off in known and unknown items (pretest 2: 6% of the

items from pretest 1 were forgotten and 7% were regained). The variance in scores was large (from 8 to 52 items out of 60, SD 11.8). In order to compare the score developments for both treatments, the pretest mean score had to be divided by 2 (32,2 : 2 = 16.1).

4.1.2 Posttest 1 and 2 (with KCR)

Posttest 1 for WLs was given 21 days after pretest 1 and for VT between day 14 and 44. It revealed a clearly larger increase in mean scores for VT than for WL. This difference is statistically significant, (Wilcoxon, Z: -4.669, $p < 0.001$). Posttest 2 (given 2 days later for WL only) shows a steep increase in scores. Compared to the new score, the VT- posttest 1 mean score is still significantly higher, though by a smaller margin, (Wilcoxon, Z: -2.686, $p < 0.007$).

4.1.3 Posttest 3 and 4 (with KCR)

Posttest 3 **for WL** was administered 52 days after pretest 1. It revealed a slight decrease **in recall rates**, whereas the separated tests on VT (day 60 and 78) yielded modest gains, (Wilcoxon, Z: -4,454, $p < 0.001$). Posttest 4 was only administered for WL (87 days after pretest 1) and resulted in a further slight reduction of scores.

4.1.4 Posttest 5 (with KCR)

This final trial was conducted 135 days after pretest 1 for WL and between 120 and 145 days later for VT. While VT-scores

decreased slightly, WL-scores were reduced by a higher margin, (Wilcoxon, $Z: -4.684$, $p < 0.001$).

The overall development in scores shows a steep increase of scores on posttest 1

for the vocabulary trainer and on posttest 2 for word lists. After that the range of differences between trials were less than $M = 2$ items for both methods. Table 2 summarizes the referenced results:

Table 2 : Mean Recall Scores on VT and WL over 6 Months

<i>Source</i>	<i>Posttest 1</i>	<i>Posttest 2</i>	<i>Posttest 3</i>	<i>Posttest 4</i>	<i>Posttest 5</i>
<i>VT</i>	26.2	-	27.9	-	27.6
<i>WL</i>	20.0	25.1	24.4	24.7	23.5

Note. Dashes indicate data not obtained Mean Pre test 16,1

Wilcoxon Signed Rank tests revealed statistically significant differences for recall scores on all levels within the study period, ($p < 0.001$, one-tailed). All obtained Wilcoxon T -values were clearly below the critical, one-tailed Wilcoxon T -values. Similar effects on significance can be seen for the Wilcoxon T -values translated into standardized Z -values. The effect size of score difference was medium in the first posttest, (Cohen's $d = 0.48$) and large in posttest 3, (Cohen's $d = 0.90$) and posttest 5, (Cohen's $d = 0.92$). The results were very robust throughout the study period, as 28 participants scored higher on VT, while the remaining four high memory achievers yielded approximately the same results on both methods. Thus, the research hypothesis

seems to be supported by the outcome of the study.

4.2 Other Noteworthy Findings

4.2.1 The Impact of 'Low-First' - Trials for VT - Scores

In a series of consecutive 'Low-First'-trials on the vocabulary trainer method, an average of nearly 70% (241 out of 345 items) of previously forgotten items could be retrieved. As can be seen from table 2, these extra scores lifted the mean score for the vocabulary trainer method remarkably over the WL approach, as reported in table 3.

Table 3 : Mean Recall Scores on VT- + LF-Trials and WL-Trials

Source	Trial 1	LF 1	Trial 2	LF 2	Trial 3	LF 3
VT	23.9	+ 2.3	25.5	+ 2.4	25.1	+ 2.5
WL	20.0	-	24.4	-	23.5	-

Note. LF-Trials 2-7 days after previous VT- Trial. Dashes indicate data not obtained

However, a qualitative analysis of these VT- results shows a trade-off between former recall failures and previous correctly recalled items:

- Around 80% of previous recall errors or failures presented in LF-trials were successfully recalled again in the following trial;
- a similar number of previous accurately recalled items were forgotten in the next trial, keeping the overall number of

correctly recalled items quite stable. Conversely, for the word list method higher forgetting rates and fewer regains of items are to be reported.

4.2.2 WL and VT Scores for Different Levels of Long-Term Memory

As table 4 shows, the scores for different groups of memory performers were as follows:

Table 4 : Final WL- and VT- Mean-Scores for Participants with Different Score Levels on Pretests (PreT)

Source	Level A	Level B	Level C
VT	29.8	28.0	24.1
WL	28.0	23.7	20.4
PreT	23.2	17.7	10.1

Note. A: > 75% correctly recalled item B: 50-75% C: < 50%

Participants with ‘good’ long-term memory scores on pretests ($>75\%$) gained modest increases on both methods with a slight advantage for VT, (Wilcoxon, Z - 2.207, $p < 0.027$).

The increase of recall scores for ‘medium’ performers (50% - 75% on pretests) was nearly 100% higher on the VT-method than the WL- method, (Wilcoxon, Z -3.113, $p < 0.002$). Participants with ‘weak’ long-term memory results on pretests ($<50\%$ on pretest) yielded the highest growth in scores on both methods, but again with noticeably higher results on the VT-method, (Wilcoxon, Z -2.805, $p < 0.005$).

4.2.3 Results Distinguished by Sorts of Error

As can be seen from table 5, the number of three means of error is noticeably lower for VT. The effect is most remarkable on ‘error fossilization’, both for systemic recall failures and persistently wrong recall (Wilcoxon, Z : 4.189, $p > 0.001$), but also noteworthy on interference induced mistakes (Wilcoxon, Z : 4.060, $p > 0.001$) and errors of omission (Wilcoxon, Z : 4.843, $p > 0.001$).

Table 5 : Mistakes on all Trials for VT and WL Distinguished by Means of Error

source	<i>Fossilization</i>	<i>Interference</i>	<i>omission</i>
VT	5	202	292
WL	85	348	674

Note. *Correct lexical definitions, containing the wrong part of speech (widow vs. widowed) or omissions in multi-word units were not included in the count for fossilization.*

Persistent errors were counted once as error fossilization and as omission or interference error on each additional occurrence.

The overall number of reviewed items was on WL 4800 and on VT 3746.

All of the above reported results lend additional support to an acceptance of the research hypothesis.

5. Discussion

As mentioned in the Method section, the study was designed as a repeated measures, longitudinal case study, hence, the independent variable was not strictly controllable over a period of 6 months. Moreover, the study sample was rather small, not normally distributed and thus does not allow over simplistic generalizations. Additionally, the assumption could be challenged that the study succeeded in creating two lists containing equally difficult items for each participant. The L3 – L2 translation direction might also be a limitation, as it does not allow inferences to mother tongue - second language structures. One has also to keep in mind that nearly all of the participants were Thai nationals. In Thailand there is a long, living tradition of rote learning, thus, the results might not be transferable to other cultural contexts. Despite these limitations, the research outcome seems to be worth conducting an in depth analysis.

5.1 The Remarkable, but Smaller than Expected Gap in Scores

As mentioned, the results for word lists are significantly lower than for the vocabulary trainer. However, scores on word lists are higher than previous research on long-term effects would suggest. The difference between expected and obtained data for word lists might be related to the applied procedures for both methods

(multiple test trials over 6 months, retrieval practice, expanding time lags, knowledge of correct response', cue sentences etc.), as these procedures are considered to facilitate long-term memory. Conversely, a crude pair-associate list, deprived of all benefiting elements might have yielded lower results. One has also to keep in mind, that retrieving previously encountered items requires other, probably less difficult tasks than encoding completely new items.

Design related carry over effects also cannot be ruled out.

5.2 The Impact of the 'Low-First Method' (Mizuno)

The most important influence on recall scores seems to be the implementation of the 'Low- First Method' for the vocabulary trainer. More frequent encounters with items of personal difficulty seem to have yielded significantly higher scores for the vocabulary trainer method (ca. 70% of inaccurately or not recalled items in the previous trial could be recalled in the follow up 'Low-First'- trial), whereas repeated word list-trials neither led to a remarkable increase nor a decrease of recall scores. Another indication of possible effects of this method is the fact that nearly 80% of items which had to be reviewed in 'Low-First'- trials were also accurately recalled in the following trial, slowing down obviously the 'speed of activation decay' (Mizuno, 1999). This

outcome accords with research findings that a flexible two-tiered system for rehearsal is advantageous over consistently expanding time lags, irrespective of the ability or failure to recall an item. Furthermore, the single, remarkable increase in recall rates by word lists occurred between posttest 1 and 2, with a gap of only 48 hours in between. The short time lag, similar to the Low-First method, could be partially responsible for this unusual improvement, while the word list-rates for longer gaps were either quite stable or decreasing.

5.3 The Impact of Noticing

Two main factors might have contributed to the remarkable difference in scores on the first posttest in which knowledge of correct response was given. Due to the 'Low-First Method', participants had to concentrate on lower numbers of flashcards (M: 13.9 vs. 30 items on word lists). This might have reduced possible influences of fatigue and diminished concentration. Another factor that could have been instrumental in affecting the results is the conscious raising technique of writing the items down. The participants had not only to write cue item and target item, but also to design the entire front side and back side of the flash cards (lay-out for cue item and cue sentence, shape, size, color of letters etc.) and by this possibly a deeper level of awareness and noticing could be

achieved, opposed to a more superficial reading-only approach for the word list items. This outcome accords with research findings that 'knowledge of correct response' is a memory enhancing technique, particularly if the item was wrongly recalled in the previous test.

Noticing and awareness might have also been influential in the remarkable rise of word list scores between posttest 1 and 2 (+ 5.1 items). Posttest 2 was scheduled 48 hours after posttest 1 and the correct response was given for both tests 90 minutes in advance. In all follow up tests, administered in longer intervals, margins of difference were remarkably lower, probably because no knowledge of correct results was given before the trial.

5.4 Results Distinguished by Performance Levels

The differences between the two methods in the group of 'good' long-term memory concerning vocabulary are modest. This outcome is likely due to 'ceiling effects' for participants with good long-term memory capacities, as they are usually not depending on a single memory strategy and are using learning strategies in a more effective way. On the other hand, all participants with 'medium'- and 'weak' long-term memory capacities yielded noteworthy higher scores on the vocabulary trainer method. Margins for retrieval failure and

regain of previously forgotten items between test trials are larger in these groups than for good performers, revealing volatile long-term memory structures which need to be addressed by a more appropriate rehearsal strategy than word lists can provide.

5.5 Results Distinguished by Sorts of Error

The most remarkable distinction for both methods can be seen in the phenomenon of '*error fossilization*', both systemic recall failures and persistent wrong recall. The root of this distinction seems to be related to the investigated methods. If insufficient encoding several months before the start of the study would be the main source of persistent retrieval failure, it should have occurred with similar frequencies on both methods, as for both treatments the equal distribution of forgotten or wrongly recalled items and the equal application of 'knowledge of correct response' was very strict. However, the systematically implemented 'Low-First'-sessions for the vocabulary trainer appear to have hindered stamping in of errors. The opposite seems to be true for some word list items, mainly low-frequency words and parts of speech like adverbs, which do not offer pictorial annotations. Additionally, rehearsal intervals might have been inappropriately scheduled. The fixed 'serial order' of the lists could have also contributed to this tendency. Some

participants' awareness of previous recall failures might have been insufficient, as successful recalls and persistent retrieval failures were not separated, leading to a rigid two-tier system of known items on one side, and imperfectly known or unknown items on the other side. The lower occurrence of single, not persistent errors of omission for flash cards might be partially due again to the referenced 'Low-First Method' and awareness raising strategies. However, the distinction to word lists is less remarkable, revealing limits of this learning device, particularly in case of volatile memory structures with high forgetting rates. Moreover, the roots of errors of omission are particularly difficult to trace and so either attrition, lack of retrieval cues, intrusion by interference, lack of concentration, and avoidance to put down inaccurate definitions could be involved. In difference to the referenced tendencies, results on interference induced errors reveal error rates on the vocabulary trainer method by smaller margins. One reason for this outcome could be that frequently forced retrieval of items could have induced '*spread of activation*' of lexical or semantically similar and thus competing lexis. Particularly the last trials (after a pause of about 8 weeks) lend support to this hypothesis. While errors of omission were rising on the less frequently tested word lists, interference induced errors were slightly increasing on the more often encountered

flash card items. Another reason for this outcome could be that the '*minimal context design*' (Laufer, 1997b) of one cue sentence is susceptible to interference, because it might not be sufficiently rich and distinctive in context. Particularly noteworthy is the occurrence of intra-list errors (both lexical and morphological). Those intra-list errors are higher on word lists, but are also found with the vocabulary trainer method, corroborating research findings on negative impacts of learning semantic sets and alphabetically ordered lexis.

5.6 The Effects on Time Efficiency

In the vocabulary trainer method, items were not presented for recall any longer if they could be recalled in three consecutive trials, while the word list was constantly given as a whole. This procedure resulted in an overall number of 3746 items for the vocabulary trainer vs. 4800 items on word lists. After three trials around 30% of all participants did not have a single flash-card item left to review, while the remaining average was two items per participant. This indicates not only higher effectiveness (25% lower review load), but also better time efficiency for this method. However, overoptimistic assumptions about long-term memory as a '*permastore*' (Bahrick, 1984) might not be appropriate, because in a last trial 6 months after the start a slight decrease of scores on flash cards was observed.

5.7 The Effects of Cue Sentences

The impact of the cue sentences is hard to determine. As confirmed by a post-study survey, many participants read cue sentences only if they could not recall an item instantly. However, the study design does not allow inferences as to whether a cue sentence led to a correct definition, which otherwise would not have been produced. Moreover, cue sentences were in some cases even the root of the error, as can be seen from specific misplaced definitions. Additionally, the low vocabulary of the participants did not always allow the formulation of powerful, interference inhibiting retrieval cues.

6. Conclusions

The results of this study indicate that reviewing previously learned, but forgotten items by methods of direct, intentional learning facilitates long-term memory of foreign language vocabulary. This general outcome accords with research findings both in applied linguistics and cognitive psychology (see literature review section). Moreover, the study shows advantages of the vocabulary trainer method over word lists, mainly due to the implementation of the 'Low-First Method'. Particularly weak memory performers seem to benefit from this rehearsal strategy, as it appears to address volatile long-term memory more appropriately than word lists. Furthermore, self production of items seems to be an important memory

enhancing strategy. In contrast to this, prefabricated, semantically or alphabetically ordered word lists appear to be prone to 'error fossilization', trace decay and interference and thus, should be avoided. However, the vocabulary trainer method also is susceptible to attrition and interference, proving that there is no magic cure against forgetting, forcing both language instructors and learners to carefully and patiently deal with it.

Another shortcoming of the vocabulary trainer method, as assigned in the on-hand-study, is the inflexible rehearsal practice for both good and weak memory capabilities. The increase of interference errors, particular committed by weaker performers in the last test trial on the vocabulary trainer show that an expanded rehearsal schedule, disregarding the retention capabilities of participants is not suitable. A more personalized scheduling of review trials should be applied.

Further research could also show whether computerized versions of the vocabulary trainer method could have an advantage over file card boxes, as they could easily add visual features, a variable set of contextual clues, a listening mode, linked learning -, reviewing - and testing modules, an individually tailored system of rehearsal intervals and personalized statistics about recall success and failure. A software version would also relieve learners from the plight

to decide whether their answers were fully correct, partially correct or incorrect.

A single memorizing strategy like vocabulary trainers cannot claim to solve long-term retention problems alone. Recent research favors broadly a mixed approach, comprising different strategies like extensive reading, and oral and written production as well as building up a large vocabulary by means of direct learning. However, the results of the study do suggest that a vocabulary trainer could provide more effective techniques of reinforcement and consolidation than traditional word lists do.

References

Baddeley, A. (1990). *Human Memory*. London: Lawrence Erlbaum Associates.

Bahrick, H.P. (1984). Semantic memory content in permastore: Fifty years of memory for Spanish learning in school. *Journal of Experimental Psychology, 113*, 1-29.

Bahrick, H.P. and Phelps, E. (1987). Retention of Spanish Vocabulary over 8 years. *Journal of Experimental Psychology, Learning, Memory, and Cognition, 13*, 344-349.

Bahrick, H.P., Bahrick, L.E., Bahrick, A.S., and Bahrick, P.E. (1993). Maintenance of foreign Language vocabulary and the spacing effect. *Psychological Science, 4*(5), 316-321.

Bahrick, H.P. and Hall, L.K. (2005). The importance of retrieval failures to long-term retention: A meta cognitive explanation of the spacing effect. *Journal of Memory and Language, 52*(4), 566-577.

Carter, R. and McCarthy, M. (Eds.). (1988). *Vocabulary and language teaching*. London: Longman.

Clipperton, R. (1994). Explicit Vocabulary instruction in French immersion. *The Canadian Modern Language Review/La Revue Canadienne des Langes Vivantes, 50*(40), 736 – 749.

Cobb, T. (1998). Breadth and Depth of lexical acquisition with hands-on concordancing, *Computer Assisted Learning, 12*(4), 345-360. Retrieved from <http://www.er.uqam.ca/nobel/r21270/cv/breadth/html>.

Crow, J.T. and Quigley, J.R. (1985). A semantic field approach to passive vocabulary acquisition for reading comprehension. *TESOL Quarterly, 19*, 497 – 513.

Dempster, F.N. (1987). Effects of variable encoding and spaced presentation on vocabulary learning. *Journal of Educational Psychology, 79*, 162 – 170.

Dempster, F.N. (1988). The Spacing Effect: A Case Study in the Failure to Apply the Results of Psychological Research. *American Psychologist, 43*(8), 627-634.

Ellis, N. (1997). Vocabulary acquisition: word structure, collocation, word class and meaning. In Schmidt, N. and McCarthy, M. (Eds.). *Vocabulary: Description, Acquisition and Pedagogy* (pp.122-139). Cambridge: Cambridge University Press.

Folse, K. (2004). *Vocabulary Myths*. Ann Arbor: University of Michigan Press.

_____. (2006). The Effect of Type of Written Exercise on L2 Vocabulary Retention. *TESOL Quarterly, 40*(2), 273 – 293.

Groot, P. J. M. (2000). Computer Assisted Second Language Vocabulary Acquisition, *Language Learning & Technology, 4*(1), 60-81. Retrieved from <http://llt.msu.edu/vol4num1/groot/defaukt.html>.

Hedge, T. (2001). *Teaching and Learning in the Language Classroom*. Hongkong: Oxford University Press.

Hermann, F. (2003). Differential Effects of Reading and Memorization of Paired Associates on Vocabulary Acquisition in Adult Learners of English as a Second Language, *Teaching English as a Second or Foreign Language, 7*(1), 1-15. Retrieved from <http://teslej.org/ej25/a1.html>.

Judd, E. L. (1978). Vocabulary Teaching and TESOL: A Need for Re-evaluation of Existing Assumptions. *TESOL Quarterly*, 12(1), 71-76.

Krashen, S. (1982). *Principles and practice in second language acquisition*. New York: Prentice-Hall.

_____. (1989). We acquire vocabulary and spelling by reading: Additional evidence for the input hypothesis. *Modern Language Journal*, 73, 440-464.

Landauer, T.K. and Bjork, R.A. (1978). Optimum rehearsal patterns and name learning. In Gruneberg, M. M., Morris, P. E. and Sykes, R. N. (Eds.), *Practical Aspects of Memory* (pp.625-632). London: Academic Press.

Laufer, B. and Osimo, H., (1991). Facilitating Long-Term Retention of Vocabulary: The Second-Hand Cloze. *System*, 19(3), 217-224.

Laufer, B. (1997a). The lexical plight in second language reading. Words you don't know, words you think you know, and words you can't guess. In Coady, J. and Huckin, Th. (Eds.), *Second Language Vocabulary Acquisition* (pp.20-34). Cambridge: Cambridge University Press.

Laufer, B. and Shmueli, K. (1997b). Memorizing new words. Does teaching have anything to do with it? *RELC Journal*, 28, 89 – 108.

Laufer, B. (2003). Vocabulary Acquisition in a Second Language: Do Learners really acquire most vocabulary by reading? Some empirical evidence. *Canadian Modern Language Review/ La Revue Canadienne des Langues Vivantes*, 59(4), 567-587.

Leitner, S. (1972). *So lernt man lernen: Der Weg zum Erfolg*. Herder, Freiburg.

Lewis, M. (1993). *The lexical approach*. Hove: Language Teaching Publications.

McCarthy, M.J. (1990). *Vocabulary*. Oxford: Oxford University Press.

Meara, P. (1980). Vocabulary Acquisition: A neglected aspect of language learning. *Language Teaching and Linguistics: Abstracts*, 13(4), 221-246.

_____. (1995). *The Importance of an Early Emphasis on L2 Vocabulary*, Retrieved from <http://www.jalt-publications.org/tlt/files/95/feb/meara/html>.

Mizuno, R. (1999). *A test of the Effectiveness and Efficiency of the Low-First Method Derived from a Reactivation Theory of Spacing Effects*. Retrieved from www.jcss.gr.jp/iccs99OLP/p3-31/p3-31.htm.

Mondria, J.-A. and Mondria-de Vries, S. (1994). *Efficiently memorizing words with the help of word cards and "hand computer": Theory and Applications*, *System*, 22, 47-57.

Nation, I. S. P. (1982). Beginning to learn foreign vocabulary: a review of research. *RELC Journal*, 13(1), 14-36.

_____. (2001). *Learning Vocabulary in another language*. Cambridge: Cambridge Applied Linguistics.

Pashler, H., Cepeda, N. J., Wixted, J. T. and Rohrer, D. (2005). When does feedback Facilitate Learning of Words?. *Journal of Experimental Psychology, Learning, Memory and Cognition*, 31(1), 3-8.

Pimsleur, P. (1967). A memory schedule, *Modern Language Journal*, 51, 73-75.

Prince, P. (1995). Second language vocabulary learning. The role of context versus translations as a function of proficiency. *The Modern Language Journal*, 80(4), 478-493.

Rivers, W. and Temperley, M. (1978). *A Practical Guide to the Teaching of English*. New York: Oxford University Press .

Schmitt, N. (1997). Vocabulary Learning Strategies. In Schmitt, N. and McCarthy, M. (Eds.), *Vocabulary: Description, Acquisition and Pedagogy*. Cambridge: CUP.

Schmidt, N. and Schmidt, D. (1995). Vocabulary Notebooks: theoretical underpinnings and practical suggestions. *ELT Journal*, 49(2), 133 -143.

Selinker, L. (1972). Interlanguage. *IRAL*, 10 (3), 209 – 231.

Thalheimer, W. (2006). *Spaced Learning Events Over Time: What the Research Says. A Work-Learning Research*. Retrieved from <http://www.work-learning.com>.

Thomas, M. H. and Dieter, J. N. (1987). The positive effects of writing practice on integration of foreign words in memory. *Journal of Educational Psychology*, 79, 249-253.

Thornbury, S. (2002). *How to Teach Vocabulary*. Malaysia: Longman-Pearson Education Ltd.

Tinkham, T. (1997). The effects of semantic and thematic clustering on the learning of second language vocabulary. *Second Language Research*, 13, 138-163.

Tulving, E. (1983). *Elements of episodic memory*. Oxford: Oxford University Press.

Waring, R. (1997). *A study of receptive and productive vocabulary learning from word cards*. Retrieved from <http://www1.haren.net.ne.jp/~Waring/papers/wordcard.html>.

Waring, R. (2002). *Why should we build up a Start-up vocabulary quickly?* Retrieved from <http://www1.haren.net.ne.jp/~waring/vocab/principles/early.htm>.

Waring, R. (2004). *In defence of learning words in word pairs – but only when doing it the ‘right’ way!* Retrieved from <http://www1.haren.net.ne.jp/~waring/vocab/principles/systematic learning.htm>.

Wesche, M. and Paribakht, T.S. (1996). Assessing second language vocabulary knowledge: Depth versus breadth. *Canadian Modern Language Review/Revue Canadienne des Langes Vivantes*, 53(1), 13-40.

Acknowledgements

I gratefully acknowledge the financial support provided by a research grant from Mahidol University International College in Nakhon Pathom, Thailand. I am thankful to Mr. James Whitlam for his comments on an earlier draft of this paper. I am thankful to Associate Professor Yaowalark Sukthana, Associate Professor Thaweeratana Siwadune, Ms. Barbara Ekamp, Ms. Ute Bäuchl, Assistant Professor Sarayut Nathaphan, Mr. Takayoshi Fujiwara, and Mr. Michael Mergy Krause for their valuable assistance in various manners.