Language Learning 50:1, February 2000, pp. 1–56

What Is Hard to Learn Is Easy to Forget: The Roles of Word Concreteness, Cognate Status, and Word Frequency in Foreign-Language Vocabulary Learning and Forgetting

Annette M. B. de Groot and Rineke Keijzer
University of Amsterdam

We looked at foreign-language (FL) vocabulary learning and forgetting in experienced FL learners, using a paired-associate training technique in which native-language words were paired with pseudowords. The training involved 6 presentations of the same 60 translation pairs, followed by a test after the 2nd, 4th, and 6th presentation round. A retest followed 1 week after training. The stimulus materials were manipulated on word concreteness, cognate status, and word frequency, and both productive and receptive testing took place. Cognates and concrete words were easier to learn and less susceptible to forgetting than noncognates and abstract words. Word frequency hardly affected performance. Overall, receptive testing showed better recall than productive testing. Theoretical accounts of these findings are proposed.

An important component in the early stages of many foreign-language (FL) training programs is paired-associate learning of

---

Annette M. B. de Groot and Rineke Keijzer, Department of Psychology.

Part of this work was presented at the 39th Annual Meeting of the Psychonomic Society, Dallas, TX, November 1998.

Correspondence concerning this article may be sent to A. M. B. de Groot, University of Amsterdam, Department of Psychology, Roetersstraat 15, 1018 WB Amsterdam, The Netherlands. Internet: pn_groot@macmail.psy.uva.nl

the native-language words on the one hand and their translations in FL on the other hand. By pairing the words of the two languages, the FL words adopt the meanings of the corresponding native-language words; in other words, a known concept gets a new name attached to it. Because the two words in a translation pair do not generally share meaning completely, there is obviously more to learning FL words than simply attaching new labels to the meanings of corresponding native-language words. The meaning assigned initially to the FL word by pairing it with its translation in the native language will subsequently have to be refined in FL contexts. However, the fit between the meanings of the two words in a translation pair is typically good enough to get the beginning FL learner started with simple sentences composed of FL words learned with the paired-associate technique, and also good enough for even quite proficient learners to go on using a translation dictionary whenever they encounter an unfamiliar FL word the meaning of which cannot be reliably inferred from its context.

The goal of the present investigation was to get a more complete picture than is currently available of the factors that affect, on the one hand, such "word-association" *learning* of FL vocabulary, and, on the other hand, the *retention* of the FL vocabulary thus acquired. Our choice of the word-association procedure as FL vocabulary training method to be employed in this study was primarily motivated by the fact that this procedure does not constrain the choice of materials to be presented for learning the way the picture-association technique (where the FL words are paired with a picture depicting the word's meaning) and the keyword method do. Unlike the latter two techniques, the word-association technique is equally suited to study the FL learning of abstract and concrete words, a fact that enabled us to look at the role of word concreteness in FL vocabulary learning and forgetting. Another reason to use the word-association procedure was implicitly mentioned above: This FL learning procedure constitutes an important component of most FL training programs. A detailed understanding of the factors that determine memory performance following learning according to this procedure will inform these training programs.

For the rather experienced FL learners tested in the present series of experiments, the word-association method also happens to be a more efficient method than both the imagery-based keyword method and the picture-association method, as was demonstrated by two recent studies performed in our laboratory (Lotto & De Groot, 1998; Van Hell & Candia Mahn, 1997). Interestingly, for less experienced FL learners the superiority of the word-association method may be less pronounced (e.g., it only occurs in terms of retrieval time, not in terms of percentage-recall scores; Van Hell & Candia Mahn, 1997), or another learning method such as object-association learning (e.g., Wimer & Lambert, 1959) or imagery-based methods may be superior (e.g., Atkinson, 1975; but see Wang & Thomas, 1995, and Ellis & Beaton, 1993a, for qualifications; the superiority of the imagery-based method may only hold when the test immediately follows learning, Wang & Thomas, 1995, or only with receptive testing of the newly learned FL words and not with productive testing, Ellis & Beaton, 1993a).

In addition to word concreteness (or, more precisely, word imageability; see below), two other word characteristics were manipulated across the present series of experiments, namely, cognate status and word frequency. The cognate status variable involves differences between words in terms of the form relation with their translation in the target FL. Cognate words share (parts of) their orthographic and/or phonological form with their translations, whereas noncognate words are dissimilar in form to their translations. The word frequency variable concerns differences between words in how often they occur in language comprehension and production. These three word characteristics—concreteness, cognate status, and frequency—have, separately or in pairs, already been manipulated in quite a few "bilingual-representation" studies, that is, studies intended to find out how lexical knowledge is organized in bilingual memory. Strong effects of all three variables have been shown in these studies (e.g., De Groot, Dannenburg, & Van Hell, 1994), effects that have often been attributed to different bilingual-memory structures for different types of words.

In contrast to the representation studies, relatively few FL vocabulary learning studies have looked at the effect of word-type manipulations in FL learning, forgetting, or both, part of the reason possibly being that in many of these studies the training set involved very few words, too few to split them out over more than one experimental condition on the basis of a word-type manipulation and obtain reliable effects of this manipulation (see, e.g., Cheung, 1996; Papagno, Valentine, & Baddeley, 1991; and Wimer & Lambert, 1959, where only three, eight, and nine words were learned, respectively). Exceptions are the learning studies run in our laboratory, in which up to 80 words were presented and concreteness (Van Hell & Candia Mahn, 1997) or cognate status and frequency (Lotto & De Groot, 1998) were orthogonally manipulated, and a correlational study by Ellis and Beaton (1993b) that looked at the effect of all these variables (but only with items as the unit of analysis). A study by Service and Craik (1993) also manipulated word type, but here word concreteness and word frequency were confounded. Analogous to the finding in studies on first language learning that concrete words are acquired earlier than abstract words (e.g., Brown, 1957; Schwanenflugel, 1991), the studies by Van Hell and Candia Mahn (1997) and Ellis and Beaton (1993b) found that concrete words were learned considerably better than abstract words. Furthermore, Lotto and De Groot (1998) and Ellis and Beaton (1993b) demonstrated that cognates were easier to learn than noncognates. As compared to the effects of concreteness and cognate status, the role of word frequency in word-association learning is inconclusive: Whereas in Lotto and De Groot (1998) high-frequency words were learned slightly better than low-frequency words, Ellis and Beaton (1993b) obtained a weak negative correlation between recall performance and frequency, suggesting that word-association learning is slightly better for low-frequency words.

In the present investigation, all three of these word-type variables—concreteness, cognate status, and frequency—were manipulated across a set of four experiments that all employed exactly the same learning procedure and that all tested participants drawn from the same population. As a result of using the same participants and learning procedure across the experiments, the effects of word type (and of the remaining variables) that were obtained are directly comparable. Of the three word-type variables to be manipulated, concreteness is of particular interest because its effect in FL vocabulary learning may inform models of bilingual lexical representation. The reason it may do so is that some of the bilingual memory models that have been proposed make specific claims about the involvement of conceptual memory in bilingual processing, claims that can be tested by looking at the present concreteness effect and how it responds to the experimental manipulations. For instance, the "revised hierarchical model" of bilingual memory proposed by Kroll and Stewart (1994) asserts that processing from L1 to L2/FL (for instance, translating L1 words into L2) implicates conceptual memory more than processing from L2 to L1 does. One of the resulting predictions of this model (see the General Discussion section for more detail) is that larger concreteness effects should occur in "productive" testing (where the native-language, L1, words are presented during testing and the corresponding newly learned words have to be produced) than in "receptive" testing (where the newly learned words serve as the stimuli presented at test and the corresponding L1 words have to be produced). As in studies by Ellis and Beaton (1993a, 1993b) and Griffin and Harley (1996), both types of testing occurred in the present study, in our case in a between-participants design and always following learning of native-language–foreign-language, L1-L2, pairs. These earlier studies showed that receptive testing produces the better recall performance.

As mentioned, the goal of this study was to learn more about the factors that determine both the (word-association) learning and the retention of foreign-language words. In all four experiments, the development of learning was tracked over three learning phases, each of which consisted of two (visual) presentations of 60 translation pairs that, after the second presentation round, were followed by either a productive test or a receptive test. Unlike in the majority of FL learning studies, where typically only correct-recall scores serve as the dependent variable, we also looked at performance in terms of retrieval time, that is, the time it took to come up with the translation of the presented test word. To look at both is relevant because to be able to use a foreign language fluently, not only must the words to be produced or understood be known, but access to them must also be fast; if it is not, working memory will be overloaded and performance will break down. The three learning phases together constituted a single learning session that took about 135 min to complete. Forgetting was measured in a retest that took place 1 week after the learning session. The retest involved either a productive test or a receptive test in which either all native-language words of the 60 trained translation pairs or all FL words of these pairs were presented just once, without further training.

A final noteworthy feature of the present investigation is that the FL words to be trained were not actually words in an existing natural language, but were letter strings that we made up ourselves; but all were orthographically and phonologically legal in the native language (Dutch) of our participants. The reasons for having the participants learn such "pseudowords" rather than actual words in some or other natural language were: first, we could rule out the possibility that any of the participants had any knowledge of the new language; second, the groups of cognate materials could be very systematically created according to a couple of stringent construction criteria (see below); and third, we could be certain that all of the new "words" to be learned were easily pronounceable. Taking account of the items' pronounceability is of particular importance, because of the important role of short-term phonological memory in FL vocabulary learning (e.g., Ellis & Beaton, 1993a; Papagno & Vallar, 1995; Service, 1992), especially in the early stages of learning the new language (Cheung, 1996), when the FL words to be learned are not readily semantically associated with native-language words (Papagno et al., 1991), and when the learners are older adults

(Service & Craik, 1993). The extent to which phonological memory can become implicated in FL vocabulary learning will depend upon the ease with which phonological codes can be generated for the FL words to be learned. By exclusively using phonologically legal, easily pronounceable sound sequences in Dutch as the FL vocabulary to be trained, phonological memory may be expected to be implicated to the same extent in all experimental conditions. This guarantees that the emerging effects of the critical experimental manipulations will not be contaminated by differences in pronounceability of the materials used in the various experimental conditions.

However, a side effect of exclusively training phonological forms that can be pronounced easily is that the training effects may be overestimated as compared to the effects obtained if the training words had been taken from a natural language phonotactically dissimilar from the native language of the learners, and therefore less easily pronounceable. Service and Craik (1993) manipulated this variable and indeed obtained higher learning scores for pronounceable pseudowords than for words in a phonotactically unfamiliar language.

To summarize, this series of experiments investigated the learning and retention, over a 1-week interval, of FL vocabulary using a paired-associate learning technique in which native-language words and FL words (the latter in fact being pseudowords) were visually presented in pairs. Both percentage recall and retrieval time were measured, and performance was assessed with both productive and receptive testing. All participants were highly practiced FL learners. The learning materials varied on three word characteristics: word concreteness (or word imageability; see below), word frequency, and cognate status.

Method: Experiments 1a and 1b (Concreteness and Cognate Status)

Participants

Forty participants took part. Twenty of them were tested on productive knowledge of the newly learned vocabulary (Experiment 1a); the remaining 20 were tested on the corresponding receptive knowledge (Experiment 1b). All 40 participants (as well as the participants of Experiments 2a and 2b) were 1st-year psychology students from the University of Amsterdam. They were also all native speakers of Dutch with a considerable amount of FL learning experience. They had received formal training in English for about 3 hr a week all through secondary school, from about age 12 till age 18. Their current university training involved the reading of mostly English textbooks. At secondary school they had received training in other languages as well, mostly at least in German and French, although typically not as much as in English.

Following the third test phase of the experiment (see below) the participants were asked to fill in a language questionnaire. They were asked to list all the languages they had any knowledge of, and to assess their comprehension and production skills in these languages on a 7-point scale (where 7 indicated equal proficiency to the corresponding skill in Dutch and 1 indicated that the skill was completely lacking). By far the majority of the participants listed English, French, and German. Most of the other languages that were listed were mentioned just once across the participants. The average production-skill ratings for the participants in Experiment 1a were 5.4, 3.6, and 3.5 for English, German, and French, respectively. The corresponding ratings for the participants in Experiment 1b were 5.4, 3.7, and 3.4. The average comprehension-skill ratings for the participants in Experiment 1a were 6.1, 4.8, and 4.1, for English, German, and French, respectively. The corresponding ratings for the participants in Experiment 1b were 6.0, 4.2, and 4.0, respectively.[1]

Materials

The stimulus materials for the learning and test phases of Experiments 1a and 1b were based on a set of 60 words selected from a corpus of 440 Dutch words for which we had earlier collected scores on a number of word characteristics, including word imageability (De Groot et al., 1994). Thirty of the selected words had received high imageability ratings, ranging from 6.20 to 6.96 on a 7-point scale; the remaining 30 words had received low ratings, ranging from 1.19 to 3.50. Because imageability and concreteness are highly correlated (typically $r > .90$), for the sake of convenience words with high- and low-imageability ratings will henceforth be called concrete and abstract, respectively; for consistency's sake the imageability variable will be called the concreteness variable. Half of the words in the groups of concrete and abstract words were paired with a cognate translation in the new (pseudo)language, the remaining half with a noncognate translation. A constraint in thinking up translations for the selected Dutch words was that the length of the Dutch words and their translations should be about the same across the four concreteness-by-cognate-status conditions. A second constraint was that the Dutch words in the four cells should be matched on language frequency. Frequency scores were taken from the CELEX corpus, which is based on a count of 42.5 million printed Dutch words (see Burnage, 1990, for a description). The matching on word frequency was based on the logarithm of the total number of occurrences of each word in the CELEX corpus.

In making up the cognate translations, three guidelines were followed: (a) The first letter of the cognate translation should be identical to that of the paired Dutch word, (b) the length difference between the cognate translation and the paired Dutch word should maximally be one letter, and (c) the overlap between the cognate translation and the paired Dutch word should vary between 40% and 75%. Percentage overlap was calculated by dividing the number of common letters in the same position by the total number of letters. Letters common to the stimuli in a translation pair but occurring in different positions were assigned half the weight. Letters in corresponding positions that were different but that had the same or very similar pronunciations in L1 and L2 were also assigned half the weight. Noncognates were made up such that they were not obviously similar to the paired Dutch words in any way (although the same letters in different positions were permitted to occur).

We performed a set of *t* tests to verify that the matching of the materials on word frequency and length in the four cells had succeeded, and that the materials in the cognate and noncognate cells were matched on concreteness. This turned out to be the case: The *p* values of these *t* tests varied between .52 and .63.

Table 1 shows the means and standard deviations of the scores for concreteness (imageability), log word frequency (in the CELEX corpus), and length in the four concreteness-by-cognate-status conditions. The corresponding length information on the FL words is also provided. Appendix A presents the concreteness and frequency information for each individual Dutch word. It also shows the translation made up for each word.

Table 1

Means and standard deviations of the variables concreteness (CON), log frequency (FR), length of the Dutch words (LEN-D), and length of the translations (LEN-T) per stimulus type in Experiments 1a and 1b

Stimulus type		CON	FR	LEN-D	LEN-T
Concrete cognates	*M*	6.52	3.25	5.60	5.47
	SD	0.19	0.67	0.74	0.74
Concrete noncognates	*M*	6.59	3.37	5.47	5.60
	SD	0.20	0.71	0.92	0.83
Abstract cognates	*M*	2.38	3.41	5.40	5.40
	SD	0.60	0.77	1.12	0.63
Abstract noncognates	*M*	2.41	3.24	5.67	5.60
	SD	0.62	0.67	1.11	1.30

Apparatus and Procedure

The experiment was run on an Apple Macintosh computer. The stimuli were presented in black lowercase letters on a light gray background in the center of the screen. A Pascal program controlled the stimulus presentation during learning and testing as well as the recording of the response times in the test phases. A microphone that activated a voice-operated switch registered the participants' responses in the test phases. All participants were tested individually.

Practice phase. To familiarize the participants with the use of the voice switch in the test phases of the experiment, prior to the first learning-then-test session they were first presented with 20 common Dutch words (in Experiment 1a) or with the English translations of these words (in Experiment 1b) for practice. These stimuli appeared one after the other in random order, a new random order for every subsequent participant. The participants were asked to produce the English translations of the Dutch words (Experiment 1a) or the Dutch translations of the English words (Experiment 1b) and to do so as rapidly as possible while making as few errors as possible. They were asked to remain silent in case they did not know the test word's translation, and to try to avoid making any sounds other than the intended response so that the voice switch would not be triggered inadvertently.

Prior to each test word a fixation stimulus (an asterisk) appeared in the middle of the screen for 1 s, just above the position of the subsequent test word. When the fixation stimulus disappeared, the test word appeared and remained on the screen until the voice switch registered the onset of the participant's response. Response time (RT) was measured from the onset of the test word. The experimenter typed the response (what was being typed did not appear on the screen) and then started the next trial by touching the return key. The experimenter also monitored the workings of the voice switch. Premature or delayed voice switch triggers were recorded. The maximum presentation time of the test word was 5 s. If no response was provided within 5 s, the test word disappeared and the next one was presented, again preceded by a fixation stimulus.

Learning-then-test sessions. The practice phase was followed by the first of three learning-then-test sessions. On each learning trial a translation pair (i.e., a Dutch word and its made-up translation) appeared on the screen for 10 s, preceded by a fixation stimulus (an asterisk), which lasted 1 s. The Dutch word and its translation were always presented simultaneously and next to each other—the Dutch word on the left, its translation on the right, and two hyphens connecting the two. The 60 translation pairs were presented in random order (a different order for each participant). After 30 pairs had been presented, the participant was allowed a brief rest before the next 30 pairs were presented. The participants did not receive any instruction as to what learning procedure to adopt. Immediately following the presentation of the second group of 30 translation pairs, all 60 pairs were presented for learning a second time, using exactly the same procedure as during the first round.

After all the translation pairs had appeared a second time, the test phase started. The procedure during testing was the same as in the practice phase (see above), except that the test word remained on the screen for 10 s instead of 5 s and that, of course, the stimulus set differed from the one presented during practice: In Experiment 1a (testing productive vocabulary knowledge) the 60 Dutch words from the trained translation pairs served as stimulus materials, whereas in Experiment 1b (testing receptive knowledge) the corresponding made-up translations of the Dutch words constituted the materials.

Two more such cycles of learning (twice) and testing immediately followed the first cycle. In all then there were six learning trials per translation pair. The translation pairs were always presented in random order, so that a participant saw the pairs in a different order in the different learning sessions. Following the third learning-then-test session each participant filled in a language-proficiency questionnaire (see above). This complete series of events took about 135 min.

Retest. All participants came back to the laboratory a week later to be tested once more, without first receiving the translation pairs again for learning. The participants of Experiment 1a, who had received the Dutch words as the test stimuli the week before, were again presented with the Dutch words. Similarly, the participants of Experiment 1b, who had received the newly learned translations as the test stimuli the week before, were again presented with these newly learned words.

Results: Experiment 1a (Productive Testing)

Recall Scores

For each participant we calculated 16 recall scores, namely, one for each of the 16 concreteness-by-cognate-status-by-session conditions. A score reflected the percentage of trials on which the participant had correctly recalled the response to the corresponding stimulus word in that particular condition. We also calculated a recall score for each item in each of the four sessions, collapsing across the 20 participants. On these recall scores two sets of analyses were performed. The goal of one of these sets was to track the learning process. It concerned two 2 (concreteness) × 2 (cognate status) × 3 (session) ANOVAs on the recall scores for the three test sessions that each immediately followed a learning phase, one ANOVA by subjects and a second by items. The purpose of the second set of analyses, again one by subjects and a second by items, was to focus on retention after the 1-week test-retest interval. It concerned two 2 (concreteness) × 2 (cognate status) × 2 (session) ANOVAs on the recall scores of the third of the three test sessions preceded by a learning episode and of the retest session a week later. The results of these two sets of analyses will be presented under the Learning and Forgetting sections, respectively, throughout this article. The upper part of Table 2 shows all 16 cell means of the learning analyses (Sessions 1 through 3) and the forgetting analyses (Sessions 3 and 4) of Experiment 1a.

Learning. All three main effects were significant (*F1* for the analysis by subjects; *F2* for the analysis by items; throughout this article, all tests adopt a $p < .05$ criterion for significance; we will not report the *p* values for the separate analyses): concreteness: $F1(1, 19) = 64.1$, and $F2(1, 56) = 75.8$; cognate status: $F1(1, 19) = 105.3$, and $F2(1, 56) = 48.8$; session: $F1(2, 38) = 118.3$, and $F2(2, 112) = 412.2$. The recall score for concrete words was 20% higher than for abstract words and it was 16% higher for cognates than for noncognates. Finally, the recall score increased from 44% in Session 1 to 75% in Session 2 to 86% in Session 3.

The two-way interactions between concreteness and session and between cognate status and session were all significant: concreteness and session, $F1(2, 38) = 5.2$, and $F2(2, 112) = 8.7$; cognate status and session, $F1(2,38) = 16.4$, and $F2(2,112) = 16.5$. The cell means for these interactions showed that the increase in recall scores over the three test sessions was larger for abstract words than for concrete words, and larger for noncognates than for cognates. The larger increase for abstract words and noncognates than for concrete words and cognates presumably reflects a ceiling effect in Sessions 2 and 3 for concrete words and cognates.

Table 2

Recall scores and response times for all concreteness-by-cognate-status-by-session conditions (Experiment 1a)

	Recall scores (%)							
	Session 1		Session 2		Session 3		Session 4	
	C	NC	C	NC	C	NC	C	NC
Concrete	69.7	43.7	91.7	79.7	93.4	92.0	78.4	69.0
Abstract	42.7	19.0	74.7	53.4	86.0	72.7	68.4	36.0
	Response times (ms)							
	Session 1		Session 2		Session 3		Session 4	
	C	NC	C	NC	C	NC	C	NC
Concrete	1,880	2,556	1,318	1,823	1,080	1,643	1,833	2,597
Abstract	2,529	3,641	2,130	3,148	1,663	2,183	2,333	3,412

Note. C = Cognates. NC = Noncognates.

Finally, the three-way interaction between concreteness, cognate status, and session was significant both by subjects and by items, $F1(2, 38) = 6.3$, and $F2(2, 112) = 3.1$. It demonstrated that in Session 1, after only two learning trials per translation pair, performance was especially poor in the case of abstract noncognates (19% correct only; see Table 2).

Forgetting. The main effects of concreteness and cognate status were significant, showing better overall recall for concrete words and cognates than for abstract words and noncognates: concreteness: $F1(1, 19) = 28.7$, and $F2(1, 56) = 56.3$; cognate status: $F1(1, 19) = 76.4$, and $F2(1, 56) = 36.8$. The main effect of session was also significant, $F1(1, 19) = 58.1$, and $F2(1, 56) = 215.3$. The recall score in Session 3, immediately after learning, was 23% higher than at retest (Session 4).

The two-way interactions between concreteness and session and between cognate status and session were significant: concreteness and session, $F1(1, 19) = 5.5$, and $F2(1, 56) = 6.7$; cognate status and session, $F1(1, 19) = 68.5$, and $F2(1, 56) = 18.4$. The top part of Table 3 presents these interactions. It shows that more forgetting occurred for the types of words that were the most

Table 3

Recall scores (in percentages) immediately following learning (Session 3) and at retest (Session 4) for abstract vs. concrete words and for cognates vs. noncognates (Experiment 1a: productive testing; Experiment 1b: receptive testing)

	Productive testing				
	Concrete	Abstract		Cognates	Noncognates
Session 3	92.7	79.3	Session 3	89.7	82.3
Session 4	73.7	52.2	Session 4	73.3	52.5
Effect	19.0	27.1	Effect	16.4	29.8
	Receptive testing				
	Concrete	Abstract		Cognates	Noncognates
Session 3	97.7	91.5	Session 3	98.5	90.7
Session 4	88.9	64.9	Session 4	90.7	63.0
Effect	8.8	26.6	Effect	7.8	27.7

difficult to learn: More forgetting occurred for abstract words than for concrete words; similarly, more forgetting occurred for noncognates than for cognates. The three-way interaction between concreteness, cognate status, and session was also significant by subjects, $F1(1, 19) = 4.4$, and marginally significant by items, $F2(1, 56) = 3.1$, $p = .09$, thus qualifying the above first-order interactions. As shown in the recall data of Table 2, Session 3 versus Session 4, the words that were the most difficult to learn, that is, the abstract noncognates, were most susceptible to forgetting.

Response Times

For each participant we calculated four mean RTs for each of the four test sessions, namely, one mean RT for each of the four conditions formed by the two levels of the concreteness and cognate status variables, collapsing across items. We also calculated a mean RT for each of the 60 words in each of the four test sessions, collapsing across the 20 participants. In calculating these means only the RTs for trials where the correct response had been produced were included. On these mean RTs the same two sets of analyses were performed as on the recall data. The lower part of Table 2 shows the cell means of all conditions.

Learning. All three main effects were significant both by subjects and by items: concreteness: $F1(1, 19) = 103.4$, and $F2(1, 56) = 49.7$; cognate status: $F1(1, 19) = 59.6$, and $F2(1, 56) = 46.9$; session: $F1(2, 38) = 32.7$, and $F2(2, 112) = 59.1$. Responses to concrete words were 832 ms faster than to abstract words, and responses to cognates were 732 ms faster than to noncognates. Finally, retrieval time decreased from 2,652 ms in Session 1 to 2,105 ms in Session 2 to 1,642 ms in Session 3. Of the interactions only the one between concreteness and cognate status was significant, but only in the analysis by subjects, $F1(1, 19) = 5.3$, and $F2 < 1$. It demonstrated particularly slow retrieval for abstract noncognates.

Forgetting. The main effects of concreteness and cognate status were significant, showing faster retrieval for concrete words and cognates than for abstract words and noncognates: concreteness: $F1(1, 19) = 37.4$, and $F2(1, 56) = 30.9$; cognate status: $F1(1, 19) = 35.5$, and $F2(1, 56) = 45.7$. The main effect of session was also significant, $F1(1, 19) = 55.6$, and $F2(1, 56) = 134.1$. Retrieval was 902 ms faster immediately after the third learning phase than at retest a week later. None of the interactions were significant.

Results: Experiment 1b (Receptive Testing)

The same analyses as reported for Experiment 1a were performed on the data of the present receptive-testing experiment. Table 4 shows the mean percentage-recall scores and the mean RTs for all 16 cells of the experiment.

Recall Scores

Learning. All three main effects were significant: concreteness: $F1(1, 19) = 42.7$, and $F2(1, 56) = 56.4$; cognate status:

Table 4

Recall scores and response times for all concreteness-by-cognate-status-by-session conditions (Experiment 1b)

	Recall scores (%)							
	Session 1		Session 2		Session 3		Session 4	
	C	NC	C	NC	C	NC	C	NC
Concrete	94.7	61.7	99.7	89.3	99.7	95.7	98.0	79.7
Abstract	68.3	34.3	91.3	70.7	97.3	85.7	83.3	46.3
	Response times (ms)							
	Session 1		Session 2		Session 3		Session 4	
	C	NC	C	NC	C	NC	C	NC
Concrete	1,422	2,442	1,005	1,719	936	1,488	1,382	2,189
Abstract	2,109	3,540	1,504	2,283	1,247	2,046	1,780	2,985

Note. C = Cognates. NC = Noncognates.

$F1(1, 19) = 35.3$, and $F2(1, 56) = 84.3$; session: $F1(2, 38) = 94.2$, and $F2(2, 112) = 265.0$. The recall score for concrete words was 16% higher than for abstract words, and the recall score for cognates was 19% higher than for noncognates; the recall score increased from 65% in Session 1 to 88% in Session 2 to 95% in Session 3.

Two two-way interactions were significant both by subjects and by items: concreteness and session, $F1(2, 38) = 28.3$, and $F2(2, 112) = 29.8$; cognate status and session, $F1(2, 38) = 38.6$, and $F2(2, 112) = 47.1$. As in Experiment 1a, the cell means for these interactions showed that the increase of the recall score over the three test sessions was larger for abstract words than for concrete words, and larger for noncognates than for cognates. The larger increase for abstract words and noncognates than for concrete words and cognates again presumably reflects a ceiling effect in Sessions 2 and 3 for concrete words and cognates.

The two-way interaction between concreteness and cognate status was significant in the analysis by subjects, $F1(1, 19) = 7.5$, but not in the analysis by items ($p > .10$). It demonstrated a particularly poor recall of abstract noncognates. The three-way interaction between concreteness, cognate status, and session was not significant.

Forgetting. The main effects of concreteness and cognate status were significant, showing better overall recall for concrete words and cognates than for abstract words and noncognates: concreteness: $F1(1, 19) = 35.3$, and $F2(1, 56) = 75.3$; cognate status: $F1(1, 19) = 34.1$, and $F2(1, 56) = 106.6$. The main effect of session was also significant, $F1(1, 19) = 60.3$, and $F2(1, 56) = 180.1$. The recall score was 18% higher immediately after learning than at retest.

The two-way interactions between concreteness and session and between cognate status and session were significant: concreteness and session, $F1(1, 19) = 38.0$, and $F2(1, 56) = 43.8$; cognate status and session, $F1(1, 19) = 25.7$, and $F2(1, 56) = 56.2$. The cell means of these interactions are presented in the bottom part of Table 3. Once again more forgetting occurred for the types of words that were the most difficult to learn: More forgetting occurred for abstract words and for noncognates than for concrete words and cognates. These two interactions were qualified by the three-way interaction between concreteness, cognate status, and session, which was significant in the analysis by subjects, $F1(1, 19) = 5.5$, but only marginally significant by items, $F2(1, 56) = 3.8$, $p = .06$. This interaction shows that the drop in recall scores was particularly dramatic for the words that were the most difficult to learn, namely, for the abstract noncognates, and that hardly any loss occurred for the words easiest to learn, the concrete cognates (see Table 4, Session 3 vs. Session 4).

Response Times

Learning. The three main effects were all significant both by subjects and by items: concreteness: $F1(1, 19) = 75.5$, and $F2(1, 56) = 37.0$; cognate status: $F1(1, 19) = 89.8$, and $F2(1, 56) = 83.0$; session: $F1(2, 38) = 55.8$, and $F2(2, 112) = 91.6$. Responses to concrete words were 619 ms faster than to abstract words; responses to cognates were 883 ms faster than to noncognates; retrieval time decreased from 2,378 ms in Session 1 to 1,627 ms in Session 2 to 1,429 ms in Session 3.

The interaction between concreteness and session was significant, but only marginally so in the analysis by items, $F1(2, 38) = 9.3$, and $F2(2, 112) = 2.4$, $p = .10$. The interaction between cognate status and session was significant both by subjects and by items, $F1(2, 38) = 8.3$, and $F2(2, 112) = 6.1$. The cell means for these interactions showed that the decrease of the RT over the three test sessions was larger for abstract words than for concrete words, and larger for noncognates than for cognates. These interactions mimic the analogous interactions in the recall data of Experiment 1 and presumably reflect a floor effect in Sessions 2 and 3 for concrete words and cognates. The interaction between cognate status and concreteness and the three-way interaction were nonsignificant.

Forgetting. The main effects of concreteness and cognate status were once again significant, with shorter RTs for concrete words and cognates than for abstract words and noncognates:

concreteness: $F1(1, 19) = 26.0$, and $F2(1, 56) = 39.7$; cognate status: $F1(1, 19) = 96.5$, and $F2(1, 56) = 108.5$. The main effect of session was also significant: $F1(1, 19) = 51.0$, and $F2(1, 56) = 181.3$. Retrieval was 655 ms faster immediately after the third learning phase than at retest a week later. The interaction between cognate status and session was significant, $F1(1, 19) = 6.0$, and $F2(1, 56) = 9.1$. It showed that the retention interval had slowed down noncognates more than cognates. This finding mirrors the analogous interaction in the recall data of Experiment 1a and the present experiment, and, again, suggests that the retention interval had a relatively large effect on words that were especially difficult to learn in the first place. None of the remaining interactions approached significance.

Overall Analyses of Experiments 1a and 1b

A comparison of the data of the productive testing experiment (Experiment 1a) and the receptive testing experiment (Experiment 1b) suggests that receptive testing produces the best performance, both in terms of retrieval time and in terms of percentage-recall scores (cf. Tables 2 and 4). To provide statistical support for this observation, the same analyses as reported for the separate experiments were again performed, but now including the combined data of Experiments 1a and 1b. These analyses thus included the additional variable type of testing (productive vs. receptive).

The results of these analyses indeed substantiated this observation: All analyses showed a significant main effect of type of testing, with receptive testing always showing the higher recall scores and the shorter retrieval times. The sizes of the effect are presented in the left part of Table 5. On the whole very few interactions with the type of testing variable occurred. This finding suggests that the superior performance with receptive testing is a robust effect that holds in general, irrespective of what types of words are being learned.[2]

Table 5

Main effects of type of testing (productive vs. receptive) in the analyses of the learning data (Sessions 1 through 3) and the forgetting data (Sessions 3 and 4), and for Experiments 1a and 1b and 2a and 2b

	Learning			
	Experiments 1a and 1b		Experiments 2a and 2b	
	Recall	RT	Recall	RT
Productive	68.2	2,133	50.0	2,513
Receptive	82.3	1,812	68.3	2,160
Effect	14.1	321	18.3	353
	Forgetting			
	Experiments 1a and 1b		Experiments 2a and 2b	
	Recall	RT	Recall	RT
Productive	74.5	2,093	58.2	2,554
Receptive	85.7	1,754	73.1	2,416
Effect	11.2	339	14.9	138

Discussion: Experiments 1a and 1b

Even though no fewer than 60 foreign-language words were trained, the recall scores were high, increasing from 44% after two presentations of the translation pairs (Session 1) to 75% after four presentations (Session 2) to 86% after six presentations (Session 3) with productive testing (Experiment 1a), and from 65% to 88% to 95% with receptive testing (Experiment 1b). These scores were of comparable magnitude to those obtained in two studies that tested participants from the same population as used in the present study; these two studies were also comparable to the present study in terms of the number of FL words that were trained. However, in contrast to the present study they used natural languages as the FL to be trained (Lotto & De Groot, 1998; Van Hell & Candia Mahn, 1997, Experiment 1). After two presentations of 60 Dutch-Spanish word pairs in a “rote rehearsal” condition comparable to the present word-association procedure, Van Hell and Candia Mahn obtained a recall score of 69% (noncognates only, receptive testing). Lotto and De Groot obtained recall scores of 71% and 93% after three and six presentations, respectively, of 80 Dutch-Italian word pairs (both cognates and noncognates, productive testing). The fact that the success rate with pseudoword training is comparable to that of training words from an existing language suggests, first, that using made-up pseudowords phonologically legal in the learners' native language as the training materials does not *generally* result in an overestimation of the effects of training as compared to studies that train words from a natural language (although, as pointed out in the introduction, it is likely to overestimate learning as compared to training a phonotactically unfamiliar native language; see Service & Craik, 1993). Second, it suggests that the concern of other researchers (Papagno et al., 1991, p. 342) that learners might not be motivated to learn nonsense material, and that the use of pseudowords in FL learning studies might therefore not be a good practice, is not warranted.

The development of the recall scores over the first three test sessions shows that most of the learning took place during the first four learning trials and then levelled off. The interactions, in the learning analyses, of session on the one hand and concreteness and cognate status on the other hand show that especially the easier words, cognates and concrete words, are close to ceiling after three to four learning trials. The gain of presenting the translation pairs a fifth and a sixth time (in Session 3) is relatively high for noncognates and abstract words. The RT data on the whole parallel the percentage-recall data, the higher recall scores being associated with the shorter retrieval times.

The study by Van Hell and Candia Mahn (1997; see also above) suggests that our participants' recall performance not only is good in an absolute sense, but may also compare favorably to that of other learner groups. These authors compared FL vocabulary learning of two groups of participants that differed from each other in terms of the amount of earlier FL learning experience and that otherwise did not differ from each other in any obvious way.

The more experienced learners were drawn from exactly the same population as we used in the present series of experiments. The recall scores were considerably higher for the experienced learners than for the inexperienced learners. In a similar study, Papagno and Vallar (1995) obtained the same result, using, as we did here, words from a pseudolanguage as the FL words to be learned. It thus seems that previous experience in FL learning may facilitate the subsequent learning of a vocabulary in yet a new language. A reason it may do so is that relevant earlier experience promotes the development of efficient learning strategies. Another reason may be that earlier FL learning experience enhances the capacity of phonological short-term memory, which, in turn, promotes FL vocabulary learning (e.g., Papagno & Vallar, 1995).

Our results clearly show that some types of foreign-language words are easier to learn than others, and that there are differences between types of words in how well they are retained over time. As to learning, more cognate and concrete translations are learned than noncognate and abstract translations, respectively, and the former are retrieved from memory more rapidly than the latter, thus replicating the results of earlier studies (see Ellis & Beaton, 1993b; and Van Hell & Candia Mahn, 1997, for the effect of concreteness, and Ellis & Beaton, 1993b; and Lotto & De Groot, 1998, for the effect of cognate status). The present concreteness effect is reminiscent of the concreteness effect in studies on first-language acquisition (e.g., Brown, 1957; Schwanenflugel, 1991), but the cause of the concreteness effect in first and second language acquisition may be different (see the General Discussion section).

The analyses intended to assess forgetting over time showed that the words that were the easiest to learn in the first place also left the more permanent traces in memory: The recall scores for cognates and concrete words decreased less over the 1-week test-retest interval than those for noncognates and abstract words. The corresponding RT data showed the analogous pattern: The increase in retrieval time after the 1-week test-retest interval was relatively small for cognates and concrete words.

Receptive testing produced the best performance, in terms of both recall scores (see also Ellis & Beaton, 1993a, 1993b; Griffin & Harley, 1996) and retrieval time, and few interactions of the type of testing variable with the remaining variables were significant. The superiority of receptive testing thus appears a general effect that holds for all types of words. We will discuss the possible sources of this effect in the General Discussion section.

Finally, retrieval time was less sensitive to the experimental manipulations than was percentage recall. In many cases statistically significant effects were obtained in terms of percentage recall, but not in terms of retrieval time. Nevertheless, whenever effects on retrieval time *were* obtained, they were consistent with the corresponding effects on percentage recall (fast retrieval accompanied by high recall scores; slow retrieval accompanied by lower recall scores), providing additional support.

The results so far have demonstrated that cognates and concrete words are easier to learn and harder to forget than noncognates and abstract words; that receptive testing produces better performance than productive testing; that our participants are very skillful FL vocabulary learners who learn a lot during a single learning session, especially during the first three or four presentations of the translation pairs; and that most of the formed traces are still accessible (in receptive testing) and retrievable (in productive testing) a week after learning. The next two experiments are exactly the same as Experiments 1a and 1b, except that instead of cognate status the word frequency variable was manipulated. Lotto and De Groot (1998) showed an effect of this manipulation on learning FL words, but the effect was rather small, substantially smaller than the effects of concreteness and cognate status. It therefore seems worthwhile to find out whether it is replicable.

Method: Experiments 2a and 2b (Concreteness and Frequency)

Participants

Forty new participants, drawn from the population described in the Participants section of Experiments 1a and 1b (see also Note 1), took part. Again 20 were tested on productive knowledge of the new vocabulary (Experiment 2a) and the remaining 20 on receptive knowledge (Experiment 2b). The average production-skill ratings for the participants in Experiment 2a were 5.8, 3.4, and 3.2 for English, German, and French, respectively. The corresponding ratings for the participants in Experiment 2b were 5.4, 3.2, and 3.1. The average comprehension-skill ratings for the participants in Experiment 2a were 6.2, 4.0, and 3.5, for English, German, and French, respectively. The corresponding ratings for the participants in Experiment 2b were 6.2, 3.9, and 3.7, respectively.

Materials

The stimulus materials were based on a set of 60 Dutch words selected from the same source of materials as used for Experiments 1a and 1b. This time the goal was to obtain four groups of words that were orthogonally manipulated on word imageability (concreteness) and word frequency, and that were matched on word length. Thirty of the selected words were concrete words, having received imageability ratings that ranged from 6.20 to 6.84 on a 7-point scale; the remaining 30 were abstract words, with imageability ratings that ranged from 1.19 to 3.69 on this scale. We categorized 15 of the words in both of these groups as high-frequency words, with log frequencies in Dutch (based on the CELEX corpus; see Experiments 1a and 1b) ranging from 3.60 to 4.58 across all 30 "high-frequency" words. We assigned the remaining 15 words within both groups to the low-frequency category. Their log frequencies ranged from 1.89 to 2.98. Each of the 60 selected Dutch words was paired with a made-up translation that was orthographically and phonologically dissimilar to the corresponding Dutch word. In other words, this time all translation pairs consisted of noncognates. A constraint in constructing the translations was that their mean length should be about the same across the four concreteness-by-frequency conditions.

We performed a set of *t* tests to verify that the matching on length of the Dutch words and the translations across the four cells had succeeded, that the materials in the concrete and abstract conditions were matched on word frequency, and that the materials in the high-frequency and low-frequency conditions were matched on concreteness. The *p* values of these *t* tests varied between .57 and .98, indicating that our criteria for selection were met.

Table 6 shows the means and standard deviations of the scores on the relevant stimulus characteristics in the four concreteness-by-frequency conditions. Appendix B presents the concreteness and frequency information for each of the Dutch words separately. It also shows the translation made up for each word.

Table 6

Means and standard deviations of the variables concreteness (CON), log frequency (FR), length of the Dutch words (LEN-D), and length of the translations (LEN-T) per stimulus type in Experiments 2a and 2b

Stimulus type		CON	FR	LEN-D	LEN-T
Concrete frequent	*M*	6.51	3.89	5.00	5.27
	SD	0.15	0.34	0.93	0.96
Concrete infrequent	*M*	6.52	2.63	5.53	5.73
	SD	0.18	0.12	1.13	1.28
Abstract frequent	*M*	2.96	3.89	5.67	5.27
	SD	0.57	0.24	1.59	1.03
Abstract infrequent	*M*	2.96	2.66	5.80	5.60
	SD	0.70	0.30	1.37	0.99

Apparatus and Procedure

The same apparatus and procedure were used as in Experiments 1a and 1b. Experiment 2a tested the participants' productive knowledge of the new vocabulary. Experiment 2b tested receptive knowledge of the new vocabulary.

Results: Experiment 2a (Productive Testing)

The same sets of analyses as reported for Experiments 1a and 1b were performed (except, of course, that the earlier cognate status variable was replaced by the frequency variable). Table 7 shows the mean percentage-recall data and the mean RT data for all 16 cells of Experiment 2a.

Recall Scores

Learning. The main effects of concreteness and session were significant: concreteness: $F1(1, 19) = 82.9$, and $F2(1, 56) = 25.1$; session: $F1(2, 38) = 159.6$, and $F2(2, 112) = 459.5$. The recall score

Table 7

Recall scores and response times for all concreteness-by-frequency-by-session conditions (Experiment 2a)

	Recall scores (%)							
	Session 1		Session 2		Session 3		Session 4	
	HF	LF	HF	LF	HF	LF	HF	LF
Concrete	33.3	29.7	66.7	63.3	80.0	76.7	60.3	51.3
Abstract	14.3	19.7	42.3	48.3	61.3	64.3	34.0	37.3
	Response times (ms)							
	Session 1		Session 2		Session 3		Session 4	
	HF	LF	HF	LF	HF	LF	HF	LF
Concrete	2,824	2,572	1,974	2,082	1,979	2,009	2,708	2,512
Abstract	2,946	3,260	2,788	2,691	2,508	2,524	3,140	3,053

Note. HF = High log frequency. LF = Low log frequency.

for concrete words was 17% higher than that for abstract words, and the recall score increased from 24% in Session 1 to 55% in Session 2 to 71% in Session 3. The main effect of frequency was not significant ($F < 1$ in both cases). The two-way interaction between concreteness and frequency was significant in the analysis by subjects, $F1(1, 19) = 10.7$, but not by items ($p > .10$). It showed a slightly larger concreteness effect for high-frequency words than for low-frequency words. None of the remaining interactions were significant.

Forgetting. The main effect of concreteness was significant: $F1(1, 19) = 46.9$, and $F2(1, 56) = 30.4$. The main effect of session was also significant: $F1(1, 19) = 40.6$, and $F2(1, 56) = 210.8$. The recall score was 25% higher immediately after learning than at retest. Neither the main effect of frequency nor the interactions between session and the remaining variables were significant (but concreteness by session: $p = .11$ by subjects, and $p = .18$ by items, providing a hint of the familiar pattern that the words hardest to learn are the first to be forgotten; the decrease in percentage recall over time was 27% for abstract words and 22% for concrete words; for all remaining interactions: $F < 1$).

Response Times

Learning. The main effects of concreteness and session were significant: concreteness: $F1(1, 19) = 57.7$, and $F2(1, 56) = 10.6$; session: $F1(2, 38) = 12.3$, and $F2(2, 112) = 37.7$. Responses to concrete words were 546 ms faster than to abstract words, and retrieval time decreased from 2,901 ms in Session 1 to 2,384 ms in Session 2 to 2,255 ms in Session 3. Neither the main effect of frequency nor any of the interactions were significant.

Forgetting. The main effect of concreteness was significant, $F1(1, 19) = 33.1$, and $F2(1, 56) = 18.5$, concrete words being retrieved faster than abstract words. The main effect of session was also significant: $F1(1, 19) = 9.7$, and $F2(1, 56) = 90.3$. Responding took on average 598 ms less immediately after learning than at retest. Neither the main effect of frequency nor any of the interactions were significant.

Results: Experiment 2b (Receptive Testing)

The same sets of analyses were again performed. Table 8 shows the mean recall and RT data for all 16 cells of Experiment 2b.

Recall Scores

Learning. The main effects of concreteness and session were significant: concreteness: $F1(1, 19) = 87.2$, and $F2(1, 56) = 35.2$; session: $F1(2, 38) = 112.7$, and $F2(2, 112) = 596.2$. The mean recall score for concrete words was 17% higher than for abstract words, and the recall score increased from 43% in Session 1 to 74% in Session 2 to 87% in Session 3. The main effect of frequency was significant in the analysis by subjects, $F1(1, 19) = 5.6$, with 70% recall for high-frequency words and 67% recall for low-frequency words. In the analysis by items this effect was, however, not significant. Finally, the two-way interaction between concreteness and session was significant, $F1(2, 38) = 6.5$, and $F2(2, 112) = 19.1$. The cell means for this interaction showed that, as in Experiments 1a and 1b, the increase of the recall scores over the three test

Table 8

Recall scores and response times for all concreteness-by-frequency-by-session conditions (Experiment 2b)

	Recall scores (%)							
	Session 1		Session 2		Session 3		Session 4	
	HF	LF	HF	LF	HF	LF	HF	LF
Concrete	57.0	53.7	86.0	79.0	92.3	90.7	72.3	70.0
Abstract	33.3	27.0	66.3	65.7	81.7	83.7	47.7	46.3
	Response times (ms)							
	Session 1		Session 2		Session 3		Session 4	
	HF	LF	HF	LF	HF	LF	HF	LF
Concrete	2,282	2,591	1,670	1,906	1,456	1,638	2,510	2,762
Abstract	2,729	3,131	2,250	2,208	2,006	2,052	3,512	3,297

Note. HF = High log frequency. LF = Low log frequency.

sessions was larger for abstract words than for concrete words. None of the remaining interactions were significant.

Forgetting. The main effect of concreteness was significant, $F1(1, 19) = 35.6$, and $F2(1, 56) = 50.5$, with better recall for concrete words. The main effect of session was also significant, $F1(1, 19) = 76.2$, and $F2(1, 56) = 346.8$: Session 3: 87%, and Session 4: 59%. The interaction between concreteness and session was also significant, $F1(1, 19) = 20.1$, and $F2(1, 56) = 26.0$. As in Experiments 1a and 1b, the decrease in recall score was much larger for abstract words (a decrease of 36%; Session 3: 83%; Session 4: 47%) than for concrete words (a decrease of 21%; Session 3: 92%; Session 4: 71%). Thus once again, more forgetting occurred for the words that were the most difficult to learn (see also Table 3). None of the remaining interactions involving session as one of the variables, nor the main effect of frequency, was significant.

Response Times

Learning. The main effects of concreteness and session were significant: concreteness: $F1(1, 19) = 26.6$, and $F2(1, 56) = 15.9$; session: $F1(2, 38) = 36.4$, and $F2(2, 112) = 76.1$. Responses to concrete words were on average 472 ms faster than those to abstract words, and retrieval time decreased from 2,683 ms in Session 1 to 2,009 ms in Session 2 to 1,788 ms in Session 3. The main effect of frequency was significant in the analysis by subjects, $F1(1, 19) = 11.2$, with the mean retrieval time 188 ms shorter for high-frequency words than for low-frequency words. This effect was, however, nonsignificant in the analysis by items. None of the interactions were significant in either analysis ($p > .10$ in all cases).

Forgetting. The main effect of concreteness was significant, $F1(1, 19) = 27.6$, and $F2(1, 56) = 20.5$, with shorter RTs for concrete words. The main effect of session was also significant, $F1(1, 19) = 52.3$, and $F2(1, 56) = 163.2$. Retrieval was 1,232 ms faster immediately after the third learning phase than at retest a week later.

None of the interactions were significant, nor was the main effect of frequency.

Overall Analyses of Experiments 2a and 2b

The same sets of analyses as reported for the combined data of Experiments 1a and 1b were performed on the data of Experiments 2a and 2b. All analyses except one showed a significant main effect of type of testing, the exception being the forgetting analysis by subjects in the RT data. In all cases receptive testing produced the better results. The size of the type of testing effects is presented in the right part of Table 5.

As in Experiments 1a and 1b, very few interactions of the type of testing variable with the remaining variables occurred. Therefore, again the conclusion is warranted that the advantage of receptive testing over productive testing is a robust effect that holds for practically all test conditions.

Discussion: Experiments 2a and 2b

The overall percentage-recall scores were lower than in Experiments 1a and 1b, especially in the productive-testing condition. At the end of training they were 71% for productive testing and 87% for receptive testing (as compared to 86% and 95%, respectively, in Experiments 1a and 1b). There is one reason why this comparison of overall performance in the two sets of experiments is not altogether correct: Whereas in Experiments 1a and 1b both cognate and noncognate translations were trained, Experiments 2a and 2b only included noncognates. A fairer comparison of performance in the two sets of experiments therefore is one that focusses only on the noncognates. Doing so still shows the better performance in Experiments 1a and 1b (cf. Tables 2, only the noncognates, and 7, all word groups, and Tables 4, only the noncognates, and 8, all word groups), although the difference was much more pronounced with productive testing than with receptive testing. No obvious reason for this difference between the two sets of experiments presents itself. One reason might be that the removal of one group of easily learned words, the cognates, in Experiments 2a and 2b exerted an overall negative effect on performance, especially in the hardest test condition (productive testing). A second reason to consider is that the pseudowords used in Experiments 2a and 2b may inadvertently have turned out somewhat less easily pronounceable than those used in Experiments 1a and 1b, so that phonological memory could not become implicated to the same extent as in Experiments 1a and 1b. This possibility is, however, not very likely, given the fact that the noncognate materials in the two sets of experiments were developed in exactly the same way. But whatever the cause of the overall difference in recall performance between the two sets of experiments, it can be concluded that also in Experiments 2a and 2b the performance levels were high and they once again testify to the good FL vocabulary learning skills of our population.

The development of the learning scores over the three test sessions was as in Experiments 1a and 1b: Most of the learning took place during the first four learning trials and then levelled off. With productive testing the recall scores increased from 24% to 55% to 71% over the three learning sessions. The corresponding recall scores with receptive testing were 43%, 74%, and 87%, respectively. The RT analyses mirrored this pattern. The receptive-testing condition once again showed that this levelling off especially concerned the easier words, that is, the concrete words. In the case of abstract words, considerable learning still takes place during the fifth and sixth presentation of the translation pairs.

Word-type effects also again materialized, but mainly concerning the concreteness variable: Concrete words were learned better than abstract words, as demonstrated by higher recall scores and faster retrieval for the former type of words. A small frequency effect on learning showed up in both the recall data and the RT data of the receptive-testing condition, showing faster retrieval and higher percentage-recall scores for high-frequency words than for low-frequency words. In both cases, however, the frequency effect did not generalize over items. The forgetting analyses of the receptive-testing condition again showed differential forgetting for concrete and abstract words, with more abstract than concrete words forgotten a week after training. However, in the productive-testing condition this effect failed to be significant (although it approached marginal significance: $p = .11$ by subjects; $p = .18$ by items; with one-tailed testing both effects would in fact be marginally significant).

Type of testing again had a large effect, both in terms of recall scores and retrieval time, and the fact that only few interactions between this variable and the remaining variables occurred suggests that the superior performance with receptive testing generally holds, irrespective of word type.

General Discussion

Why is it easier to learn an FL translation for concrete words and cognates than for abstract words and noncognates, and why did word frequency not have an effect in this study? Why are the words that are the easiest to learn retained best? What causes receptive testing to be superior to productive testing? Following we will provide possible explanations of our results.

Cognate Status

The effect of cognate status replicates the effect of this variable in Lotto and De Groot (1998), where Dutch undergraduates drawn from the same population as tested here learned Italian vocabulary. At the time we suggested three possible causes of the effect, two that localized the effect in the learning stage and one that attributed it to an advantage of cognates over noncognates in the recall stage: (a) Due to the form overlap between cognate translations and the absence of such overlap between noncognate translations, in the case of cognate learning there is in fact less to learn than when noncognates are presented for learning. (b) Whereas in noncognate learning new entries have to be created in memory, cognate learning may only involve adding new information to, or adapting, memory representations that already existed in memory prior to learning. The former process may be more demanding than the latter, thus causing the disadvantage for noncognates. This account tallies with the view that cognates but not noncognates share a representation in bilingual memory (e.g., Kirsner, Lalor, & Hird, 1993; Sánchez-Casas, Davis, & García-Albea, 1992). (c) Due to the form overlap between cognate translations, a cognate test stimulus will be a strong cue for the retrieval of the corresponding translation.

An informative additional finding in the study by Lotto and De Groot (1998) was that the cognate effect occurred not only when the present word-association learning procedure was employed, but also, and to the same extent, when during training the Italian FL words were paired with the pictures representing the words' meanings. Yet, pictures do not share aspects of their form with the corresponding FL words. Because it most likely is the form relation between cognate translations that underlies the cognate advantage during learning, from the occurrence of the effect following picture-association learning and word-association learning alike it follows that the participants in picture-association FL learning generated the corresponding forms. Lotto and De Groot argued that the forms involved are the phonological forms, not the orthographic ones. They concluded that their data squared with the support for the role of phonology in FL vocabulary learning as provided by other researchers (e.g., Ellis & Beaton, 1993b; Papagno et al., 1991; Papagno & Vallar, 1995; Service & Craik, 1993; see also the introduction).

Concreteness

Effects of concreteness abound in studies on the learning and representation of verbal materials, and indeed a couple of demonstrations of the specific present concreteness effect—in the paired-associate learning of familiar, native-language words paired with previously unfamiliar letter sequences—have also been reported before (Ellis & Beaton, 1993b; Van Hell & Candia Mahn, 1997). As to the source of these effects, different hypotheses have been advanced and tested. Dual-coding theory (e.g., Paivio, 1986; Paivio & Desrochers, 1980) poses that concrete words are stored in both a verbal system and an image system, whereas most abstract words are only stored in the verbal system. The additional representation in the image system is thought to benefit performance on concrete words in a multitude of tasks, both tasks that focus on the learning of verbal materials and tasks that focus on their representation. In terms of dual-coding theory, the present concreteness effect might result from the fact that a new name for a concept can be attached to two memory representations in the case of concrete words (namely, to the image representation and the L1 verbal representation) but to only one when the new name is paired with an abstract word (the L1 verbal representation). In other words, when the new name is paired with a concrete word there is more opportunity to anchor it in memory than when it is paired with an abstract word.

A competitor of dual-coding theory is context-availability theory, which holds that comprehension processes are supported by the addition of contextual information—from the linguistic environment, memory, or both—to the materials that are to be understood, and that comprehension suffers if the retrieval of contextual information fails or falters (e.g., Bransford & McCarrell, 1974; Schwanenflugel, 1991; Schwanenflugel & Shoben, 1983). Concrete and abstract words usually differ from each other in terms of how easy it is to think of a context for them, as demonstrated by the high correlation that typically occurs between the concreteness and the context availability of words. (A word's context availability is usually assessed in rating studies where participants indicate on a 7-point scale how easy or difficult they think it would be to think up a context for the word.) Differences in context availability may therefore underlie the concreteness effect. That this is indeed the case is strongly suggested by the results of a number of studies—using various monolingual (e.g., lexical decision) and crosslanguage (e.g., word translation) tasks—that have attempted to dissociate the effects of these two variables (De Groot, 1992; Schwanenflugel, Harnishfeger, & Stowe, 1988; Schwanenflugel & Shoben, 1983; Van Hell & De Groot, 1998). The typical result was that the common concreteness effect materialized when concreteness and context availability were confounded with each other, but not when concrete and abstract words were matched on context availability (that is, on the above ratings of how easily a context can be thought up for them).

In one study by Sjarbaini (1998), concreteness and context availability were disentangled in an FL vocabulary learning experiment. When the common confounding with context availability occurred, a small concreteness effect in learning FL vocabulary occurred (the percentages recall were 65% and 61% for concrete and abstract words, respectively; receptive testing was used). When concrete and abstract words were matched on context availability, the concreteness effect was not significant (63% and 62%, respectively). The context-availability account may thus also apply to concreteness effects in FL vocabulary learning. That it may indeed also apply to the concreteness effects in the present study (though note that these were much larger than the concreteness effect obtained by Sjarbaini) is suggested by a set of a posteriori ANOVAs that we performed on the context-availability scores of our experimental words and that was run to determine whether and how the context-availability variable covaried with the word-type manipulations focussed on in this series of experiments. The context-availability scores for the Dutch words used in our experiments were derived from the corpus of De Groot et al. (1994). As expected, our materials also contained the natural confound between concreteness and context availability: The context-availability scores for the concrete and abstract words used in Experiments 1a and 1b were 5.63 and 3.38 (scores on a 7-point scale); for the concrete and abstract words in Experiments 2a and 2b they were 5.52 and 3.79. In both cases the difference was highly significant ($p < .00001$). Cognates and noncognates (in Experiments 1a and 1b) and high-frequency and low-frequency words (in Experiments 2a and 2b) did not differ in context availability (cognates: 4.48 vs. noncognates: 4.53; high-frequency words: 4.67 vs. low-frequency words: 4.65; $p > .65$ in both cases).

It is unclear what participants actually do in a task where they have to rate words on how easy or difficult they think it would be to think up a context for them, but it is not at all unlikely that their ratings are in fact based on the amount of information in the underlying memory representations. According to this view, when rating words on context availability, the participants access the corresponding memory representations and assign a word a high rating when the representation is informationally dense, whereas they assign it a low rating when the representation is informationally thin (see also Van Hell, 1998). Some researchers (De Groot, 1989; Kieras, 1978) have argued that the memory representations of concrete words contain more information than those of abstract words. This is suggested by the fact that in a "continued free word association" task more associations (in other words, larger *m* scores; Noble, 1952) are obtained for concrete stimulus words than for abstract stimulus words (De Groot, 1989). These *m* scores presumably directly reflect the amount of information in the word representations. Given this difference between the representations of concrete and abstract words, the present view of what participants actually do when they provide context-availability ratings would explain why concreteness and context availability are usually confounded: Context-availability ratings are usually higher for concrete than for abstract words because concrete words are stored in memory representations that are informationally more dense than are the representations of abstract words. In terms of the present views, matching concrete and abstract words on context availability amounts to matching them on the number of information units in the underlying memory representations.

This difference between concrete and abstract words—that, as hypothesized here, underlies the difference in context availability between these two types of words—may also be the source of the present concreteness effect in FL vocabulary learning: The larger number of elements stored in the memory representations of concrete words provides relatively many opportunities to anchor the new FL words in memory. This account of the concreteness effect in FL vocabulary learning is ultimately very similar to the one in terms of dual-coding theory suggested earlier. In both cases richer representations for concrete words than for abstract words are thought to favor the process of attaching new names to the representations of concrete words. The critical difference, however, is that dual-coding theory assumes the comparative informational riches of the representations of concrete words is due to (modality-specific) image representations that are typically only stored for concrete words. The present view is less specific in that it only assumes that more information is stored for concrete words than for abstract words, irrespective of the nature of the stored elements. These may well be all amodal, for concrete and abstract words alike (see also De Groot, 1989). In other words, unlike dual-coding theory the present account does not assume qualitative, but only quantitative differences between the representations of concrete and abstract words (see also Van Hell, 1998).

The present view of concreteness effects in FL vocabulary learning cannot account for the different acquisition rates for concrete and abstract words in first-language acquisition (e.g., Brown, 1957; Schwanenflugel, 1991): Word representations in memory are presumably built up from scratch and differences between word representations for concrete and abstract words will therefore be the outcome of the acquisition process, not the starting point. In other words, at the onset of first-language acquisition the representations of concrete and abstract words are both informationally thin. What presumably causes the concreteness effect in first-language acquisition is that the learning of concrete words, but not of abstract words, is supported by the perceptual presence of these words' referents in the child's environment.

Implications for models of bilingual memory? The finding that word concreteness affects FL vocabulary acquisition not only in the later stages of learning, but also from the onset of training onward does not fit comfortably with the "developmental" model of bilingual memory (Chen & Leung, 1989; De Groot & Hoeks, 1995; Kroll & Curley, 1988). This model states that at lower levels of FL proficiency translation pairs are stored in memory in "word-association" structures and that translation comes about via direct (word-association) connections between the form representations of the words in a translation pair; conceptual memory (where meaning is stored) is assumed not to be involved. In contrast, for higher levels of FL proficiency translation pairs are assumed to be stored in "concept-mediation" structures, and translation is thought to involve conceptual memory (see, e.g., Chen & Leung, 1989; and Potter, So, Von Eckardt, & Feldman, 1984, for details). According to the developmental model, effects of word concreteness should only occur at the higher levels of FL proficiency (the reason being that word concreteness involves an aspect of the meaning, not the form, of words, and that effects of manipulating this variable thus suggest the processing of the word's meaning, and, by implication, the involvement of conceptual memory, during task performance). The present concreteness effects, full-blown right from the onset of FL learning, thus appear to provide a challenge for the developmental model.

Altarriba and Mathis (1997) and Yang (1997) also provided data that are problematic for the developmental model. Altarriba and Mathis showed that immediately after having been trained on a set of previously unknown FL words, the learners demonstrated semantic interference effects in a translation-recognition task on crosslanguage pairs that included the newly trained words. Similarly, already after 10 hr of training in the artificial language "Keki," Yang (1997) obtained a semantic priming effect of Keki target words preceded by L1 English primes. In fact this crosslanguage priming effect was equally large as the effect obtained for English targets preceded by English primes. Both of these findings suggest that conceptual memory is already implicated during early stages of FL acquisition.

Furthermore, the fact that word-concreteness effects occur with productive and receptive testing alike is a result that appears problematic for the "revised hierarchical model" of bilingual memory (Kroll & Stewart, 1990, 1994). This model posits that translation from L1 to L2 (here, "productive testing") predominantly implicates conceptual memory, whereas translation from L2 to L1 ("receptive testing") relatively often comes about via the direct connections between the translations' form representations (in other words, via the "word-association" connections mentioned above). Under the assumption that an effect of the semantic variable concreteness reflects the involvement of conceptual memory, the model would thus predict a larger concreteness effect with productive testing than with receptive testing. The finding that the effect is not modulated by type of testing therefore suggests that conceptual memory might be implicated to the same extent in both directions of translation (see also De Groot & Poot, 1997; La Heij, Hooglander, Kerling, & Van der Velden, 1996).

Translation from L1 to L2 is often slower than translation from L2 to L1 (Kroll & Sholl, 1992), an effect that was also obtained in the present study. This finding played a major role in the development of the revised-hierarchical model (see Kroll & Stewart, 1994). The model attributes this effect to the different memory routes assumed to be implicated in the two directions of translation: The (longer) route through conceptual memory, used in L1 to L2 translation, takes more time than the (shorter) direct "word-association" route used primarily in L2 to L1 translation. If, however, as the present concreteness effects suggest, productive and receptive testing (in other words, L1 to L2 translation and L2 to L1 translation) both involve access to conceptual memory, the present RT difference between productive and receptive testing must have another cause. A number of causes for the superiority of receptive testing—in terms of both retrieval time and percentage recall—will be presented below.

It remains to be seen, however, whether translation processes immediately following FL learning are the same as those involved when bilinguals are tested on FL knowledge that was already established in memory before they entered the laboratory. The models of bilingual memory discussed in this section were based on experiments that tested the latter type of knowledge. Because of this difference between the training experiments on the one hand and the "representation" experiments on the other hand, caution should be exercised in blindly using results from the former to inform models based on the latter.

Word Frequency

Word frequency played a marginal role in this study. An effect of this variable only occurred in the learning analyses of Experiment 2b (on both the RT and percentage-recall data), and in both cases it was only significant in the analyses by subjects. Furthermore, as compared to the effects of cognate status and concreteness, the frequency effect was very small (an effect of 188 ms in the RT analysis and of 3% in the analysis of percentage recall). The effect was also small in Lotto and De Groot (1998; 100 ms and 7%, respectively) and, in fact, the small size of the effect at the time motivated the reinclusion of the frequency manipulation in the present study. It seemed worthwhile to find out whether the small effect obtained then was replicable and should, thus, be accounted for.

A study by De Groot (1989) provided support for the idea that the memory representations of frequent words contain (at least sometimes) *slightly* more information than those of less frequent words. Lotto and De Groot (1998) therefore suggested that the small frequency effect they obtained might have been caused by a small difference in information density between their frequent and less frequent materials (cf. our interpretation of the concreteness effect; see above). If that account of the frequency effect was correct, it does not have to come as a surprise that we generally did not obtain the effect here: We have suggested that context-availability ratings of words reflect the amount of information in the words' memory representations. The analyses on the context-availability ratings of the present materials (see above) showed that the present high- and low-frequency words did not differ in context availability, and thus (as assumed) did not differ in information density either. It thus seems that, unlike concreteness, word frequency is not generally confounded with information density.

Word-Type Effects on Retention

Our finding that words that are the easiest to learn also leave the more permanent memory trace is consistent with the results of Bahrick and Phelps (1987), who tested the residual Spanish vocabulary of a group of participants trained on English-Spanish translation pairs 8 years earlier. These authors observed that retention after 8 years was best for the Spanish words that at the time had required the fewest learning trials to obtain criterion performance. Bahrick and Phelps (1987) looked at the relation between learning and forgetting at a global level, that is, without focussing on the role specific word characteristics might play in learning and forgetting. Our results thus extend their finding by showing that a word's concreteness and cognate status turn out to greatly influence whether (or how soon) it is mastered and how well it will be retained over time.

Atkinson (1972) distinguished between three states in which an FL word can be during or immediately after training: P (for "permanent"), T (for "temporary") and U (for "unknown"). A word in State P is known and the knowledge concerned is permanent. A word in State T is only temporarily known, so that subsequent learning of other words will cause interference, resulting in forgetting of this word known previously. A word in State U is unknown. Immediately after training, FL words in States P and T will lead to a correct response and words in State U will, of course, lead to an error or to no response whatsoever. Retesting after a delay will only lead to a correct response for State P items, and not only State U items but also State T items will lead to failure. In other words, forgetting will have occurred for words previously in State T, but not for words previously (and still) in State P.

In terms of Atkinson's model, the present finding that recall performance for noncognates and abstract words deteriorated more over the test-retest interval than did the recall of cognates and concrete words indicates that relatively many of the non-cognates and abstract words were in State T immediately after training—or vice versa, that relatively many of the cognates and concrete words had already reached State P at the conclusion of the training phase. Note however that also for cognates and concrete words forgetting occurred, indicating that not all of them had reached State P either.

The Superiority of Receptive Testing

A number of possible causes of the superior performance with receptive testing have been advanced. Horowitz and Gordon (1972) held the difference in availability between native-language words and foreign-language words, the former being more available than the latter, responsible for the effect. A difference in availability is likely to reflect a difference between words in how well they are established in memory. If true, the "availability" account is consistent with another account of the superiority of receptive testing, namely, that there is an inherent difference between comprehension and production tasks: Comprehension tasks are easier and can be performed on the basis of memory traces that are less well consolidated than can production tasks (e.g., Griffin & Harley, 1996). Receptive testing of newly learned FL words only requires the comprehension of the latter, whereas productive testing requires their production. The newly established memory representations will therefore often be consolidated well enough to lead to successful performance in receptive testing but too poorly to do so in productive testing.

A similar account of the effect, provided by Ellis and Beaton (1993a), focusses on a difference between productive and receptive testing in terms of the activation patterns produced by native-language words on the one hand and newly learned FL words on the other hand. The representation of a native-language word in memory has many connections to other word representations in the native-language system; in addition, it is connected to the representation of the newly learned FL word. In contrast, the representation of the newly learned FL word is only linked up with the representation of its translation equivalent in the native language and is not yet embedded in a rich FL memory network. As a result, the presentation of the native-language word of a translation pair will give rise to a large amount of activation in the native-language system. This activation predominates the activation the representation of the newly learned translation equivalent will receive via the crosslanguage memory link between the representations of the words in the translation pair. As a consequence, the FL word will often not become available. Because of the absence of connections to other elements in the FL language system, no such predominance of activation in a non-targeted part of the memory system will take place when the newly learned FL word is presented for translation in the native language. Consequently, the representation of the native-language word will receive all the activation sent off by its FL translation, and the corresponding word will be readily available.

Finally, the fact that only productive testing requires the production of an articulation pattern that is unfamiliar was suggested as yet another possible cause of the superiority of the receptive-testing condition (see Ellis & Beaton, 1993a). Given the fact that these accounts do not appear to be mutually exclusive, several of them may hold and these may jointly produce the observed effect.

A relevant component of Horowitz and Gordon's (1972) account of the superiority of receptive testing (see above) is the "principle of associative symmetry," which holds that the presentation of a stimulus pair A-B leads not only to a bidirectional association between the two elements in the pair (rather than to a unidirectional association), but also to one that is of equal strength in both directions. That the associations formed are indeed bidirectional is supported by the present study as well as by other studies (e.g., Griffin & Harley, 1996): Despite the fact that during learning only L1-L2 pairs and no L2-L1 pairs were presented, recall performance is often successful when an L2 word is the test stimulus (and in fact it is even more successful). However, Horowitz and Gordon's (1972) assumption that the association is equally strong in both directions has been invalidated by Griffin and Harley (1996), who demonstrated that the presentation of A-B pairs leads to stronger "forward" associations, from A to B, than "backward" associations, from B to A. Other things being equal, this results in better performance with the A-terms of the pairs presented during testing than with presentation of the B-terms. Given this result, our finding of a large and consistent advantage for receptive testing, with L2 words as the test stimuli, is even more impressive. It occurred despite the fact that learning was always with L1-L2 pairs, and the connection from L2 to L1 was, therefore, the weaker one of the two.

Implications for FL Training Programs

Our findings lead to several implications and recommendations for FL training programs. First, if a goal of the course designers is to introduce words that are relatively easy to learn before more difficult ones, they should include relatively many cognates and concrete words in the initial phases of training. One reason for wanting to introduce easy words before more difficult ones might be that the ensuing learning success will be highly motivating for the learners. A second reason could be that the initial teaching of easy FL words will relatively soon provide the learners with a large enough basic vocabulary to continue learning independently of the teacher (by reading FL texts on their own). This holds especially if concrete words and cognates with a high usage frequency in the FL will be selected for inclusion in the initial stages of training (see, e.g., Nation, 1993, for the relation between word frequency and text coverage).

A note of warning regarding the use of cognates in FL training programs is in order though. In a discussion of the work on second language acquisition and cognates and the practical implications of this work for the teaching of FL vocabulary, Meara (1993) distinguished four patterns of "cognacy" relations between pairs of languages, each of them demanding a different type of vocabulary teaching method. The present suggestion to introduce many cognates in the initial phases of training in order to speed up initial vocabulary learning may only suit the cognacy relation that was implicitly assumed in the present study (namely, the one where the native and foreign language share many cognates and where the words in cognate pairs share roughly the same meaning and are used in the same situations in the two languages). If one of the other cognacy relations between the L1-FL language pair exists, the excessive exploitation of cognates in the training program may either not be an option (because few cognates exist), or it may ultimately lead to contextually inappropriate or even offensive and ridiculous FL production (Meara, 1993).

The ultimate goal of FL vocabulary training is permanent storage of the new vocabulary, not temporary storage. Therefore, our finding that abstract words and noncognates lead to more transient FL memory traces than do concrete words and noncognates leads to the recommendation that FL training programs be designed such that abstract and noncognate translation pairs are retrained more than concrete and cognate translation pairs. A relatively simple way to effectuate this might be to group the presentation of different types of translation pairs in the textbooks (that is, to present concrete cognates, concrete noncognates, abstract cognates, and abstract noncognates in separate groups), at the same time marking the level of difficulty of the various groups visually, and to train the learners to pay more attention to and relearn more often the groups marked to be difficult than the groups marked to be relatively easy to learn.

Finally, the finding that receptive testing (from L2 to L1) produces better recall than productive testing (from L1 to L2), even though the L1-L2 presentation order of the translation pairs during training will have led to stronger links from L1 to L2 than from L2 to L1, prompts the recommendation (see also Griffin & Harley, 1996) that FL vocabulary training programs present the translation pairs in the order L1-L2 rather than in the order

L2-L1. This way the more difficult skill, that is, FL production, will be boosted.

Limitations of This Study

Extensive as this study is, the present experimental manipulations have ignored two important facts regarding FL vocabulary learning. The first concerns the important role of phonological memory in many language processes (see Gathercole & Baddeley, 1993, for an overview), including vocabulary learning (Cheung, 1996; Ellis & Beaton, 1993b; Gathercole & Baddeley, 1990; Papagno et al., 1991; Papagno & Vallar, 1995; Service, 1992; Service & Craik, 1993). As pointed out earlier, the fact that we only used pseudowords phonologically legal in the participants' native language as the FL words to be learned (thus promoting the involvement of phonological memory in learning) is likely to have overestimated the effect of training as compared to the learning that would have taken place had we trained nonwords phonologically illegal in Dutch, or actual words from a natural language with a phonological structure very different from Dutch (Service & Craik, 1993). Therefore, our results probably inform the process of learning natural languages phonologically similar to L1 more than they inform the learning of phonologically unfamiliar languages.

A second limitation is that we have tested a rather restricted view of FL vocabulary learning. The goal of FL vocabulary learning is to associate the new FL word forms with their meanings, not to associate them with the corresponding word forms in the native language. Yet, we have not *directly* tested whether the participants indeed assigned meaning (namely, the meaning of the paired Dutch forms) to the FL word forms. However, the ubiquitous presence of the concreteness effect (a semantic effect), and specifically its occurrence when the newly learned FL words were presented as the test stimuli, strongly suggest that meaning is processed in L1-FL paired-associate learning and indeed assigned to the FL words. Nevertheless, because the meanings of FL words and their translations in L1 do not generally overlap completely, to reach the ultimate goal of contextually appropriate use of the newly learned vocabulary, the paired-associate FL training technique will have to be supplemented with other training techniques, such as presenting the FL words in appropriate FL contexts.

Revised version accepted 23 July 1999

Notes

[1] To be able to compare directly the results from all experiments of this study, it is important that the participants of all four experiments shared the same characteristics and were equally proficient foreign-language users. To find out whether this requirement was met, a 4 (experiment) by 2 (type of skill: comprehension vs. production) by 3 (language: English, German, and French) ANOVA was performed on the participants' production and comprehension ratings in each of these three languages. The results of this analysis indicated that indeed comparable groups had participated in the four experiments: The main effect of experiment was not significant ($F < 1$). The overall proficiency scores varied between 4.44 and 4.66 across the four experiments. Furthermore, the interactions between experiment on the one hand and type of skill and language on the other hand were not significant (experiment by type of skill: $F[3, 67] = 1.6, p > .10$; experiment by language: $F[6, 134] = 1.5, p > .10$). The analysis furthermore showed that the participants' comprehension skill in these three foreign languages was better than their production skill (mean ratings of 4.83 and 4.20, respectively; $F[1, 67] = 64.2$), and that the participants had better knowledge of English than of German, the latter in turn being better than their knowledge of French. The mean ratings for these three languages were 5.84, 4.08, and 3.63, in the above order. Type of skill did not interact significantly with language, $F(2, 134) = 1.2, p > .10$, a finding that indicates that the proficiency difference between comprehension and production was the same for all three languages. Finally, the second-order interaction between the three variables did not approach significance ($F < 1$). These results support the conclusion that similar groups of participants were tested in the four experiments of this study.

[2] The successful retrieval of a response in an early test session may affect the retrieval of this same response in a later test session. If true, the better recall in the later test sessions for pseudowords paired with concrete words and cognates than with abstract words and noncognates may at least in part result from the fact that the former types of stimuli have more often led to successful *recall* than the latter have, rather than from differences in *learning*. The results of one further analysis on a subset of the data of Experiments 1a and 1b were consistent with this idea: The RT data of Session 2 in both

Experiments 1a and 1b were split into two parts: (1) The RTs for correct responses to items that were also correctly responded to in Session 1, and (2) the RTs for correct responses to items that were not yet correctly responded to in Session 1. On these data a 2 (concreteness) by 2 (cognate status) by 2 (earlier recall: yes vs. no) by 2 (type of testing: productive vs. receptive) ANOVA was performed. In addition to the familiar effects of concreteness, cognate status, and type of testing, a significant effect of the variable earlier recall occurred. Items recalled successfully for the second time were retrieved 706 ms faster on average than items recalled correctly for the first time (1,508 ms vs. 2,214 ms; $p < .0001$). The interaction between type of testing and earlier recall was not significant, $F < 1$, indicating that the effect of earlier recall was statistically equally large with productive and receptive testing. The interpretation of the faster retrieval of words successfully recalled before is, however, not at all unequivocal. The shorter retrieval times may simply reflect the fact that these items are easier to learn in the first place (as evidenced by the fact that, unlike other items, they were already recalled in Session 1) rather than resulting from previous successful retrieval. But importantly, whatever its source, this effect does not invalidate our conclusion that concrete words and cognates are easier to learn than abstract words and noncognates, because the advantage for the former types of words already materializes in the first test session, where none of the responses has been recalled before.

References

Altarriba, J., & Mathis, K. M. (1997). Conceptual and lexical development in second language acquisition. *Journal of Memory and Language, 36,* 550–568.

Atkinson, R. C. (1972). Optimizing the learning of a second-language vocabulary. *Journal of Experimental Psychology, 96,* 124–129.

Atkinson, R. C. (1975). Mnemotechnics in second-language learning. *American Psychologist, 30,* 821–828.

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

Bransford, J. D., & McCarrell, N. S. (1974). A sketch of a cognitive approach to comprehension: Some thoughts about understanding what it means to comprehend. In W. B. Weimer & D. S. Palermo (Eds.), *Cognition and the symbolic processes* (pp. 189–229). Hillsdale, NJ: Lawrence Erlbaum Associates.

Brown, R. W. (1957). Linguistic determinism and the part of speech. *Journal of Abnormal and Social Psychology, 55,* 1–5.

Burnage, G. (1990). *CELEX: A guide for users.* Nijmegen, The Netherlands: SSN.

Chen, H.-C., & Leung, Y.-S. (1989). Patterns of lexical processing in a non-native language. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 316–325.

Cheung, H. (1996). Nonword span as a unique predictor of second-language vocabulary learning. *Developmental Psychology, 32,* 867–873.

De Groot, A. M. B. (1989). Representational aspects of word imageability and word frequency as assessed through word association. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 824–845.

De Groot, A. M. B. (1992). Determinants of word translation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 1001–1018.

De Groot, A. M. B., Dannenburg, L., & Van Hell, J. G. (1994). Forward and backward word translation by bilinguals. *Journal of Memory and Language, 33,* 600–629.

De Groot, A. M. B., & Hoeks, J. C. J. (1995). The development of bilingual memory: Evidence from word translation by trilinguals. *Language Learning, 45,* 683–724.

De Groot, A. M. B., & Poot, R. (1997). Word translation at three levels of proficiency in a second language: The ubiquitous involvement of conceptual memory. *Language Learning, 47,* 215–264.

Ellis, N., & Beaton, A. (1993a). Factors affecting the learning of foreign language vocabulary: Imagery keyword mediators and phonological short-term memory. *Quarterly Journal of Experimental Psychology, 46a,* 533–558.

Ellis, N. C., & Beaton, A. (1993b). Psycholinguistic determinants of foreign language vocabulary learning. *Language Learning, 43,* 559–617.

Gathercole, S. E., & Baddeley, A. D. (1990). The role of phonological memory in vocabulary acquisition: A study of young children learning arbitrary names of toys. *British Journal of Psychology, 81,* 439–454.

Gathercole, S. E., & Baddeley, A. D. (1993). *Working memory and language.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Griffin, G., & Harley, T. A. (1996). List learning of second language vocabulary. *Applied Psycholinguistics, 17,* 443–460.

Horowitz, L. M., & Gordon, A. M. (1972). Associative symmetry and second-language learning. *Journal of Educational Psychology, 63,* 287–294.

Kieras, D. (1978). Beyond pictures and words: Alternative information-processing models for imagery effects in verbal memory. *Psychological Bulletin, 85,* 532–554.

Kirsner, K., Lalor, E., & Hird, K. (1993). The bilingual lexicon: Exercise, meaning, and morphology. In R. Schreuder & B. Weltens (Eds.), *The* *bilingual lexicon* (pp. 215–248). Amsterdam/Philadelphia: John Benjamins.

Kroll, J. F., & Curley, J. (1988). Lexical memory in novice bilinguals: The role of concepts in retrieving second language words. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory: Current research and issues: Vol. 2. Clinical and educational implications* (pp. 389–395). Chichester: John Wiley & Sons.

Kroll, J. F., & Sholl, A. (1992). Lexical and conceptual memory in fluent and nonfluent bilinguals. In R. J. Harris (Ed.), *Cognitive processing in bilinguals* (pp. 191–204). Amsterdam: North-Holland.

Kroll, J. F., & Stewart, E. (1990, November). *Concept mediation in bilingual translation.* Paper presented at the 31st Annual Meeting of the Psychonomic Society, New Orleans, LA.

Kroll, J. F., & Stewart, E. (1994). Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language, 33,* 149–174.

La Heij, W., Hooglander, A., Kerling, R., & Van der Velden, E. (1996). Nonverbal context effects in forward and backward word translation: Evidence for concept mediation. *Journal of Memory and Language, 35,* 648–665.

Lotto, L., & De Groot, A. M. B. (1998). Effects of learning method and word type on acquiring vocabulary in an unfamiliar language. *Language Learning, 48,* 31–69.

Meara, P. (1993). The bilingual lexicon and the teaching of vocabulary. In R. Schreuder & B. Weltens (Eds.), *The bilingual lexicon* (pp. 279–297). Amsterdam/Philadelphia: John Benjamins.

Nation, P. (1993). Vocabulary size, growth, and use. In R. Schreuder & B. Weltens (Eds.), *The bilingual lexicon* (pp. 115–134). Amsterdam/Philadelphia: John Benjamins.

Noble, C. E. (1952). An analysis of meaning. *Psychological Review, 59,* 421–430.

Paivio, A. (1986). *Mental representations: A dual-coding approach.* New York: Oxford University Press.

Paivio, A., & Desrochers, A. (1980). A dual-coding approach to bilingual memory. *Canadian Journal of Psychology, 34,* 388–399.

Papagno, C., Valentine, T., & Baddeley, A. (1991). Phonological short-term memory and foreign-language vocabulary learning. *Journal of Memory and Language, 30,* 331–347.

Papagno, C., & Vallar, G. (1995). Verbal short-term memory and vocabulary learning in polyglots. *Quarterly Journal of Experimental Psychology, 48A,* 98–107.

Potter, M. C., So, K.-F., Von Eckardt, B., & Feldman, L. B. (1984). Lexical and conceptual representation in beginning and proficient bilinguals. *Journal of Verbal Learning and Verbal Behavior, 23,* 23–38.

Sánchez-Casas, R. M., Davis, C. W., & García-Albea, J. E. (1992). Bilingual lexical processing: Exploring the cognate/noncognate distinction. *European Journal of Cognitive Psychology, 4,* 293–310.

Schwanenflugel, P. J. (1991). Why are abstract concepts hard to understand? In P. J. Schwanenflugel (Ed.), *The psychology of word meanings* (pp. 223–250). Hillsdale, NJ: Lawrence Erlbaum Associates.

Schwanenflugel, P. J., Harnishfeger, K. K., & Stowe, R. W. (1988). Context availability and lexical decisions for abstract and concrete words. *Journal of Memory and Language, 27,* 499–520.

Schwanenflugel, P. J., & Shoben, E. J. (1983). Differential context effects in the comprehension of abstract and concrete verbal materials. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9,* 82–102.

Service, E. (1992). Phonology, working memory, and foreign-language learning. *Quarterly Journal of Experimental Psychology, 45a,* 21–50.

Service, E., & Craik, F. I. M. (1993). Differences between young and older adults in learning a foreign vocabulary. *Journal of Memory and Language, 32,* 608–623.

Sjarbaini, L. (1998). *Het concreetheidseffect bij het leren van vocabulaire in een tweede taal* [The concreteness effect in learning a vocabulary in a second language]. Unpublished master's thesis, University of Amsterdam.

Van Hell, J. G. (1998). *Cross-language processing and bilingual memory organization.* PhD Thesis. University of Amsterdam.

Van Hell, J. G., & Candia Mahn, A. (1997). Keyword mnemonics versus rote rehearsal: Learning concrete and abstract foreign words by experienced and inexperienced learners. *Language Learning, 47,* 507–546.

Van Hell, J. G., & De Groot, A. M. B. (1998). Disentangling context availability and concreteness in lexical decision and word translation. *Quarterly Journal of Experimental Psychology, 51,* 41–63.

Wang, A. Y., & Thomas, M. H. (1995). The effect of imagery-based mnemonics on the long-term retention of Chinese characters. In B. Harley (Ed.), *Lexical issues in language learning* (pp. 167–183). Ann Arbor, MI/Amsterdam/Philadelphia: John Benjamins.

Wimer, C. C., & Lambert, W. E. (1959). The differential effects of word and object stimuli on the learning of paired associates. *Journal of Experimental Psychology, 57,* 31–36.

Yang, L. (1997). Tracking the acquisition of L2 vocabulary: The Keki language experiment. In J. Coady & T. Huckin (Eds.), *Second language vocabulary acquisition. SSLA, 19,* 125–155.

Appendix A: Stimulus Materials (L1-FL) of Experiments 1a and 1b

	Concreteness	Log frequency
Concrete cognates		
broer-breur	6.32	3.73
brood-brodde	6.38	3.47
gitaar-gettar	6.50	2.42
jongen-jange	6.44	4.18
lepel-lippel	6.84	2.89
lichaam-lechem	6.46	4.09
naald-nadel	6.38	2.82
paraplu-parredu	6.50	2.59
rivier-riffer	6.44	3.47
schaar-skare	6.96	2.52
stier-stire	6.36	2.74
trein-trinen	6.77	3.54
varken-vark	6.46	2.98
viool-vialo	6.42	2.70
vrouw-vrauk	6.50	4.58
Concrete noncognates		
bruid-maffel	6.20	2.66
handdoek-bodelei	6.56	2.84
kaars-roegen	6.69	2.98
kikker-edoe	6.62	2.57
krant-morees	6.60	3.69
meisje-wakkel	6.32	4.18
moeder-sluup	6.56	4.40
muis-zapel	6.80	2.95
paard-ipseel	6.65	3.82
rijst-hasser	6.76	2.39
stoel-blisme	6.84	3.81
tijger-duunze	6.68	2.47
vader-toker	6.24	4.39

vinger-bumt	6.76	3.82
vogel-tufeen	6.58	3.61
Abstract cognates		
domein-domaan	3.36	2.80
dwang-diank	2.00	2.71
eerbied-eerned	2.00	2.84
geval-fals	1.96	4.36
gunst-gonste	2.36	2.98
indruk-indark	2.12	3.82
invloed-invlen	1.85	3.91
inzicht-inzipt	1.92	3.63
kans-kents	1.88	3.93
reden-reude	2.16	3.98
spraak-spreik	3.20	2.38
tijd-teits	3.08	4.66
vorm-vorim	3.50	4.15
wrok-wroek	1.88	2.52
zwakte-zwokt	2.36	2.46
Abstract noncognates		
biecht-snog	2.69	2.34
blaam-spakje	1.19	1.89
deugd-muper	1.31	2.84
ding-pardaan	2.65	4.20
gerucht-evoliek	2.76	2.96
gevaar-stroek	2.85	3.69
herstel-groop	2.80	2.99
manier-bisdalf	1.73	4.20
oorzaak-wots	1.69	3.67
poging-alake	2.28	3.67
ruil-vanort	3.08	2.81
waarde-mulp	2.56	3.83
wanhoop-rief	2.68	2.98
wraak-klaspert	3.12	2.88
zorg-buikel	2.68	3.71

Appendix B: Stimulus Materials (L1-FL) of Experiments 2a and 2b

	Concreteness	Log frequency
Concrete / more frequent		
bloem-snalik	6.68	3.60
boer-kalla	6.44	3.63
haar-kodiel	6.60	3.61
hoofd-assel	6.24	4.36
koffie-plam	6.52	3.67
krant-morees	6.60	3.69
lichaam-evoliek	6.46	4.09
meisje-wakkel	6.32	4.18
moeder-sluup	6.56	4.40
muur-elne	6.46	3.79
stoel-blisme	6.84	3.81
tuin-mave	6.56	3.70
vogel-tufeen	6.58	3.61
vrouw-toker	6.50	4.58
vuur-bijn	6.35	3.64
Concrete / less frequent		
bruid-maffel	6.20	2.66
citroen-kars	6.72	2.68
gitaar-buikel	6.50	2.42
kalf-wumsel	6.27	2.49
kikker-edoe	6.62	2.57
kraan-geschak	6.48	2.80
mier-bumt	6.80	2.49
naald-bodelei	6.38	2.82
paraplu-mift	6.50	2.59
peer-nufrijg	6.80	2.63
potlood-rufoen	6.58	2.72
stier-muper	6.36	2.74
tijger-duunze	6.68	2.47

viool-stroek	6.42	2.70
vlinder-kodeisje	6.54	2.65
Abstract / more frequent		
afstand-bekkel	3.50	3.73
deel-gevolt	3.23	4.22
ding-pardaan	2.65	4.20
eeuw-spoden	2.77	3.99
geluk-schom	3.69	3.65
geval-wots	1.96	4.36
inzicht-groop	1.92	3.63
leeftijd-snouk	2.77	3.64
plan-fuls	3.60	3.93
regel-plark	3.65	3.77
stilte-ploks	3.04	3.63
toekomst-bekaar	2.92	3.74
verschil-breefje	2.62	3.87
vorm-rief	3.50	4.15
waarde-mulp	2.56	3.83
Abstract / less frequent		
biecht-snog	2.69	2.34
blaam-spakje	1.19	1.89
boosheid-kotiem	3.38	2.56
daling-scharf	3.73	2.56
domein-roegen	3.36	2.80
dreiging-filk	2.80	2.82
erfenis-moof	3.42	2.72
gunst-bisdalf	2.36	2.98
misdaad-hasser	3.69	2.96
raadsel-ipseel	3.56	2.89
reuk-zapel	2.96	2.36
ruil-vanort	3.08	2.81
stank-wirpel	3.23	2.82
wraak-klasper	3.12	2.88
wrok-alake	1.88	2.52