Fluency in second language writing: the effects of enhanced speed of lexical retrieval

Snellings, P.J.F.

Citation for published version (APA):
CHAPTER 2

LEXICAL RETRIEVAL: AN ASPECT OF FLUENT SECOND LANGUAGE PRODUCTION THAT CAN BE ENHANCED

ABSTRACT
The efficiency of lexical retrieval, an essential sub-process of productive language skills, is crucial in fluent writing and speaking. We examine the feasibility of an experimental computerized training for fluent lexical retrieval in the second language in a classroom setting, applying techniques previously restricted to laboratory use. Results on both a lexical decision and a written lexical retrieval task show that students in each trained words condition have higher accuracy scores and superior reaction times on the trained words in comparison to students who were not trained on these words, with medium to large effect sizes. We discuss implications of these findings for the study of productive language skills.

INTRODUCTION
For fluent language production lexical retrieval is an essential sub-process. It involves selection of lexical concepts (resulting in so-called lemma's that represent the word's syntax) that are subsequently encoded (morphologically, phonologically and phonetically) to be either articulated or written down (Levelt, 1989; Levelt, Roelofs, & Meyer, 1998). Fluency is important for speaking, so that the speaker may hold the floor, and also for writing, especially in an educational context with time restrictions imposed. Although there has always been an interest in language fluency (Schmidt, 1992), research in the acquisition of vocabulary has mainly focused on the acquisition of new words (but see Henriksen, 1999; Laufer, 1998) for studies concerning the nature of productive vocabulary). Recently however, the level of fluency in accessing items in the lexicon has also received attention in this field of enquiry (Meara, 1996). Chenowith & Hayes (2001) have shown that second language (L2) language courses increase fluency (words per minute) in writing. In their small cross-sectional study (N=13), students at different course levels were asked to write a short essay while verbalizing the text. Although these results are promising, it would be interesting to see whether similar results are obtained when more experimental control is exerted over the training and language input and with more students involved. Moreover, in the Chenowith & Hayes study the "words-per-minute" measure was not taken under conditions which required speed, so it does not reflect how fast lexical retrieval can be when speed is required.

Although fluency is considered to be a worthwhile goal in education, there has been no focused attention on lexical retrieval in L2 educational practice. Until now, a wide range of recommendations for improving language production have been made, from stimulating interaction for improving speaking skills (Damhuis, 1993) to modeling of text forms, composing processes and socio-cultural purposes and functions of writing in the L2 (Cumming, 1995). However, explicitly paying attention to lexical retrieval as a way to achieve fluency is clearly outside the objectives of current L2 teaching practice.

Still, even if we realize that lexical retrieval deserves a focused training of its own, there is as yet no evidence as to whether an experimental intervention aiming to enhance the fluency of lexical retrieval in L2 can actually be effective in an educational context.

Lexical retrieval is an essential process in both oral and written language production but research into the ways we might enhance this process has nevertheless been lacking.

Although Levelt (1989) has given a detailed theoretical account of the speech process in the first language (L1), including lexical retrieval, the emphasis was on describing the processes involved. Since then there has been no systematic attempt to enhance this sub-process experimentally. Poulisse (1997) has carried out a critical analysis of the different theories concerning speech production of L2 learners. She describes the process of lexical retrieval and offers explanations but makes no reference to research attempting to alter the process experimentally.

In their influential writing model, in which the emphasis lies on planning and monitoring, Hayes and Flower (1980) also distinguish text production processes and lexical retrieval (“translating”). Kellogg (1996), in his description of the components of working memory (based on Baddeley, 1986) involved in written text production, stresses the importance of lexical retrieval in writing as well. In an experimental study, Kellogg (1994) further demonstrates that lexical retrieval is highly effortful, even in the case of native speakers. Although planning and reviewing caused the most reaction time (RT) interference, it turned out that translating content into sentences (i.e. lexical retrieval and syntactic processing) was also highly effortful and slowed RTs about 350 ms over baseline times. So Kellogg underlined the importance of lexical retrieval and showed that the process of lexical retrieval puts considerable demands on working memory. McCutchen and her colleagues (McCutchen et al., 1994) once more stated the importance of lexical retrieval. They looked at the relationship between lexical retrieval (the fluency of translating processes) and writing skills. They demonstrated that skilled L1 writers (across grades 3, 4, 7 and 8) showed better performance on a reading and speaking span task by having both larger spans (they remembered more last words or were able to generate grammatical sentences for longer lists) and by producing longer sentences. In another experiment they found shorter latencies and higher accuracy rates on lexical decision tasks for the skilled writers. As a consequence, they stress the importance of lexical retrieval, especially for the less skilled writer’s planning and reviewing abilities in writing. In their opinion, fluent lexical retrieval is important because more attentional control needed for text production processes such as lexical retrieval may result in less attention that can be devoted to planning, monitoring or other cognitive operations used in writing (Hayes, 1996; Kellogg, 1996). Nevertheless, even though the importance of efficient lexical retrieval for both spoken and written skills has been stressed repeatedly, there is no evidence from experimental interventions aimed at enhancing lexical retrieval.

As the aim of the present study is to investigate whether enhancing written lexical retrieval in a classroom setting is feasible and leads to gains in retrieval speed, we have looked both at related research into word recognition and reading and at findings from psycholinguistic research and psychological theories on instruction.

Whereas lexical retrieval has been neglected, there is some research on the highly related process of “lexical access” (“word recognition”). In lexical retrieval, concepts are transformed into linguistic forms, whereas in lexical access, the forms have to be recognized to arrive at the correct lexical concept. Although the order in which the components of the process are active is reversed in lexical access and lexical retrieval, most components are interchangeable (Kroll & Stewart, 1994; Levelt, Roelofs, & Meyer, 1999). Therefore, it is interesting to have a look at the findings from research on word recognition to guide us in our current aims. Not only has there been an emphasis on the relationship between word recognition and reading (Favreau & Segalowitz, 1983; Hahne & Friederici, 2001; Koda, 1996; Perfetti, 1985), but there have also been specific suggestions on how to train word recognition. Segalowitz, Poullsen and Komoda (1991) have suggested that to develop reading proficiency, a training in lexical look-up, associating a print/phonological pattern with a meaning, may be useful in reducing demands on working memory. It has also been suggested
that care has to be taken that the same meaning of the words used is trained consistently. If the aim is to train line in a story about fishing this should always be used in the fishing sense and not in both a fishing and telephone sense, for example. This is what Segalowitz et al. (1991) call “constant practice”. Unfortunately, other research on the training of lexical access is either restricted to learning new letters (LaBerge & Samuels, 1974) or acquiring artificial words in the laboratory (Yang, 1997). Still, apart from a focus on speed, LaBerge and Samuels (1974) pointed to the importance of a criterion baseline. In addition, there is extensive research on the training of reading fluency (for an overview see Kuhn & Stahl, 2000). Although this research involves either children with reading problems (e.g. Van der Leij, 1998) or children who are beginning to read in their native language, it nevertheless shows once more that if the aim is to increase speed, an explicit focus on speed is helpful.

For a training in lexical retrieval to be effective, it is essential that we focus on the right words, and we have to realize that the demands of lexical retrieval are not the same for all words. Words with a lower printed frequency take more time in lexical decision tasks than frequent words (Whaley, 1978). In naming tasks, both those involving the naming of pictures and those involving the naming of written words, this frequency effect has also been found (Forster & Chambers, 1973). Gernsbacher (1984) has shown that it is not printed frequency itself but “experiential” familiarity (as rated by subjects) that is the underlying variable that causes the effects. Bilinguals with a slight L1 predominance also show slower RTs to neutral L2 targets than to neutral L1 targets in lexical decision tasks (Favreau & Segalowitz, 1983). Obviously, this difference cannot be the result of the printed frequency of the words in both languages, and it is likely that the larger RTs are the result of differences in familiarity. Hence, to increase speed of lexical retrieval, the learner’s familiarity with the words has to be increased. It seems worthwhile to focus on those words that are known but not yet that familiar, so as to maximize the impact of the training. Unfortunately, as aptly pointed out by Wolter (2001), word-frequency ratings, even if based on large corpora, cannot give conclusive evidence in predicting which words are known or not known by a particular individual. Moreover, most larger corpora consist of native speaker data and are therefore not applicable to L2 learners. To get a closer approximation of the words that are familiar to a particular group of L2 learners, it seems therefore wise to look at the vocabulary produced by other learners who have had a similar amount of classroom instruction in English.

In a discussion of the role of formulaic language in L2 acquisition, Weinert (1995) has stressed the importance of formulas, both for reducing cognitive processing and as a production strategy (see also Wray & Perkins, 2000). In view of these considerations, a focus on single words and on combinations of words seems warranted.

Morton (1979), in his logogen model, argues that repeated exposure lowers a threshold that has to be reached for identification to take place. Because repeated exposure to words, either through experience or through production, can lower their thresholds and increase familiarity, repeated exposure may also reduce the cognitive effort involved in lexical retrieval in L2 contexts. As beginning L2 learners are far less familiar with the words in the language and have slower retrieval times than more advanced learners or native speakers have for the same words, increasing the familiarity of words is likely to be of most benefit to this group.

On the basis of learnability theories, we assume that words have to be known receptively before they can be used productively. In the present context, this means that exercises focusing on lexical access should precede exercises requiring lexical retrieval. The former exercises should require only recognition of the words and their properties, whereas the latter exercises can be more demanding, involving actual production of the target words by learners. Moreover, as Daneman and Green (1986) have pointed out, processing skill differences in working memory are most likely highly task specific. These authors
demonstrated that efficiency of sentence production processes, rather than efficiency of sentence comprehension processes, determines working memory capacity when individuals have to produce vocabulary in context. Hence, to train lexical retrieval it is essential that learners ultimately be trained with a task that relies on productive processes. In addition, we expect to find larger effects on a direct measure of production (a lexical retrieval task) than on an indirect measure tapping comprehension (a lexical decision task). As has been argued both in psychological theories (Ausubel, 1968) and in theories about L2 acquisition (Krashen, 1981), a strong focus on the meaning of the words and expressions used in the training leads to better retention. Therefore, care has to be taken that correct completion of all tasks crucially depends on access to the meaning of the words targeted.

Our goal in the research presented here was to investigate whether an experimental intervention aiming to enhance the fluency of lexical retrieval in L2 can actually be effective in an educational context. In the following sections we will give a detailed description of the experiment we conducted, including the tasks used in the experimental computer program, the experimental findings, and the implications for further research into L2 production skills.

METHOD

Participants
A total of 100 Dutch students participated in this study, all in the beginning of their third year of secondary education (grade 9, students from 14 to 15 years old). So that we would be able to generalize our results across different tracks, students were sampled from four different classes, two classes from lower general secondary education (n=46) and two classes from pre-university education (n=54). As beginning learners are likely to benefit the most from instruction, we focused on L2 learners with only a few years of experience in the target language. They had received at least two full years of formal English tuition (two or three lessons per week) at secondary education level, and a 2-year introductory course at the elementary level focusing mainly on basic oral communication skills (maximum one lesson per week). We asked them two questions about their language background: whether they spoke Dutch with their parents and whether they had acquired Dutch as their LI. The results showed that 85 students were native speakers of Dutch. The remaining 11 students (only 96 students filled in the questionnaire) were L2 speakers of Dutch and were not native speakers of English (although one student who had acquired Dutch as his first language spoke English with his father and Dutch with his mother, and one other student spoke Dutch with her parents but had acquired both English and Dutch as an LI). The L2 speakers were divided evenly across the two conditions.

Design
In a true experimental counterbalanced design, students from four classes were randomly assigned to one of two treatments that both involved speed training. Each class was split in half, and the resulting eight groups were trained concurrently over a 4-week period by two experimenters. Each treatment used a different set of stimuli words, either word set A or word set B. Students in the “trained A words condition” were trained on word set A. Students in the “trained B words condition” were trained on word set B. Students in one group were not trained in the other group’s word set. It is important to note that differences in the two sets of words are not an issue in this research design.

The dependent variables are lexical access and lexical retrieval. The tests consisted of words from word set A, word set B and word set C. The additional set of words (“control words” or “C words”) was not trained in either condition and was included as a covariate.
Materials

Stimulus words
As we wanted to increase the retrieval fluency of words instead of teaching new words, the words had to be familiar. As pointed out before, word frequency ratings may not be the best indication of familiarity, especially because there are no learner corpora for productive vocabulary of Dutch L2 learners. To obtain a selection of familiar (productive) words, we elicited words in a writing test from a group of students similar to those in the study. We conducted a pilot study in which cartoon-based stories were written by untrained students of the same age and similar levels of English ability as the participants in the current study. We used cartoons because they enabled us to elicit a restricted range of words and provided us with an insight into the words that could be used productively. An additional reason for using cartoon tasks is their promise for future research, as their restricted vocabulary makes it possible to investigate the effects of enhanced lexical retrieval on actual written production. We selected only words that were used by the students and had straightforward translations in Dutch. In making the selection we also used teacher intuitions about word difficulty. The words belonged to different grammatical classes (e.g. noun, article, determiner, adjective, adverb, verb). In view of the relevance of collocations for reducing cognitive effort, some words were used on their own, others in combination with each other.

Tasks and procedures
First, a vocabulary test to assess language skills was administered. Then students were trained in class for 4 weeks. To eliminate experimenter effects, each group trained by the same experimenter contained students from both conditions. All students received the same type of computer training (but with different words) and the same amount of computer-assisted exercises. The treatment was administered using specially designed software1 that could be used outside the laboratory in a classroom setting, training a large group of learners at the same time.

Although both lexical decision and picture-naming tasks can be assumed to tap lexical access, only the picture-naming task is a productive task and proceeds from meaning to word form (instead of the other way round). In addition, we had to take into account that the set of words in which we are interested (including nouns, pronouns, articles, verbs, adverbs, adjectives and combinations of these categories) is difficult to depict unambiguously with pictures. Because we wanted a clear focus on these words, we considered word translation tasks. They tap processes similar to those involved in picture-naming tasks in that both picture-naming and translation tasks involve concept mediation (for similar tasks in testing, also see De Groot et al., 1994; Potter, So, von Eckardt, & Feldman, 1984). The word translation task we employed has learners fill in the correct translation within a meaningful context, a situation that matches lexical retrieval in written production most closely.

Because our objective is to increase retrieval speed, it seemed sensible to encourage students to speed up in the training and provide them with a baseline for reference to their progress. Moreover, as the words in the present study were most likely already familiar to the learners, we assumed that working under time pressure instead of mere exposure would be a motivating factor as well.

---

1 This software was a joint development within the current study and a reading intervention study by Annegien Simis and Jan Hulstijn. Although the exercises in both studies differed, the framework was identical and was programmed by Boudewijn de Roovere. For a more detailed description, click the Nelson tab on the following site: [http://members.ams.chello.nl/b.roovere/](http://members.ams.chello.nl/b.roovere/)
Each student encountered each word 10 times and received immediate feedback on correctness as well as on both their current speed and their speed on the same item the last time they responded to it (the criterion baseline). “Correct” feedback followed when the student had correctly indicated which constituent followed in a sentence (block exercise), detected the semantic anomaly (detection exercise) and replaced it (correction exercise), or wrote down the correctly spelled translation (translation exercise). We consistently trained the same meaning of the words used. All students did the same number of exercises. Students were trained in a school setting in scheduled English classes. Standard duration of each training was 50 min (except for one training of 45 min) and there was a 1-week interval between each training session. After the treatment, students completed the lexical decision and lexical retrieval task in one session in the same week as the last training session (because of school time tables, one group sat for the test on the same day as the last training session, two groups the following day, and one group 2 days after the last training). Having collected data on the lexical decision and lexical retrieval tasks, we also administered a typing task and asked the students questions about their language background.

Each specific word was trained with four different types of exercise, moving from receptive skills to the target production skills: a block exercise, a detection exercise, a correction exercise, and a translation exercise.

**Block exercise**
The aim of the block exercise was to both strengthen the receptive knowledge of the words and to train syntagmatic knowledge (collocational restrictions). Students had to complete a short sentence as fast as possible by choosing from two constituents, one of which resulted in an incorrect word order, as in the following example:

*After some time*

woke up she

Z M

Students had to press either Z or M for the constituent they thought would be the first one following *After some time*.

**Detection exercise**
To strengthen the connections between the concepts and the lexical forms, the detection exercise had a strong focus on the meaning of the words involved. Learners had to decide whether the meaning of the words and expressions in the specific context presented was appropriate. Students had to focus on the meaning of the words in a short sentence and detect semantic anomalies as fast as possible, indicating which word caused the anomaly, as in the following example:

*We went to the pool while(Z) we arrived(M) home.*

Students had to indicate whether *while* (Z) or *arrived* (M) makes the sentence odd. Here, the word *while* makes the sentence odd as a word like *after* is more likely.

**Correction exercise**
In the correction exercise, students were asked to judge words for appropriateness of use and if the context required it, to choose a better substitute from two alternatives. This exercise also aimed at enhancing the connections between the concepts and the lexical forms. Students again had to focus on the meaning of the words in a short sentence. They also had to detect semantic anomalies as fast as possible, but this time they had to indicate which of two words
would make the sentence logical (or accept the present word marked “X”), as in the following example:

Sue felt very ill. After some ages (X) she felt better though.

<table>
<thead>
<tr>
<th>time</th>
<th>centuries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z</td>
<td>M</td>
</tr>
</tbody>
</table>

Students had to indicate whether ages (X) is acceptable, or whether replacing it with time (Z) or centuries (M) would yield a more semantically acceptable sentence. Here, time (Z) would be more likely.

**Translation exercise**

To train lexical retrieval we ultimately needed a task that relies on productive processes. We chose a translation task because such a task makes a focus on more words at the same time (in contrast to picture tasks) possible. The translation exercise we employed has learners fill in the correct translation within a meaningful context, a situation that matches lexical retrieval in written production most closely. Students had to translate Dutch words in simple English carrier sentences. In addition to the ordinary format we developed a “disappear version” in which the Dutch words and the English sentence disappeared from the screen after 15 s, stimulating the students to give faster answers. The following example is illustrative of the task:

voordat the bridge collapsed he had reached the other side

Translation: before

The order of the items in all exercises was randomized each time before they commenced an exercise. In the block exercise the order of the two constituents was also randomized. The exercises were tested in a small-scale pilot study in order to establish not only that both the software and hardware functioned properly and also that the exercises were adequate for our aims. On the basis of this pilot study, we clarified the instructions and put more emphasis on the speed aspect, as we wanted all students to have the same focus. Problems in the computer program we were using and in the design of the exercises were solved in the final version employed in the experiment. Table 1 shows the order of the exercises in the final experiment.

<table>
<thead>
<tr>
<th>Lessons</th>
<th>Exercises</th>
<th>Exercises</th>
<th>Exercises</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lesson 1</td>
<td>Block</td>
<td>Block</td>
<td>Correction</td>
</tr>
<tr>
<td>Lesson 2</td>
<td>Correction</td>
<td>Detection</td>
<td>Detection</td>
</tr>
<tr>
<td>Lesson 3</td>
<td>Translation</td>
<td>Translation</td>
<td></td>
</tr>
<tr>
<td>Lesson 4</td>
<td>Translation</td>
<td>Translation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(disappear version)</td>
<td>(disappear version)</td>
<td></td>
</tr>
</tbody>
</table>

**Dependent measures**

To tap lexical access and lexical retrieval, we used a Lexical decision task and a written productive translation task (WPTT), respectively.
**Lexical decision task**

In the lexical decision task, students were presented with 120 letter strings (plus 20 practice items) on the screen of a laptop computer. Of these strings, 60 were real words, whereas the other 60 were pseudowords. The distribution of the length of morphologically correct pseudowords was identical to that of the real words. In the 60 real words, there were 20 A words, 20 B words, and 20 C (control) words. The control words were easy words, most of which were derived from Woutersen (1997). The grammatical classes used were nouns, adjectives, adverbs, pronouns, prepositions, numerals and verbs. The order of pseudowords and real words was randomized but the same for all students.

**WPTT**

In the WPTT, students had to write down the English translations of Dutch words as fast as possible. The Dutch words were in simple English carrier sentences to make them unambiguous. There were 20 A words, 20 B words, and 20 C (control) words (plus 10 practice items). The control words were easy words, similar in grammatical class to the A and B words. There was no overlap between the words used in the WPTT and the lexical decision task. The grammatical classes of the words were nouns, pronouns, articles, verbs (including prepositional verbs and infinitives with to), proforms (e.g., something), prepositions, numerals, adverbs, adjectives, and combinations of these (see Appendix A for the tested WPTT words). Responses were saved into log files; we recorded the time the first key was hit, whether the first key was correct, when the final Return response was given, whether the answer was correct, and finally whether students corrected their answers. We looked at the time it took students to type their answer and press the Return key. By using the Return key at the end of the word instead of the first letter, we deviated from the procedure most commonly used in naming tasks. One reason not to look at the first letter was the fact that many items consisted of more than one word; e.g. “a beautiful waterfall” or “called for help”. Clearly, the time it took to type the first letter “a” or “c” does not necessarily give a good indication of the time required to retrieve the whole item. In addition, students could type the first letter and hesitate about what had to follow before going on and hitting the Return key. To take these effects into account we decided to focus on responses of the whole word or combination of words.

**Covariates**

To control for initial differences we used covariates. We had several candidate measures for covariates: a vocabulary measure, a typing-speed measure, and the control words (scores on lexical decision control words, for controlling scores on the trained [italics added] lexical decision words and scores on WPTT control words for controlling scores on trained [italics added] WPTT words). The vocabulary and control words measures could be used as a control for initial differences on both dependent measures. The typing-speed measure was intended only for controlling initial differences on the WPTT, because students had to use their typing ability in this task, which was irrelevant for the purpose of the study. Ideal covariates are variables that correlate significantly with the dependent variable and have low correlations among themselves (Stevens, 1996). We started by selecting covariates that correlated .30 or higher with the dependent measures.

In the case of the lexical decision scores, the correlations with the vocabulary score were less than -.10 (-.07 for A words and -.03 for B words) and around .80 with the C (control) words (.83 for A words and .79 for B words). As a result, we selected control words as the covariate for lexical decision.

In the case of the WPTT scores, the correlations with the vocabulary score were again less than -.10 (-.08 for A words and -.05 for B words), around .40 for the typing-speed score
(.40 for A words and .42 for B words), and around .80 for control words from the WPTT (.80 for A words and .76 for B words). As the correlation between the vocabulary score and the WPTT scores is less than .30, vocabulary scores were not selected as a covariate.

The correlations of both the typing-speed score and control words with the WPTT scores are sufficiently high to be included as a covariate, but there is also a considerable intercorrelation (.49) between typing speed and control words from the WPTT. In view of these intercorrelations, it is important to establish the unique contribution of each candidate covariate in explaining variance of the WPTT. We carried out a multiple regression analysis to establish these unique contributions. This analysis shows that the unique contribution of typing speed to variance explained was zero for A words and near zero for the B words. In contrast, the control words had a unique contribution of 48.2% to the variance in the A word scores and 41.2% to variance in the B word scores. Consequently, control words were selected as covariate for the WPTT scores.

Data collection
Apart from the paper-and-pencil vocabulary test, all tests were taken on laptop computers. To avoid practice effects arising from words in the carrier sentences of the lexical retrieval task (context words) recurring in the stimulus words of the lexical decision task, the lexical decision task was administered first. Because the lexical decision task did not involve carrier sentences and hence involved fewer words, we could preclude recurrence of words in the lexical retrieval task. Students started with the vocabulary test, and after the training period, they first completed the lexical decision task, second the WPTT, and finally the typing test and the questions about language background.

Analyses
Means, standard deviations, and estimates of reliability (Cronbach’s α) were calculated. The hypotheses were tested by two one-way multivariate analyses of covariance (MANCOVAs) on the lexical decision and lexical retrieval scores, respectively. The independent variable was condition (A or B). The dependent variables in the first MANCOVA were lexical decision A words and lexical decision B words. Results of evaluation of the assumptions of independent observations, a linear relationship between the dependent variables and the covariate, homogeneity of the regression slopes and reliability of the covariate were satisfactory (Stevens, 1996). The sampling distributions are not multivariate normal, but MANCOVA is robust with respect to this violation. Population covariance matrices for the dependent variables are not equal either, but because the group sizes are approximately equal (n-ratio=53:47, which is less than 1.5:1), MANCOVA remains robust. In the case of the second MANCOVA, the dependent variables were WPTT A words and WPTT B words. Results of evaluation of the assumptions (Stevens, 1996) were satisfactory.

Scoring
In the speed tests two scores are relevant: RT and accuracy. Mean RTs for lexical decision and lexical retrieval were computed across the “hits”(i.e. correct responses on the items). Since we are not interested in differences in linguistic knowledge but solely in speed of processing, we used only items with sufficiently high hit rates. We had to set cut-off percentages, because not all students were trained. Items with low hit rates (especially in the untrained groups) prevent a fair experimental comparison of RTs, because they would be based on different sets of words as the distribution of hit scores differs between students.

In the lexical decision task, words with a hit rate lower than 87.5 % in at least one of the two groups were removed. This percentage takes into account a chance level of 50% correct guesses. Because we believed that the stimuli had to be rather easy, we set a minimum
level of 75%, uncorrected for guessing. The correction produces the cut-off limit of 87.5%. In the WPTT, words with a hit rate lower than 50% in one of the two groups were removed. This percentage takes into account that correction for guessing in this task is not necessary and that the stimuli are more difficult to answer correctly than in the lexical decision task. Hits in the WPTT were correct written responses or “correct except for spelling” (to be counted as a hit, the answer given had to make clear which word was intended and it had to be in the correct grammatical form). For the WPTT too slow answers (more than 20 s) were recoded into a missing value. In the lexical decision task individual responses that were unrealistically quick (RTs < 250 ms) were also recoded into missing values. The above procedures resulted in 15 A words, 13 B words and 16 C words for the lexical decision task. In the WPTT, 9 A words, 13 B words and 17 C words were included.

**Missing values**

Estimation of missing values was necessary to arrive at a full data matrix for all students. Without such a full matrix, experimental comparison of the two trained conditions on the same set of words is not possible. Moreover, analysis of reliability is seriously hampered if few students are left with valid scores on all words. To prevent missing value estimation on the basis of too few data for individual students, we had to set a criterion for a minimal amount of hits.

Missing values were estimated according to the Expectation-maximization (EM) algorithm of the Statistical Package for the Social Sciences (SPSS, Acock, 1997; Hox, 1999). For lexical decision it appeared that all students had a sufficient amount of hits for each set of words. The items of the WPTT were more difficult, and therefore there were some students with very few hits. The minimal amount of hits for inclusion in the analysis was set at 50% for each word set. Three and five students, respectively, in the trained A words and in the trained B words condition did not meet this requirement and were dropped from the analyses.

**RESULTS**

In this section we will first present descriptive measures of the tests in the experiment. We will then proceed with a discussion of the results for our two dependent measures, the lexical decision task and the WPTT.

**Descriptive and psychometric statistics**

<table>
<thead>
<tr>
<th>Words</th>
<th>Trained A words condition</th>
<th>Trained B words condition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n=47</td>
<td>n=53</td>
</tr>
<tr>
<td></td>
<td>k</td>
<td>M</td>
</tr>
<tr>
<td>A words</td>
<td>15</td>
<td>712</td>
</tr>
<tr>
<td>B words</td>
<td>13</td>
<td>772</td>
</tr>
<tr>
<td>C words</td>
<td>16</td>
<td>762</td>
</tr>
</tbody>
</table>
In Table 2 we give an overview of the unadjusted scores on our dependent measures, without taking into account initial differences (the covariate). We see that in the case of the A words on the lexical decision task, students in the trained A words condition score around 700 ms, and the trained B words condition is slower, with scores around 900 ms. In the case of the B words, both the trained A words condition and the trained B words condition score around 770 ms. In Table 2 we also see that the score of the students in the trained A words condition on the control words is still in the 700 ms range, whereas the score from students in the trained B words condition is in the 800 ms range. This suggests that the students in the trained A words condition are faster from the start. Reliability of the lexical decision task was high for both groups.

Table 3. Descriptives WPTT: number of items (k), unadjusted mean RT scores (in ms), standard deviation and reliability (Cronbach's alpha).

<table>
<thead>
<tr>
<th>Words</th>
<th>Trained A words condition</th>
<th>Trained B words condition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n=44</td>
<td>n=48</td>
</tr>
<tr>
<td>k</td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>A words</td>
<td>9</td>
<td>7,486</td>
</tr>
<tr>
<td>B words</td>
<td>13</td>
<td>8,346</td>
</tr>
<tr>
<td>C words</td>
<td>17</td>
<td>5,763</td>
</tr>
</tbody>
</table>

As Table 3 shows, in the WPTT the trained A words condition scores in the 7,000 ms range on the A words, whereas the trained B words condition is slower, scoring in the 9,000 ms range. In the case of the B words, the trained B words condition scores just within the 6,000 ms range and the trained A words condition scores in the 8,000 ms range. From Table 3 it also becomes clear that both conditions score in the 5,000 ms range on the control words but again the trained A words condition is faster. Reliability of the WPTT was high.

Lexical Decision: Results for the A and B words
Looking at hit scores on all lexical decision items, we expected a higher hit score for the trained words condition on the trained words when compared to students that were not trained on these words. In contrast to the analysis of speed differences, in which we only wanted easy items, we retained difficult items in our analysis of hit scores. As expected, on the A words the trained A words condition approached the maximum score of 20 (M=19.62, SD=0.71) and the trained B words condition scored slightly lower (M=17.49, SD=1.75). On the B words we see the same pattern; again the trained B words condition has superior hit scores if we look at all items (trained B words condition: M=19.60, SD=0.69, trained A words condition: M=16.04, SD=1.57). As with the A and B words, which were trained in one of the experimental conditions, we chose easy words for the control words. The fact that both groups had relatively high and similar hit scores (trained A words condition; M=18.45, SD=1.32, trained B words condition; M=18.85, SD=1.05) shows that the control words were familiar to both groups.
Figure 1. Mean Word Recognition RT (in ms) on the A and B words in the lexical decision task for each condition after adjusting for scores on the control words.

To find out whether the training had an effect on speed of lexical decision, we used a MANCOVA with speed on the control words as a covariate. Figure 1 shows that the trained A words condition ($M=740, SE=13.32$) is faster on the A words than the trained B words condition ($M=868, SE=12.53$). There is a significant main effect for both condition, with $F(1,97)=48.020, p<.001, \eta^2 =.33$, and the covariate control words with $F(1,97)=256.042, p<.001, \eta^2 =.73$. Figure 1 further shows that in the case of the B words, the trained B words condition ($M=740, SE=13.47$) is faster than the trained A words condition ($M=191, SE=14.31$). Again there is a significant main effect for both condition, with $F(1,97)=8.552, p<.01, \eta^2 =.081$, and the covariate control words with $F(1,97)=177.440, p<.001, \eta^2 =.65$.

It can be concluded that the training in speed on the A words and the B words results in superior speed among students on words trained in their condition in comparison to students who were not trained on these words. Moreover, in the case of the A words, 33% of the variance in scores is due to the training. In terms of Cohen’s (Cohen, 1988) effect size, a small effect is an $f$ value of .10 ($\eta^2=0.010$), a medium effect is an $f$ value of .25 ($\eta^2=0.0588$) and a large effect is an $f$ value of .40 ($\eta^2=0.345$). In this case we can therefore speak of a medium effect. In the case of the B words, the condition effect is also medium.

**WPTT: Results for the A and the B words**

In the WPTT we also expected students in the trained words condition to have a higher hit score on all trained items when compared to students that were not trained on these words. In contrast to the analysis of speed differences, we retained difficult items in our analysis of hit scores (the same procedure was used in the case of the lexical decision results). As this was a productive test, we expected lower hit scores in general, but especially in the case of the untrained words. Results show that in the case of the A words the trained A words condition still has superior hit scores on all words (trained A words condition; $M=15.89, SD=3.35$, trained B words condition; $M=7.98, SD=3.77$), yet not as high as was the case in the lexical decision task. In the case of the hit scores on the B words, the trained B words condition scored fairly high ($M=16.45, SD=2.54$) and the trained A words condition scored lower ($M=10.53, SD=3.31$). We chose easy words for the control words, and both groups had scores
similar to scores obtained by trained words conditions on trained words (trained A words condition $M=15.51$, $SD=2.74$, trained B words condition; $M=15.64$, $SD=2.41$).

![Figure 2. Mean Lexical Retrieval RT (in ms) on the A and B words in the WPTT for each condition after adjusting for scores on the control words.](image)

In the case of the speed on the A and B words in the WPTT, we used a MANCOVA with speed on the control words as a covariate. Figure 2 shows that the trained A words condition ($M=7,591$, $SE=160.03$) is faster on the A words than the trained B words condition ($M=9,244$, $SE=153.20$). There is a significant main effect for both condition, with $F(1,89)=55.595$, $p<.001$, $\eta^2=.38$, and the covariate control words with $F(1,89)=263.678$, $p<.001$, $\eta^2=.75$. Figure 2 also shows that on the B words the trained B words condition ($M=6,861$, $SE=110.79$) is faster than the trained A words condition ($M=8,428$, $SE=115.72$). Again there is a significant main effect for both condition, with $F(1,89)=95.475$, $p<.001$, $\eta^2=.52$, and the covariate control words with $F(1,89)=308.782$, $p<.001$, $\eta^2=.78$. It can be concluded that also in the case of the WPTT the training in speed on both the A and the B words results in superior speed among students on the words in their particular condition as compared to students that were not trained on these words. Moreover, in the case of the A words, we see that the $\eta^2$ for condition is .38, so that the training resulted in a large effect; 38% of the variance in scores is the result of the training. In the case of the B words, the $\eta^2$ for condition is even larger (.52), so we can conclude that the effect of the training on these words was very large. As an additional control on the EM procedure, we also carried out our analyses on the WPTT data without EM imputation and reached the same conclusions.

**DISCUSSION**

In this study we have reported on the results of an experimental training aimed at reduction of the cognitive demands of lexical retrieval. Lexical retrieval is an important sub-process in both oral and written language production that has so far been neglected in research on L2...
instruction. In the present research we have applied methods in a classroom setting that have until now been restricted to laboratory use. In particular, we have investigated whether training lexical retrieval in a classroom setting leads to gains in retrieval speed.

Effects on lexical access
The results for the lexical decision task show that the training resulted in a speed up of lexical access. Moreover, all students in the trained words condition had higher hit scores on the words on which they were trained than students that were not trained on those words, even though the mean hit scores on control words were similar in both the trained A words condition and the trained B words condition. This means that the familiarity of the trained words has increased. As for RTs, in the case of both sets of words (the A and B words), all trained students had scores on the words on which they were trained that were superior to the untrained students on these words. Moreover, in the case of the A words, the size of the effect was medium, with 33% of the variance in scores due to the training. In the case of the B words there was also a medium effect. Here 8% of the variance in scores is due to the training; existing differences in speed that are captured by our covariate explain most of the variance (65%). If we take into account that students encountered each word only 10 times over a 4-week period (with forgetting taking place between two consecutive training sessions), this is a remarkable result. We see that in all comparisons the experimental training resulted in better hit scores and in faster RTs on the lexical decision task. This is an important finding because the lexical decision test uses a different format from the exercises in the training. Consequently, this result cannot be due to increased experience with the test format. As we pointed out above, some components involved in lexical access and lexical retrieval are similar (Levelt, 1989). As a result, we should expect effects of our training in lexical retrieval to surface on a measure of lexical access as well, although the effects may be less pronounced than on a direct measure of lexical retrieval. In light of the effects found on the lexical decision test, which uses a different format than was used in the training, we can therefore conclude that the underlying process has been influenced by the training.

Effects on lexical retrieval
If Daneman and Green (1986) are correct in stating that processing skill differences in working memory are highly task specific, it follows that a measure that resembles the trained task most closely should show the largest increase in the processing skills involved. In the case of the WPTT, we therefore expected larger effects than on the lexical decision test. Similar to what we found on lexical access, our results on the WPTT show a significant difference on both types of words, the trained students being both more correct and faster on the words trained in their condition in comparison to students that were not trained on these words. In addition, the size of the speed effects is extremely large, especially in the case of the B words. In the case of the A words, 38% of the variance in scores is due to the training, and in the case of the B words, the effect is even stronger, with 52% of the variance explained. We can conclude that a 4-week, highly focused training aimed at speedup of responses results in both higher hit scores and faster RTs on lexical retrieval. It is therefore likely that we have succeeded in reducing the cognitive demands of lexical retrieval on working memory. In other words, the current study has demonstrated that a training focusing on speed and providing immediate feedback on both speed and correctness can effectively alter and improve an essential cognitive sub-process involved in L2 production.

Future Inquiry
Now that we have shown that fluent lexical retrieval can be experimentally induced, it would be interesting to see whether there is an effect of increased speed of lexical retrieval on
speaking skills, writing skills, or both. In view of the importance of efficient lexical retrieval for speaking (Levelt, 1989; Poulisse, 1997), the effect of experimental training on speaking deserves further attention. Clearly, we have demonstrated in this study that the methodology used to train learners can be implemented in a classroom setting. Nevertheless, an approach aimed at speaking would require the development of new measures for speaking skills that could be used outside the laboratory.

In the case of writing skills, previous research (Kellogg, 1994; McCutchen et al., 1994) has demonstrated that there is a relation between working memory and writing, but there is little experimental evidence on the nature of this relationship. Given the necessity of lexical retrieval for the writing process, we expect that increased retrieval speed on a set of trained words should have an influence on writing tasks requiring retrieval of the same words. We predict a direct effect on writing, in the sense that it is likely that students who retrieve words more easily will use these words more often in a writing task than students who have difficulty in retrieving them. In view of the predicted trade-offs among processes (Kellogg, 1996), more attentional control needed for text production processes such as lexical retrieval may result in less attention that can be devoted to other cognitive operations such as planning, executing, or editing. Consequently, we would expect effects on the quality of texts.

From a different perspective, Kintsch (1998) argued that with efficient retrieval structures, knowledge from long-term memory (LTM) can be used. He further argued that retrieval structures need strong and well-practiced knowledge. Recently, McCutchen (2000) has applied Kintsch’s (1998) model of working memory to writing. She claims that writers can form retrieval structures and retrieve knowledge from LTM only once encoding processes such as text generation become sufficiently fluent. From the overview in McCutchen (2000), it becomes clear that genre knowledge is related to writing skills and that topic knowledge results in higher quality texts. Consequently, if speeded retrieval gives access to knowledge about genres and topics in LTM, it ought to exert considerable influence on the quality of texts. In the case of L2 learners, however, the processes involved may be slightly different. Assuming a non-language specific conceptual system (e.g. Kroll & Stewart, 1994), we think it plausible that L2 learners can access LTM via their L1 structures. Nevertheless, we do feel that enhanced lexical retrieval might have other benefits. In our view, once the knowledge has become strong and well practiced, efficient retrieval structures may make it possible to retrieve related L2 words from LTM. It is an interesting question whether L2 learners can access other relevant words on the same topic and become able to generate more detailed content, resulting in better texts in the L2.

On the basis of the results in the present study we are justified in carrying out further research into the effects of training lexical retrieval on both writing and speaking. At the same time we have only laid the groundwork for further research. It is crucial to gain more insight into the processes involved. Ultimately, we need to reach a better understanding of which words are most suitable for training and which are also valuable in more general text types and in speaking contexts. Because of the restricted range of the vocabulary elicited in cartoon tasks, these tasks provide a direct way of testing the effects of enhanced lexical retrieval on writing. Once effects can be demonstrated on cartoon tasks, we can investigate the effects on other writing tasks. From a theoretical point of view it is important to know more about the relation between a training with a focus on speed and characteristics of the words trained, such as word type and familiarity. To maximize the impact of the training for practical purposes, inclusion of words used in a particular teaching context is also warranted. The current research justifies investigating these issues. In addition, RTs that were restricted to research contexts and laboratory settings have now become available for classroom use and have been shown to be effective in improving an essential underlying process of L2 productive skills.