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

Snellings, P. J. F.

Citation for published version (APA):
CHAPTER 4

THE EFFECT OF ENHANCED LEXICAL RETRIEVAL ON L2 WRITING

ABSTRACT
Lexical retrieval is an essential sub-process of language production and in view of the limited capacity of Working Memory, its efficiency is crucial in writing. We developed an experimental computerised training for fluent lexical retrieval, applying techniques previously restricted to laboratory use in a classroom setting and examined the effects on L2 writing. In a true experimental counterbalanced design, two sets of words were trained for productive use. Previous analyses of a selection of the data (Snellings et al., 2002, Chapter 2) have shown that students in each trained words condition attained superior skills in lexical retrieval on the trained words in comparison to students that were not trained on these words. The current analysis involving additional data makes clear that these skills transfer to narrative writing. The trained students used more words appropriate to the narratives. In one of the two trained groups students also showed significant improvements in their ability to express essential content elements. No effects were found on global communicative quality. Results are discussed in the context of theories of Working Memory in text production. We also consider the implications of these results for education.

INTRODUCTION
Lexical retrieval ("translating") and text production processes were already distinguished by Hayes and Flower (1980) in their influential writing model. Moreover, Kellogg (1996), in his description of the components of Working Memory (WM, see Baddeley, 1986) involved in written text production, stresses the importance of lexical retrieval. Lexical retrieval 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 (Levelt, 1989; Levelt et al., 1998) or written down. Lexical retrieval is an essential process in both oral and written language production as it is an important factor in attaining fluency. In addition, spending too much attention on lexical retrieval may have adverse effects on the writing process because WM capacity is severely limited (for a discussion of attention limits see Kellogg, 1999; Lea & Levy, 1999). As a result, more attention for lexical retrieval results in less attention that can be devoted to planning, monitoring or other cognitive operations used in writing (Hayes, 1996; Kellogg, 1996). Already in 1980, Flower and Hayes (Flower & Hayes, 1980) suggested that by making sentence production processes more efficient (e.g. with sentence-combining exercises), the writer has time to concentrate on other important constraints of writing. We would like to argue that the effort involved in lexical retrieval, which is part of task difficulty, is an important factor in determining to what extent cognitive capabilities are taxed. If lexical retrieval proceeds relatively effortlessly, which is not inconceivable in a first language (L1), WM capacity will be less heavily taxed. In contrast, when a second language (L2) is involved, lexical retrieval becomes more problematic for all students. For this reason, research on the relationship between WM and writing in a second language is all the more relevant. Sofar, most studies on this relationship have been concerned with L1 writing. The current study will specifically address the effects of WM capacity on L2 writing.

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 reaction times about 350 ms over baseline times. Kellogg underlined the importance of lexical retrieval for writing and showed that the process of lexical retrieval put considerable demands on Working Memory.

The idea that due to attention limitations cognitive resources dedicated to one sub-process preclude attention being given to other important sub-processes and that this may even reflect on the quality of the process as a whole has a well-established tradition in reading research. In studies on reading, evidence has been found for a relation between the limited capacity of WM and the understanding of text (Favreau & Segalowitz, 1983; Just & Carpenter, 1992; Koda, 1996; Perfetti, 1985).

In the following sections we shall first discuss research in the field of writing that shows a relationship between WM and writing proficiency (for a review see McCutchen, 1996). Second, we will discuss studies showing negative effects on writing of increased cognitive demands on WM. Finally, we will discuss studies that have demonstrated effects of decreasing WM load on writing.

Daiute (Daiute, 1981) showed that syntax errors in writing resulted from memory limitations. She argued that a large number of words before error onset, strong perceptual clauses or complex syntactic environments constitute a burden to short-term memory. Once important grammatical information from prior clauses has become inaccessible, it becomes difficult to complete sentences correctly. Benton, Kraft, Glover, and Plake (1984) found that on some cognitive-processing tasks good writers were better able to hold and manipulate both small and larger units of information than poor writers. Swanson and Berninger (1996) showed that Working Memory contributes unique variance to writing independent of reading skill and that WM measures best predicted the text generation sub-process. Furthermore, only WM measures that reflect executive processing (sentence span measures that required “process monitoring”) significantly predicted writing fluency, micro-organization and quality. Ransdell and Levy (1996; 1999) found a relationship between a writing span measure and writing. Finally, McCutchen, Covill, Hoyne and Mildes (1994), specifically stated the importance of lexical retrieval (the fluency of translating processes) for writing. In a correlational study 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. They followed Flower and Hayes (1980) in that writers have to handle constraints simultaneously and the more attention devoted to translating and lexical retrieval, the less can be devoted to planning, generating knowledge or reviewing. As a consequence, they stressed the importance of lexical retrieval, especially for the less skilled writer’s planning and reviewing abilities in writing. Once writers can retrieve words fluently, they can devote attention to other processes and this enables them to write better texts.

Nevertheless, although the evidence for a relationship between WM and writing has been convincing, we have to be careful in drawing the conclusion of a causal relationship between WM and writing on the basis of correlational studies. Therefore, we will now consider some experimental studies that have focused on the effects of increasing demands on WM (for a review, see Alamargot & Chanquoy, 2001). In these experimental studies researchers have systematically increased the cognitive load in writing tasks. By tapping the influence of this load on writing they have shown the importance of WM. Fayol, Largy and Lemaire (1994) carried out an experiment in which increased memory load in writing due to concurrent tasks resulted in agreement errors. Penningroth and Rosenberg (1995) showed that
narrative writing with a higher information-processing load negatively affected ratings of coherence but not holistic quality. In addition, Lea and Levy (1999) demonstrated that a secondary task negatively affected both writing fluency as well as writing quality. Finally, in a study focusing on lexical retrieval, Brown, McDonald, Brown and Carr (1988) showed that when lexical material had to be retrieved from memory and could not be copied, lexical retrieval usurped attention from execution processes resulting in reduced legibility and more errors. Interestingly, Scardamalia, Bereiter and Goelman (1982) only found a relationship between WM constraints and text length but not with text quality. They showed that for children (grade 4 and 6) slow rate of output did not lead to a decline in text quality and neither did the interference of mechanical demands. On the other hand, in another experiment (cf. Scardamalia et al., 1982) they showed that when children were forced to retain information in short-term memory, there was sometimes information loss between forecasts and actual writing, suggesting that a slow writing rate may also have a disadvantage in that children forget what they were going to say. This would be reflected in terms of a loss in richness of content, but all depending on the phase of writing they are in.

Some experimental studies have taken another approach by showing how reduced demands on WM during the writing task facilitate writing. Although research in this direction has been scarce, this approach is promising because it may have important practical implications for education as well. As far as the reflection process is concerned, Kellogg (1990) showed that when writers only received a topic to write about, making plans before starting to write reduced the demands during the actual writing process and resulted in better texts. Bereiter and Scardamalia (1987) showed that helping children to identify local problems in their texts resulted in better evaluations, diagnoses and revisions. Van Gelderen (1997) showed that error detection and revision activities on the macro-level of the text could be considerably improved by several facilitations of the revision process, such as improving someone else’s text and providing standardized phrases for evaluation and reformulation. Recently, Chenowith & Hayes (2001) have shown in a small cross-sectional study that L2 language courses increased fluency (words per minute) in writing. Finally, Glynn, Britton, Muth and Dugan (1982) showed that when teachers in a graduate course did not have to pay attention to the logical sequence of the ideas expressed, or the incorporation of the ideas into sentences, these teachers’ preliminary drafts improved. In particular, if they could suffice with using ordered or unordered propositions they were able to use more persuasive arguments that argued pro a certain position in their preliminary drafts. As most arguments in final drafts were transferred from the preliminary drafts these effects persisted in the final drafts. In the final drafts they also used more arguments per sentence if they had not been asked to consider mechanics, the logical sequence of the ideas expressed nor the incorporation of the ideas into sentences in their preliminary drafts. Interestingly, a second experiment involving undergraduates of average and low verbal ability (as measured on the Scholastic Aptitude Test) showed that argument production in both preliminary and final drafts only increased when structure demands were absent for writers with average verbal ability. Writers with low verbal ability could not profit from the relaxed demands on structure, suggesting that positive effects on quality not only depend on cognitive demands on WM but also on verbal ability.

In summary, even though the importance of cognitive capacity has been stressed repeatedly, most evidence has either been correlational or restricted to showing that increasing the demands on Working Memory subskills has negative effects on writing (see for different findings, Scardamalia et al., 1982). Few studies have tried to improve writing by reducing the demands on WM. Moreover, there has been no research trying to improve L2 writing by enhancing lexical retrieval skills, even though it is an essential process in language production that puts considerable demands on working memory, especially in a second language. In addition, as pointed out before, there is evidence from studies such as Glynn et al. (1982) that
positive effects on argument production appear conditional on verbal ability of the writers. Consequently, enhancing lexical retrieval may be a promising approach as it is one aspect of verbal ability. Both from a theoretical and a practical perspective it would therefore be interesting to see whether an experimental intervention aimed at enhancing lexical retrieval can have a positive effect on writing.

In an earlier study we have demonstrated that L2 lexical retrieval can be enhanced through an experimental intervention (Snellings et al., 2002, Chapter 2). We have shown that training resulted in higher accuracy scores and superior reaction times for Dutch learners of English (L2), with medium to large effect sizes. In the current study, we will perform additional analyses and we will include writing scores. In particular, we will look at the effects of enhanced lexical retrieval on several aspects of written L2 texts. We will first investigate whether the training resulted in increased use of the words in the writing tasks. Secondly, we will investigate whether trained students were better able to express the details of the story and the desired content. To examine whether enhanced lexical retrieval had an effect on global communicative quality of the text we conduct an additional exploratory analysis. Consequently, we will test the following two hypotheses:

a. Enhancing lexical retrieval of a specific word set has a positive effect on the productive use of this word set in L2 writing.

b. Enhancing lexical retrieval of a specific word set has a positive effect on the expression of essential content elements in L2 writing.

In addition, because global text quality is the result of many different aspects of language production such as style or organizing activities, and although lexical retrieval is important, it is but one aspect, we will explore whether:

c. Enhancing lexical retrieval of a specific word set has a positive effect on the global quality of L2 writing.

In the following sections we will give a detailed description of the experiment, including the tasks used in the experimental computer program, the dependent measures based on the writing assignments and the implications for further research into second language writing skills.

METHOD

Participants
A total of 103 Dutch students participated in this study, all in the beginning of their third year of secondary education (grade 9, age 14-15). Students were sampled from four different classes, two classes from lower general secondary education (n=47) and two classes from pre-university education (n=56). They had received two full years of formal English tuition (2-3 lessons a week) at secondary education level, and a two year introductory course at the elementary level focusing mainly on basic oral communication skills (maximum 1 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 first language. The results showed that 90 students were native speakers of Dutch. The remaining 11 students (only 101 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 a first language). The L2 speakers of Dutch 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 randomly split in half and the resulting 8 groups were trained concurrently over a four-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 of each group were not trained in the other group’s Word Set. After the training all students wrote two writing assignments (based on cartoons). Writing assignment A could be written using the words from Word Set A whereas the words in Word Set B could be used in writing assignment B.

**Materials**

**Stimulus words**
As we wanted to increase the retrieval fluency of words instead of teaching new words, the words had to be familiar. Word-frequency ratings may not be the best indication of familiarity, especially because there are no learner corpuses for productive vocabulary of Dutch EFL learners. To obtain a selection of familiar (productive) words that could be used in the stories, we conducted a pilot study in which 9th grade students had to write stories based on the same cartoons as in the current study. We selected only words that were used by the students and had straightforward translations in Dutch. In the selection we also used teacher intuitions about word familiarity. 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 (Weinert, 1995), we also selected some word combinations.

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

Because our objective was increasing retrieval speed, students were encouraged to speed up in the training. The program provided them with a baseline (speed of previous reaction on the same word) for reference to their progress. Moreover, as the words 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.

Each student encountered each word ten times and received immediate feedback on correctness and current speed. “Correct” feedback followed when the student had correctly responded. We consistently trained the same meaning of the words used so that mappings of stimuli to responses were consistent rather than varied (Schneider & Shiffrin, 1977; Segalowitz & Gatbonton, 1995). Students were trained in a school setting in scheduled English classes. Standard duration of each training was 50 minutes (except for one training of 45 minutes). There was a one week interval between each training session. Having collected data on lexical decision and lexical retrieval tasks to establish whether lexical retrieval was
enhanced, we administered the writing assignments in the following week (all students first did the assignment belonging to the words they were trained in).

Each specific word was trained with four different types of exercise, moving from receptive skills to the target production skills:

**Block exercise**
The aim of this exercise was to both strengthen the receptive knowledge of the words and training 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.

After some time
woke up she
Z M
Students had to press either Z or M for the constituent they thought is the first one following After some time.

**Detection exercise**
To strengthen the connections between the concepts and the lexical forms, this 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 this specific context 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.

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, they were asked to judge the words for appropriate 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”).

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

time centuries
Z M
Students had to indicate whether ages (X) is acceptable, or whether it had better be replaced by time (Z) or centuries (M) to yield a 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. Picture naming tasks can be assumed to tap lexical access. In addition, the process tapped proceeds from meaning to word form just as in language production in writing. Unfortunately, the set of words we are interested in (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 a word translation task. This task taps similar processes as in picture naming, in that both picture naming and translation tasks involve concept mediation (Potter et al., 1984). For similar tasks in testing, also see (De Groot et al., 1994; Potter et al., 1984). The word translation task has learners fill in the correct translation within a meaningful context, a situation that matches lexical retrieval in written production most closely. Another reason to choose a translation task was that this makes a focus on more words at the same time (in contrast to picture tasks) possible.

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 seconds, stimulating the students to give faster answers.

**voordat** the bridge collapsed he had reached the other side

Translation: *before*

The order of the items in all exercises was randomized each time the exercise was done. 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 the proper functioning of both software and hardware but also the adequacy of the exercises. On the basis of this test we clarified the instructions and put more emphasis on speed as we wanted all students to have the same focus. Problems in the program and in the design of the exercises were solved in the final version used in the experiment. In the final experiment students began with the block exercises, followed by the correction, detection and translation exercises.

**Dependent Measures**

After the training in lexical retrieval and establishing an increase in retrieval speed, we administered two writing assignments (Cartoon tasks). 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. Students wrote two cartoon-based stories (writing assignment A and B). One story was about two boys who go fishing, the other about a man who goes walking in the mountains and encounters a tiger. The pictures were in chronological order. The dependent variables are measured either in writing assignment A or B. They are “Number of trained words used”, “Expression of content elements” and “Global text quality”.

**Number of trained words used**

We counted words appearing in the students’ texts that were trained in isolation or as part of a collocation. In addition, we looked at the occurrence of collocations themselves. Because our aim was to look into the effects of more efficient lexical retrieval and reduced demands on Working Memory, we only considered the use of words and collocations that were trained in that exact form.

**Expression of content elements**

To define content elements that were essential for the story we used the definition of “narrative schemata” as described by Stein and Trabasso (1982). In their view, schemata “specify the components of a story that should be included in a text, the temporal sequence of the narrative events, and the type of causal relations that should connect the events”. Crucially, these schemata enable us to establish whether certain parts of a story have been omitted. As pointed out by Stein and Trabasso, the underlying representation of a story in
memory can be inferred from a text. As a result, even though some parts may be deleted from
the text structure (e.g. “internal responses”), they should be inferable from the text. Stein and
Trabasso divide the stories in a “Setting” and the “Episode”. The latter consists of five
categories that contain specific information and serve a different function in the schema. The
five categories are Initiating event, Internal response, Attempt, Consequence and Reaction.
These categories are psychological entities and not linguistic. An “Initiating event” can be an
action that causes the protagonist to respond emotionally and formulate a goal. “Internal
response” is the emotional reaction of the protagonist that causes him or her to initiate action.
The “Attempt” is the overt action to attain a goal, “Consequence” is the event indicating goal
attainment and “Reaction” is the emotional response of the protagonist to the outcome of
his/her actions.

Only content elements that occurred in the appropriate part of the story were scored.
The combination of narrative schema and content elements makes this possible. Content
elements are represented by “propositions” or predicates and their arguments (Kintsch, 1998).
Because we did not aim for an all-encompassing analysis of the text into propositions but only
used them as a tool to isolate content elements, we restricted the predicates in our analysis to
verbs (e.g. “boys are lucky”). On the basis of the narrative schemata the elements were
identified and put into a check-list. We divided the text into propositions before scoring.

The division of the text into propositions and the check-list with content elements
allowed us to judge whether the students had expressed the content elements in their correct
position in the story. Scorers indicated whether each proposition contained an essential
content element, added other or irrelevant information. Each proposition normally contains
one content element but occasionally a proposition can contain more content elements (we
only scored this for essential content elements, e.g. the proposition “they saw three big fish”
contains the content elements “they saw fish” and “the fish were big”). The content elements
can be expressed by various linguistic forms. On an abstract level the content element should
be the same, e.g. “they see boys” is equivalent to “they meet boys” because “meeting” implies
“seeing” each other. Whenever a content element is not linguistically realized but is implied
by the other words it is also scored (e.g. “the owner warns the man” implies “there is an
owner”).

Global text quality
The writing assignments instructed students to use the correct register for their peers, to
induce interest value, and finally, to be coherent in the absence of the cartoons themselves.
The texts were rated on their global quality by two independent raters according to a “primary
trait” instruction, i.e. they had to give a general rating of the quality of the text considering its
primary trait, “how well written is this story”. Raters were instructed to concentrate on
content and organization, and to consider linguistic errors only to the extent that they
interfered with the understanding of the story. Besides these general instructions the raters
were provided with benchmark texts as a reference. Five benchmarks were selected on the
basis of a separate analysis with six raters according to a scaling procedure (Blok, 1986). The
benchmarks represented a very weak (10th percentile), a weak (25th), an average (50th= score
of 100), a good (75th) and a very good text (90th).

Covariate
To control for initial differences in word knowledge we used scores on a vocabulary test as a
covariate. The vocabulary test [an adapted version of a test used in the NELSON project:
(Schoonen et al., 2002) consisted of multiple-choice items, each item having a neutral carrier
sentence with a word in bold print. Students had to choose between four alternative Dutch
translations. Words were selected from the 5000 most frequent words in the COBUILD
corpus (Cobuild, 2000). On the basis of a pretest we removed items that were too easy (a p-value >.90), to obtain a more discriminative test. In addition, we removed items with negative Item-Total correlations. The final test had 64 items. Scores on the vocabulary test correlated significantly (p<.05, one-tailed) with all three dependent variables and the correlations ranged from .19 to .40.

Data collection
Both the vocabulary test and the writing assignments were paper and pencil tests. The teacher administered the vocabulary test before the start of the training. Following the training period, a trained research assistant administered the writing assignments.

Analyses
Means, standard deviations, and estimates of reliability (percentage of rater agreement for Use of content elements and Cronbach's α for the Primary trait score) were calculated. The hypotheses were tested by two One-way MANCOVA's on the measures we derived from the writing assignments, with Condition (A or B) as independent and Number of trained words used, Expression of content elements and Global text quality as dependent measures, and vocabulary knowledge as a covariate.

In the case of the MANCOVA's on the writing assignments, results of evaluation of the assumptions of independent observations, equal population covariance matrices for the dependent variables, 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 only exception was the assumption of multivariate normality of dependent variables in each group for the MANCOVA on the writing assignment B measure. In the case of “Expression of content elements” in the Trained B words condition, the Kolmogorov-Smirnov test was significant, indicating that multivariate normality did not hold here. However, because this is the only group and dependent variable for which this assumption was not valid, the effect on power is small (Stevens, p.247). In addition, MANCOVA is rather robust regarding a violation of multivariate normality.

Scoring
In the case of “Number of trained words used” we used the summed scores of the isolated words, words used in isolation that were trained as part of a collocation and the collocations themselves. We did not include derivations or inflected words because applying conjugalional or inflectional rules may be effortful and hence an effect on quality measures is less likely. In the case of “Expression of content elements”, one scorer scored all the texts after we had established agreement between the scorer and the principal investigator. Scorer reliability was established by looking at the percentage of agreement between the scorer and the principal investigator on twenty texts of each writing assignment. The identically scored content elements were 89.9% for writing assignment A and 87.4% for writing assignment B.

Finally, in the case of “Global text quality” two raters scored each text independently. Panel reliability (Cronbach’s α) was .84 for “writing assignment A” and .88 for “writing assignment B”.

Missing values
In the analysis of both writing assignment A and writing assignment B only trained students were included that had done all exercises and the writing assignment (writing assignment A: 41, writing assignment B: 42). For inclusion of data from the assignment with untrained words, presence at all exercises was of course not required (writing assignment A: 52, writing
assignment B: 50). As a result, the analysis of writing assignment A was based on scores from 93 students, the analysis of writing assignment B on 92 students.

RESULTS

In this section we will first present descriptive measures of the tests used in this experiment. Next, we will discuss the experimental effects on our dependent measures for both writing assignment A and B.

Descriptive and psychometric statistics

Dependent measures

Table 1. Descriptives writing measures A: unadjusted mean scores (maximum possible score between brackets) and standard deviation.

<table>
<thead>
<tr>
<th>Words</th>
<th>Trained A words condition</th>
<th>Trained B words condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Used A words</td>
<td>41.05 (71)</td>
<td>28.87 (71)</td>
</tr>
<tr>
<td>Expressed content elements A</td>
<td>25.39 (47*)</td>
<td>22.54 (47*)</td>
</tr>
<tr>
<td>Primary trait score A</td>
<td>100.62</td>
<td>96.62</td>
</tr>
</tbody>
</table>

*if elements are expressed more than once the maximum possible score can be higher.

In Table 1 we give an overview of the observed scores for the writing assignment A measures. We see that in the case of the Used A words, students in the Trained A words condition’s mean score is 41.05 words and the Trained B words condition’s mean score is lower with a mean score of 28.87 words. Concerning Expressed content elements A, the Trained A words condition’s mean score is 25.39, also higher than the Trained B words condition’s score of 22.54. Finally, the Global text quality score for students in the Trained A words condition is 100.62, again higher than the mean score of 96.62 for students in the Trained B words condition.
Table 2. Descriptives writing measures B: unadjusted mean scores (maximum possible score between brackets) and standard deviation.

<table>
<thead>
<tr>
<th>Words</th>
<th>Trained A words condition ( n=50 )</th>
<th>Trained B words condition ( n=42 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( M )</td>
<td>( SD )</td>
</tr>
<tr>
<td>Used B words</td>
<td>22.06 (69)</td>
<td>8.45</td>
</tr>
<tr>
<td>Expressed content elements B</td>
<td>16.12 (31*)</td>
<td>4.40</td>
</tr>
<tr>
<td>Primary trait score B</td>
<td>95.38</td>
<td>11.72</td>
</tr>
</tbody>
</table>

*if elements are expressed more than once the maximum possible score can be higher.

Table 2 shows the observed scores on the writing assignment B measures. Students in the Trained A words condition’s mean score on the Used B words is 22.06 words and therefore lower than the Trained B words condition’s mean score of 31.19 words. Concerning Expressed content elements B, the Trained A words condition’s mean score of 16.12 is again lower than the Trained B words condition’s score of 17.12. Finally, the Global text quality score for students in the Trained A words condition is 95.38, nearly identical to the score of 95.40 for students in the Trained B words condition.

Covariate

On the vocabulary test the mean score was 45.19 (\( N=103 \), maximum possible score=64, \( SD=8.74 \)) and the reliability was .87. So, on average the vocabulary test was not too difficult and its internal consistency good.

Experimental effects writing assignment A

We predicted the scores to be higher for students in the Trained A words condition for both number of Used A words and Expressed content elements A. The analysis of text quality scores has been added for exploration. Our first measure derived from writing assignment A is the “Used A words” score. In our analysis we used the summed scores of isolated words and collocations.
Figure 1. Scores on writing measures A: Used A Words.

Results of the MANCOVA show a significant main effect for condition with $F(1, 90)=32.032, p<.001, \eta^2=.26$. Figure 1 shows that students in the Trained A words condition ($M=40.83, SE=1.56$) used more A words than students that were not trained on these words ($M=29.04, SE=1.38$). In the case of Expressed content elements there is again a significant main effect for condition with $F(1, 90)=5.567, p<.05, \eta^2=.06$.

Figure 2. Scores on writing measures A: Expressed Content Elements A.
Figure 2 shows that students in the Trained A words condition obtained a higher score on Expressed content elements A ($M=25.30, SE=.85$) than students that were not trained ($M=22.61, SE=.76$). Results of the MANCOVA show no significant effect of the training on Global text quality. Nevertheless, adjusted scores are in the expected direction with students in the Trained A words condition scoring higher on the Primary trait score A ($M=100.39, SE=1.79$) than students that were not trained ($M=96.80, SE=1.58$).

**Experimental effects writing assignment B**

On the writing assignment B measures, we expected superior scores for students in the Trained B words condition on number of Used B words and Expressed content elements B. Again, the analysis of Global text quality scores was exploratory.

![Figure 3. Scores on writing measures B: Used B Words.](image)

There is a significant main effect for condition with $F(1, 89)=27.574, p<.001, \eta^2=.24$. Figure 3 shows that students in the Trained B words condition ($M=31.32, SE=1.31$) used more B words than students that were not trained on these words ($M=21.96, SE=1.20$).
Figure 4. Scores on writing measures B: Expressed Content Elements B.

Although Figure 4 shows that students in the Trained B words condition obtain a slightly higher score on Expressed content elements B (M=17.21, SE=.56) than students that were not trained (M=16.05, SE=.52) this difference is not significant with F(1, 89)=2.307, p=.13.

Finally, results from the MANCOVA show that the differences on Global text quality are not significant. Students in the Trained B words condition score nearly identical on the Primary trait score B (M=95.64, SE=1.63) to students that were not trained (M=95.19, SE=1.49). Because the interaction between the covariate vocabulary and condition (homogeneous regression slopes assumption) was exactly at the significance level (p=0.05) for the dependent variable Global text quality, we also carried out an analysis without covariate. The results remained the same.

DISCUSSION

Lexical retrieval is an important sub-process in written language production that has been neglected in research into second language writing instruction. So far, research into cognitive demands on Working Memory and writing has either been correlational or has demonstrated the importance of Working Memory by showing negative effects of increasing cognitive demands during the writing process. In contrast, research on reducing cognitive demands has been limited. In the current study, we have investigated a new approach to decreasing the cognitive effort in L2 writing. We have looked at the effects of increased lexical retrieval skills on the productive use of words, the expression of content elements and in an exploratory analysis, we also looked at the effects on Global text quality.

Effects on Number of trained words used

The results for both the Used A words and the Used B words show that in each case students in the Trained words condition can use more words productively than students that were not trained. In terms of Cohen’s effect size (Cohen, 1988), 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 the case of the A words the size of the effect was medium with 26% of the...
variance in scores due to the training. In the case of the B words there was also a medium effect. Here 24% of the variance in scores is due to the training.

**Effects on Expression of content elements**
Students in the Trained A words condition used more content elements A than the untrained students. A total of 6% of the scores can be accounted for by the training. Although the results of the content elements B were not statistically significant, the pattern of differences was similar to that of the content elements A. Again, the trained B group had a higher mean than the trained A group. These results suggest that when students have more attention to spare because lexical retrieval of the trained words proceeds effortlessly, it is beneficial to other writing processes. Attention can be devoted to a search for more appropriate wording, even though these words themselves are not literally trained. In addition, students can monitor their output and examine whether all essential content elements are expressed. We will finally offer a tentative explanation why the effects, although in the right direction, were not as prominent in the case of writing assignment B. One possibility could be that the link between the trained words and the content to be expressed in writing assignment B was less strong. Even though students in the trained group used more of the trained words, the content elements could also be expressed using other words, thus reducing the influence of the trained words. Another possibility could be that writing assignment B was more difficult, even if a selection of the necessary words was easy to retrieve. This meant that the content of the B cartoons was more difficult to express, so that even though students in the Trained words condition could use more words it did not help them in expressing the content in a more detailed way.

**Effects on Global text quality**
Besides testing the predictions on used words and expressed content elements, we explored whether there were any effects of the training on the scores of Global text quality. On the basis of these results, we can conclude that even though the results were in the expected direction in the case of writing assignment A, the means of students in the Trained words conditions on both tasks were not significantly higher. It appears that the effect of enhanced lexical retrieval was not large enough as to influence scores on a global text level. This result qualifies predictions by Hayes and Kellogg (Hayes, 1996; Kellogg, 1996), who assumed that more attentional control needed for text production processes such as lexical retrieval results in less attention that can be devoted to planning, monitoring or other cognitive operations used in writing. In the next section we will discuss possible reasons and suggest ways to investigate this issue.

**Implications for a theory of lexical competence, suggestions for future research and educational relevance**
In the current experiment, words that were trained in a computerised setting had to be transferred and applied in a writing task. From the research discussed by Anderson (2000), it becomes clear that even transfer involving the same expertise is not without problems if this expertise has to be applied in different contexts. The finding that the trained students used more words from the trained word set implies that the effect of the training transfers to a real writing task, leading to a change in the writing process that can be detected in the writing product. This finding is also relevant for theories of lexical competence. DeKeyser has argued that understanding sentences in a language quickly does not mean that one can produce them fast (DeKeyser, 2001). Henriksen (1999) provided a theoretical account of the different dimensions that play a role in the development of vocabulary. She argued that there are three possible causes for a difference between receptive and productive competence. Either a difference between input and output specifications (encoding orthography and sound, see Ellis,
1995), the level of “automaticity” (Meara, 1996), or the type and quality of the meaning representation. In our view, the learning of output specifications through practice and extensive repetition may eventually lead to automaticity so that these two factors could be subsumed under one factor labeled “retrieval speed”. The current findings have implications for such a theory of productive competence. The difference in word use after the speed training shows that even in the case of relatively easy words, regardless of the quality of the meaning representation, speed in accessing lexical items in the mental lexicon (retrieval speed) is an important factor in attaining productive competence. Although we have demonstrated that increasing speed increases productive competence, it remains an issue for further research whether actual “automaticity” (see for a review of this concept, DeKeyser, 2001) can be trained and established. It may well be that effects are more profound in that case.

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. Now that effects have been demonstrated on cartoon tasks, we can investigate the effects on other writing tasks and in other limited domains, for instance in English for specific purposes. To maximize the impact of the training for practical purposes, inclusion of words used in a particular teaching context is also warranted. Ultimately, we need to get a better understanding of which words are most suitable for training and which are also valuable in more general text types. From a theoretical point of view it is also important to know more about the relation of a focus on speed and characteristics of the words trained, such as word type and familiarity.

Apart from the increased use of trained words, we have demonstrated that enhanced lexical retrieval has effect on the expression of detailed content in L2 writing. A think-aloud study into L2 writing by Zimmermann (2000) may clarify how these local effects come about. He showed that tentative formulations and simplified tentative formulations uttered before the actual act of writing are relatively frequent. Tentative formulations are formulations that are rejected and replaced by other formulations in the actual written text. Once lexical retrieval proceeds effortlessly, it could facilitate writing in two different ways. In the first place it may help students to retrieve the necessary vocabulary directly, without having to simplify the wording of the concepts they want to express. This suggests that the quality of the first formulation improves and attention does not need to be dedicated to tentative formulations that have to be rejected later (see also, Gelderen & Oostdam, in press). Alternatively, the training provides effortless retrieval of a tentative formulation that can be changed later. In this case, it will not be reflected in the used words because the tentative formulation has been replaced in the final version. In both scenarios, struggling for the first preliminary formulation becomes unnecessary and students do not get stuck at finding tentative formulations. As a result, more attention can be dedicated to finding a more precise and better formulation.

Although we did find effects at a local text level, we did not find significant effects on global writing quality. It could be that the link between the trained words and the content to be expressed was not strong enough. However, on the basis of our data, we assume that on a global text level, other skills such as organising and revising ability obscure the differences in efficiency of lexical retrieval and may compensate for less efficient retrieval processes. These results are in accordance with previous research, as both Penningroth and Rosenberg, as well as Scardamalia et al. (Penningroth & Rosenberg, 1995; Scardamalia et al., 1982) have argued that the effects of reduced cognitive demands only operate at the local text level. Research by Scardamalia et al. (1982) suggested that due to memory loss children failed to write all the words in their forecasts. Typically, this resulted in reduced richness of content. Penningroth and Rosenberg (1995) found only an effect on coherence but not on holistic quality.
Another reason may be the fact that we did not measure cognitive load directly. The research by Penningroth and Rosenberg (1995) showed that even though cognitive demands were experimentally increased, the overall level of cognitive effort as measured by Reaction time on a secondary task (pressing a key as fast as possible when a tone occurred during writing) did not increase. This was because writers coped with a high WM load by changing their writing strategies. Interestingly, thinking-aloud reports made clear that the undergraduate students changed the distribution of planning, translating and reviewing across the writing process. Under a high processing load students did not translate in the first phase of writing ("the first third of the writing session") but seemed to be forced to catch up in the third and final phase of the writing process. In the first phase they were already reviewing, a process that is normally reserved for the final phase. Future research should therefore take up these matters more directly, either by using RT's on a secondary task during writing, or by establishing individual differences in Working Memory on a complex span measure (see for example Ransdell & Levy, 1996). It is reasonable to assume that effects may be stronger for students with a smaller span as students with a larger span are better able to cope with the higher demands of lexical retrieval on WM capacity.

Finally, it has to be noted that even though the sub-process of lexical retrieval has become fast and efficient, other essential processes in writing definitely are not. In particular, experienced writers' self reports demonstrate the amount of effort even these writers have to invest in the writing process (see for a discussion Torrance & Jeffery, 1999). Crucially, and in contrast to beginning writers in the L2, expert writers show a far greater metacognitive control. Although it is safe to assume that some of our writers (grade 9, age 14-15) possess some metacognitive skills in their mother tongue that may be transferred and used in the L2 when the demands on WM are not too high, other writers may lack these skills altogether so there is nothing to transfer. Glynn et al. (1982) raised a similar argument when their research showed that only undergraduates with average verbal ability could profit from reducing the demands on WM. They found no effects in the case of students with low verbal ability and concluded that "additional capacity can provide writers only with the opportunity to increase their pool of persuasive arguments; it cannot remediate deficiencies in ideational fluency". In this case, enhanced lexical retrieval may provide the opportunity to use metacognitive skills but cannot make up for the lack of it. Future research should therefore directly measure the metacognitive skills in the L1.

The results in this study do have significant implications for writing instruction. The increase of the productive use of words and the more detailed expression of essential content elements are important aspects of writing. As the discussion by Henriksen (1999) showed, it is by no means clear how to establish productive competence. These findings show that enhancing lexical retrieval is one approach that effectively increases production in actual writing. The implication of this finding is that in teaching, attention should be focused on this dimension as well. Simply teaching words until their meaning is known may not be sufficient. Only when students can retrieve words effortlessly, teaching has succeeded in the sense that students will be able to use the words productively. On a practical level, if we take into account that students encountered each word only ten times over a four week period, the time on task required seems reasonable in an educational context. In addition, it seems advisable to provide spaced presentations rather than massed presentations. Apart from the effective use of spaced presentations in the current research, psychological learning theories also suggest that spaced presentations are particularly effective (see for discussion Ellis, 1995). In view of the importance that has recently been attributed to collocations for reducing cognitive effort (Weinert, 1995), it might be useful for educational and research purposes to put a stronger focus on combinations of words. As pointed out before, metacognitive skills remain an important aspect in writing high quality texts. For students with sufficient metacognitive skills,
training in lexical retrieval could be sufficient whereas students who lack these skills may be better served by a training that combines metacognitive skills with a training in enhanced lexical retrieval. Nevertheless, we would like to stress that a focus on enhanced lexical retrieval is not a plea for a repetitive training of words through drills. The current training does not focus on structures but on the meaning of utterances in a meaningful context. In this respect our approach is similar to the communicative approach to automaticity along the lines suggested by Gatbonton & Segalowitz (1988).

This is the first study that has explicitly focused on the effects of enhanced lexical retrieval on L2 writing and these results provide new insights into the process of writing and the factors that influence writing quality. We have found that enhancing lexical retrieval of a specific set of words has positive effects on both the number of words from this set that can be used productively as well as on the level of detail in which essential content can be expressed. In an exploratory analysis, we found that reducing cognitive demands on Working Memory by enhancing lexical retrieval did not result in texts of higher global quality. To clarify this issue, we suggest that research is needed that looks more directly at both the metacognitive skill of the L2 writers and the WM load during the writing process.