Developing second-language listening comprehension: Effects of training lower-order skills versus higher-order strategy.
Poelmans, P.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Chapter 5  Training study

5.1. Introduction

As stated already in the previous chapters, one of the requirements for L2 learners to communicate adequately with native speakers of the L2 is to understand what these speakers say. To achieve this, it is necessary to recognize words in continuous speech. Traditionally, listening training in second language courses focuses on the higher-order comprehension part of the listening comprehension skill. Less attention is paid to training – let alone automatization – of lower-order skills such as word recognition. As described previously (see Chapter 1), many students fail, for example, the listening comprehension part of the national exam Dutch as a second language; this suggests that the effect of the ‘traditional training method’ is not optimal. A more effective method might entail a training of a combination of lower-order and higher-order component skills. The assessment of the effectiveness of such a combined approach would, however, diminish the chances of determining the contribution of each of the components. For research purposes therefore, it appears to be more appropriate to test the effect on listening comprehension of each component separately. This is what we aimed to do in this study. Thus, the experiment reported in this chapter should not be regarded as a comparison of two methods competing for educational relevance but as a comparison of two theoretically relevant component processes. The result of this theoretically oriented experiment may, however, well have implications for language education.

The main goal of the research described in this thesis, was to compare the effect on listening comprehension of a training of lower-order components with a training of higher-order strategies. The research question of the main experiment of our study can therefore be formulated as follows:
What is the relative effect of two training methods to improve listening comprehension performance of intermediate L2 learners,

(i) a method focusing on improvement of lower-order word recognition skills, and

(ii) a method consisting of the assignment of higher-order global comprehension tasks?

To answer this research question, a training study was set up, involving two experimental groups following different training programs. The first group, the so-called Recognition group, trained listening by focusing on the recognition of individual words in concatenated speech and on the identification of typically Dutch speech sounds. The second group, named the Comprehension group, trained listening in a more ‘traditional way’ by focussing on comprehension. A third group was formed as the Control group, which was excluded from the training programs. Before the actual training started there was a test session to measure participants’ pre-treatment performance. After the training, there was a second test session to test post-treatment performance in order to investigate the effect of the training. In summary, the set-up of the study was as presented in Figure 5.1.

Event: Test stage 1 → Training: → Test stage 2
- focused on understanding
- focused on recognition

Duration: 1 week 7-hours in four weeks 1 week

Figure 5.1: Set-up of the training study.

During the two test stages, tests to establish participants’ knowledge of Dutch grammar and vocabulary were administered as well as tests the results of which were expected to indicate a relation between these tests and general listening comprehension skills (e.g. memory tests). However, the most important were the tests that focused on the listening comprehension process itself (the listening comprehension test) and its component processes: word recognition (Lexical Decision test) and sentence processing (Sentence Verification test). Because these tests measure processes that were the focus of the training (e.g., recognizing words, comprehension of speech), they were expected to be sensitive to the effect of the training as well; they are defined as the dependent variables.
Our investigation aimed to follow the idea propounded by Segalowitz and Segalowitz (1993) that, for successful listening comprehension in a second language, recognizing (most of) the words spoken is conditional. This implies that to improve general listening comprehension, training lower-order ‘recognition’ skills such as word recognition would be more cost effective than training higher-order ‘understanding’ skills. In other words, the participants of the Recognition group will benefit more from the training than does the Comprehension group; this will be expressed amongst others by a better result on the listening comprehension test. In the light of this reasoning the following hypothesis can be formulated:

Students participating in the Recognition group will exceed students in the Comprehension group on the general Listening Comprehension test.

If no significantly difference is found in the performance between the Recognition group and the Comprehension group on the Listening Comprehension test it is expected that such a difference, in favour of the Recognition group, will be found in the performance on the Sentence Verification test or at least in the performance on the Lexical Decision test since these tests measure sub-processes that are specifically trained in the Recognition group.

The next sections describe in detail the method (section 5.2) and the results of the training study (section 5.3). In the last section of this chapter (section 5.4) the results will be interpreted and discussed.

5.2. Method

The study is designed as a pretest – treatment – posttest experiment with three groups of participants: two experimental groups (Recognition and Comprehension) and a no-training Control group.

5.2.1. Participants

All participants in this study were adult learners of Dutch as a second language between the ages of 20 and 37. They had various mother tongues: Afrikaans (1), Arabic (21), Bengali (1), Chinese (8), Czech (2), English (10), Estonian (2), Farsi (4), Finish (1), French (2), Georgian (1), German (9), Hebrew (3), Hungarian (1), Icelandic (1), Indonesian (4), Italian (2), Japanese (2), Moroccan Arabic (2), Norwegian (1), Persian (2), Polish (2),
Portuguese (1), Romanian (2), Russian (10), Serbian (2), Croatian (3), Spanish (11), Swedish (6), Turkish (5), Ukrainian (2), and Vietnamese (1); their length of stay in the Netherlands ranged from 5 months to 5 years at the time of testing.

Participants were students or prospective students at the University of Amsterdam, at the Vrije Universiteit of Amsterdam or at Universiteit Leiden. They were either student of Dutch Studies, which is a department offering foreign students a Master’s program in Dutch as a second language, or they were recruited in obligatory language courses for students with insufficient proficiency in Dutch who want to study at a Dutch university. Participants were recruited class wise, i.e., they participated in the study as a group during the class-hours of the language course they took. The assignment of students to classes was done by the teachers of the different universities at the beginning of the course. The fact that the study was conducted during class hours meant that it was not possible to assign students randomly to experimental conditions. It also implies that there was no control over the group size. However, classes were randomly assigned to experimental conditions. Five intact classes participated in the study, two in each of the experimental conditions and one in the control condition. Students of the Control group were recruited only at the University of Amsterdam at a later point in time in comparison with the participants of the other groups. The students of the Control group were comparable with the students of the experimental groups in the sense that they were all on the same level of language courses, furthermore they were taught Dutch in the same way as the other students at the University of Amsterdam, and by the same teachers. The students in the Control condition was offered the opportunity to follow the training that the Recognition group had followed, after the second stage of testing as the students in the Control group were not eager to participate in the test stages without participating in the training study. All students were paid for their participation.

Not all students participated in all tests; Table 5.1 gives an overview of the number of students in the various experimental stages broken down by participant group.

---

1 There was a Control group that was tested at the same time as the experimental groups. However, due to circumstances the test phases of the Control group were not completely under our control, we therefore decided to compile a new Control group.
Table 5.1: Participant numbers broken down by Condition and by Test stage.

<table>
<thead>
<tr>
<th></th>
<th>Recognition</th>
<th>Comprehension</th>
<th>Control</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stage 1</td>
<td>52</td>
<td>52</td>
<td>21</td>
<td>125</td>
</tr>
<tr>
<td>Stage 2</td>
<td>36</td>
<td>44</td>
<td>19</td>
<td>99</td>
</tr>
<tr>
<td>Complete data set</td>
<td>28</td>
<td>37</td>
<td>18</td>
<td>83</td>
</tr>
</tbody>
</table>

5.2.2. Test stages

The study consisted of two stages: Test stage 1 (pretraining sessions) and Test stage 2 (posttraining sessions). Both stages included three sessions: two group sessions and one individual session. During these sessions several tests were administered. We make a distinction between the following four kinds of tests:

1. **Selection tests** which served to determine whether participants did indeed possess a basic knowledge of Dutch, as required. In this category fall the Basic Vocabulary test and the Basic Grammar test, which were both administered during Test stage 1.

2. **Mediating tests** measuring individual differences in knowledge and skills, which might attenuate the effect of training on listening comprehension:

   1) The Vocabulary Size test, administered in Test stage 2, aimed to obtain a gross estimation of L2 lexical proficiency.

   2) The Listening Span test, administered in Test stage 1, aimed to measure processing and storage capacity of linguistic information at the sentence and lexical level.

   3) The Serial Recognition test, administered in Test stage 1, aimed to measure processing and storage capacity of linguistic information at the phonological level.

These three tests were included for two reasons. First, we wanted to ascertain whether interindividual differences in lexical knowledge as well as in processing and storage capacity for verbal information might attenuate the impact of the two training methods on listening comprehension. For this purpose, performances on the three tests were entered as covariates in the analyses comparing performance of the three groups on the three dependent variables (Listening Comprehension, Sentence Verification, and Lexical
Decision). Second, the inclusion of these variables would allow us to investigate an additional research question, not central to the issue of this study, but yet interesting enough to pursue. This issue pertains to the relative contribution of various forms of knowledge and skills on listening comprehension performance. This additional exploratory research question was formulated as follows:

_What is the relative contribution to listening comprehension of the following predictors: word recognition, sentence verification, vocabulary size, and verbal memory capacity at the sentence and lexical level (listening span) and at the phonological level (serial recognition)?_

For this purpose, we aimed to use the Vocabulary Size, Listening Span, and Serial Recognition tests, in addition to the Lexical Decision and Sentence Verification tests as independent variables in multiple regression analyses with Listening Comprehension (in pretest and posttest versions) as dependent variables.

3. **Control test** The Reading Comprehension test was administered to investigate whether participants performing poorly on the Listening Comprehension test would also perform poorly if this test was administered in a written input format, i.e., as a reading comprehension test. This Reading comprehension test was administered in Test stage 2, after the administration of the Listening Comprehension test.

4. **Pre and posttests**, which served to answer our main research question. The following measures were taken both in Test stage 1 and in Test stage 2, i.e., before as well as after training: Listening Comprehension, Lexical Decision, and Sentence Verification. As will be described in section 5.2.2.2, the post-treatment versions of these three measures consisted largely of the same items and partly of new items as those in the pre-treatment version.

**5.2.2.1. Test stage 1**

Figure 5.2 gives an overview of the tests administered in Test stage 1. Tests in italics were also administered after training, during Test stage 2.
Figure 5.2: Overview of the tests used in Test stage 1.

Selection tests
The selection tests can be categorised as off-line tests. These can be defined as tests that allow students to reflect before giving a response; students can react to stimuli without any time pressure. The Basic Vocabulary test and the Basic Grammar test that are used in the present training study as Selection tests, are the Knowledge tests of Pilot Study II. The design of these tests is described in detail in Chapter 4 section 4.2.1.1.

Both the Basic Vocabulary and the Basic Grammar test were group administered as paper-and-pencil tests. Students had to complete each test in one hour. Results were computed by counting the correct answers.

Mediating tests
Verbal memory tests were administered which aimed to test participants’ capacity to process orally presented sentences containing common words and store the last word from these sentences (Listening Span test) and their capacity to process and store orally presented strings of meaningless morphemes constructed in accordance with Dutch phonology (Serial Recognition test).

Performance on the memory test used in Pilot Study II, which was a reading span task, did not show a significant correlation with the comprehension tests used in that study while such positive cross-validation would be expected on the basis of the literature (e.g., Daneman and Carpenter 1980). As described in the discussion of Pilot Study II (section 4.4) this lack of correlation might be due to a pronunciation problem on the part of the participants. It was reasoned that the task of pronouncing the sentences, might have been too demanding so that there might not have been enough capacity left for remembering the last word of each sentence. We therefore decided to use a Listening Span test instead of a Reading Span test.
in the training study. In this way, no cognitive capacity is taken up by motor skills.

In addition to the visual versus auditory mode in which the span task is conducted, the length of the items as well as the language (L1 or L2) of the items may influence the results of the test. To accommodate the potential effect of word (language) familiarity, a Serial Recognition test was included. This task is not linked to an existing lexicon of a language in the sense that the experimental items are all invented words. However, the items of this Serial Recognition test were constructed following the phonology of Dutch and, in this sense, they are language specific.

1) Listening Span test
This test is an adapted version of the listening span test described in Daneman & Carpenter (1980). The test comprises 54 unrelated sentences, 10 to 13 syllables in length. The sentences used in this listening span task were the same as the sentences of the Reading Span test used in Pilot Study II. The only difference between the Reading and Listening Span test – other than the mode of presentation – was the order in which the sentences were presented. The test was computer administered and there was no overlap between the target words of this test and the word items of the lexical decision test.

Participants were asked to listen to the orally presented sentences and to remember the last word of each sentence. They were told to expect the number of sentences per trial to increase over the course of the test. There was a pause of 2000 ms between the sentences. After each set, a star appeared on the computer screen, which was the signal for participants that the set was finished and that they had to write down on paper the last word of each sentence they had heard in the set. After writing down the words, participants had to press a button to continue the test, which implies that the time between two series was variable between participants. All participants heard all sentences, so the test was not terminated when a student failed to remember the words of a trial. The administration lasted approximately 15 minutes, including instruction time. The scores were computed by counting the correctly remembered (written) words. Each correctly reproduced final word was credited with one point regardless of the number of sentences per set.

The Listening Span test was administered using a notebook computer and headphones.
2) Serial Recognition test

The Serial Recognition test used is an adapted version of Gathercole’s Serial Recognition test (Gathercole, Pickering, Hall, and Peaker, 2001). Participants listened to two consecutive series of nonwords. Both series comprised the same nonwords but they did or did not appear in the same order. The participants had to indicate whether the order of the nonwords was the same by ticking *gelijk* ‘same’ or *verschillend* ‘different’ on their answer sheet. The test consisted of three blocks of seven pairs of nonword strings. The first block consisted of seven pairs of four nonwords each, the second block contained seven pairs of five nonwords and the last block comprised seven pairs of six nonwords. All nonwords were phonologically legal by Dutch standards. They were all monosyllabic CVC strings and there was considerable phonological overlap within each series: only the V or the coda C varied as can be seen in (1) and (2), respectively.

(1) geep gip guup goop gep /γep/ /γip/ /γyp/ /γop/ /γep/
geep gip goop guup gep /γep/ /γip/ /γop/ /γyp/ /γep/

(2) huf hul hur hug hus /huf/ /hol/ /hor/ /hoy/ /hos/
huf hul hur hug hus /huf/ /hol/ /hor/ /hoy/ /hos/

A small pilot experiment indicated that this overlap of phonemes was necessary to increase the test’s difficulty and discriminative power. In this pilot experiment four native speakers and one second language learner of Dutch did the test without overlap in the items. The maximum error score for these participants was 1. Therefore it was decided to increase the difficulty of the test by making the items more similar. There were seven items (string pairs) in each set to avoid a gambling strategy that can occur with an equal number of items due to an (expected) equal number of yes/no responses. Response to the first pair of each block was not scored; it was considered as a practice item, participants, however, were not aware of this. This means that there were 18 experimental items and 3 control items. There was a pause of 1.5 s between the first and second series, and 750 ms between the items within a series.

The Serial Recognition test was administered group wise and took approximately ten minutes. Scores were computed by counting the correct answers. Each correct response was credited with one point, regardless of the number of nonwords in a string.
Pre-training measures

1) General Listening Comprehension

The Listening Comprehension test used in this study consists of sections of the national exam Dutch as a second language (Staatsexamen NT2). The test was, at the moment of our study, not part of the course materials of the language courses followed by the participants; participants were therefore not familiar with this test.

The Listening Comprehension test was divided into a pretest part, a posttest part and a reading comprehension test part. This division of the passages is based on the results of Pilot Study II (see Chapter 4). These results were used in a Linear Discriminant analysis (LDA) discriminating good and poor listeners and readers (see Chapter 3 section 3.4 for a description of the technique). Students that participated in Pilot Study II, were considered to be representative for the language learners participating in the training study. Therefore, we can use their scores as an indicator for the degree of difficulty of each fragment. Learners passed the test if they had a score of at least 65% correct. Table 5.2 gives an overview of the results for each speech passage.

Table 5.2: Overview of results used to categorize speech passages into different experimental tests based on the results of Pilot Study II. Scores indicate the percentage correctly categorized subjects. Standard deviation is given between parentheses.

<table>
<thead>
<tr>
<th>Passages (number of questions)</th>
<th>Discriminant analyses</th>
<th>Percent correct</th>
<th>Predicted p-values (CITO)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Listening</td>
<td>Reading</td>
<td>Listening</td>
</tr>
<tr>
<td>1 (7)</td>
<td>65</td>
<td>65</td>
<td>54 (15)</td>
</tr>
<tr>
<td>2 (10)</td>
<td>80</td>
<td>90</td>
<td>54 (21)</td>
</tr>
<tr>
<td>3 (6)</td>
<td>75</td>
<td>75</td>
<td>66 (20)</td>
</tr>
<tr>
<td>4 (6)</td>
<td>80</td>
<td>70</td>
<td>66 (27)</td>
</tr>
<tr>
<td>5 (7)</td>
<td>65</td>
<td>90</td>
<td>53 (19)</td>
</tr>
<tr>
<td>6 (4)</td>
<td>75</td>
<td>80</td>
<td>64 (22)</td>
</tr>
</tbody>
</table>

Table 5.3 shows the categorization that is based on these results. The number of questions in the pretest and the posttest is equal. Furthermore, we made sure that there was a passage in all three tests that discriminates well between good and poor listeners/readers. The balance between the pretest and the posttest is optimal.
Table 5.3: Deviation of the speech passages to the different tests.

<table>
<thead>
<tr>
<th>Fragment</th>
<th>Pretest: listening</th>
<th>Posttest: listening</th>
<th>Posttest: reading</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>6</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>

The contents of the passages of the pretest were: an interview with a Dutch singer (ten questions), an interview with a designer of Dutch money (six questions), a passage of a Dutch radio show about wearing a bicycle helmet (seven questions) and tips about the purchase of a computer (four questions).

2) Auditory Lexical Decision
The Lexical decision test that was used in the present study, is identical to the LD test in Pilot Study II. Like the items of the Listening Comprehension test the items of the Auditory Lexical Decision test were divided into a pretest part and a posttest part. This categorization was, again, based on a discriminant analysis discriminating between good and poor listeners and on the mean percent correct and mean reaction time. Both the pretest (60 items) and the posttest (100 items) contained a number of one, two and three-syllable words and nonwords. As discussed under Pilot Study I (see Chapter 3), the randomisation of the items presented may have resulted in rather long reaction times. Therefore, items were blocked in Pilot Study II such that the participants heard only overarticulated items in the first part of the test and only underarticulated stimuli in the second part (or vice versa). The blocked design was used in both the pretest and the posttest.

3) Sentence Verification test
The results of Pilot Study II (described in Chapter 4) suggested the necessity for a test with items that bridge the gap between the difference in length of the single words of the Lexical Decision test and the large speech units of the Listening Comprehension test. The items of the Sentence Verification Test meet this requirement. Participants heard a sentence of which they had to decide whether it expressed a truth or a falsehood. They were asked to answer as quickly and accurately as possible. The maximum response time allowed before time out was set to 8000 ms, starting from the onset of the sentence. The test consists of 50 items, 25 true and 25 false. All words used in the sentences were chosen from the Basiswoordenboek Nederlands.
The length of the sentences ranged from 10 to 15 syllables. An experienced male speaker of standard Dutch recorded the items in a sound insulated booth. Two examples of sentences used are given in (3); a true sentence is shown in 3 (i), while 3 (ii) shows a false sentence.

(3)  
(i)  De maand oktober komt na de maand september.  
‘The month of October follows the month of September.’

(ii)  Vlees kan je het best bij de kapper kopen.  
‘Meat can best be bought at the hairdresser’s’

There was a practice session that included sample stimuli that were different from the experimental items. Midway during the test there was a short break. The test lasted approximately ten minutes, including instruction time and practice.

By analogy to the items of the Lexical Decision test and the Listening Comprehension test, the items of the Sentence Verification test were divided into a pretest and a posttest part. The sentences were semi-randomly selected for these two versions of the test: the distribution of true and false sentences was equal in pre- and posttest. Appendix F lists all the sentences.

5.2.2.2. Test stage 2

Test stage 2 was administered three weeks after Test stage 1. In the first week of Test stage 2 there was still one training session, this session of course preceded the first test session. The tests used in this stage are presented in Figure 5.3.

![Figure 5.3: Overview of the tests used in Test stage 2.](image-url)
The test battery of this session included a Mediating test, a Control test and three posttests. The Mediating test was a Vocabulary Size test, a Reading Comprehension test was the Control test, while a Lexical Decision test, a Sentence Verification test, and a Listening Comprehension test formed the posttests (post-training dependent variables). The posttests were partly repetitions of the pretests as is shown in Table 5.4.

<table>
<thead>
<tr>
<th>Tests</th>
<th>Pretest</th>
<th>Posttest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexical Decision</td>
<td>X (60)</td>
<td>X (60) + Y (40)</td>
</tr>
<tr>
<td>Sentence Verification</td>
<td>X (30)</td>
<td>X’ (10) + Y (20)</td>
</tr>
<tr>
<td>Listening comprehension</td>
<td>X (27)</td>
<td>X’ (14) + Y (13)</td>
</tr>
<tr>
<td>Reading comprehension</td>
<td></td>
<td>X” (13)</td>
</tr>
</tbody>
</table>

The pretest part (X) of both the Lexical Decision test and the Sentence Verification test were repeated completely in the posttest, a second part (Y) was added. The pretest part of the Listening Comprehension test (X) was partly repeated in the posttest (X’), part Y was added. Yet another part of the pretest version of the Listening Comprehension test was repeated as the Reading Comprehension test (X”).

**Mediating test**

The Mediating test that was administered in Test stage 2 is the Vocabulary Size test.

**Vocabulary Size test**

This test is based on a vocabulary test developed by Hazenberg (1994). Hazenberg developed a multiple-choice vocabulary test based on word frequency in order to measure the size of participants’ L2-vocabulary. The lemmas (23,550 in the original Hazenberg & Hustijn –list (H&H)) were divided into four frequency groups (following the INL-corpus): (i) $x \geq 500$, (ii) $100 \leq x < 500$, (iii) $25 \leq x < 100$, (iv) $x < 25$. The original test of Hazenberg comprised 140 items. In this study selecting every third word of the Hazenberg list reduces this number to 50 items; the selection follows the frequency categorization of the original test. We reduced the number of

---

2 The frequencies are based on the INL-corpus linked with *Basiswoordenboek Nederlands* of Van Dale. INL stands for *Instituut voor Nederlandse Lexicografie* ‘ Institute of Dutch Lexicography’.

3 After the first selection round 47 items were selected (140/3), to add three items the selection procedure started again with the second item of the Hazenberg-list.
items because of time constraints. Each item of the test exists of a short sentence with one underlined word, the meaning of which has to be indicated. Participants can chose from four possible descriptions/definitions or indicate that they have no idea about the meaning of the underlined word. This option was included to prevent a forced choice. Due to the relative low frequency of the test items none of them can be found in *Basiswoordenboek Nederlands* (1983). However, the words that were used to describe or define the four options of the target words’ meanings were all taken from this (high-frequency word) dictionary. In this way we made sure that participants could understand the meaning of the descriptions. There were test items about nouns, verbs, adjectives and adverbs. An example of a stimulus is given in (4).

(4) *Hij spaart voor een auto*  He saves money for a car  

1) *Hij stopt voor een auto langs de weg*  1) He stops for a car along the road  
2) *Hij is onder een auto gekomen*  2) He was run over by a car  
3) *Hij verzamelt geld om een auto te kopen*  3) He collects money to buy a car  
4) *Hij verzamelt foto’s van auto’s*  4) He collects pictures of cars  
5) *Ik weet het echt niet*  5) I really do not know

Appendix G lists the items we used referring to the original item numbers of the Hazenberg test (1994).

**Control test**
The Control test that was administered was the Reading Comprehension tests.

**Reading Comprehension test**
The Reading Comprehension test was the written version of (a part of) the Listening Comprehension test that was administered in Test stage 1, before training. The purpose of this test was to check whether incorrect answers given on the Listening test were due to a problem in parsing oral input or to a problem in deriving the meaning from the correctly parsed input. If errors were caused by a comprehension problem one could expect faults on the Reading test as well. However, if listening comprehension errors were due to an intelligibility problem one would expect fewer or no incorrect answers on the Reading Comprehension test. In order to reduce the possibility of a repetition effect, the order of the alternatives as well as the order of the six speech parts was different in each test. The passages of the Reading Comprehension test consisted of an interview with a designer of Dutch money (six questions) and a passage of a Dutch radio show about wearing a
bicycle helmet (seven questions). The Reading Comprehension test was administered after the Listening Comprehension test.

**Post-training dependent measures**

1) **Listening Comprehension test**
   As discussed earlier (see also figure 4) the Listening Comprehension test of the pretest was partly repeated in the posttest, either as listening material or as reading material. The passages used as the posttest Listening test were: an interview with a town and country planner (seven questions), an interview with a Dutch singer (ten questions), a passage of a radio show about living in a house that is built on polluted soil (six questions), and a passage with recommendations about the purchase of a computer (four questions). The number of questions in the pretest and the posttest was equal.

2) **Lexical Decision test**
   The posttest version of the Lexical Decision test was the extended Lexical Decision test of the pretest; in addition to the 60 items of the pretest, 40 items were added, so the Lexical Decision posttest contained 100 items in total.

3) **Sentence Verification test**
   The posttest version of the Sentence Verification test is partly the same as the Verification test of the pretest. It consisted of 30 sentences, ten of which are repetitions of sentences in the pretest. The 20 additional sentences were constructed along the same lines as those in the pretest. As in the pretest, half of the sentences were true, half were false.

5.2.2.3. **Design of the study in summary**

In summary the following enumeration can be given:

The main measures of the study are:
   - Listening Comprehension test (pre and post)
   - Lexical Decision test (pre and post)
   - Sentence Verification test (pre and post)

The independent variable of the study, Condition, has the following three levels: Recognition, Comprehension, and Control.

Other measures used in the study are:
   - Selection tests
     - Basic Vocabulary test (stage 1)
     - Basic Grammar test (stage 1)
Mediating tests
- Vocabulary Size test (stage 2)
- Working memory: Listening Span (stage 1)
- Phonological Loop: Serial Recognition test (stage 1)

Control test
- Reading Comprehension (stage 2)

The Sentence Verification test and the Lexical Decision test have a double function in the present study. On the one hand they function as dependent measures; results on these tests can give information on the effect of the treatment. On the other hand will they be used as predictor tests in a regression approach, to investigate the predictive value of these tests for general listening comprehension.

5.2.2.4. General procedure in the test sessions
The off-line tests (Listening Comprehension, Selection Tests, the Serial Recognition test, and Control test) were administered group wise, while the on-line tests (Lexical Decision, Sentence Verification, and the Listening Span test) were run on an individual basis.

During the first meeting of Test stage 1, all participants did the Listening Comprehension test and the Basic Vocabulary test. During the second meeting, they took the Basic Grammar test and the Serial Recognition test. The individual on-line tests were performed after the first meeting of Test stage 1, but before the start of the training. Time slots of 45 minutes in total were allocated for the Lexical Decision test, the Sentence Verification test and the Listening Span test per individual. Test stage 2 took place after the training. As in Test stage 1, in Test stage 2 the off-line tests were done group wise and the on-line tests on an individual basis. At the first meeting of Test stage 2, all participants did the Listening Comprehension test and the Reading Comprehension test. Test stage 1 lasted approximately 3h30 in total, Test stage 2 approximately 2h30.

All off-line tests were conducted without special equipment. For the on-line tests notebook computers and headphones were used. The on-line tests were developed with and run by E-prime software (Psychology Software Tools http://www.pstnet.com/e-prime/e-prime.htm). E-Prime forms an experimental environment that is often used in experimental psychology (see Poelmans & de Jong, 2002 for a more detailed description of the software).
5.2.3. Training

5.2.3.1. Material

Most exercises of the training study are based on materials taken from the national exam of Dutch as a second language, level II and I. The difference between the levels II and I is twofold. First of all, the difference can be described in relation to the topics covered by the speech passages: level I deals with concrete topics, while level II consists of subjects of a more abstract nature; moreover, level II contains more contemplative and argumentative texts. A second difference can be found in the restrictions concerning the words used in the speech passages. There are no restrictions on word usage in level II whereas level I is marked by relatively high-frequency words: all words used in the multiple choice questions at level I can be found in De Kleijn & Nieuwborg’s *Basiswoordenboek Nederlands* ‘Basic dictionary of Dutch’. Additionally, in level I there are no questions about specific words unless these words are included in this basic dictionary, or unless they are explained in the speech fragment in question. Similarities between the levels are:

- at both levels, the topics of the speech passages are non-fictional
- speech passages are meant for a general audience
- the speech passages are authentic or they have the characteristics of an authentic text
- passages are informative, instructive, contemplative or argumentative
- speech passages consist of spontaneous speech or of read aloud text
- speakers do not address non-native speakers in particular.

Furthermore, receptive knowledge of Dutch dialects or sociolects is never tested. This, however, does not mean that only speakers of standard Dutch have been recorded; speakers may exhibit regional accents or sociolects (Staatsexamencommissie 1993).

Exercises that were not based on the

4 The exams we used were: 1995 I & II, 1996 I & II, 1997 I & II, 1998 I & II and 1999 I

5 An exception is made for words that cannot be found in the dictionary but can be assumed to be known, at least in a receptive way, an example is *computer* ‘computer’.

6 The specification of the materials can be found in the regulation printed in: *Staatsblad van het Koninkrijk der Nederlanden* 1993, 569, Besluit van 14 oktober 1993, Houdende vaststelling van het Staatsexamenbesluit Nederlands als tweede taal.
national exam were based on recordings of an experienced female native speaker of Dutch. The recordings were made in a soundproof booth.

The described materials were used for the training of both experimental groups, i.e., both groups were confronted with the same kind of speech, with the same level of word frequency and with the same grammatical constructions. The kind of input was similar, the amount of input and the kind of exercises, however, were not. With the PRAAT speech processing software the passages were edited into separate sentences or other listening chunks that were necessary for the exercises. In sections 5.2.3.3.1 and 5.2.3.3.2 all exercises will be described. First the training programme of the so-called Comprehension group will be presented. This group was trained specifically in higher order understanding skills; the training was focused on understanding the message of what was said. Secondly, the exercises of the Recognition group will be given. This group was trained with special emphasis on lower-order recognition skills. The training was focussed on the recognition of individual words in the speech stream, and on the recognition of Dutch phonemes.

5.2.3.2. Training procedure
Both groups participated in a seven hour training program, spread over four weeks. Because of the personal pace of participants and the different treatments (in the two experimental groups), the participants did not receive the same amount of input during the individual sessions. At the beginning of each session, participants received a book with the exercises for that particular session. During the individual sessions (three for the comprehension group, six for the recognition group), participants worked with an individual tape in a personal booth in a language laboratory during exactly one hour. Students were free to listen to the speech fragments as often as they wanted. Books of the individual sessions contained the keys to the exercises together with the training material. This way, students were able to correct their own work and to check whether they did well or not. During the group sessions (one for the Recognition group, four for the Comprehension group), students listened to speech fragments that were played to them over a central loudspeaker. Participants’ responses in these sessions were corrected and evaluated group wise. The students saw on an overhead slide the sentence on which the correct answer was based; the speech passages, however, were not repeated during these group sessions.

During all training sessions, each exercise started with both an auditory and a written instruction. This dual mode presentation minimized the chances of misunderstanding the instructions. In addition, the dual instruction functions as an extra exercise in listening comprehension; the students could check whether they recognized the spoken words by
comparing what they heard with what they saw. In the instructions, a test item is given with the correct answer and an explanation why the given answer is correct. One male and one female native speaker of Dutch recorded the auditory version of the instructions.

The first training session started with an informative letter that explained lower-order and higher-order listening strategies. Like the instructions of the training, this letter was presented orally as well as visually. One male and one female native speaker of Dutch recorded the letter. Lower-order information units that were described were:

- discrepancy between what one hears and what one sees
- phonetic processes, like assimilation, deletion, insertion and reduction.

Higher-order information units that were discussed were:

- activating schema’s
- coping with signal words
- listening attitudes
- dealing with context
- dealing with different kinds of texts.

Appendix H shows the complete text of the instruction letter.

5.2.3.3. Experimental groups
Two experimental groups participated in the training study: the Comprehension group and the Recognition group. In addition to these two experimental groups there was a Control group involved in the study. Students of this Control group participated in both test sessions but were excluded from training. In section 5.2.3.3.1 the exercises of the Comprehension group will be described. Section 5.2.3.3.2 shows an overview of the exercises of the Recognition group.

5.2.3.3.1. Comprehension group
The listening comprehension training of the Comprehension group focused on higher-order skills; participants of this group trained understanding. The method used for this group is comparable to the way listening comprehension is trained in most standard language courses. A paradigm that is used very often in this kind of training is the multiple-choice test. Students listen to a speech passage and have to answer questions concerning the passage by choosing the correct option from several possibilities. Following this standard method, five complete examples of the national
exam in Dutch as a second language were included in the training material of the Comprehension group. Students filled in – part of – these five tests divided over five training sessions, following the procedure described in section 5.2.3.2. During the two remaining training sessions, students made other exercises that focused on understanding speech. The speech passages of these exercises were also based on the national exam. The exercises of these two individual sessions will now be presented and explained. The transcription of the (Dutch) speech the subjects heard is given, including hesitations and repetitions.

**Definition exercise**

The definition exercise consists of a speech passage and two, visually presented, definitions of an expression that was used in this fragment. Participants have to choose the correct definition of the two. The expressions questioned cannot be found in *Basiswoordenboek Nederlands* (1996), so they could not be assumed to be known by the students. During the construction of the exercise it was ascertained that comprehension of all essential information is needed in order to mark the correct definition. An example is given in (5).

(5) Participants hear:

*Formeel zijn er geen maxima of minima in de regeling die bij ons van kracht is, aangegeven. Maar de gewone werktijden, die lopen eigenlijk tot zes uur ’s avonds en uren die mensen daarna zouden willen werken, dat zijn eigenlijk overuren. Dat zijn uren waar ons bedrijf dan een extra vergoeding tegenover stelt. In tegenstelling dus, tot die glijuren, die je gewoon zelf kunt compenseren en één uur, is één uur. Maar bij overuren dan is er een overwerktoeslag en afhankelijk of het in de avonduren of op zaterdag of op zondag is, wordt die toeslag kleiner of groter.*

‘Officially, no minima or maxima have been specified in the arrangement that is in effect here. But the ordinary working hours run until six PM, in fact, hours that people want to work after six are overtime. Overtime hours are hours that are paid extra by our company. This is in contrast, then, to flexible hours. People can just compensate these flexible hours themselves and one hour is one hour. But when it comes to overtime then there is an overtime bonus and how much the bonus is depends on the time the work has been done: in the evening, on a Saturday or on a Sunday.’
Participants read:

*een overuur is:*

1 *een uur dat men meer werkt dan nodig*  
   ‘an hour that one works more than necessary’
2 *een uur dat men te laat op het werk is*  
   ‘an hour that one arrives too late at work’

*Proposition exercise*

The proposition exercise consists of a speech passage and a, visually presented, proposition concerning its content. Participants have to indicate whether they think the proposition is true or false. If they think the proposition is true, they have to circle *waar* ‘true’, if they think it is false, they have to circle *niet waar* ‘not true’. Again, the propositions are constructed in a way that comprehension of the essential information is necessary in order to determine the correctness of the proposition. Propositions are never based on the recognition of specific words. An example is given in (6).

(6) Participants hear:

*U heeft waarschijnlijk wel eens gehoord van een ondernemingsplan. Wat is dat nu en waar heeft u het voor nodig? Nou, een ondernemingsplan is *eh* niets anders dan het systematisch opschrijven van de keuzes die u maakt hè, de keuze van moet ik nou kiezen voor een luxe pand dat dan minder goed bereikbaar is voor de leveranciers of moet ik dan kiezen voor een andere vestigingsplaats. Daarnaast de de financiële paragraaf, de investeringen-begroting dus wat moet ik allemaal aan gaan schaffen voordat ik überhaupt kan gaan starten. Nou die keuzes en de redenen waarom schrijft u in zo’n ondernemingsplan. En bovendien is het ook een geheugensteuntje. Want als u moet gaan uitvoeren, dan komt weer dat keuzemoment. Eh misschien niet over de vestigingsplaats maar over andere zaken. Wat hadden we toen bedacht? Oh ja, wij hebben toen gezegd dat wij willen gaan adverteren in de plaatselijke krant om maar een zijstraat te noemen. Eh dat ondernemingsplan maak je eventueel voor de gemeente, om jezelf te presenteren. Maar ’t is vooral voor jezelf. En d’r bestaan allerlei standaard ondernemingsplannen die je kunt gebruiken. Daarvoor kun je terecht bij de Kamers van Koophandel.*
'You have probably heard of a business plan. What is it and what do you need it for? Well, a business plan is nothing else than a systematic write-up of the choices that you make. The choice of, do I choose a luxury building that is not very accessible to deliverers or do I choose another business location? Another point is the the financial annex, the investments budget, so what do I have to purchase before I can start? Well, these choices and the arguments why, is what you write down in a business plan. It is also a memory support. Because when you have to perform, the moments of choices come again. Maybe not about the location of the business but about other things. What did we decide? Oh yeah, we said that we want to advertise in a local newspaper, to mention a side path. Possibly you make such a business plan for the local authorities, just to present yourself. But most important of all you make it for yourself. There are all kinds of standard business plans that you can use. You may obtain them from the Chamber of Commerce.'

Participants read:

Een ondernemingsplan is een overzicht van de gemaakte keuzes.

Waar

Niet Waar

‘A business plan is an overview of choices that were made.’

True

Not true

Summary exercise
The summary exercise consists of a speech passage and three possible written summaries. Participants have to indicate which summary is correct. The visually presented summaries differed in such a way that the essential information of the entire passage has to be understood in order to be able to choose the correct alternative. An example is given in (7).

(7) Participants hear:

Je ziet de mensen dus aankomen, dan ga je dus niet als een idioot de showroom in hollen, wat ze misschien wel elders doen, maar gewoon, je kijkt een beetje waar ze heen lopen en dan zie je vanzelf wel waar ze naar kijken. Gebruikt of nieuw. En dan ga je vragen of je ze verder kan dienen met wat informatie.
‘So you see people coming, so then you’re not going to run into the showroom like a fool, which is what they may do in other places, but, well, you just look and see where they go to and then you cannot help seeing what they are looking at. Used or new. And than you go and ask if you can help them with some information.’

Participants read:

1. *Anne vertelt dat het belangrijk is de mensen niet te laten wachten wanneer ze de showroom binnenkomen.*
   ‘Anne says that it is important that you do not let people wait when they enter the showroom.’

2. *Anne geeft de mensen eerst de tijd om even rond te kijken omdat ze vaak niet weten wat ze willen.*
   ‘Anne gives people some time to look around because they often do not know what they want.’

3. *Anne vindt dat de klanten de tijd moeten krijgen om eerst zelf rond te kijken. Als autoverkoper kan je de mensen daarna de nodige informatie geven.*
   ‘Anne thinks that people deserve some time to look around. As a salesman you can ask people after a while what they want.’

5.2.3.3.2. Recognition Group

The listening training of the Recognition group focused on lower-order skills. Participants of this group were trained in recognizing words and Dutch phonemes. The method used for this group is only partially based on existing methods; the listening part of one exam Dutch as a second language is used in one of the training sessions. During the six remaining sessions each participant practised with an individual training package (including a tape and an exercise book). Materials for these individual sessions were pretested by three native speakers, who were asked to perform the task without listening to the speech fragments. Their failing to perform the tasks correctly was taken as evidence that the construction of the exercises was successful; the aim was that it is only possible to complete the exercises if every word of the spoken sentences is recognized. The group session followed the same procedure as the group sessions of the comprehension group. During the individual sessions several exercises were made, the order of which changed over the different sessions. The exercises of the individual sessions are described below.
Choice-of-words exercise
Participants hear a sentence while reading the written version of the sentence. In the transcription of the sentence a choice between two possible words was offered. Participants had to decide which of the two possible, visually presented, words was the correct transcription of the speech fragment. Here, ‘possible’ can be defined as both semantically and syntactically correct. Half the time the two words were phonologically similar, half the time they were semantically similar. A phonologically similar example is given in (8i) and a semantically similar example is given in (8ii).

(8) Participants hear:
(i) *En dat deden ze met hun handen.*
   ‘And they did it with their hands.’

(ii) *De brandweer en politie van Valkenburg hebben in de nacht van woensdag op donderdag urenlang gezocht naar vijf toeristen die verdwaald waren in de grotten.*
   ‘The fire department and the police of Valkenburg looked for hours during the night between Wednesday and Thursday for five tourists who had gotten lost in the caves.’

Participants read:

(i) *En dat deden ze met hun handen/tanden.*
   ‘And they did it with their hands/teeth.’

(ii) *De brandweer en politie van Valkenburg hebben in de nacht van woensdag op donderdag urenlang gezocht naar vijf toeristen die verdwaald waren in de bossen/grotten.*
   ‘The fire department and the police of Valkenburg looked for hours during the night between Wednesday and Thursday for five tourists who had gotten lost in the woods/caves.’

Fill-in exercise
In the fill-in exercise participants had to write down a missing word in a written rendition of the passage. The distance between the blanks ranged from 15 to 20 words. The words to be filled in were never names or words that we expected to be unknown to the students, e.g., technical terms. The
gaps could only be filled in correctly if the students recognized all the words of the sentence. An example is given in (9).

(9) Participants hear:

_Welkom in speelgoedatelier Berend Botje. Wij maken en verkopen hier houten speelgoed. Mijn naam is Betty de Vries. Ik zal jullie eerst even iets vertellen over hoe het atelier begonnen is. Daarna krijgen jullie een rondleiding._

‘Welcome in toys studio Berend Botje. We make and sell wooden toys here. My name is Betty de Vries. I will first tell you something about the history of the studio. Then you can do a guided tour.’

Participants read:

_Welkom in speelgoedatelier Berend Botje. Wij maken en verkopen hier houten speelgoed. Mijn naam is Betty de Vries. Ik zal _______________ eerst even iets vertellen over hoe het atelier begonnen is. Daarna krijgen jullie een _______________._

‘Welcome in toys studio Berend Botje. We make and sell wooden toys here. My name is Betty de Vries. I will first tell _______________ something about the history of the studio. Then you can do a _______________.’

_Count-and-write exercise_

The count-and-write exercise forces students to recognize every single spoken word. Participants hear a sentence and they are asked to write down a certain word of the sentence, for example the third word. The words that had to be written down were neither names nor words that were expected to be unknown to the students. Target words ranged from the first till the seventh word in the sentence. An example is given in (10), where the first number is the number of the sentence; the second number indicates which word the participants should write down.

(10) Participants hear

_Bel dan naar de receptie van de fabriek._

‘Then call the reception desk of the factory.’
Participants read:
1 5 ...........................................

Participants write:
1 5 receptie

Phoneme identification exercise
The phoneme identification exercise trains students in recognizing typical Dutch phonemes. One male and one female native speaker of Dutch recorded the material used for this exercise in a sound proofed booth. They recorded target words by reading aloud a list of citation forms. The exercises were constructed after the recording; this way, the items were completely neutral as far as intonation is concerned. The items of the phoneme recognition exercises consists of minimal pairs such as man - maan ‘man - moon’. The goal of the exercises is to improve the identification of the Dutch phonemes, for example the difference between short and long vowels, which can give words a different meaning. Most of the exercises were based on the vowel distinctions between: /a/-/a/, /e/-/e/, /o/-/o/, and /i/-/i/. A few items were based on the distinction between /o/-/y/ and /o/-/ø/. There are three versions of the phoneme identification exercise:

- Participants hear two words and have to indicate whether these words are the same or not. If they think the words are the same, they have to circle gelijk ‘same’, if they think the words are not the same, they have to circle verschillend ‘different’ on their answer sheet.

- Participants hear a word and see two words on their paper. They have to circle the word they think they heard.

- Participants hear two words and they see one word on their answer sheet. They have to indicate whether they heard the word on the answer sheet as the first or as the second word by circling 1 or 2.

Examples are given in (11i), (11ii), and (11iii), respectively.

(11) Participants hear:

(i) zien zin
    ‘see’    ‘sentence’
Participants read:

(i) GELIJK  VERSCHILLEN
   ‘Same’  ‘Different’

(ii) zoon  zon
       ‘son’  ‘sun’

(iii) veel  vel
       ‘much’  ‘skin’

Listen-and-read exercise

The listen-read exercise provides students with the transcription of a speech fragment they hear. There are some discrepancies between the auditory and the visual version of the fragment. Participants have to mark the differences between the spoken and the written versions of the text by underlining the position were they notice dissimilarity. There are three kinds of differences: (i) word additions, (ii) word substitutions, and (iii) word deletions. An example is given in (12). The lines underneath three phrases show the difference between the auditory and visual fragments. They are included for clarification matters in this example; they were not present in the real exercise.

(12) Participants hear:

Ik werk op de stadsredactie van De Gelderlander. Dat is dus een redactie die aandacht besteed aan het lokale nieuws, in dit geval van de stad Nijmegen. En daar werken wij met tien mensen. En die bestrijken de verschillende terreinen wat betreft de stad. Dat kan van sport variëren tot woningbouw, tot sociale kwesties tot culturele zaken enzovoort. En ik houd me bezig met politieke verslaggeving, zaken die in de gemeenteraad spelen en sociale dingen. En dat is heel breed.

‘I am a member of the town news editorial staff of De Gelderlander. That is the editorial staff that works on the local news, in this case of the town of
Nijmegen. The staff consists of ten people who work on the different aspects of the town news. These aspects vary from sports to house building, to social affairs to cultural affairs, and so on. I work on political affairs, issues that concern the city council and social matters. It is a very large area.’

Participants read:

*Ik werk op de stadsredactie van de Gelderlander. Dat is dus een redactie die aandacht besteed aan het nieuws van de stad, in dit geval van de stad Nijmegen. En daar werken wij met zeven mensen. En die bestrijken de verschillende terreinen wat betreft de stad. Dat kan van sport variëren tot woningbouw, tot sociale kwesties tot culturele zaken enzovoort. En ik houd me bezig met verslaggeving, zaken die in de gemeenteraad spelen en sociale dingen. En dat is heel breed.*

*Word count exercise*

The word count exercise makes it necessary to recognize every single word. Participants hear a sentence and have to count the number of its words. The sentences contained high-frequency words, i.e., words used in every day speech. The exercise was constructed so that participants do not hear two successive sentences of the same speaker. Sentences ranged in length from 3 to 13 words; only 14 of the 420 sentences that occurred in this exercise, had the maximal length of 13 words. An example of the sentences used is given in (13).

(13) Participants hear:

*Zij zijn de enige krant in deze regio.*

‘They are the only local newspaper in the region.’

Participants read:

1 ...................................................................

Participants write:

1 ..............................8...................................

It should be noted that the exercises described in this subsection, were ‘exercises’ not ‘tests’. That is students were provided with the correct response. They could then listen to the sentence or passage as often as they
wanted to verify what was said word-by-word. By relistening while focusing on the recognition of every word, students were given the means of improving their skills in recognizing words in concatenated speech.

5.3. Results

The analyses given in this section are a first step in answering the research questions. In summary, the research questions were: What kind of training method shows the most positive influence on general listening comprehension: the Recognition method or the Comprehension method? Hand in hand with this main question go the following sub-questions: What is the effect of training on processing information at the word and sentence levels (measured with the Lexical Decision and Sentence Verification tests respectively)? Two additional questions concern the influence of the methods on the automatization of the word recognition process and the relation between working memory span and listening comprehension.

First, the procedure of the data selection that preceded the analyses is described (section 5.3.1), then the actual analyses are presented (section 5.3.2).

5.3.1. Selection of data

As can be seen in Table 5.1 of section 5.2.1, we collected a complete set of data from 83 participants. This means that 83 students participated in all test sessions and that all students in the Recognition and Comprehension group participated in the training sessions. For the statistical analyses, data of 64 students were selected applying the following two criteria:

1. Students had to have a score on the pretest version of the Listening Comprehension test of less than 90% correct, i.e., their score had to be less than 24/27. This criterion was used because students with a higher score would not have much room to exhibit an improvement of their listening comprehension skill. Fifteen students did not meet this criterion. Thirteen of these students had a Germanic mother language (English, German or Swedish), while the two other mother tongues involved were Croatian and Russian. Of the 15 students whose data were excluded from analysis, seven students participated in the Recognition group, six students in the Control group and two students in the Comprehension group.
Students had to meet the criterion of 60% correct on both the Basic Grammar test and on the Basic Vocabulary test. Four students did not meet this criterion; they failed either the Basic Grammar test or the Basic Vocabulary test\(^7\). All four students participated in the Comprehension group. The rationale to use the criterion, set by the developers of the test, in this study is that students with a score lower than 60% correct are deemed not to have enough knowledge to work with the training material in a satisfactory way.

Table 5.5 shows an overview of the group sizes of the three groups before and after selection.

**Table 5.5 Stepwise selection of students whose data are analysed.**

<table>
<thead>
<tr>
<th>Condition</th>
<th>N before selection</th>
<th>Failures criterion 1</th>
<th>Failures criterion 2</th>
<th>N after selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recognition</td>
<td>28</td>
<td>7</td>
<td></td>
<td>21</td>
</tr>
<tr>
<td>Comprehension</td>
<td>37</td>
<td>2</td>
<td>4</td>
<td>31</td>
</tr>
<tr>
<td>Control</td>
<td>18</td>
<td>6</td>
<td></td>
<td>12</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>83</strong></td>
<td><strong>15</strong></td>
<td><strong>4</strong></td>
<td><strong>64</strong></td>
</tr>
</tbody>
</table>

**5.3.2. Results of Test stages 1 and 2**

In this section, descriptive results of performance on the tests administered in Test stage 1 and 2 are presented. The results of the Selection tests, the Control test and the Mediating tests are given first, then the results on pre and posttests will be described and finally a comparison of these results will be made.

We present no separate analyses for the students of the different universities. Since we are interested in the differences between the training methods a comparison between the students of the two universities is not relevant. Appendix I presents a description of the off-line test scores by university.

**5.3.2.1. Selection tests**

The selection tests were included in the study to determine the language proficiency level of the participants at the onset of the study. As described in section 5.3.1, results of participants were only included in the statistical

\(^7\) Three of these students did not have an average mean score of 60% correct on these selection tests; the fourth student had a mean score of 62.5% correct taken over both selection tests but scored less than 60% on the Basic Grammar test.
analyses of the study if the score on both Selection tests was at least 60% correct. Table 5.6 shows performances on the Basic Vocabulary test and on the Basic Grammar test.

Table 5.6: Performance on the Selection tests. Standard deviation is given between parentheses.

<table>
<thead>
<tr>
<th>Condition</th>
<th>N</th>
<th>Basic Vocabulary</th>
<th>Basic Grammar</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mean (max = 60)</td>
<td>Percent correct</td>
</tr>
<tr>
<td>Recognition</td>
<td>21</td>
<td>48 (6)</td>
<td>81 (10)</td>
</tr>
<tr>
<td>Comprehension</td>
<td>31</td>
<td>47 (6)</td>
<td>78 (9)</td>
</tr>
<tr>
<td>Control</td>
<td>12</td>
<td>47 (5)</td>
<td>78 (8)</td>
</tr>
</tbody>
</table>

The differences between the groups are not significant for either of the two Selection tests as a univariate one-way ANOVA shows: Basic Vocabulary $F(2, 61) = .47, p = .63, \eta^2 = .015$, Basic Grammar $F(2, 61) = .26, p = .77, \eta^2 = .008$.

5.3.2.2. Control test

The Reading Comprehension test that was administered during Test stage 2 was included to diagnose possible intelligibility problems. The differences between the scores of the three groups (Recognition group 79% correct (SD 14, exact scores: 10/13 (SD 1.8)), Comprehension group 75% correct (SD 14, exact scores: 9.7/13 (SD 1.9)), Control group 75% correct (SD 15, exact scores: 9.8/13 (SD 2)) were not significant: Reading Comprehension $F(2, 61) = .63, p = .54, \eta^2 = .020$.

5.3.2.3. Mediating tests

As has been mentioned in section 5.2.2, the two memory tests, administered in Test stage 1, aimed to measure processing and storage capacity of linguistic information at the sentence and lexical level (Listening Span test) and at the phonological level (Serial Recognition test). The Vocabulary Size test that was administered in Test stage 2 was used to find possible individual differences in L2 vocabulary size. Table 5.7 shows the performances on the Mediating tests (mean scores and percent correct).
Table 5.7: Performance on the Mediating tests. Standard deviation is given between parentheses.

<table>
<thead>
<tr>
<th>Condition</th>
<th>N</th>
<th>Listening Span</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mean (max = 54)</td>
<td>Percentage correct</td>
<td></td>
</tr>
<tr>
<td>Recognition</td>
<td>21</td>
<td>39 (8)</td>
<td>72 (14)</td>
<td></td>
</tr>
<tr>
<td>Comprehension</td>
<td>31</td>
<td>37 (8)</td>
<td>69 (14)</td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>12</td>
<td>37 (4)</td>
<td>69 (8)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>Serial Recognition</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mean (max = 18)</td>
<td>Percentage correct</td>
<td></td>
</tr>
<tr>
<td>Recognition</td>
<td>21</td>
<td>12 (3)</td>
<td>67 (15)</td>
<td></td>
</tr>
<tr>
<td>Comprehension</td>
<td>31</td>
<td>13 (2)</td>
<td>70 (13)</td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>12</td>
<td>13 (2)</td>
<td>71 (14)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>Vocabulary Size test</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mean (max = 50)</td>
<td>Percentage correct</td>
<td></td>
</tr>
<tr>
<td>Recognition</td>
<td>21</td>
<td>31 (5)</td>
<td>62 (9)</td>
<td></td>
</tr>
<tr>
<td>Comprehension</td>
<td>31</td>
<td>29 (5)</td>
<td>58 (11)</td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>12</td>
<td>26 (7)</td>
<td>52 (14)</td>
<td></td>
</tr>
</tbody>
</table>

The only significant difference between groups concerns Vocabulary Size \( F(2, 61) = 3.5, p = .037, \eta^2 = .10 \) as a univariate one-way ANOVA shows. A Tukey posthoc test shows a significant difference between the scores of the Recognition group and the Control group whereby the Recognition group had the higher score. The other differences between conditions are not significant, with \( F < 1 \) for the other two measures.

Preliminary observation:
There are no differences between the groups at the pre-training level as the performances on the Selection tests made clear. The lack of significant differences between the groups on the Control test and the memory test also indicate that the groups are comparable in their language proficiency and memory span as far as tested here. The only difference that was found was in the performances on the Vocabulary Size test, which was administered in test Stage 2, after training.

5.3.2.4. Performances on pre- and posttests
Three tests were conducted during both test stages, i.e. before and after training: Listening Comprehension, Lexical Decision and Sentence Verification. Comparing the results of the pre- and posttests makes it possible to answer the research question concerning the effectiveness of the training programs.
The Lexical Decision test and the Sentence Verification test have two functions in this study. On the one hand, they will be used, in this section, as dependent variables to determine the effect of the training programs; on the other hand they will be used in a multiple regression analysis as predictors of general listening comprehension (section 5.3.3).

In the following subsections, performances on each test are presented. The first test that will be described is the Listening Comprehension test, followed by the on-line tests: Lexical Decision and Sentence Verification.

5.3.2.4.1. Listening Comprehension

Table 5.8 gives an overview of the mean performance on the Listening Comprehension test in both pretest and posttest versions. It also shows the difference scores (delta or Δ values), i.e. posttest performance minus pretest performance. The results are given in mean scores as well as in percents correct.

<table>
<thead>
<tr>
<th>Condition</th>
<th>N</th>
<th>Pretest</th>
<th>Posttest</th>
<th>Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mean</td>
<td>Percent</td>
<td>Mean</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(max = 27)</td>
<td>correct</td>
<td>(max = 27)</td>
</tr>
<tr>
<td>Recognition</td>
<td>21</td>
<td>17 (4)</td>
<td>63 (14)</td>
<td>19 (5)</td>
</tr>
<tr>
<td>Comprehension</td>
<td>31</td>
<td>16 (4)</td>
<td>59 (14)</td>
<td>17 (5)</td>
</tr>
<tr>
<td>Control</td>
<td>12</td>
<td>18 (3)</td>
<td>68 (11)</td>
<td>19 (4)</td>
</tr>
</tbody>
</table>

A univariate one-way ANOVA shows that the groups do not differ at the pre-training test moment (pretest): $F(2, 61) = 1.68, p = .20, \eta^2 = .052$.

As can be seen in Table 5.8 the highest delta value is for the Recognition group. A univariate ANCOVA analysis (on mean scores) with performance on the posttest as the dependent variable and performance on the pretest as covariate shows, however, that the differences between the groups are not significant: $F(2, 60) = .44, p = .65, \eta^2 = .014$. Thus, the Recognition group did not improve more than the Comprehension group, nor did the latter improve more than the former, nor indeed did the two training groups improve more than the no-training Control group. The Pearson correlation between pretest and posttest performance was $.67, p < .001 (N = 64)$.

Additional ANCOVA’s were conducted in which additional covariates were included. First, Vocabulary Size was included as a covariate in addition to
pretest Listening Comprehension. This did not result in a significant group effect \( F(2,59) = .56, p = .58, \eta^2 = .019 \). Then, the two memory capacity tests (Listening Span and Serial Recognition) were entered as covariates in addition to pretest Listening Comprehension. This did not result either in a significant group effect \( F(2,58) = .34, p = .71, \eta^2 = .012 \). Finally, all four tests mentioned were entered as covariates (pretest Listening Comprehension, Vocabulary Size, Listening Span, and Serial Recognition). Again, no significant Group effect was found \( F(2,57) = .54, p = .59, \eta^2 = .018 \). Thus, even if individual differences in vocabulary and verbal memory capacity were taken into account, the training methods did not significantly affect listening comprehension.

In summary, we can say that the effect of training on the Listening Comprehension test is not as expected. The performances on the Listening Comprehension test indicate that there are no significant between-group differences in either of the two test sessions.

### 5.3.2.4.2. Sentence Verification

In this section the results on the Sentence Verification test are presented. For the analyses offset RTs were used.

Table 5.9 presents the mean results on the Sentence Verification test.

<table>
<thead>
<tr>
<th>Condition</th>
<th>N</th>
<th>Pretest</th>
<th></th>
<th></th>
<th>Posttest</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>RT</td>
<td>Percent correct</td>
<td>RT</td>
<td>Percent correct</td>
<td></td>
</tr>
<tr>
<td>Recognition</td>
<td>21</td>
<td>1300 (499)</td>
<td>86 (5)</td>
<td>1257 (416)</td>
<td>92 (7)</td>
<td></td>
</tr>
<tr>
<td>Comprehension</td>
<td>31</td>
<td>1249 (430)</td>
<td>85 (6)</td>
<td>1183 (431)</td>
<td>91 (9)</td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>12</td>
<td>1103 (274)</td>
<td>84 (8)</td>
<td>1146 (295)</td>
<td>83 (17)</td>
<td></td>
</tr>
</tbody>
</table>

A univariate one-way ANOVA showed that there was no significant difference between the groups at the first test stage (pretest): \( F(2,61) = .82, p = .45, \eta^2 = .026 \).

Univariate ANCOVAs with performance on the posttest as the dependent variable and performance on the pretest as covariate show that the differences between the groups at Test Stage 2 (posttest) were not significant, neither for the reaction times \( F(2,60) = .72, p = .49, \eta^2 = .023 \).
nor for the percentage correct \( (F(2,60) = .94, p = .40, \eta^2 = .030) \). Breaking the data down in true versus false sentences did not result in significant differences. A significant group effect was neither found when the data were broken down in new versus overlapping data.

In summary, we can say that the effect of training on the Sentence Verification test is not as expected. The results of the Sentence Verification task indicate that there are no significant between-group differences in either of the two test sessions.

5.3.2.4.3. Auditory Lexical Decision

Between-group analyses were done to investigate the impact of the two training methods in comparison with each other. Which training resulted in the best improvement? For the analyses, offset RTs were used. Report on performance in the Lexical Decision tests is restricted to analyses showing significant effects; an exception is made for the overall results (i.e., the results not broken down by lexicality or context). Results of analyses conducted on performance on the Lexical Decision test are reported in the following order:

(i) analyses with general data
(ii) analyses per lexical condition, i.e. words versus nonwords
(iii) analyses on overlapping and new items
(iv) analyses on overarticulated versus underarticulated items;

(i) General data

The general data are given in Table 5.10. Analyses are based on clean data, i.e. data without outliers, RT’s on correct decisions only.

Table 5.10: Performance on Lexical decision pretest and posttest versions: reaction times (in ms), and percentage correct by condition. Standard deviations are given between parentheses.

<table>
<thead>
<tr>
<th>Condition</th>
<th>N</th>
<th>Pretest</th>
<th>Posttest</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>RT</td>
<td>Percent correct</td>
</tr>
<tr>
<td>Recognition</td>
<td>21</td>
<td>950</td>
<td>74 (21)</td>
</tr>
<tr>
<td>Comprehension</td>
<td>31</td>
<td>907</td>
<td>75 (8)</td>
</tr>
<tr>
<td>Control</td>
<td>12</td>
<td>820</td>
<td>81 (9)</td>
</tr>
</tbody>
</table>
Results of a univariate one-way ANOVA show that the differences between conditions are not significant in the pretest version of the test: \( RT \ F(2,61) = 1.4, p = .26, \eta^2 = .04 \), percent correct \( F(2,61) = 1.9, p = .15, \eta^2 = .06 \).

A univariate ANCOVA with the RTs on the posttest as dependent variables and the RTs of the pretest as a covariate shows that the differences in RT between the conditions are also not significant after training \( RT \ F(2,60) = 2.5, p = .09, \eta^2 = .08 \), percent correct \( F(2,60) = .24, p = .79, \eta^2 = .008 \).

(ii) Results broken down by lexicality

The results of a univariate ANCOVA show a significant effect between the conditions on the nonwords: \( RT \ F(2,60) = 6.2, p = .003, \eta^2 = .17 \). Tukey post-hoc tests show a significant difference between the Recognition group and the Control where the participants of the Control group react significantly faster than the participants of the Recognition group. The differences between percent correct were not significant, neither in the reactions to the words, nor in the reactions to the nonwords.

(iii) Overlapping versus new items

Table 5.11 presents the reactions on the overlapping and new items.

<table>
<thead>
<tr>
<th>Words</th>
<th>Condition</th>
<th>N</th>
<th>Overlapping items</th>
<th>New items</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>RT</td>
<td>% correct</td>
<td>RT</td>
<td>% correct</td>
<td>RT</td>
<td>% correct</td>
</tr>
<tr>
<td></td>
<td>Recognition</td>
<td>21</td>
<td>716 (178)</td>
<td>94 (9)</td>
<td>671 (219)</td>
<td>93 (9)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Comprehension</td>
<td>31</td>
<td>593 (149)</td>
<td>94 (10)</td>
<td>541 (124)</td>
<td>94 (5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>12</td>
<td>540 (71)</td>
<td>93 (7)</td>
<td>498 (87)</td>
<td>92 (7)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Nonwords</th>
<th>Condition</th>
<th>N</th>
<th>Overlapping items</th>
<th>New items</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>RT</td>
<td>% correct</td>
<td>RT</td>
<td>% correct</td>
<td>RT</td>
<td>% correct</td>
</tr>
<tr>
<td></td>
<td>Recognition</td>
<td>21</td>
<td>1280 (365)</td>
<td>59 (22)</td>
<td>1315 (318)</td>
<td>58 (23)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Comprehension</td>
<td>31</td>
<td>1062 (292)</td>
<td>60 (18)</td>
<td>1075 (248)</td>
<td>62 (18)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>12</td>
<td>919 (253)</td>
<td>72 (20)</td>
<td>922 (260)</td>
<td>72 (21)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The table shows that there is no large difference between reacting on overlapping or new items; the pattern between the groups remains the same. They also show that there is, however, a large difference between reacting to words and reacting to nonwords.
The results of a univariate ANCOVA show that the between-group differences in reactions are significant for the reactions to new words (RT $F(2,60) = 4.1$, $p = .022$, $\eta^2 = .12$), new nonwords (RT $F(2,60) = 6.7$, $p = .002$, $\eta^2 = .18$) and to overlapping nonwords (RT $F(2,60) = 7.3$, $p = .002$, $\eta^2 = .20$). Tukey post-hoc tests for the reactions to the overlapping nonwords showed that the Control group responded significantly faster than the Recognition group. For the reactions to the new words and the new nonwords, the post-hoc test shows a significant difference again between Recognition group and Control group, but also between Recognition group and Comprehension group; the participants of the Recognition group were slower than the participants of the other conditions. None of the differences between percentages correct were significant.

(iv) Overarticulated versus underarticulated items
Finally, the scores on overarticulated and underarticulated items are analyzed. Tables 5.12a and 5.12b show the performances on the overarticulated and underarticulated items.

Table 5.12a: Performance on overarticulated items: reaction times (ms), and percentage correct by condition. Standard deviations are given between parentheses.
Table 5.12b: Performance on underarticulated items: reaction times (ms), and percentage correct by condition. Standard deviations are given between parentheses.

<table>
<thead>
<tr>
<th>Condition</th>
<th>N</th>
<th>Overlapping items</th>
<th>New items</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RT</td>
<td>% correct</td>
<td>RT</td>
</tr>
<tr>
<td>Recognition</td>
<td>21</td>
<td>598 (130)</td>
<td>92 (16)</td>
</tr>
<tr>
<td>Comprehension</td>
<td>31</td>
<td>546 (142)</td>
<td>95 (5)</td>
</tr>
<tr>
<td>Control</td>
<td>12</td>
<td>488 (80)</td>
<td>89 (13)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Condition</th>
<th>N</th>
<th>Overlapping items</th>
<th>New items</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RT</td>
<td>% correct</td>
<td>RT</td>
</tr>
<tr>
<td>Recognition</td>
<td>21</td>
<td>1305 (321)</td>
<td>60 (22)</td>
</tr>
<tr>
<td>Comprehension</td>
<td>31</td>
<td>1036 (341)</td>
<td>60 (18)</td>
</tr>
<tr>
<td>Control</td>
<td>12</td>
<td>765 (220)</td>
<td>72 (24)</td>
</tr>
</tbody>
</table>

The figures show that there is no large difference between the reactions to overarticulated versus underarticulated items. The between-group difference in reactions to the underarticulated overlapping nonwords is significant ($F(2,60) = 4.07, p = .022, \eta^2 = .12$) as is shown by a univariate ANCOVA. Tukey posthoc test showed that participants of the Recognition group react significantly slower than participants of the Control group. The difference between the reactions to the underarticulated overlapping words shows a trend ($p = .07$).

For the differences in reactions to the new items, results of a univariate ANCOVA show significant results for the overarticulated new nonwords ($F(2,60) = 10.01, p << .001, \eta^2 = .25$), and the underarticulated new nonwords ($F(2,60) = 6.18, p = .004, \eta^2 = .17$). A post-hoc Tukey test shows again a difference between the Recognition group and the Control group and between the Recognition group and the Comprehension group for the reactions to the overarticulated new nonwords where the reactions of the Recognition group were significantly slower. This latter pattern is also found for difference in reactions to the underarticulated new nonwords.

**Summary of the results**
The between-group analyses show the following results:
There is no significant between-group difference on pretest performance.
There is a significant difference between performances on the posttest whereby the performances of the Recognition group were slower than the performances of the other groups.
There is a significant between-group difference for overlapping and new items (RT). RTs on overlapping nonwords are significantly faster for the Control group than for the Recognition group. For the reactions on the new words and the new nonwords, the Recognition group is significantly slower than the other two groups.

There is a significant difference between the performances of the Recognition group and the other two groups on both new overarticulated and underarticulated nonwords; RTs of the Recognition group were significantly slower. For the underarticulated overlapping nonwords, there was a difference between the performances of the Recognition group and the Control group whereby the Recognition group was the slowest.

In summary, we can say that the effect of the training on the Lexical Decision test is not as expected: participants of the Recognition group reacted significantly slower than the participants of the other two groups on several conditions. Furthermore, participants of the Control group performed unexpectedly well in comparison with the two experimental groups, at least on this auditory lexical decision task.

5.3.3. Correlations and regression analyses

5.3.3.1. Correlations

As described in section 5.3.2.3, the Mediating tests were included to give additional information on the results of the three dependent measures: Listening Comprehension, Lexical Decision and Sentence Verification. The potential explanatory role of the Mediating tests in performance on the dependent variables was investigated with correlational analyses. Although, the results of correlational analyses between the Mediating tests and performances on Lexical Decision and Sentence Verification are not really necessary to answer the research question we will present these analyses since they give useful information for the regression analysis.

First correlations are presented; then the results of regression analysis are given. Correlations are displayed in Tables 5.13a, 5.13b, and 5.13c, respectively. In these tables, the correlations with Reading Comprehension, which was defined as a control test, and the correlation with the pretest version of the variable are also mentioned.
Table 5.13a: Correlations between Listening Comprehension, Mediating tests and Control test (N = 64).

<table>
<thead>
<tr>
<th></th>
<th>Listening</th>
<th>Reading</th>
<th>Voc. size</th>
<th>Listening</th>
<th>Serial rec.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>post</td>
<td>comp.</td>
<td></td>
<td>span</td>
<td></td>
</tr>
<tr>
<td>Listening Pre</td>
<td>.670*</td>
<td>.424*</td>
<td>.425*</td>
<td>.013</td>
<td>.097</td>
</tr>
<tr>
<td>Listening Post</td>
<td>.502*</td>
<td>.375*</td>
<td></td>
<td>.046</td>
<td>.005</td>
</tr>
<tr>
<td>Reading Comp.</td>
<td></td>
<td></td>
<td>.478*</td>
<td>.077</td>
<td>−.023</td>
</tr>
<tr>
<td>Vocabulary Size</td>
<td>−.134</td>
<td>.183</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Listening Span</td>
<td></td>
<td></td>
<td></td>
<td>.234</td>
<td></td>
</tr>
</tbody>
</table>

* p < .001

As can be seen in Table 5.13a, there are no significant correlations between the two memory tests and the Listening Comprehension tests. There are, however, relatively high correlations between the Vocabulary size test and Listening Comprehension. This latter result is as expected as a large vocabulary size heightens the chance on success on integrative skills as reading and listening. These correlations are in line with the results of Pilot Study II.

Table 5.13b: Correlations between Lexical Decision (Reaction times (RT) and Percentage correct (%)) and Mediating tests (N = 64).

<table>
<thead>
<tr>
<th></th>
<th>% pre</th>
<th>RT post</th>
<th>% post</th>
<th>Read comp</th>
<th>Voc. size</th>
<th>List. span</th>
<th>Serial rec.</th>
</tr>
</thead>
<tbody>
<tr>
<td>RT pre</td>
<td>−.279*</td>
<td>.583*</td>
<td>−.321*</td>
<td>−.107</td>
<td>−.235</td>
<td>−.225</td>
<td>−.112</td>
</tr>
<tr>
<td>% pre</td>
<td>−.227</td>
<td>.670*</td>
<td>.156</td>
<td>.008</td>
<td>.154</td>
<td>.212</td>
<td></td>
</tr>
<tr>
<td>RT post</td>
<td>−.118</td>
<td>.097</td>
<td>.060</td>
<td>−.106</td>
<td>−.212</td>
<td></td>
<td></td>
</tr>
<tr>
<td>% post</td>
<td></td>
<td></td>
<td></td>
<td>.246*</td>
<td>.254*</td>
<td>.265*</td>
<td>.160</td>
</tr>
</tbody>
</table>

* p < .05

Note that Vocabulary Size and Reading Comprehension correlate modestly and significantly with the correct scores of the Lexical Decision test when performed the second time but not when performed the first time. The substantial correlations between the two versions of the Lexical Decision test are as expected.
Table 5.13c: Correlations between Sentence Verification and Mediating tests (N=64).

<table>
<thead>
<tr>
<th></th>
<th>% pre</th>
<th>RT pre</th>
<th>% post</th>
<th>RT post</th>
<th>Read comp</th>
<th>Voc. size</th>
<th>List. span</th>
<th>Serial rec.</th>
</tr>
</thead>
<tbody>
<tr>
<td>% pre</td>
<td>−.094</td>
<td>.703*</td>
<td>−.374*</td>
<td>.011</td>
<td>−.403*</td>
<td>−.147</td>
<td>−.203</td>
<td></td>
</tr>
<tr>
<td>% post</td>
<td>−.112</td>
<td>.149</td>
<td>−.055</td>
<td>.181</td>
<td>.233</td>
<td>.111</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* *p < .05

A remarkable result of Table 5.13c is the lack of relation between the percents correct of the Sentence Verification tests and the reaction times of these tests. The lack of relation between the percents correct of the pretest version of the Sentence Verifications task and the Vocabulary size test is also remarkable.

As stated before, the function of the Lexical Decision test and the Sentence Verification test is twofold. They function as dependent variables but also as predictors for the Listening Comprehension test. The results that go together with their function as dependent variables were given in tables a-c; Table 5.13d shows the correlation coefficients between the two on-line tests and the Listening Comprehension test.
Table 5.13d: Correlation between Listening Comprehension (LC) and on-line lexical decision (LD) and Sentence Verification (SV) tests in pretest (top panel) and posttest version (bottom).

<table>
<thead>
<tr>
<th></th>
<th>LC post</th>
<th>LD RT pre</th>
<th>LD % pre</th>
<th>SV RT pre</th>
<th>SV % pre</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listening Comp. Pre</td>
<td>.459*</td>
<td>-.095</td>
<td>.033</td>
<td>-.400*</td>
<td>.002</td>
</tr>
<tr>
<td>Listening Comp. post</td>
<td>-.079</td>
<td>.199</td>
<td>-.196</td>
<td>-.033</td>
<td></td>
</tr>
<tr>
<td>LD RT pre</td>
<td>-.279*</td>
<td>.584*</td>
<td>-.114</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LD % pre</td>
<td>-.277*</td>
<td>.093</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RT SV pre</td>
<td></td>
<td></td>
<td>-.094</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>LC post</th>
<th>LD RT post</th>
<th>LD % post</th>
<th>SV RT post</th>
<th>SV % post</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listening Comp. pre</td>
<td>.459*</td>
<td>-.291*</td>
<td>.079</td>
<td>-.407*</td>
<td>.148</td>
</tr>
<tr>
<td>Listening Comp. post</td>
<td>-.005</td>
<td>.294*</td>
<td>-.365*</td>
<td>.074</td>
<td></td>
</tr>
<tr>
<td>LD RT post</td>
<td>-.456*</td>
<td>.412*</td>
<td>-.189</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LD % post</td>
<td>-.226</td>
<td>.340*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RT SV post</td>
<td></td>
<td>-.203</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* p <.05

The significant correlations between the on-line tests (RT or percentage correct) and the Listening Comprehension task indicate that there is indeed a relation between the tests. Remarkable is the lack of relation between percents correct of the pretest version of the Sentence Verification task and the Listening Comprehension tests.

5.3.3.2. Regression analyses

Another way of investigating the relations between the dependent variables and the Mediating tests is the use of regression analyses. In our study we were not interested in a maximally efficient model, what we wanted to know is how well the results on a listening comprehension test can be predicted based on all predictors.

The results of an Entry Regression analysis are given below. Table 5.14 shows the result of a Multiple Regression analysis with the posttest version of the Listening Comprehension test as dependent variable. The analysis was done with the Mediating tests, and the other dependent variables as predictors (posttest version). Altogether the predictors explain 52% of the variance in the Listening Comprehension test.
Table 5.14: Regression analysis for variables predicting the performance on the Listening Comprehension test (posttest version).

<table>
<thead>
<tr>
<th>Predictor variable</th>
<th>β</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vocabulary size</td>
<td>.274</td>
<td>2.15</td>
<td>.036</td>
</tr>
<tr>
<td>Lexical decision RT</td>
<td>.105</td>
<td>.782</td>
<td>.438</td>
</tr>
<tr>
<td>Lexical decision % correct</td>
<td>.228</td>
<td>1.770</td>
<td>.082</td>
</tr>
<tr>
<td>Listening Span</td>
<td>−.006</td>
<td>−.050</td>
<td>.959</td>
</tr>
<tr>
<td>Serial Recognition</td>
<td>−.020</td>
<td>−.152</td>
<td>.880</td>
</tr>
<tr>
<td>Sentence Verification RT</td>
<td>−.303</td>
<td>−2.190</td>
<td>.032</td>
</tr>
<tr>
<td>Sentence Verification % correct</td>
<td>−.112</td>
<td>−.839</td>
<td>.405</td>
</tr>
</tbody>
</table>

The results shown in Table 5.14 indicate that the contributions of Reaction time of the posttest version of the Sentence Verification task and the Vocabulary Size test are substantial. Together they explain 46% of the variance.

Since the results of regression analysis with the Lexical Decision test and the Sentence Verification test as dependent variables do not contribute to answering the main research question these are not given here.

5.3.4. Coefficient of variation

As described in Chapter 2, Segalowitz (1993, 1998, 2000) uses the coefficient of variation ($CV_{RT}$, computed by dividing $SD_{RT}$ by the mean RT) as a measure of automatisation. He argues that the combination of a positive correlation between this coefficient and the mean RT and a declining $CV_{RT}$ reflect automatisation. In the present study the CV-concept is therefore adopted as our measure of automatisation. Section 5.3.4.1 describes $CV_{RT}$ for the results on the Sentence Verification test, while section 5.3.4.2 presents $CV_{RT}$ for the Lexical Decision test.

5.3.4.1. Coefficient of variation for the Sentence Verification test

Table 5.15 shows the $CV_{RT}$ of the Sentence Verification test. The correlation coefficient between $CV_{RT}$ and mean RT is given in the same table in parentheses. As in all analyses above, the offset RTs are used.
Table 5.15: CV<sub>RT</sub> of Sentence Verification by group, for both test moments, with correlation coefficient between CV<sub>RT</sub> and mean RT in parentheses.

<table>
<thead>
<tr>
<th>CV&lt;sub&gt;RT&lt;/sub&gt;</th>
<th>Comprehension</th>
<th>Recognition</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>CV&lt;sub&gt;RT&lt;/sub&gt; pretest</td>
<td>.34 (.57*)</td>
<td>.40 (.52*)</td>
<td>.24 (.28)</td>
</tr>
<tr>
<td>CV&lt;sub&gt;RT&lt;/sub&gt; posttest</td>
<td>.36 (.55*)</td>
<td>.33 (.37)</td>
<td>.24 (.39)</td>
</tr>
<tr>
<td>CV&lt;sub&gt;RT&lt;/sub&gt; overlapping items</td>
<td>.45 (.35)</td>
<td>.29 (.42)</td>
<td>.25 (.04)</td>
</tr>
</tbody>
</table>

As can be seen in Table 5.15 there is only a decline in the CV<sub>RT</sub> for the Recognition group, the difference between the pretest and posttest CV<sub>RT</sub> is, however, not very large. There is no positive significant correlation between CV<sub>RT</sub> posttest and mean RT, it would therefore be wrong to conclude that the process investigated here has become more automatised.

5.3.4.2. Coefficient of variation for the Lexical Decision test

Table 5.16 shows the CV<sub>RT</sub> of the Lexical Decision test. For the analyses given in Table 5.16, the reaction times on all conditions are used (over- and underarticulated items, words and nonwords alike)

Table 5.16: CV<sub>RT</sub> of Lexical Decision by group, for both test moments, with correlation coefficient in parentheses.

<table>
<thead>
<tr>
<th>CV&lt;sub&gt;RT&lt;/sub&gt;</th>
<th>Comprehension</th>
<th>Recognition</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>CV&lt;sub&gt;RT&lt;/sub&gt; pretest</td>
<td>.21 (.22)</td>
<td>.30 (.68*)</td>
<td>.21 (.32)</td>
</tr>
<tr>
<td>CV&lt;sub&gt;RT&lt;/sub&gt; posttest</td>
<td>.20 (.50*)</td>
<td>.20 (.43)</td>
<td>.13 (.16)</td>
</tr>
<tr>
<td>CV&lt;sub&gt;RT&lt;/sub&gt; overlapping items</td>
<td>.22 (.30)</td>
<td>.20 (.30)</td>
<td>.13 (.04)</td>
</tr>
</tbody>
</table>

Within all groups the CV<sub>RT</sub> diminishes as expected as the process is expected to be more efficient a second time the task is done. This is in this case expressed by a change in the speed of the process. However, a speedup of a process is not enough to conclude that the process became more automatised. For that a positive correlation with the mean RT is needed. As Table 5.16 shows, there are no positive significant correlations between the CV<sub>RT</sub> and the RTs in the second test session for the Recognition group and the Control group. We cannot therefore conclude that the process of word recognition became more automatised due to the training, despite a decline in CV<sub>RT</sub>. There is a significant correlation between the CV<sub>RT</sub> and the mean RT in the posttest for the Comprehension group. However, since the difference between the CV<sub>RT</sub>s of the pretest and the posttest is minimal it is
not likely that the word recognition process of participants is this group became automatised.

5.3.5. General summary of the results

In general, we can say that the effect of the training is not as straightforward as we had expected. The differences between the two experimental groups and the control group are not large and when there are differences they are often not in the expected direction. For example, the participants of the Recognition group reacted more slowly and less accurately on the posttest version of the Lexical Decision test than did the other two groups despite the fact that the former participants received a training that focused on word recognition. The results will be further interpreted and discussed in the following section.

5.4. Discussion and conclusion

Many tests were administered to answer the main research question of this study: What kind of training is more beneficial to general listening comprehension, a training that focuses more on lower-order recognition skills or a training that focuses on higher order comprehension skills? The previous section gave an overview of the results of these tests; in this section the relation between the results and the research question will be discussed. The discussion of the results on the different kinds of tests is followed by a general discussion.

Based on the performances on the Selection tests, Mediating tests, and the dependent measures of Test stage 1 (see section 5.3.3.1, section 5.3.3.2, and section 5.3.3.4 respectively), we can state that the three groups at the beginning of the study were equal; there were no significant differences between the groups on any of the tests before training. This implies that differences that are found in the performances on the three dependent measures (Listening Comprehension test, Lexical Decision test, and Sentence Verification test) at Test stage 2 were expected to reflect an effect of training.

5.4.1. Dependent Variables

Listening Comprehension test
The listening comprehension test is the first dependent measure of which the results will be discussed. This test is the most important of the dependent variables. The results in Table 5.8 suggest that the Recognition group shows
the largest improvement, with a difference between posttest and pretest results of 6% (for the other two groups the difference between posttest and pretest results was 4%). However, concluding that the Recognition method (thus training lower-order skills) has the best effect on general listening comprehension would not be legitimate since the small differences between the performances of the groups (i.e. the difference between 6% and 4%) are not significant.

In the introduction of this chapter the hypothesis was formulated that if a training effect was not found in the performances on the general listening comprehension test, an effect was expected to be found in the performances on the Sentence Verification test or at least in the performances on the Lexical Decision test.

**Sentence Verification test**

At first sight the results of the Sentence Verification task do not show a clear training effect either. As is presented in Table 5.9 in section 5.3.2.4.2, the performance between the Comprehension group, the Recognition group, and the Control group are not significantly different on the posttest version of the test.

The conclusion that can be drawn is that the results are not a clear cut answer to the question which training method has the better influence on the Sentence Verification test, even though the results of the Comprehension group seem better than those of the other groups. Yet, training in listening, independently of the method, seems to have a positive effect on the development of the listener’s sentence verification skill. The results on the Sentence Verification test (see Figure 5.9) show that the participants in the Control group generally reacted faster on the stimuli than the participants in the experimental groups, but that their percentage correct is lower than that of the two other groups. Even though the differences between the groups are not significant, one could speak of some kind of a dormant training effect. Training seems to have an effect on the accuracy of the reactions; it seems as if the control group focuses on the speed of the reaction while the participants in the two experimental groups seem more conscious of the fact that it is also important that responses are correct. During the training, the accuracy of the responses was stressed in that correcting the exercises was an important part of the training. It seems as if participants of both experimental groups did the tests more consciously (resulting in slower but more accurate reactions) than the participants of the Control group. In that case, the training can be described as a process of awareness raising.
Lexical Decision test

Again, we have to conclude that the effect of the different training methods is not unequivocal for this test. The Recognition method, which we expected to have a positive (if not the best) influence on the Lexical Decision test, seems to have almost no effect. Moreover, the results of the Recognition group seem to be the worst in almost all conditions (see for example section 5.3.2.4.3). This is remarkable since the students of the Recognition group were the only participants that were trained in lower-order skills such as word recognition, i.e., the skill that is tested in the Lexical Decision test. The results of the Control group on this test are unexpectedly good. In several conditions the participants of the Control group reacted better (faster and more accurate) than the participants of the experimental groups as the between-group comparisons show. The within-group analyses, however, indicate that an important condition, in which the participants of the experimental groups show significant improvement in contrast to the participants of the Control group, is in the reactions to underarticulated (overlapping) words. This could indicate that any training whether emphasizing lower-order or higher-order skills, does have a positive effect on the word recognition process. The students in the experimental groups participated in a training in which exercises were given that were based on fluent speech. Participants in these groups were exposed systematically to underarticulated speech; they listened to underarticulated speech units and made exercises based on these units. It suggests that working with this material (receiving training) improves the recognition of words in this speech quality.

5.4.2. Mediating tests

Several Mediating tests (Vocabulary Size test and Memory tests) were included in the assumption that the results on these tests could give additional insight on the results of the dependent variables.

The results, presented in Table 5.7, show that there was only one significant difference between the groups in the reactions on the Mediating tests, namely in the Vocabulary Size test. Participants of the Recognition group turned out to possess a somewhat larger vocabulary size than the participants of the two other groups. The difference in performances do not allow the conclusion that there was a training effect in favour of the Recognition group.
5.4.3. Automatisation

An important aspect on which the present study was based, is concerned with the issue of automatisation. The training effect of the Recognition group was expected to lead to an improvement of the automatisation of word recognition processes. To monitor the change in level of automatisation we followed the method as proposed by Segalowitz & Segalowitz (1993). As described in Chapter 2 this method is based on a decline of the Coefficient of Variability ($CV_{RT}$) together with a positive correlation of this measure with the mean reaction times. Tables 5.15 and 5.16 of the previous section (section 5.3.4.1 and section 5.3.4.2), show no stable pattern of declining $CV_{RT}$’s in combination with positive correlations. We cannot therefore conclude that training (independent of the specific method) positively affected automatisation. The result of the $CV_{RT}$ analyses does not meet our expectations.

5.4.4. General discussion

As described above we did not find a clear training effect. Before we give a few possible reasons we would like to argue that the lack of a training effect is not likely due to errors in the design, administration or scoring of the variables measured. All dependent and independent variables were measured with tests containing a sufficient number of items, exhibiting no ceiling or floor effects, without abnormal, or extremely skewed distributions. The distribution of the items in the pretest and posttest versions of the three dependent variables (Listening Comprehension, Sentence Verification, and Lexical Decision) was based on the results of Pilot study II (Chapter 4), using discriminant analyses. Furthermore, even in retrospect, we are confident of the construct validity of the tests as they were constructed following principles well established in the literature.

The most plausible explanation for not finding a training effect is the lack of overlap between the contents of the materials and exercises used in the training and the contents of the materials and tasks used in the tests of the three dependent variables. The lower-order processes that were trained in the Recognition group could not be successfully employed in the posttests as these consisted of material too much different from the materials in the training. The necessity of overlap between what is trained and what is tested is known in the literature as the ‘transfer appropriateness’ of learning (Bransford et al. 1979; Morris et al. 1977). We may have been too optimistic in assuming that the similarity between the cognitive processes that occur at the time of learning and processes that occur at the time of testing (or at the time that the learned knowledge or skill is put into practice) would be
sufficient to allow successful transfer of training. However, the linguistic materials used in the training and tests probably did not exhibit sufficient overlap. Of course, there was some overlap, as some of the words (in the context of concatenated speech) used in the training also occurred in the texts used in the Listening Comprehension posttest (although not in the same phonological and prosodic context), but the sentences, passages, and texts were different. This is true for both training conditions. Thus, the Comprehension group was equally unable as the Recognition group to transfer what they had learned to do with the materials and tasks in the training to the materials and tasks of the posttest version of the Listening Comprehension test. This lack of overlap also pertained to the other two dependent variables, Sentence Verification and Lexical Decision.

An additional reason why we did not find effects of training may lie in the conditional nature and function of word recognition skills. As we have seen in Chapter 2, most theories regard word recognition skills as a prerequisite for text comprehension in both reading and listening. Perhaps, our participants did already know a sufficient number of words and possessed sufficient skills to recognize the words they knew in concatenated speech, to free their mental capacity to process information at the higher levels of understanding the meaning expressed at the sentence, and text levels. In other words, the training in the Recognition group may have targeted the phoneme and word recognition skills not at their prerequisite levels but beyond. Thus, although it was not the case that participants in the Recognition group performed their exercises without making any errors, they may have possessed sufficient skills in recognizing a sufficient number of words before training commenced to comprehend the texts in the pretest and posttest versions of the Listening Comprehension test. However, the Comprehension group did not perform the exercises in their training program without errors either, nor did they exhibit progress in listening comprehension after training. The lack of progress in both groups, and the far from perfect performance on the Listening Comprehension test both before and after training (between 59 and 72 percent across the three groups) suggests that participants were at an intermediate level in at least two relevant dimensions: (1) knowledge of vocabulary and grammar, and (2) skill in the recognition of normal, frequent words, and perhaps as well in the third relevant dimension of the skill namely in applying higher-order comprehension strategies. Neither the recognition nor the comprehension training program changed this state of affairs essentially.

It could be argued that the lack of training effects was caused by the short duration of the training programs. It is possible that if training programs had lasted for a longer period, they would have affected listening comprehension positively. This may seem a simple and rather self-evident
reasoning. However, it raises another, more pertinent question. If training of a longer duration would positively affect listening comprehension, what element or elements of the training program would then cause this effect? There would be various options for the implementation of an extended recognition training program, ranging from repeating the tasks with the same materials implemented in the program of this study, from using other tasks with new materials. It could be argued that the inclusion of more new materials in an extended program, comprising more words and more contexts, would increase the chances of linguistic overlap with the materials used in the test used to measure the dependent variable, listening comprehension. Thus, extending the program in such a way would logically increase the transfer appropriateness of the program and hence diminish the potential gap between training and test.

One might argue that another reason for not finding a clear training effect could be the characteristics of the items themselves used in the dependent measures. Let us take as an example the words used in the Lexical Decision test. The words of the test were all highly frequent and known by the participants. One could argue that if words were used in the test (and in the training) with a lower frequency, the effects might have been clearer. However, we believe that the characteristics of high-frequency words argue in favour of their use rather than against. High-frequency words are more predictable (given a certain context) than low-frequency words (as is shown by Grosjean 1980). Moreover, Lieberman (1963) showed that words that were predictable given their context (e.g., nine in the English expression A stitch in time saves nine) were pronounced less carefully than were the unpredictable words (e.g., nine in The next number is nine). Jurafsky et al. (2001) also found evidence in this direction. Their results in favour of the Probabilistic Reduction Hypothesis stated that word forms (of functions words as well as of content words) are more reduced as they have a higher probability. These results lead to the conclusion that high-frequency words are pronounced in a more sloppy way than low-frequency words. Besides being pronounced less carefully, high-frequency words are also spoken faster than low-frequency words (Aylett & Turk 2002). These two features of the speech, speed and speech quality (underarticulation), are two characteristics that make the recognition of words more difficult, they make it hard to recognise words as is also shown in the results on the Lexical Decision test of this training study (see section 5.3.2.4.3). Learning to recognise high-frequency words, which by definition are used very often in daily language, is therefore more efficient for language learners than focusing on low-frequency words that are pronounced more carefully and more slowly.
A final remark concerns the results of the Control group. On several tests in the present study the results of this group were significantly better than the results of the two experimental groups. A reason for this could be the difference in consciousness between the Control group and the experimental groups while executing the task. It seems that the participants of the Control group, for example in reacting to the stimuli of the Lexical Decision test, focused mainly on speed, while the participants of the experimental groups subscribed more to the importance of the accuracy of a decision. It should be noted that the Control group was not a real no-treatment group in that the participants of this group like these of the experimental groups received L2 speech input during their classes and everyday live. However, the Control is a real no-treatment group in that the participants of this group did not participate in an intensive specific training.

To summarize this discussion, we argue that the most likely reason why neither the training of word recognition skills nor the training of higher-order comprehension strategies had a positive impact on listening comprehension of intermediate L2 learners is probably due to the fact that, although the training programs could be regarded as transfer appropriate in terms of processes, they were not transfer appropriate in terms of linguistic content, as there was hardly any overlap between linguistic elements in the texts used in the training and the linguistic elements in the texts used in the test. The most fruitful perspective for future research appears to be to focus on the truly conditional role of lower-order and higher-order skills. In the present study, lower-order and higher-order skills may have played a rather uncritical role leaving room for forms of compensating knowledge and skills. A study of the conditional role of either kind of skills would require a more rigorous experimental manipulation of materials and tasks in both training and test.