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

Citation for published version (APA):
4.1. Introduction

In second language research and second language education it is often necessary to determine the language level of L2-learners. Usually, off-line tests are used for this purpose. As described in Chapter 1 section 1.4.1, these can be defined as tests that allow participants to think before giving a response (e.g., a grammar test). Off-line tests do not give information about the processes underlying a language skill; they only indicate the language user’s general command of the target language. If, for example, a language learner fails a listening comprehension test, which is an off-line test, one cannot say whether this is due to a lack of language knowledge or to insufficient automatisation of this knowledge. To create a more detailed profile of the language user it is necessary to use on-line tests. These can be defined as time-critical tests that require the participants to respond while the ongoing stimulus is still being processed. The performance on these tests reflects the status of the processes underlying a language skill (Rietveld & Van Heuven 2001).

The present pilot study demonstrates the necessity of on-line tests as a complement to off-line tests in compiling a detailed profile of a language learner. We investigated the extent to which performance on listening comprehension tests of adult learners of Dutch as a second language can be explained by knowledge tests on the one hand and tests that measure the status of word recognition on the other. The research question was:

Is there empirical evidence that one can subdivide second language learners into three categories (Table 4.1)?
Table 4.1: Listening comprehension as predicted by language knowledge and speed of recognition.

<table>
<thead>
<tr>
<th>Status of word recognition process</th>
<th>Knowledge of the language</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good</td>
<td>Good (category #1)</td>
</tr>
<tr>
<td>Good</td>
<td>Good (category #1)</td>
</tr>
<tr>
<td>Poor</td>
<td>Poor (category #2)</td>
</tr>
<tr>
<td>Poor</td>
<td>Poor (category #3)</td>
</tr>
</tbody>
</table>

Learners in category 1 have a good knowledge of the language, and, if necessary, are able to use their knowledge under time pressure. These students can be assumed to have a near native command of the L2. Learners in category 3 have poor results on the knowledge tests. By definition, automatisation of knowledge is therefore impossible. Learners in this category can be seen as beginners. Finally, learners in category 2 have sufficient language knowledge, as learners in category 1, but they lack the aural word recognition skills necessary to successfully use their lexical/grammatical knowledge within the real-time constraints of the on-line listening tasks. We therefore predict that they have a poor listening comprehension skill. We assume that the listening comprehension skills of learners in this category could be improved by intensive training; therefore category 2 is the crucial category in our study. The possibility to divide language learners into strict categories based on language knowledge and language-processing skills implies that besides an overview of the learners’ abilities also an overview of the learners’ shortcomings can be made. As a result of this detailed language profile a language training method can be developed to fill in the needs of the learner to improve the general language command. The results of the present study can then be used as a precedent for the selection of students who will participate in the training study (described in Chapter 5).

As described in Chapter 2 section 2.5 one of the individual’s characteristics that might influence the performance on language tasks is the memory span. We therefore included in the present study a memory test. The results of this test were expected to allow us to create a language profile of the participants, not only based on language specific tasks, but also based on the individual characteristics of the language learners.

In summary, this pilot study examined whether it is possible to divide adult learners of Dutch as a second language into discrete categories, or into profiles based on language knowledge and word recognition status. A second aim of the study is to demonstrate the added value of on-line tasks in compiling a detailed language profile.
4.2. Method

In this section a detailed description of the materials we used is given, as well as a report of the procedure we followed.

4.2.1. Overview of tests

In order to answer the research question a number of tests were administered. As is described in the introduction, it is possible that in addition to knowledge of the language (measured by off-line tests) and the time that is necessary to activate that knowledge (measured by on-line tests), individual characteristics, like memory span, can influence people’s performance. We therefore included a memory test in the test battery. An overview of the tests used in the study is given in Figure 4.1.

\[\text{Test battery} \quad \text{Off-line tests} \quad \text{On-line tests} \quad \text{Memory test}\]

- Knowledge tests
- Skill tests
- Visual LD
- Auditory LD
- Reading Span

- Basic Vocabulary
- Listening Comprehension
- Basic Grammar
- Reading Comprehension

*Figure 4.1: Overview of the tests used in the study.*

The tests we used can be divided into three categories: Off-line tests, On-line tests and a Memory test. The tests are described in detail in the following sections.

4.2.1.1. Off-line tests

The first category that will be described is the Off-line tests. As can be seen in Figure 4.1, this category can be split into Knowledge tests and Skill tests. A description of these two kinds of language test is given below.

Knowledge tests
Both the Basic Vocabulary test and the Basic Grammar test are developed at the department NT2 (Dutch as a second language) of the Vrije Universiteit
of Amsterdam, in order to determine the language level of the students with respect to the vocabulary and grammar knowledge of Dutch.

**Basic Vocabulary test**
This test aims to examine receptive knowledge of 2000 highly frequent words of Dutch. Each item of this 60-item multiple choice test consists of a sentence from which one word, the target word, is omitted, followed by four words, namely three distracters and the omitted word. The distracters are existing Dutch words belonging to the same word class as the omitted word; they are not used as the correct word in any of the subsequent items. The target words as well as the distracters of this test are chosen from *Basiswoordenboek Nederlands* (1983). This dictionary is an earlier version of *Basiswoordenboek Nederlands* (1996) that is described in section 3.2.3.2 of the previous chapter. All target words in the test are sampled from the 1,000 least frequent words of this dictionary (based on Hazenberg, 1994) subdivided into three frequency groups of equal size according to the following proportional division: ten nouns, three adjectives, five verbs, one adverb, and one item from another word class. An example of an item is given in (1):

(1) **Dutch**
- *Ga je vanavond met me naar de film?*  
- *Nee, ik ……………….. lekker thuis.*

**English**
- *Would you like to go to the movies tonight?*  
- *No, I’ll be ………………………at home.*

A) begin  
B) blijf  
C) praat  
D) zoek

The item in English is not an exact translation of the Dutch version, but it gives an idea of the item’s characteristic features. This remark holds for all other glosses in this thesis.

**Basic Grammar test**
This 53-item multiple-choice test aims to assess participants’ receptive knowledge of basic, frequent structures in Dutch morphology and syntax.

---

1 In the edition of 1996, 18 lemmas of the original dictionary were removed, while 110 lemmas were added. These new lemmas were terms that had become highly frequent since the first publication of the dictionary (in 1983), e.g., *computer* 'computer'.
The test follows the structure of the Basic Vocabulary test. Each item consists of a sentence lacking one word, and offers four response alternatives, i.e., three distracters and the correct response. There are six items on the use of adjectives, four on the use of determiners, twelve on the use of pronouns, four on the use of specific particles (e.g., er ‘there’, daar ‘over there’, hier ‘here’), four on the use of negation, twelve on the conjugation of verbs and eleven on word order. An example of an item is given in (2):

(2) Dutch
Kijk eens:
Ik…………. mijn nieuwe schoenen aan.
Vind je ze mooi?

A) hebt
B) hebben
C) heeft
D) heb

English
Look:
I .................... my new shoes.
Do you like them?

A) weared
B) wears
C) wore
D) am wearing

Procedure and grading
Both the Basic Vocabulary test and the Basic Grammar test were group-administered as paper-and-pencil tests during a Dutch class. Students had to complete each test in one hour; in practice, each test lasted approximately 40 minutes. Results were computed by crediting each correct answer with one point.

Skill tests
The skill tests comprise parts of the national exam Dutch as a second language (Staatsexamen NT2). Every year, there is a release of a test similar to the annual official national exam NT2. These releases consist of parts taken from this official exam. The parts we used in this study were released as Voorbeeldexamen NT2 II 1999. At the time of testing, the parts were not public and participants could therefore not have been familiar with this test. This way a repetition-effect could be excluded. The listening comprehension part of the official exam focuses on general listening comprehension, specific listening and selective listening. We used a speech editor (PRAAT, Boersma & Weenink (1996)) in order to create a test similar to previous editions of the national exam

Listening Comprehension test
The Listening Comprehension test used in this study consists of 6 parts each with a different subject. There are two interviews, three sections of a radio
show and one part with recommendations concerning the purchase of a computer. Each part consists of several speech fragments with multiple-choice-questions (40 questions in total) and is introduced by a voice-over. At the beginning of the test, the participants were told to read the question carefully before listening to the passage. There was enough time between units to answer the question and to read the following question.

**Reading Comprehension test**
The Reading Comprehension test was the written version of the Listening Comprehension test. Our purpose with this test was to check whether incorrect answers given on the listening test were due to an intelligibility problem or to a comprehension problem. If errors were caused by a comprehension problem one could expect faults on the reading task as well. However, if listening comprehension errors were due to an intelligibility problem one would expect fewer or no incorrect answers in the reading mode. In order to reduce the possibility of a repetition effect, the order of the response alternatives as well as the order of the test parts was different in both modalities.

**Procedure and grading**
Participants filled out both Skill tests during Dutch classes. The order of testing was fixed; all students first performed the Listening Comprehension test before filling out the Reading Comprehension test. Each test lasted approximately one hour. There was an interval of three days between the administration of the Listening Comprehension and the Reading Comprehension tests. Results were obtained by crediting each correct answer with one point.

**Apparatus**
The listening speech materials for the comprehension test were played using an ordinary cassette-recorder with internal speakers. No further special equipment was involved in the administration of the tests.

**4.2.1.2. On-line tests**
An auditory and a visual lexical decision test were run to get an impression of the individuals’ degree of automaticity of the processes underlying reading and listening. In this section both decision tasks are described.

**Auditory Lexical Decision task**
The items used in this experiment were the same as those used in Pilot study I (see Chapter 3): these were words and nonwords in an overarticulated and an underarticulated version. In Pilot study I the appearance of the stimuli of
the different conditions was randomized, as described in the discussion of that study, it is possible that the rather long reaction times were due to this randomization. For this reason the items were blocked according to the speech quality in the present study, meaning that the participants heard only overarticulated speech in the first part of the session and only underarticulated items in the second part, or vice versa.

**Visual Lexical Decision task**
The stimulus material of this test was made up of 50 real Dutch words and 50 Dutch-like nonwords. Like the words in the auditory experiment, the words used in the visual experiment were chosen from *Basiswoordenboek Nederlands* (1996). These were monosyllabic words as well as two and three-syllabic words. The number of 1, 2 and 3-syllable stimuli was (almost) equal to the distribution used in the auditory test. There was no overlap between auditory and visual items. The stimulus list is given in Appendix D.

**Procedure**
The procedure of the Auditory Lexical Decision test was the same as described in Pilot Study I (see Chapter 3 section 3.2.4). In the Visual Lexical Decision test, participants also responded in a computer-controlled lexical decision task. On each trial, participants had to decide whether the item they saw on the computer screen was or was not an existing Dutch word by pressing a green button and a red button respectively. The item remained on the screen until the participant responded but there was a four seconds time-out. Before the experiment started there was a training session that included sample stimuli of both experimental conditions (either word or nonword). Halfway through the test there was a short break. Both the auditory and the visual lexical decision test lasted approximately 15 minutes each, including instruction time and practice.

**Apparatus**
Both lexical decision experiments were conducted using a Silicon Graphics Iris Indigo workstation and a 17"-screen. The visual items appeared in 24-point Lucida Typewriter font. Participants sat right in front of the computer screen, the distance between the participant’s eyes and the screen was approximately 50 cm.

**4.2.1.3. Memory test**
In the present study we used a Reading Span task to measure the memory span of the participants.
Material
The test, which is an adapted version of the test described by Daneman & Carpenter (1980), is composed of 54 unrelated sentences, each 10 to 13 syllables in length. The sentences are divided into four trials; each trial consists of three sets of three, four, five or six sentences. Each sentence ends in a different content word, chosen from Basiswoordenboek Nederlands (1996); there was no overlap between the words of this test and the word items in the on-line lexical decision tests. The sentence-final words are three to five characters long; the number of characters was equal across set, for example the total amount of characters for each set in the three-sentence trial was 11. An example of a three-sentence trial is given in (5):

(5) De bloemen staan in een grote gele vaas.
‘There are flowers in a big yellow vase.’

Ik vind het heerlijk om te slapen in de zon.
‘I like sleeping in the sun’

De politie kent de naam van de dief.
‘The police know the name of the thief.’

Procedure and grading
The Reading Span task was administered on an individual basis. Participants were told to read aloud the visually presented sentence and to remember the last word of each sentence; looking at the example given in (5) the words that had to be remembered: vaas, ‘vase’, zon, ‘sun’, dief, ‘thief’. Participants were warned to expect the number of sentences to increase during the course of the test. After the participants finished reading, the test administrator pressed a button; the read sentence disappeared and a new sentence appeared on the screen. Once a trial ended, the word TEST appeared on the screen. This was the signal for the students that the presentation of the items had ended and that they had to write down (on paper) the last words of the sentences they just read. After writing down the remembered words, the instructor pressed a button in order to continue the test; the time between two trials was therefore variable across participants. All participants read all sentences; the test was not terminated when a student failed to remember the words of a trial. The experiment lasted approximately 15 minutes, including instruction time. The scores were computed by crediting each correctly remembered (written) word with one point, regardless of the order in which they were produced.
Apparatus
The Reading Span task was administered using a Silicon Graphics Iris Indigo workstation and a 17” screen. As before, the items appeared in 24 Lucida Typewriter font.

4.2.2. Participants
Twenty native-speakers of Dutch and 20 L2-learners of Dutch participated in the on-line tests and the memory test. The 20 L2-learners also participated in the off-line tests, as did 5 native speakers. These 5 native speakers scored 100% correct on these off-line tests; therefore the decision was made to generalize this result to the group of native speakers instead of using real data. The native speakers were all students at Leiden University. The non-native speakers were students of VASVU, Amsterdam. Students from abroad who want to study at the Vrije Universiteit have to take a preparatory year, in order to attain the same level as the Dutch students in the field of, for example, mathematics, English and Dutch. The institute that organizes these courses is called Voorbereidend jaar Anderstalige Studenten Vrije Universiteit ‘Preparatory Year for International Students Free University’ (VASVU). The L2 participants’ mother tongue was never a Germanic language. Their length of stay in the Netherlands at the time of testing ranged from five months to five years. Participants were paid for their participation.

4.2.3. General Procedure
As far as the non-native participants are concerned, all tests were administered at the Free University Amsterdam, in a period of two weeks. The off-line tests were run groups-wise during class hours. The on-line tests as well as the memory test were administered on an individual basis. The individual sessions were run in a quiet room at Free University Amsterdam. The windows of this room were clad and there was artificial light.

4.3. Results
In this section the results of all tests in the test-battery are given. First the performance on the off-line tests are described, then follows a detailed overview of the on-line test data. At the end of the section, the correlations between performances on the different tests are presented. The analyses given in this section are based on the results of all participants; no data were excluded because of high error rates. Since we were interested in a possible
categorization of L2-learners, focus in this section is on the results of these participants. The results presented in this section are based on clean data (as defined in Chapter 3 section 3.2.5.1).

### 4.3.1. Off-line test scores

The results of the native speakers are not presented, the sub-group of the native speakers that filled out the off-line tests had a maximal score of 100%; their results were generalized.

The scores on the Knowledge tests of the L2-speakers are rather high: Basic Vocabulary 80% correct (SD 10) and Basic Grammar 81% correct (SD 10). This indicates that overall, the knowledge of basic grammar and basic vocabulary of all participants is good. There was a large difference between the performances on the skill tests: Listening comprehension 58% correct (SD 13), Reading comprehension 71% correct (SD 14). The difference between these two tests is significant as shows the result of a Paired-Samples t-test $t(19) = -3.77$, $p = .001$. This suggests that the problems of non-native speakers in performing the Listening comprehension test are due to an intelligibility problem rather than a comprehension problem. The participants do comprehend the input when they are presented in the visual mode, while they have problems with processing the parts when these are presented auditorily. An interesting detail is that 32% of the errors on the reading comprehension test were exclusive errors, i.e., errors not made in the auditory mode but only in the visual mode. This means that some correctly answered questions of the listening task were incorrectly answered in the reading task. Table 4.2 gives an overview of the error scores of both skill tests.

**Table 4.2: Overview of L2 error scores on the skill tests.**

<table>
<thead>
<tr>
<th></th>
<th>Listening Comprehension</th>
<th>Reading Comprehension</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Overlapping errors</strong></td>
<td>141 (35%)</td>
<td>141 (35%)</td>
</tr>
<tr>
<td><strong>Mode Exclusive errors</strong></td>
<td>209 (53%)</td>
<td>66 (17%)</td>
</tr>
<tr>
<td><strong>Total number of errors</strong></td>
<td>350 (88%)</td>
<td>207 (52%)</td>
</tr>
<tr>
<td><strong>Total numbers of items x subjects</strong></td>
<td>400 (100%)</td>
<td>400 (100%)</td>
</tr>
</tbody>
</table>

### 4.3.2. On-line test scores

Table 4.3 shows the mean performance of the native speakers and the non-native speakers on the Auditory and Visual lexical decision test. It also
shows the difference scores (delta or $\Delta$ values). The results are based on onset analyses.

**Table 4.3:** Mean performance: percentages correct and RT (ms) for native speakers ($N = 20$) and non-native speakers ($N = 20$). Standard deviation is given between parentheses.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Non-native speakers</th>
<th>Native speakers</th>
<th>$\Delta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Auditory LD</td>
<td>74 % (8)</td>
<td>93 % (5)</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>1426 ms (290)</td>
<td>1203 ms (383)</td>
<td>223</td>
</tr>
<tr>
<td>Visual LD</td>
<td>85 % (11)</td>
<td>98 % (2)</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>1436 ms (246)</td>
<td>1103 ms (319)</td>
<td>333</td>
</tr>
<tr>
<td>$\Delta$ %</td>
<td>−11</td>
<td>−5</td>
<td></td>
</tr>
<tr>
<td>ms</td>
<td>−10</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

Paired-Samples t-tests show that the difference in reaction times between the scores on the on-line tasks (visual vs. auditory lexical decision) is not significant for the non-native speakers, $t(19) = −.13, p = .89$, nor for the native speakers, $t(19) = 2.09, p = .05$. The difference between the percentages correct, however, are significant: non-native speakers $t(19) = −4.61, p = <.01$, native speakers $t(19) = −4.39, p = <.01$. The differences in the mean scores between the two experimental groups are all significant, as is shown by Paired-Samples t-tests. The results of these analyses are given in Table 4.4:

**Table 4.4:** Overview of differences between L1 and L2 speakers.

<table>
<thead>
<tr>
<th>Condition</th>
<th>df</th>
<th>RT</th>
<th>p</th>
<th>% correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Auditory LD</td>
<td>19</td>
<td>−2.80</td>
<td>&lt;.05</td>
<td>10.37</td>
</tr>
<tr>
<td>Visual LD</td>
<td>19</td>
<td>−4.54</td>
<td>&lt;.01</td>
<td>4.72</td>
</tr>
</tbody>
</table>

The significant difference between native and non-native speakers is, of course, as expected. Moreover, the results confirm the results of Pilot Study I: the status of the word recognition process of non-native speakers is not at the same level as is the process of native speakers.

The results of the lexical decision tests will now be described in more detail.

**4.3.2.1. Auditory lexical decision**

As described in the previous chapter, the results on on-line tasks are expressed in terms of reaction times and percentage correct. It was stated
that the reaction times could be measured from the onset of the stimulus (onset analyses) as well as from the offset of the stimulus (offset analyses). The question now is which of these possible analyses can best be used in further dealing with the test scores. A significant correlation between onset reaction times and offset reaction times of this experiment \((r = .96, p < .01)\) suggests that there will be no difference between these analyses. However, the correlation between length of the items and the onset and offset RTs suggests the opposite, as shown in Table 4.5.

*Table 4.5: Correlation between item length and different kinds of RTs.*

<table>
<thead>
<tr>
<th>Length of item</th>
<th>Onset RT</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Onset RT</td>
<td>.02</td>
<td>.22</td>
</tr>
<tr>
<td>Offset RT</td>
<td>−.11</td>
<td>&lt; .01</td>
</tr>
</tbody>
</table>

The correlations of Table 4.5 suggest that offset analyses will make a more strict distinction between the test-results. Therefore only analyses based on offset RTs are given in the following sections that concern auditory lexical decision.

**Between-group results (offset analyses)**

The results of Paired-Samples t-tests show only significant differences between native and non-native speakers in the reactions to the nonwords. The differences between the two experimental groups are significant for both the overarticulated and the underarticulated nonwords. Reacting to words is as difficult or as easy for native speakers as it is for the non-native speakers; no significant difference is found between both groups in the reactions to the words. Table IV.6 gives a detailed overview of the t-tests results.

*Table 4.6: Effect of language (L1 vs. L2) broken down by condition.*

<table>
<thead>
<tr>
<th>Condition</th>
<th>RT</th>
<th>df</th>
<th>t</th>
<th>p</th>
<th>df</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Words</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overarticulated words</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Underarticulated word</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nonwords</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overarticulated nonwords</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Underarticulated nonword</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Within-group results (offset analyses)
The difference between reaction times to words and nonwords was significant for the non-native speakers \(t(19) = -12.08, p < .01\), as well as for the native speakers \(t(19) = -7.59, p = <.01\). The difference between percentages correct on words and nonwords were also significant for both groups: native speakers \(t(19) = 3.04, p = .007\), non-native speakers \(t(19) = 11.08, p < .01\).

The results of the between-group and within-group analyses are in line with the results of Pilot Study I; again the strictest difference between native and non-native speakers is found in the reactions to the nonwords. Native speakers reacted 657 ms faster to these items than did non-native speakers (L1 881 ms and L2 1538 ms respectively). Appendix E gives an overview of the results broken down by conditions.

Automatisation criterion
In Chapter 3 a criterion was presented to distinguish the fully automatised (auditory) word recognition process of native speakers from the not completely automatised process of non-native speakers. Extended analyses on the data of the experiment described in Chapter 3 indicated that the best condition to distinct between L1 and L2 speakers was the one and two syllable overarticulated nonwords. The distribution of the mean RTs and the mean percentages correct of the participants on this condition showed an almost perfect distinction. Figure 4.3 shows the distribution of the individual’s mean results on the condition that was defined as the best separator.
As in Figure 3.3 of Chapter 3 it can be seen that there are a few participants that are categorized in an unexpected category: one native speaker reacted in the speed dimension as a non-native speakers and one non-native speaker react in both the speed and the accuracy dimension as native speakers. Based on the available information about these two participants nothing extraordinary could be detected. The non-native speaker was a 46 years old male L2-speaker from Ukraine, at the time of testing he lived 4 years in The Netherlands.

The application of the criterion developed on the data of the present experiment subscribes the validity of the criterion; the performances on one and two syllable overarticulated nonwords is a stable criterion to distinguish fully automatised L1 from not completely automatised L2 processes in the auditory mode. Figure 4.4 shows that the criterion is also useful in the visual mode.
Figure 4.4 indicates that the distribution of the mean RTs and the mean percentages correct of the participants on the nonwords showed an almost perfect distinction. This subscribes to the validity of the criterion.

**Presentation of the items**

The items and the procedure of this auditory LD experiment were similar to those of Pilot I, the order of presentation of the items, however, was not. In Pilot Study I the items of the various conditions were presented in random order; all conditions were mixed. In the present study the presentation was more structured: the participants heard first all overarticulated items and then all underarticulated items, or vice versa. The reason for this change was the rather long reaction times of Pilot Study I. The results of a Univariate One-way ANOVA of the LD results of the two pilot studies shows that the difference between the percentages correct of Pilot Study I and Pilot Study II (77% \(-\) 74% respectively) is not significant for the non-native speakers ($F (1, 42) = .69, p = .41$). The difference is, however, significant for the
native speakers (89% – 93%): $F(1, 42) = 4.84, p = .033, \eta^2 = .10$). The differences between the reaction times is significant for both group: non-native speakers (1275 ms – 1426 ms) $F(1, 42) = 5.59, p = .023, \eta^2 = .12$; native speakers (1009 ms – 1203 ms) $F(1, 42) = 5.18, p = .028, \eta^2 = .11$.

The suggestion made in the discussion section of Pilot Study 1, that the reaction times would probably be faster if the items were blocked by speech quality turned out not to be correct. The comparison of the results shows that the participants reacted significantly faster on the mixed version than on the blocked version of the experiment. The remark has to be made that the participants of the experiments were not the same, they were however comparable.

4.3.2.2. Visual lexical decision (onset-analyses)

In this section the results on the visual lexical decision task are given. As expected the native speakers have a better mean score (higher percent correct, faster RT) on the visual lexical decision task than the non-native speakers. Mean scores for the Dutch participants were 1105 ms (SD 319), 98% (SD 2) correct. For the L2-learners mean scores were 1436 ms (SD 242), 85% (SD 11) correct. Table IV.7 shows the mean scores for both experimental groups broken down by lexicality. It also shows the difference scores (delta or $\Delta$ values).

<table>
<thead>
<tr>
<th>Words</th>
<th>Nonwords</th>
<th>$\Delta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>RT</td>
<td>% correct</td>
<td>RT</td>
</tr>
<tr>
<td>L1</td>
<td>1002 (299)</td>
<td>99 (2)</td>
</tr>
<tr>
<td>L2</td>
<td>1130 (206)</td>
<td>97 (3)</td>
</tr>
<tr>
<td>$\Delta$</td>
<td>125</td>
<td>2</td>
</tr>
</tbody>
</table>

The results presented in Table 4.7 show that all differences between the reactions of both groups are significant except for the reaction times on the words. Both groups react – statistically – equally fast to these items but since the difference in percentage correct between the groups is significant we may conclude that the performance of the native speakers is better than the performance of the non-native speakers. Again the largest difference between native speakers and non-native speakers is found in the reactions to the nonwords.
Table 4.7 also shows that, the differences between percentage correct and RT on words and non-words are significant for the native speakers as well as for the non-native speakers.

4.3.3. Memory test

The native speakers had a score of 77 % (SD 9) correctly remembered words on the Reading Span task; the non-native speakers had a score of 63% (SD 14) correct. The difference between these scores is significant, as the result of a One-way ANOVA shows: F (1, 38) = 13.95, p = <.01, $\eta^2 = .27$.

4.3.4. Relations between the tests

The present pilot study was set up to investigate a possible relation between language knowledge, status of the word recognition process and general listening comprehension skills. We tried to find out whether a relation between these aspects can have a predictive character; in other words, is it possible to predict the level of general listening comprehension from the results of knowledge tests and on-line tests assessing the status of the auditory word recognition process.

Section 4.3.4.1 describes the relation between the auditory and visual tests, off-line as well as on-line. Section 4.3.4.2 gives the correlations between all tests that were used in the present study.

4.3.4.1. Results on auditory versus visual tests

The visual version of the listening comprehension test (reading test) was included in order to distinguish between intelligibility and comprehension. In addition, the visual lexical decision task was included to investigate the modality-specific characteristics of the word recognition process. Table 4.8 shows the mean results on the visual and auditory tests, both off-line and on-line. The table also gives information about the status of the differences between the two modes. The analyses are based on the results of the non-native speakers only.

Table 4.8: Mode comparisons in word recognition (Word rec.) and text comprehension tests (Comp.). Reaction times in milliseconds (RT) and percentage correct (%). Results of non-native speakers only. Standard deviation is given between parentheses.
A Paired-Samples t-test showed that the difference between the Reading Comprehension test (visual) and the Listening Comprehension test (auditory) is significant: $t(19) = -3.77$, $p = < .01$. The difference in percentages correct between the auditory and visual word recognition test is also significant: $t(19) = -4.61$, $p < .01$, the difference in reaction time, however, is not: $t(19) = -.130$ (ns.).

The significant results support the idea that the problems non-native speakers have with processing of auditory items are due to intelligibility problems rather than to comprehension problems, since their scores on the tests in the visual mode are significantly better than the scores on the auditory tests. Yet another argument for this idea is the lack of correlation between the RT scores on the visual and auditory lexical decision tests for the non-native speakers ($r = -.095$). The scores of the native speakers, whose recognition process is assumed to be fully automatised in both the visual as the auditory mode, do show a significant correlation ($r = .83$, $p = .001$). The lack of a correlation for the non-native speakers indicates that there is a difference in efficiency of processing for these participants.

4.3.4.2. Correlations between the tests used in the study
Table 4.9 shows the results of correlation analyses. The analyses are based on the test scores of the non-native speakers.
Table 4.9: *Inter test correlations based on the results of the non-native speakers only.*

<table>
<thead>
<tr>
<th></th>
<th>A RT</th>
<th>V %</th>
<th>V RT</th>
<th>R.S.</th>
<th>R.C.</th>
<th>L.C.</th>
<th>Gr.</th>
<th>Voc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audit. LD</td>
<td>−.16</td>
<td>.70*</td>
<td>−.12</td>
<td>.16</td>
<td>.21</td>
<td>.02</td>
<td>.29</td>
<td>.55*</td>
</tr>
<tr>
<td>% correct</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Audit. LD RT</td>
<td>−.03</td>
<td>−.10</td>
<td>0.05</td>
<td>−.26</td>
<td>.29</td>
<td>.08</td>
<td>−.09</td>
<td></td>
</tr>
<tr>
<td>Visual LD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% correct</td>
<td>.27</td>
<td>−.09</td>
<td>.25</td>
<td>.26</td>
<td>.34</td>
<td>.70*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Visual LD RT</td>
<td>−.06</td>
<td>−.06</td>
<td>0.08</td>
<td>−.23</td>
<td>.08</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reading span</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>−.24</td>
</tr>
<tr>
<td>Reading comp.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>−.17</td>
</tr>
<tr>
<td>Listening comp.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>−.20</td>
</tr>
<tr>
<td>Basic Grammar</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>−.36</td>
</tr>
<tr>
<td>Basic Vocabulary</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

As can be seen in Table 4.9 the correlation between the Knowledge tests (Basic Grammar and Basic Vocabulary) is significant. Both knowledge tests correlate significantly with the Skill tests (Listening Comprehension and Reading Comprehension), expressing the expected relation between knowledge and skill. The more knowledge one has, the higher is the chance of a good skill. The relation between Basic Vocabulary and the accuracy of the lexical decision tasks indicates that the knowledge of words and the vocabulary size have a positive influence on the word recognition process, again this is as expected. The lack of relation between the Skill tests amongst themselves strengthens the idea that errors on the Listening Comprehension task are not caused by a comprehension problem but by a listening problem. The lack of significant correlations between the on-line tests and the Listening Comprehension test is remarkable. We expected at least a positive correlation between the results on the Auditory lexical decision test and the Listening Comprehension test because the recognition of spoken words is an important sub-process of listening.
4.3.5. Categorization of the participants

As has been described earlier, the goal of this second pilot study was to investigate the predictability of learners’ listening comprehension skills on the basis of their knowledge of the language and the speed of the aural word recognition process. If it is possible to predict the level of listening comprehension on the basis of the measurements mentioned, than it should be possible to categorize language learners into fixed categories as displayed in Table 4.1 at the beginning of this chapter.

Table 4.10 shows the attempt to categorize the L2-learners that participated in the present study. For the Knowledge tests and the Skill tests the pass scores set by the developers of the tests, were used. This means that the participants needed to have a score of at least 75% correct on the Basic Grammar test and at least 70% correct on the Basic Vocabulary test to fall into the Good category. For the Listening Comprehension test the norm was 65% correct. The norms of the on-line tests were based on the mean scores of the L2-participants. This resulted in a criterion of a mean reaction time < 1426 ms, and at least 74% correct. When the results did not make clear at once in which category the participant belonged (for example because the criterion of the grammar test was not reached but the results did meet the criteria of the vocabulary test), the categorization was based on the relation of the result that did not meet the criterion, to the other test results. For example, a participant who did not meet the reaction time criterion but did meet all other criteria was categorized in category 1, unless the reaction time of the participant was very different in comparison with the criterion (e.g., 1570 ms). In Table 4.10 a and b are included for explanatory reasons, see section 4.4.1.

Table 4.10: Distribution of the 20 L2 participants in terms of listening comprehension level, language knowledge and speed and accuracy of word recognition.

<table>
<thead>
<tr>
<th>Status of Word recognition process</th>
<th>Knowledge of the language</th>
<th>Good</th>
<th>Poor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Listening good</td>
<td></td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Listening poor</td>
<td></td>
<td>6 a</td>
<td></td>
</tr>
<tr>
<td>Poor</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Listening good</td>
<td></td>
<td>4 b</td>
<td></td>
</tr>
<tr>
<td>Listening poor</td>
<td></td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Listening good</td>
<td></td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Listening poor</td>
<td></td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>
Table 4.10 shows that the categories as proposed in Table 4.1 in the introduction of this chapter are too strict; ten participants behaved unexpectedly. Six of them had sufficient language knowledge and the status of their word recognition process did also meet the criterion, but they had a bad listening comprehension skill. Four students also had enough language knowledge but despite that, the status of their word recognition process was poor while their listening comprehension skill did meet the criterion.

4.4. Discussion

The results described in the previous section show that the native speakers had better overall scores than the L2-speakers (see for example section 4.3.2). This is, of course, as expected. However, there were a few unexpected results, which will be discussed in this section.

4.4.1 Unexpected results

One remarkable result is that the difference between the performance on the Listening and Reading Comprehension task was not due to a bad performance on the knowledge tests or on the auditory decision test, as the scores on these tests were fairly high. The idea of the categorisation was that a lack of language knowledge stands in the way of a good result on the Listening Comprehension test. If a language learner does not have a basic level of L2-word knowledge, it would be practically impossible to understand an L2-speaker; if the L2-learner has enough knowledge of the target language but is not able to use this knowledge automatically (as expressed here by scores on the lexical decision tasks) a good performance on the general listening task would also be impossible. Nevertheless, the results show that despite a good language level and good performance on the on-line tests, performance on the listening comprehension task was rather poor. One possible explanation for this unpredicted finding might be an intelligibility problem, a view supported by the significant difference between the results on the listening and reading task, the significant differences between the visual and auditory lexical decision test, and the lack of correlation between the two decision tests (see section 4.3.3.1). However, the relatively large number of new errors made on the Reading comprehension test, which might be the result of a gamble strategy in the Listening comprehension test, does not allow us to conclude that a listening problem rather than a comprehension problem was the sole cause of the result.
4.4.2 Interpretation of inter-test relations

The correlation table (Table 4.9) shows some anticipated effects, for example, the correlation between the Knowledge tests, and the correlation between the Basic Vocabulary test and the lexical decision tests. Nonetheless, there is also a strange lack of some other anticipated effects. The most striking example of an unexpected result is the lack of a correlation between the performance on the Reading Span test and the comprehension tests since the literature (cf. Daneman & Carpenter 1980), would lead to expect such an effect in the present study. The reason for this lack may have been pronunciation difficulties experienced by the participants. The first step in the memory test, reading aloud the sentences, was not always taken without problems; participants sometimes hesitated to pronounce the words or repeated (i.e., corrected the pronunciation of) a word. It is possible that the focus on pronouncing the sentences took so much cognitive capacity that there was no capacity left for remembering the last word of the sentence. Using a recorded version of the test can solve this pronunciation problem. Therefore, in the training study we will use an aural version of the test. The significant difference between native and non-native speakers in the memory test suggests that verbal memory span is a language dependent construct. It seems to be that native speakers, with fully automatised language processes in the target language, are more capable of remembering the items than the non-native speakers, with not fully automatised language processes in the target language. These results and the idea that memory span is language dependent, or rather affected by language proficiency, confirm the results of a study done by Service et al (2002). She also found an effect of language skill on the L2 working memory span, indicating that the higher the proficiency in a language the better the results on a working memory task are. To find a more pure measure for the memory span, and thus minimise the language-effect, a specific language neutral task will be included in the training study (see Chapter 5).

Another remarkable result is the lack of relation between the performance on the Auditory Lexical decision test and the Listening Comprehension test. Since the recognition of words is a significant part of listening comprehension, we expected that when L2-learners show good performance on an auditory lexical decision test, their scores on a listening comprehension test would also be good. Unfortunately, we did not find a significant correlation between the scores of both tests. This unexpected lack of relation might be due to the rather large differences between the items of the tests (words in the LD test versus speech fragments in the listening comprehension test). It is possible that the L2-learners were able to process the single words of the lexical decision test, but that they had problems with processing the larger speech units of the listening comprehension test. The
opposite of this scenario is also plausible; it is possible that the L2-learners were able to score rather high on the listening comprehension test by making optimal use of the context, but failed the lexical decision test because they were not able to understand the individual items of this test. This means that not only item-length but also the contextual information of the items can form an explanation; it is not necessary for the L2-learners to understand every single word of the speech stream since the context makes it possible to sufficiently comprehend the speech without understanding each single word. Another difference between the Listening Comprehension test and the Auditory lexical decision test is the tested cognitive skill. A listening comprehension test is focused on comprehension (higher order processing), whereas a lexical decision test is focused on intelligibility (lower order processing). As stated above, it is possible to comprehend a speech fragment without understanding every single word, but it is not possible to react correctly to a lexical decision task without recognizing the individual item. The differences between the items of a lexical decision task and the items of listening comprehension test on the one hand and the tested skills on the other can be bridged by a Sentence Verification task. The form of the items in such a verification task, which are sentences, lies midway between the items of the LD experiment and the speech parts of the comprehension task. To be able to answer the question about the sentences in the test correctly, it is necessary for the participants to recognize the individual words and to understand the complete sentence. A Sentence Verification task will therefore be included in the test battery of the training study (see Chapter 5).

4.4.3 In summary

The goal of Pilot study II was to find out whether level of L2 listening comprehension could be predicted with scores on knowledge tests in combination with status of word recognition. Using these scores and the level of listening skill, L2-learners were divided into categories. Table 4.10 shows that it is indeed possible to categorize L2-learners on the basis of on-line and off-line test scores. This table, however, also shows that the categories used in Table 4.1 were too strict; ten L2-learners – all with a good knowledge of the language – had to be divided into new categories. Despite fast word recognition, six of these participants (category a) showed a poor result on the listening comprehension task. The other four participants showed slow word recognition, however, they did pass the listening comprehension test (category b). A different use of bottom-up and top-down strategies can be the cause of this sub-categorisation. The six L2-participants of category a are not able to process larger speech units while they are able
to process single words. The opposite is true for the four participants of category $b$.

The fact that sub-categorisation was needed in order to allocate all participants, gives an idea about the necessity of the combination of on-line and off-line tasks in order to create a correct profile of the language learners. Without the use of the Auditory Lexical Decision task one could easily draw the wrong conclusion that the participants of category $a$ are poor L2-listeners since they did not pass the (off-line) Listening Comprehension test. In a similar vein, one could wrongly conclude that the participants of $b$ are good listeners. The results of the (on-line) lexical decision task nuance these conclusions; they make clear that participants of $a$, in contrast to those of $b$, are able to process single words. In other words, reaction times and percentages correct on the on-line tasks make it possible to make rather subtle differentiations between the language skills of L2-learners resulting in a detailed language proficiency profile.