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 3
Pilot study I:
An automatisation criterion

3.1. Introduction

As described in Chapter 2 section 2.6, there are cognitive processes that are completely automatised while other processes are not. In that section the process of L1 word recognition was given as an example of an automatised process. It is assumed that word recognition is fully automatised in one’s mother tongue but not in a second language. Intensive practice and prolonged exposure can, however, improve the automatisation of the L2 word recognition process. Improvement does not simply mean speeding up the process (i.e., quantitative change) but restructuring the underlying components (i.e., qualitative change) (Segalowitz & Segalowitz 1993).

If the L2 word recognition process can become more automatised the question is whether it is possible for the L2 word recognition process to become fully automated (near-native), or whether there will always be a distinct difference in degree of automatisation between L1 and L2 speakers. In other words, will the distributions of scores obtained by L1 and L2 participants on score performance criterion overlap (Figure 3.1a), or will the two score distributions be strictly separated (Figure 3.1b)? For expository reasons we use speed as an indicator of automatisation on the figures below.

We are aware of the fact that this is a simplification of the situation.

![Fig. 3.1a Overlapping populations.](image1)

![Fig. 3.1b Non-overlapping populations.](image2)
A perfect separation of the L1 and L2 distributions of the scores, as visualized in Figure 3.1b, would mean that the levels of automatisation are qualitatively different. This would indicate that the process of L2 word recognition could become more automatised but that no L2-learner will ever reach the native level; in this case there is a ceiling effect (Cook 1997). If there is no overlap we will assume that the boundary between automatised and non-automatised word recognition lies mid-way between the fastest and most accurate L2 learner and the slowest and least accurate L1 speaker. However, overlap of the distributions, as displayed in Figure 3.1a, would indicate that the word recognition process of (highly proficient) L2-learners can become as automated as that of native speakers of the target language. In that case, finding the boundary is a less straightforward matter. One could claim that, again, the boundary lies at the lowest individual mean of the native speakers (the non-native speakers whose process is automatised are now included in the automatised area). But the overlap might also be due to the fast reactions of some non-native speakers only. It is possible that they responded very fast, but made a lot of errors. In this case, claiming that the word recognition process of the L2 speakers is automated, would be mistaken: it is, after all, possible that the process is simply speeded up. This subscribes to the statement that using speed of processing as an indicator of automatisation is a simplification.

What is therefore needed is a criterion to decide whether word recognition in L2 listeners is or is not automated. The present study describes the development of such a criterion. The experimental task we used was auditory lexical decision (LD). The participants had to decide, as fast and accurately as possible, whether the string of sounds they heard is or is not a real word. We expected that short reaction times combined with accurate decisions indicate an automatised word recognition process. In summary, the research question of the present study can be formulated as follows: is there a criterion to distinguish between automatised and non-automatised processes?

3.2. The experiment

In this section the auditory lexical decision experiment is described. First the paradigm used is explained, in section 3.2.2 the various conditions are described, and section 3.2.3 gives an overview of the method of the experiment. Section 3.2.4 describes the procedure and in section 3.2.5 the results are given.
3.2.1. The Lexical Decision (LD) paradigm

Several experimental paradigms to investigate the process of (spoken) word recognition such as gating, priming, and lexical decision can be used. These paradigms have in common that they produce results that reflect the cognitive processes involved in word recognition. A technique that is very often used is the lexical decision task\textsuperscript{1}; this is an on-line task. The assumption is that the results obtained by this kind of task are closely linked with the processing at the level that is of interest to the investigator (Zwitserlood 1989). On-line tasks are supposed to tap into the cognitive processes while they are still operating; crucial is that participants have to react under time-pressure. This is in contrast with the working of off-line tasks. Results obtained with the latter kind of tasks reflect the result of processes run without time constraints. In these kind of tasks participants can think before giving a response. Off-line tasks will be described in more detail in Chapter 4.

In this study the lexical decision paradigm is used to study the process of spoken word recognition. McCusker, Hooley-Wilcox and Hilliger introduced the auditory lexical decision task in 1979 (Goldinger 1997); before this time only the visual lexical decision task was used. During the auditory LD task, spoken words and nonwords are presented in random order. Participants have to indicate whether the string of sounds they hear is or is not a real word of the language under investigation, by pressing response buttons as quickly and accurately as possible. The delay between the onset of the stimuli and the response is measured; the result of this measure is what is called the decision latency or reaction time. Since the participants have to make a decision about the lexicality of the items, it can be said that the responses obtained reflect the auditory word recognition process. The idea is that an affirmative response is based on a positive match between the stimulus and the mental lexicon, while a no-response is the result of the lack of a representation of the stimulus in the lexicon.

3.2.2. Experimental conditions

The criterion we are looking for must be able to separate fully automatised from not completely automatised processes. Keeping in mind the distinction between L1 and L2 word recognition, this means that the criterion separates native and non-native speakers. It is reasonable to think that reacting to carefully pronounced stimuli would be easy for both populations, therefore the criterion to distinguish between the word recognition process of native

\textsuperscript{1} See Grosjean and Frauenfelder (1997) for a description of other techniques.
and non-native speakers can probably not be found in the reactions to these items. For this reason we introduced an additional contrast between stimuli, namely the distinction between stimuli from overarticulated versus underarticulated speech (cf. Lindblom 1990):

(i) overarticulated speech, i.e., words carefully pronounced in isolation

(ii) underarticulated speech, i.e., the same words electronically excerpted from fast and sloppy speech (produced by the same speaker)

It is quite common for native speakers to hear sloppy spoken speech, since people do not speak slowly and clearly in everyday life. Native speakers know the rules that allow the reconstruction of the intended forms that are distorted by, for example, elision, reduction and assimilation. Non-native speakers do not know these rules. Even though they know how to reconstruct sloppy speech in their mother tongue, they are not capable to do this in the L2 unless they are highly proficient in this language. Therefore, the expectation is that it will be more difficult for L2 speakers to react to underarticulated words than it is for the native speakers. The idea is that if recognition of underarticulated speech takes disproportionally longer for L2 learners than for L1 learners then a true sign of complete automatisation will be the (near-) absence of an effect of overarticulated versus underarticulated speech. Furthermore, one can assume there will be more errors in the responses of L2 speakers than in the L1 responses.

3.2.3. Method

In this section the method of the pilot study is described. The first subsection gives information about the participants of the study; the second subsection describes the material that is used. Section 3.2.3.3 and section 3.2.3.4 describe the apparatus and the procedure of the experiment.

3.2.3.1. Participants

Twenty-five native speakers of Dutch and 25 speakers of non-Germanic languages participated in this experiment. Most of the native speakers were students at the Universiteit Leiden. The non-native speakers were students of the Dutch Studies department at Universiteit Leiden. This department gives the opportunity to students from abroad to get a master’s degree in Dutch as a second language. The non-native speakers ranged in Dutch fluency from beginner to almost fluent. Non-natives with different proficiency levels were
used to ascertain that even the word recognition of highly skilled L2 learners is not fully automated, as our hypothesis predicts. Their length of stay in the Netherlands ranged from six months to eight years. The participants were paid for their participation.

3.2.3.2. Stimulus materials

The words used in this experiment were chosen from Basiswoordenboek Nederlands (Kleijn, de & Nieuwborg 1996). This dictionary, which contains a list of 2,143 highly frequent words of the Dutch language, is often used by developers of teaching materials for Dutch L2 courses and by Dutch L2 language teachers. It can therefore be assumed to be representative for the vocabulary knowledge that language learners are assumed to have if they reach a basic language proficiency level.²

In the experiment there were 50 real Dutch words and 50 Dutch-like nonwords. The real Dutch words were all mono-morphemic nouns; the nonwords were made following the morphological and phonological rules of the Dutch language. The stimuli consisted of one, two and three-syllable items (e.g., words: knie ‘knee’, lepel ‘spoon’, uitdrukking ‘expression’ respectively, nonwords: kloes, pagel, oeregien). The number of syllables was matched across words and nonwords. Appendix A lists the test-items.

The stimuli were recorded in a sound insulated booth by a male speaker of standard Dutch, experienced in doing this type of task. The speech was recorded on a DAT tape via an external microphone. The native speaker of Dutch recorded all stimuli once in overarticulated speech and a second time in underarticulated speech. The overarticulated version of the (non)words was recorded by reading aloud a list of citation forms. For the underarticulated version, the speaker read aloud sentences that contained the target word. Every target item (e.g., sleutel ‘key’, pagel) occurred in a sentence such that it was not accented. Two examples of the sentences recorded are given below, small caps indicate accented words, and the bold items are the target items:

\[
\text{Ik heb de \textbf{sleutel} van het \textit{HUIS} nodig, niet van de \textit{GARAGE}.}
\]

‘I need the \textbf{key} of the \textit{HOUSE}, not of the \textit{GARAGE}.’

\[
\text{Is dat de \textbf{pagel} van \textit{JAN} of van \textit{PIET}?}
\]

‘Is she the \textbf{pagel} of \textit{JAN} or of \textit{PIET}?’

The underarticulated items were cut out of the recorded sentences using the Praat speech processing software (Boersma & Weenink, 1996).

The mean duration of the words was 434 ms; mean duration of the nonwords was 424 ms. Appendix B shows the mean duration of the items broken down by lexicality, context and length. Stimuli were down sampled from 48 kHz to 16 kHz, 16 bits.

3.2.3.3. Apparatus
The experiment was run in a sound insulated booth, on a Silicon Graphics machine, using headphones. Colours marked two buttons on an ordinary keyboard as the two answer buttons; green marked the Yes-button (j-key), red marked the No-button (f-key).

3.2.4. Procedure
Participants responded in a computer-controlled auditory lexical decision task; the stimuli were presented on-line over headphones. On each trial, the participants had to decide whether the stimulus they heard was or was not an existing Dutch word by pressing a green ‘yes, it’s a word’ or a red ‘no it’s not a word’ button, as fast and as accurately as possible. There was a four seconds time limit for responding. Participants were tested one at a time in a sound insulated booth, in which the computer was placed. Before the experiment started, there was a practise session that included sample stimuli of the four experimental conditions. These practice items, of course, were different from the items offered in the experiment. The four conditions (overarticulated words, overarticulated nonwords, underarticulated words and underarticulated nonwords) were blocked such that if participants heard the overarticulated version of an item they did not hear the underarticulated version of the same item and vice versa. Participants never knew whether the next item would be overarticulated or underarticulated speech, the use of words and nonwords was, of course, also randomised. Halfway the session there was a short break. The experiment lasted approximately 15 minutes, including instruction time and practice.

After the lexical decision test, the non-native speakers filled in a vocabulary list that contained all experimental words. The participants had to translate the Dutch words into English; if they did not know the English term they were allowed to give an equivalent in their mother tongue. Results of this translation task confirmed our assumption that the L2 learners knew all the words.

---

3 J.J.A Pacilly at the phonetics laboratory of Universiteit Leiden developed the software for presentation and data collection.
3.2.5. Results

In this section the results of the experiment are given. The first sub-section describes the outlier procedure that is followed in this thesis, section 3.2.5.2 gives the within-group results.

3.2.5.1. Outlier procedure

Dealing with reaction time data means that one has to deal with outliers. Outliers are data that do not fall into the general flow. It is assumed that these extreme data do not reflect the process that the investigator targets. They can be caused by various factors such as fast guesses, guesses based on the participant’s failure to reach a decision, or the inattention of the participants (Ratcliff 1993). Because of their disruptive effect on results, outliers should be eliminated. However, what makes dealing with outliers so difficult is the overlap that often exists between outlier RTs and RTs that reflect the process under investigation. Several techniques to cope with outliers can be found in the literature; examples are elimination (i.e., throwing away the outliers), or using substitutions (e.g., replacing the outliers by the mean group result).

The method that we used for dealing with outliers in the analysis of our experiments in this thesis uses two criteria. The first criterion is that valid RTs have to be longer than 300 ms. Reaction times below 300 ms can be regarded as fast guesses or mistakes, i.e., as outliers. This criterion is based on the fact that the mean duration of a spoken syllable is 250 ms (the mean duration of a spoken word is approximately 500 ms). In view of the fact that time is needed to press the button after the recognition of the stimulus or after taking the decision that the stimulus is not a real word, it is not possible for a RT below 300 ms to reflect a complete recognition procedure. The second criterion is that RTs have to lie within the range of 2 standard deviations from both the participant and the item means in a particular condition. Reaction times that not meet these criteria are considered as outliers. Since the number of outliers will be relatively small – given the criterion of 2 SD – outliers will be eliminated from further analysis. No substitution will be attempted.

The results that are described in the following sections of this thesis are based on cleaned data. The outlier procedure described is the second step in the cleaning procedure. The first step is the removing of incorrect responses (i.e., no-responses on real words and yes-responses on nonwords). The incorrect responses are thus removed from the data set before the criteria of the outlier procedure are applied.

The RT performance in all four conditions is presented. Reactions to nonwords can give additional information above the reactions to words. Since the word recognition process is assumed to be fully automatised for
native speakers and not for the non-native speakers, a difference in reactions to nonwords can be expected. Furthermore, it is possible that the criterion to make a distinction between the L1 and L2 processes can be found in the reactions on the nonwords.

3.2.5.2. Mean results: within-groups and between-groups

The results of 48 participants were analysed; one native speaker and one L2-speaker were excluded because of the large number of errors they made (the native speaker had only 9% correct, the non-native speaker had only 53% correct). Overall, the L2-learners were slower to react and they made more errors than the native speakers did.

The mean reaction time of the native speakers was 1009 ms (SD 152) and their mean percentage correct was 89 % (SD 6), while the mean RT of the non-native speakers was 1275 ms (SD 218) and their mean percentage correct was 77 % (SD 10). Table 3.1 shows the mean results of the native and non-native speakers broken down by lexicality. It also shows the difference scores (delta or \( \Delta \) values), i.e., reactions on words minus reactions on nonwords.

Table 3.1: Mean results (reaction times in ms and percentages correct) of native and non-native speakers broken down by lexicality (SD). Differences (\( \Delta \)) that are significant are marked by ‘*’.

<table>
<thead>
<tr>
<th></th>
<th>L1</th>
<th></th>
<th>L2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RT</td>
<td>% correct</td>
<td>RT</td>
<td>% correct</td>
</tr>
<tr>
<td>Words</td>
<td>928 (148)</td>
<td>91 (6)</td>
<td>1087 (174)</td>
<td>86 (9)</td>
</tr>
<tr>
<td>Nonwords</td>
<td>1106 (189)</td>
<td>88 (13)</td>
<td>1567 (363)</td>
<td>67 (19)</td>
</tr>
<tr>
<td>( \Delta )</td>
<td>-178*</td>
<td>3 (22)</td>
<td>-480*</td>
<td>19*</td>
</tr>
</tbody>
</table>

As can be seen in Table 3.1 the results of Paired-samples t-tests show that the difference in percentage correct between words and nonwords is only significant for the non-native speakers, while the difference in RT between words and nonwords is significant for native speakers as well as for non-native speakers. The results are based on onset analyses.

Table 3.2 presents the mean results (onset analyses) of the native and non-native group broken down by lexicality and speech quality.
Table 3.2: Mean results (reaction times in ms and percentage correct) for L1 and L2 speakers broken down by lexicality and speech quality (SD). Differences (Δ) that are significant are marked by ‘*’.

<table>
<thead>
<tr>
<th></th>
<th>Overarticulated</th>
<th>Underarticulated</th>
<th>Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RT</td>
<td>% cor.</td>
<td>RT</td>
</tr>
<tr>
<td>L1</td>
<td>906 (127)</td>
<td>99 (3)</td>
<td>954 (190)</td>
</tr>
<tr>
<td>L2</td>
<td>1047 (183)</td>
<td>97 (6)</td>
<td>1134 (181)</td>
</tr>
<tr>
<td>Δ</td>
<td>−141*</td>
<td>2</td>
<td>−180*</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Overarticulated</th>
<th>Underarticulated</th>
<th>Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RT</td>
<td>% cor.</td>
<td>RT</td>
</tr>
<tr>
<td>L1</td>
<td>1136 (195)</td>
<td>89 (16)</td>
<td>1088 (218)</td>
</tr>
<tr>
<td>L2</td>
<td>1701 (369)</td>
<td>63 (21)</td>
<td>1453 (364)</td>
</tr>
<tr>
<td>Δ</td>
<td>−565*</td>
<td>26*</td>
<td>−365*</td>
</tr>
</tbody>
</table>

As Table 3.2 indicates that the only differences that are not significant are the differences between percentage correct on the overarticulated and underarticulated nonwords within groups, and the difference in percentage correct on the overarticulated words between-groups. The largest effect of language is found in the reactions to overarticulated nonwords; this holds for the reaction times as well as for the percentages correct.

A table with the results broken down by lexicality, context and length is given in Appendix C.

3.3. The coefficient of variation as a measure of automatisation

The aim of this pilot study was to find a criterion that is capable of distinguishing between fully automatised and non-automatised processes. Keeping in mind the assumption that the process of word recognition happens automatically in one’s mother tongue but not in a second language, one can conclude that the criterion must be able to separate native speakers from non-native speakers.

As described in Chapter 2, Segalowitz and Segalowitz (1993) suggest the use of the Coefficient of Variability – in combination with the correlation between this CV_{RT} and the mean RT – as a measure of automatisation. The results of the application of this measure to the data obtained in this experiment are given in Table 3.3.
Table 3.3: Results of the application of the CV<sub>RT</sub> measure for automisation. Mean reaction time (for correct decisions on words and nonwords only), standard deviation, coefficient of variation, and the correlation coefficient between mean and CV are broken down for native (N=24) and non-native listeners (N=24).

<table>
<thead>
<tr>
<th></th>
<th>Mean RT</th>
<th>SD</th>
<th>CV&lt;sub&gt;RT&lt;/sub&gt;</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>Native speakers</td>
<td>1009</td>
<td>152</td>
<td>.35</td>
<td>.28 (p = .177)</td>
</tr>
<tr>
<td>Non-native speakers</td>
<td>1275</td>
<td>218</td>
<td>.23</td>
<td>.23 (p = .276)</td>
</tr>
</tbody>
</table>

The differences between the two CV<sub>RT</sub> s is significant as was shown by a Independent Samples t-test: \( t(23) = 6.34, p < .01 \).

As can be seen in Table 3.3, the correlations between Mean RT and CV<sub>RT</sub> are not significant. In other words, although the native listeners are some 270 ms faster in their lexical decisions, the standard deviations are reduced in direct proportion to the mean. It is not the case that faster respondents are characterized by progressively shorter standard deviations, which is expected if the word recognition process is automatised. The application of the CV measure, therefore, does not reflect the idea – at least not in the present case – that the process of word recognition is automatised in one’s mother tongue but not in a second language.

Segalowitz and Segalowitz (1993), however, introduced a number of refinements in their procedure, which we will now implement here in order to determine whether symptoms of automatisation can be found in the more sophisticated analysis. Segalowitz and Segalowitz did find the predicted correlation between mean RT and CV<sub>RT</sub> when they limited the analysis to only the top and bottom one third of the listeners (accumulated over native and non-native subjects). Accordingly, we ranked our participants along a fast-slow continuum based on their mean Rats, and we excluded the one-third of the participants (N=16) in the middle of the continuum (mean RT = 1094 (SD 60), CV<sub>RT</sub> = .28). The result now shows the expected correlation between CV<sub>RT</sub> and mean RT. Table 3.4 presents the results of this re-analysis.

---

4 Apparently, extremely fast non-natives are a-typical for the group of L2-subjects. Similarly, very slow native subjects are a-typical; assuming that word recognition is always an automatised process in the mother tongue, such slow performance may be indicative of a disorder (e.g., dyslexia).
AN AUTOMATISATION CRITERION

Table 3.4: Results of the application of the $CV_{RT}$ measure for automatisation for the fastest one-third ($N=16$) and the slowest one-third ($N=16$). Further see Table 3.3.

<table>
<thead>
<tr>
<th></th>
<th>Mean RT</th>
<th>SD</th>
<th>$CV_{RT}$</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fastest one-third</td>
<td>919</td>
<td>81</td>
<td>.24</td>
<td>.544 ($p = .03$)</td>
</tr>
<tr>
<td>Slowest one-third</td>
<td>1414</td>
<td>144</td>
<td>.35</td>
<td>.122 ($p = .65$)</td>
</tr>
</tbody>
</table>

The difference between the two $CV_{RT}$s is significant as was shown by a Independent Samples t-test: $t (15) = –5.21, p < .01$.

Two participants in the fastest group were non-native speakers, and two participants of the slowest group were native speakers. Looking at the data of these ‘wrongly placed’ participants we can say that the native speakers that are categorised in the slow-group, have a-typically high percent correct scores (95%) but (in recompense) a-typically slow reaction times (1325 ms) while the non-native speakers that are categorised in the fast-group have rather low mean percentages correct (63%) but fast reaction times (983 ms). Since the results of these two natives and two non-natives are not in line with the results of the group of other native speakers (89%, 980 ms) and other non-native speakers (78%, 1302 ms) the results indicate a trade-off effect between speed and accuracy in a lexical decision test. The ‘wrongly placed’ participants did not follow the instructions in that they only focused on speed or on accuracy instead of focusing on both.

Figure 3.2 shows that using the $CV_{RT}$ in combination with the mean RT does result in a perfect separation between the fastest one-third of the group and the slowest one-third of the group. In the figure, the black triangles represent the performances of the fastest one-third, the reversed black triangles represent the slowest one-third, and the squares represent the performances of the middle one-third.
Since we had to exclude a certain part of the data in order to get a significant correlation between $CV_{RT}$ and mean RT, and using the $CV_{RT}$ in combination with the mean RT does not result in a strict distinction between native and non-native speakers, we decided to develop our own criterion (described in the following section) for automatisation. Since we believe that besides speed, accuracy is also an important indicator of automatisation, an advantage of developing our own criterion is that we can develop a criterion that is not only based on speed of reactions but also on their accuracy.

3.4. The criterion for separating L1 and L2 word recognition processes

The criterion that is based on both speed and accuracy has been found by using linear discriminant analysis (LDA). This is a statistical technique that discriminates objects (in this case subjects) into a set of pregiven categories (here L1 vs. L2 speakers) by assigning to each object a score on one or several discriminant functions. In the case of a binary categorisation only a
AN AUTOMATISATION CRITERION

A single discriminant function is used. An object’s D-score is then expressed as a linear weighted addition of discriminating variables:

\[ D_i = aX_i + bY_i + \ldots + zZ_i \]

The weighting coefficients \((a)\) for each predictor \((X_i)\) variable are determined on the basis of maximisation of between-category distance and minimisation of within-group variance. The success of the LDA can be determined post hoc by comparing predictor and actual category membership of the objects in the dataset. In a series of LDAs, we looked for the most successful combination of two predictor variables, one of which had to relate to percentage correct, and the other to speed of decision, since these two factors are indicators of automatisation. Scores in each of these two domains were entered into the procedure, e.g. all RTs, only correct RTs, \(\Delta RT\) (overarticulated – underarticulated), \(\Delta RT\) (overarticulated – underarticulated) for monosyllabic words only, etc. The most successful LDA turned out to be based on (i) RT for correct rejections of overarticulated mono and disyllabic nonword stimuli and (ii) percentage correct rejections of such strings. This means that this combination is the criterion we have been looking for. The result of the LDA confirms the results given in Table 3.2, where it was shown that the largest effect size was found in the reactions on the overarticulated nonwords. The optimal discriminant function is found in the results of a sub-set of these stimuli namely the results on the combined set of mono and two-syllable nonwords. Figure 3.3 shows the optimal separation. Each square indicates the results of one participant; the black squares represent the native speakers, and the white squares represent the non-native speakers.
As can be seen in Figure 3.3, the distinction between native and non-native speakers is not perfect. Three non-native speakers behave like the native speakers while one native speaker reacts like the non-native speakers do. The distribution of the results will be discussed in more detail in section 3.5.

3.5. Discussion

3.5.1. Reaction Times

It can be observed that the reaction times presented in section 3.2.5.2 are relatively long. One possible explanation is the randomised presentation of the items: participants never knew whether the next item would be overarticulated or underarticulated speech. This variation may have confused the listeners and made it more difficult for them to build a frame of reference within which a stimulus structure could be evaluated. As a result, they probably needed more time to decide whether the items they heard were or were not real Dutch words. If the stimuli would be blocked (i.e., presented in
homogeneous blocks containing only overarticulated items vs. underarticulated items) one would expect faster decision times: if all the overarticulated items are presented together as well as the underarticulated items, participants can build more constrained frames of reference for each of the two types of speech quality. Another possible explanation for the rather long RTs is that the RTs were measured from the onset of the stimuli. This means that the duration of the (non)words is included in the RT. However, analyses showed that the pattern of the results remains the same when the offset RTs are used (i.e., when reaction times were measured from the end of the stimulus).

3.5.2. Native versus non-native speakers

The results indicate that the overarticulated word condition was the easiest to perform in for both native as non-native speakers (lowest RT in combination with highest percent correct). Reacting on words was not difficult in general. The high frequency of the words used as stimuli might explain why it was relatively easy also for the L2-learners to react to these stimuli. Despite the good responses of the non-native speakers on the words, the differences between the native and non-native groups were significant. The fact that there was no significant difference between the groups in percent correct on overarticulated words in combination with the significant difference in reaction times indicates that speed of recognition is an important indicator of automatisation. It is possible for non-natives to react accurately on items but this accuracy costs time. The native speakers, whose word recognition process is automatised, are able to react fast and accurately on words. Reacting to the nonwords, however, was relatively difficult for both the non-native speakers and for the native speakers. It is not surprising that it was rather problematic for the non-natives to react to the nonwords, since the nonwords follow the grammatical rules of the Dutch language: all items were possible words. This construction criterion, in combination with the fact that the vocabulary of the non-natives is not so extensive, could be the reason for the poorer results on the nonwords. Native speakers determine in an automatised way the uniqueness point of words and the nonword point of nonwords; they are therefore able to decide in an early stage – and with much more confidence – that an item is a real word or not of the target language whereas non-native speakers are not.
3.5.3. The processing of nonwords as a source of information

The results indicate that the reactions on the nonwords are of great importance in this study where we tried to find a criterion to distinguish between fully automatised and non-automatised (or not yet fully automatised) processes. As described in section 3.4 and shown in Figure 3.3, the criterion is found in the reactions on 1- and 2-syllable overarticulated nonwords. That the sharpest difference between natives and non-natives is found in the reactions on the nonwords does not mean that the criterion cannot be used in research on word recognition. In Chapter 2 the word recognition process is defined as a process in which language users identify the phonemes of the word and search their mental lexicon to decide whether the item they heard (or saw) is or is not an existing word. Identification of the phonemes as well as searching the lexicon are automatised processes in native speakers, who immediately know whether the item is or is not a Dutch word. Non-native speakers are not able to take this decision in an automatised way; their knowledge of Dutch is not solid enough and their mental lexicon is not large enough. The fact that the nonwords are constructed following the rules of the Dutch grammar may aggravate the feeling of uncertainty on the part of the non-native speakers. It could be the case that the non-native speakers think that the string of sounds is a possible word that they just do not know and so draw the wrong conclusion.

This feeling of uncertainty is probably also a factor in the delay on the correct decisions to the nonword. As stated above, one can assume that the mental lexicon of non-native speakers is not as extensive as the lexicon of native speakers. Therefore, one might expect the reactions to the high-frequency stimuli to be faster for non-native speakers: as their mental word list is shorter, they need less time to go through it. The results show, however, that they need more time rather than less. It is possible that this was caused by a phoneme identification problem. In other words, the problem might not have been a lack of automatisation of the lexical search in the mental lexicon, but a lack of knowledge of the Dutch phonemes (and of the phonotactic rules of Dutch). Following this reasoning we could expect that L2 listeners would do much better and come quite close to L1 speakers, if the LD-experiment were repeated with visual presentation of the stimuli. Therefore a visual lexical decision task is included in the test battery of Pilot study II, see Chapter 4.
3.5.4. Unexpected results

Looking at Figure 3.3 one can see that there is one native speaker who acted like a non-native and three L2 speakers who are – incorrectly – grouped with the Dutch native listeners. The native speaker that behaved like a non-native was born and raised in Frisia, a northern province of The Netherlands, where Frisian is spoken. While related to Dutch this is a very different language (van Bezooijen 2002). Although Frisia is a bilingual Frisian/Dutch community (both languages are taught and spoken in primary education; Wales in the United Kingdom might be comparable) this student may have been dominantly Frisian. In that case the Frisian participant should be considered as a non-native speaker of Dutch who behaves (more) as expected. Since the result of this native speaker is extreme even if we consider the participant as a non-native speaker, it seems likely that the environment in which the participant grew up is not the only possible explanation, physical disorders can be another explanation. After the test phase, it was, unfortunately, not possible to contact this participant to investigate the physical condition. The three non-natives who reacted like the native speakers were Czech, Hungarian, and Croatian. At the time of testing they had been in the Netherlands for six months, seven years and 18 months respectively. There was nothing extraordinary in the information they filled in on the inquiry form. They had knowledge of English and German but so did the other non-native speakers. Previous research (e.g., Flege, 1987; Bongaerts et al. 1997) showed that there are people who are able to achieve a native level of proficiency in (the pronunciation of) their second language. Maybe we should conclude that the three participants simply are some of the few very talented people that are able to become near native, as far as our test results allow such a conclusion. It is true that only three non-native participants reached the L1 level, but all our L2 students had lived in the Netherlands for only a few years. It is possible that using L2 speakers who have been living in the Netherlands for many more years, would create more overlap between the native and non-native populations.

3.5.5. Automatisation as a continuum

Figure 3.3 also shows that there is an automatisation continuum. The spread of the L2 results indicate that the process of word recognition may be automated to a greater or lesser extent. The discriminant function (expressed in the figure by the discriminant line) can be used as a cut-off: any score below the cut-off will be considered the result of not fully automatised processes. This means we can conclude that the process of word recognition
for high-frequency words became automatised for the three non-native speakers whose reaction times lie in the native field.

The criterion that is found will be used in this thesis to investigate the relation between the degree of automatisation of a process of word recognition and overall listening comprehension.