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

Poelmans, P.

Citation for published version (APA):
Developing second-language listening comprehension
Effects of training lower-order skills versus higher-order strategy

The study described in this PhD thesis is the first that investigates the relationship between language knowledge, listening comprehension, and automatisation of spoken word recognition among adult learners of a second language (L2).

The development of a criterion for indexing automatisation of spoken word recognition is described as well as the attempt to predict listening comprehension skills on the basis of information about language knowledge and status of the word recognition process. The main experiment, involving 83 adult Dutch L2 learners, focuses on the relative effect of two distinct methods for the training of listening in L2:

(i) focus on lower-order skills like word recognition in connected speech
(ii) focus on higher order strategy.

Although the results are not so straightforward that recommendations are made towards a change in the teaching of L2 listening skills, the thesis draws conclusions that are of interest both for educationalists and researchers. For example, the results show the usefulness, or even the necessity, of using time-critical tests for compiling detailed language profiles of second language learners. With respect to the notion of automatisation, the results contribute to the idea that automatised and controlled processes do not form a dichotomy but that there is a continuum between the two extremes.

The detailed description of the tests and training-exercises that were used offer a wealth of information both for researchers and language teaching professionals.
Developing second-language listening comprehension: Effects of training lower-order skills versus higher-order strategy
This dissertation and the underlying research were financially supported by the Netherlands Organisation for Scientific Research (NWO) under project number 575-21-009 (principal investigators J. Hulstijn and V. van Heuven). This project was part of the larger research program ‘To learn, automatize, and control a second language: activation and control in bilingual memory’ which was coordinated by A. de Groot.

Cover illustration: produced by CreaSign
NUR 632

Copyright © 2003 Petra Poelmans. All rights reserved.
Developing second-language listening comprehension: Effects of training lower-order skills versus higher-order strategy

ACADEMISCH PROEFSCHRIFT

der verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. mr. P.F. van der Heijden
ten overstaan van een door het college voor promoties ingestelde commissie, in het openbaar te verdedigen in de
Aula der Universiteit

op dinsdag 10 juni 2003, te 12.00 uur

door

Petra Poelmans

geboren te Heusden (België)
Faculteit der Geesteswetenschappen
Universiteit van Amsterdam

PROMOTIECOMMISSIE

Promotores: prof. dr. J. Hulstijn
             prof. dr. V. van Heuven

Overige leden: prof. dr. K. de Bot
               dr. A. Dijkstra
               prof. dr. A. de Groot
               prof. dr. ir. L. Pols
               dr. R. Schoonen
voor papa
Dankwoord

In de eerste plaats wil ik mijn dank uitspreken voor mijn twee promotoren, Jan Hulstijn en Vincent van Heuven. De veelvuldige gesprekken over het verloop van het onderzoek waren steeds informatief en uitdagend. Meerdere malen had ik na zo’n gesprek, waarbij jullie aan een kant van de tafel zaten en ik aan de andere, het gevoel dat ik de verdediging reeds had doorstaan. Jan en Vincent, ik heb veel van jullie geleerd, daarvoor dank ik jullie.

De overige commissieleden, Kees de Bot, Ton Dijkstra, Annette de Groot, Louis Pols en Rob Schoonen dank ik voor de bereidwilligheid in mijn commissie plaats te nemen en voor het geven van commentaar op de leesversie van dit proefschrift.

De leden van de Staatsexamencommissie Nederlands als tweede taal en de medewerkers van het CITO wil ik bedanken voor de toestemming tot het gebruik van het staatsexamenmateriaal. Mijn dank gaat ook uit naar de docenten van INTT en VASVU; dankzij hun bereidheid lestijd af te staan kon ik groepen studenten aan diverse experimenten en de training onderwerpen.

Mijn project was een deel van het NWO-aandachtsgebied ‘Leren en automatiseren van een tweede taal’. De leden van het aandachtsgebied, Ingrid Christoffels, Ton Dijkstra, Annette de Groot, Tony Chessa, Jaap Murre en Béryl Schulpen dank ik voor de interessante discussies die we voerden tijdens de bijeenkomsten. Ingrid en Béryl, jullie waren fijne collega OIO’s, onze afspraken waren momenten om naar uit te kijken.

Ingrid Christoffels, Nel de Jong, Laura Sabourin, Annegien Simis en Marie Stevenson, jullie waren fantastische (hotel)kamergenoten.

Hoewel officieel een UVA-promovenda, was ik enkel in Amsterdam te vinden voor afspraken, lezingen of andere bijeenkomsten. Ondanks mijn slechts sporadische aanwezigheid voelde ik me er steeds thuis. Hiervoor wil ik graag de leden van leerstoelgroep Tweedetaalverwervening en de ACLC-promovendi bedanken. Het echte werk vond, dankzij de gastvrijheid van Vincent van Heuven, in Leiden plaats. Ik heb het Leidse fonetisch lab steeds als een inspirerende en warme plek ervaren. Maarten, Rob, Johanneke, Jos, Vincent, Ellen, Bert, Ruben, Antje, Hongyan en Jie bedankt! Ik zal met plezier terugdenken aan de koffie-en theepauzes, de lunches en onze labetentjes.

Naast bovengenoemden die het leven tijdens de werkuren veraangenaamden zijn er natuurlijk ook een heleboel mensen die me de overige tijd bezighielden. Belangrijk hierbij was de Leidse Harmoniekapel. De wekelijkse repetities en alles wat er uit voortkwam (etentjes,
spelletjesavonden, vakanties,...) waren een welkome afwisseling. In het bijzonder wil ik hiervoor bedanken, Marcel, Jeroen, Shanna, Marco, D., Marco H., Franc, Elvira, Johan, Hanneke, Paul, Harry, Guus en Han.


Een bijzonder woordje van dank voor mijn paranimfen, Johanneke en Simone. Johanneke, je hebt het onderzoeksproces als mijn kamergenote van zeer nabij meegemaakt met alle ups en downs. De momenten dat we onze stoelen naar elkaar toe draaiden om even stoom af te blazen of om gewoon even bij te praten zijn me erg dierbaar, net als onze bezoekjes aan het Danstheater. Simone, vijf jaar geleden ging je met me mee naar mijn eerste sollicitatiespraak voor de OIO-plaats die tot dit proefschrift heeft geleid. Vandaag sta je weer naast me en dat heb je altijd gedaan op belangrijke momenten. Dank je wel.


Mama en Wendy, weten dat jullie achter me staan betekent heel veel voor mij. Naar Houthalen komen is na al die jaren nog steeds thuiskomen. Dank aan Martijn en Dorien, twee zonnetjes in mijn leven.

Papa, ik mis je op deze dag meer dan ooit. Ik weet hoe je naar deze dag uitkeek, hoe trots je was. Ik troost me met de gedachte dat je in ons hart bij ons bent.

Lex, dank voor je geloof in mij en voor alles waar geen woorden voor zijn!
# Contents

1 Introduction
  1.1 Introduction  
  1.2 Differences between listening and reading  
  1.3 Listening in foreign-language courses  
  1.4 The present study  
    1.4.1 Research questions  
  1.5 The outline of the thesis in summary

2 Background
  2.1 Introduction  
  2.2 The listening process  
  2.3 The L2 listening process  
  2.4 Problems in listening  
  2.5 Memory  
    2.5.1 Working memory  
    2.5.2 Measuring working memory capacities  
  2.6 Automatisation  
    2.6.1 Theoretical background  
    2.6.2 Measuring automatisation

3 Pilot study I: An automatisation criterion
  3.1 Introduction  
  3.2 The experiment  
    3.2.1 The Lexical Decision paradigm  
    3.2.2 Experimental conditions  
    3.2.3 Method  
      3.2.3.1 Participants  
      3.2.3.2 Stimulus materials  
      3.2.3.3 Apparatus  
    3.2.4 Procedure  
    3.2.5 Results  
      3.2.5.1 Outlier procedure  
      3.2.5.2 Mean results  
  3.3 The coefficient of variation  
  3.4 The criterion  
  3.5 Discussion
3.5.1 Reaction Times 48
3.5.2 Native versus non-native speakers 49
3.5.3 Processing nonwords 50
3.5.4 Unexpected results 51
3.5.5 Automatisation as a continuum 51

4 Pilot study II: The predictability of listening comprehension
4.1 Introduction 53
4.2 Method 55
  4.2.1 Overview of tests 55
    4.2.1.1 Off-line tests 55
    4.2.1.2 On-line tests 58
    4.2.1.3 Memory test 59
  4.2.2 Participants 61
  4.2.3 General procedure 61
4.3 Results 61
  4.3.1 Off-line test scores 62
  4.3.2 On-line test scores 62
    4.3.2.1 Auditory lexical decision 63
    4.3.2.2 Visual lexical decision 68
  4.3.3 Memory test 69
  4.3.4 Relations between the tests 69
    4.3.4.1 Auditory versus visual tests 69
    4.3.4.2 Correlations 70
  4.3.5 Categorization of the participants 72
4.4 Discussion 73
  4.4.1 Unexpected results 73
  4.4.2 Interpretation of inter-test relations 74
  4.4.3 In summary 75

5 Chapter 5: Training Study
5.1 Introduction 77
5.2 Method 79
  5.2.1 Participants 79
  5.2.2 Test stages 81
    5.2.2.1 Test stage 1 82
    5.2.2.2 Test stage 2 88
    5.2.2.3 Design of the study in summary 91
    5.2.2.4 General procedure in the test sessions 92
  5.2.3 Training 93
    5.2.3.1 Material 93
    5.2.3.2 Training procedure 94
5.2.3.3 Experimental groups
5.2.3.3.1 Comprehension group
5.2.3.3.2 Recognition group

5.3 Results
5.3.1 Selection of data
5.3.2 Results of test stages 1 and 2
5.3.2.1 Selection tests
5.3.2.2 Control test
5.3.2.3 Mediating tests
5.3.2.4 Pre- and posttest performances
5.3.2.4.1 Listening Comprehension
5.3.2.4.2 Sentence Verification
5.3.2.4.3 Auditory Lexical Decision

5.3.3 Correlations and regression analyses
5.3.3.1 Correlations
5.3.3.2 Regression analyses
5.3.4 Coefficient of Variation
5.3.4.1 CV for Sentence Verification test
5.3.4.2 CV for Lexical Decision test
5.3.5 General summary of the results

5.4 Discussion and conclusion
5.4.1 Dependent Variables
5.4.2 Mediating tests
5.4.3 Automatisation
5.4.4 General discussion

6 Conclusions and suggestions for further research
6.1 Recapitulation of questions and main findings
6.2 Suggestions for applications and for further research

References

Appendices
Appendix A
Appendix B
Appendix C
Appendix D
Appendix E
Appendix F
Appendix G
Appendix H
Appendix I
<table>
<thead>
<tr>
<th>Samenvatting</th>
<th>169</th>
</tr>
</thead>
<tbody>
<tr>
<td>Curriculum vitae</td>
<td>177</td>
</tr>
</tbody>
</table>
Chapter 1

1.1. Introduction

Everyone who has ever learned a foreign language has probably experienced the frustrating feeling of not being able to communicate with native speakers of the language despite years of training in the target language. Often one knows the words when they are presented visually but one does not recognise them in a spoken utterance. The main cause of this communication problem is the disability of listeners to recognise the words in the pace in which they are spoken. In other words, listeners may have enough vocabulary knowledge but they may be unable to use this knowledge under time pressure. Another phenomenon familiar to many learners of a foreign language is knowing ‘what to say how’ before or after a conversation but not during the conversation itself. Again, language learners may have enough vocabulary and grammar knowledge to construct correct sentences in reply to what an interlocutor has just said but they may not be able to access and retrieve the language elements from their mental lexicon fast enough.

The two situations sketched above indicate that there is a difference between using a language with and without time constraints. It seems that language learners are often unable to use the knowledge of the language they have (e.g., vocabulary, grammar, and pronunciation) in a fluent (i.e., fast and accurate) way. This suggests that learning to use language knowledge in a fluent way is conditional on successful communication. In foreign-language courses, however, focus is traditionally placed on the acquisition of language knowledge rather than on the automatisation of knowledge. It is obvious that the attainment of language information is an indispensable first step in becoming a fluent foreign language user, but if the process of language learning is limited to knowledge appropriation, learners will never become fluent foreign language users, let alone that a near-native language level will be reached.

In a world in which international communication becomes more and more important, the necessity of good foreign language use increases. For example, for people immigrating into The Netherlands, learning the language of the host country in a quick and efficient way is necessary. The political climate in which it is incumbent upon immigrants to learn the official language of the host country, as was for example stated by the Dutch
government in the *Troonrede* ‘Speech from the throne’ on Prinsjesdag 2002, makes it important that research will be conducted to develop the best method for learning a foreign language. Since listening is an important factor in communication, the importance of research on the listening process and the way listening is best dealt with in language courses is obvious.

### 1.2. Differences between listening and reading

In the early days of foreign-language listening research, listening was mainly seen as a mirror of reading. Listening was studied as ‘reading with your ears’, as it were. Nowadays, however, there is a growing awareness of essential differences between listening and reading. Speech contains features like variation in pronunciation, hesitations in speech delivery, incomplete utterances and even flawed sentence structure. This ‘production noise’ has to be filtered out, so to speak, by the listener. Readers, on the other hand, usually deal with the final product of writing, the text, bearing no, or fewer, overt signs of online production difficulties. Furthermore, readers will have little difficulty with the recognition of the words in written text as the beginnings and endings of words are marked by inter-word spaces whereas listeners have to process a speech stream largely consisting of concatenated words. Another important difference between the two skills is concerned with the speed in which the input has to be processed. In most reading situations, readers can read at their own pace without time pressure. If they come across a difficulty in the text, they can reread the problematic sentence or phrase and pause to think. Listeners, however, have an on-line contact with speakers; listeners are dependent on the rate with which speakers deliver their speech. In all likelihood, it is especially the fact that processing speed cannot be dictated by listeners themselves that makes listening so much harder than reading for many foreign-language users.

---

1 Prinsjesdag is the annual event on the third Tuesday of September on which the queen of The Netherlands reads aloud the plans of the government for the coming year. In 2002 it was explicitly said that immigrants will be demanded to learn Dutch: *De regering zal daarom de integratie van allochtonen verder bevorderen, hetgeen van nieuwe Nederlanders vereist dat zij onze taal spreken ... ‘The government shall therefore encourage the integration of immigrants, this implies that immigrants have to learn our language.’* The complete text can be found on www.regering.nl. An English version of the text can be found on www.nrc.nl.
1.3. Listening in foreign-language courses

It is only since the early 1900s that listening was believed to have an important role in language teaching. The focus on speech perception was intensified when anthropologists started to describe the spoken languages of the world. The spirit of the times is expressed by Leonard Bloomfield (1942, in Rost 2002): ‘one learns to understand and speak a language primarily by hearing and imitating native speakers’. Despite the general role that listening was believed to have (and still is believed to have) in language acquisition, it was not until the 1970s that listening started to be treated as a separate part of the language curriculum (Rost 1990).

The idea that listening should be considered as a skill in its own right is reflected in many present-day foreign-language courses. Mendelsohn (1998, 81) describes it as follows: ‘There has been a shift from non-teaching in the Audio-Lingual period (‘They’ll pick it up by osmosis’), to haphazard listening to texts (many being readings of written language) followed by comprehension questions, to a ‘strategy-based approach’ in which students are taught strategies - that is, they are taught how to listen’. The present focus in foreign-language pedagogy on listening comprehension strategies is what we will call in this thesis the traditional way of listening teaching.

In most language courses, listening comprehension is dealt with by emphasising the comprehension of the overall meaning of the message. Less focus is being paid on the recognition of phonemes and words - one of the first steps of listening. Results of for example Staatsexamen NT2 II, ‘National exam Dutch as a second language level II’ indicate that the effect of the traditional method of focussing on global comprehension is not optimal; many candidates fail the listening part of this exam (Kerkhoff 1997). It is the opinion of several experts that focussing on lower-order processes of the listening skill and becoming familiar with the characteristics of spoken language, will be helpful in processing speech efficiently (Matter 1986, Koster 1987, Mendelsohn 1994). There are, however, also researchers that do not believe in the effectiveness of such a low-order approach. Buck (1995, cited in Mendelsohn 1998, 87) for example, calls the ‘pre-

---

2 Berne (1998) gives the results of a questionnaire study amongst second language (L2) instructors. She concludes that the language instructors believe that the development of listening skills is crucial to overall L2 development but they also show that there is some resistance to teaching listening as a separate skill. The study shows a gap between listening research and listening in teaching practice.

3 This exam will be described in more detail in Chapter 4 of this thesis.

4 As an example: only 39% of the candidates in 2001 (June) succeeded for the listening exam II. In Comparison, in March 2002 36% succeeded for the listening exam II (Kerkhoff 2002).
communicative practice’ (i.e., sound discrimination, recognizing reduction, assimilation, word boundary distortion) a ‘necessary but insufficient condition for success’. In summary, we can say that while focussing on what is called lower-order processes like the recognition of sounds and words seems useful, the impact of such an approach is still unclear.

1.4. The present study

The idea that focussing on lower-order processes might be beneficial to listening comprehension has not been empirically tested. We have no knowledge of (training) studies that investigated the effect of focussing on the lower-order processes of listening comprehension in comparison to the effect of the more traditional method of training higher-order comprehension skills. This lack of empirical information formed the main motivation to conduct the present study.

As has been just said, there is no previous research that compares the effect of a lower-order training with the effect of a higher-order training on general listening comprehension. In this sense the present study is unique. Another interesting aspect of the study is that it investigates empirically a claim made by Segalowitz & Segalowitz (1993) that automatisation of aural word recognition is conditional on successful listening comprehension. The idea of focussing on lower-order processes, like word recognition, to improve general listening comprehension skills is strongly related to this automatisation hypothesis. As has been said at the beginning of this chapter, it is mandatory that language users be able to use their knowledge of the language under time pressure to communicate adequately. In this sense, automatisation of the use of knowledge will improve the listening process.

The study described in the present thesis, is the first that investigates the relationship between language knowledge, listening comprehension, and automatisation of spoken word recognition. Insight in this relationship will be useful for instance for course developers and language teachers.

If this study were to find that language knowledge is of little impact on listening comprehension unless word recognition takes place
automatically, it would probably be a good idea to adapt the way listening is trained.\(^5\)

1.4.1. Research questions

In the previous sections it was mentioned that the main goal of the present study is to investigate the effect of lower-order training on general listening comprehension, and that a positive effect could provide empirical support for the automatisation hypothesis of Segalowitz and Segalowitz (1993), mentioned in the previous section. On the basis of this hypothesis one could argue that adult native speakers have no difficulty in comprehending L1 speech, as word recognition processes in L1 have become automatic. However, when at intermediate, or perhaps even at rather advanced levels of L2 learning, their word recognition processes in L2 have not yet become fully automatised. This raises the question whether non-native speakers can reach native levels of aural word recognition. Is it then possible to draw a strict boundary between L1 and L2 word recognition in terms of speed? To address this issue, Pilot Study I was set up. The research question of this first pilot study was:

\[ \text{Research question in Pilot Study I} \]
\[ \text{Is it possible for non-native speakers to recognize words as fast and accurate as native speakers, or will there always be a distinct difference between L1 and L2 speakers?} \]

To answer this question, the Lexical Decision paradigm was applied. Native and non-native speakers of Dutch participated in an experiment in which they had to decide, as fast as possible, whether strings of sounds they heard were or were not real Dutch words. Both accuracy and speed of their responses was measured. In this way it was possible to determine the status of the participants’ word recognition process. The development of the criterion is described in Chapter 3.

\(^5\) The focus of this study is on the listening comprehension skill of second language learners. The difference between learning a foreign language (FL) and acquiring a second language (L2) lies in the context in which the language is learned. A second language is learned in the community in which the target language is the medium of communication (e.g., learning Dutch in The Netherlands or Flanders); one speaks of foreign language learning if the language is learned is not the official language of the country (e.g., learning Dutch in The United States). It is unlikely that second and foreign language learning are essentially different with respect to the psycholinguistic processes of listening. In the remainder of this thesis, ‘second language’ (L2) will be used throughout.
In second language research and second language education it is often necessary to determine the language-proficiency level of L2-learners. Usually, so-called off-line tests are used for this purpose. Off-line tests can be defined as tests that allow participants to think before giving a response, as in a traditional paper-and-pencil grammar test, administered without time pressure. While it is probably correct to argue that performance on off-line tests indicate a language user’s general command of the language in question, it does not give information about the processes underlying the skills uses in such tests. For example, if a language learner fails a listening comprehension test, one cannot say whether this is due to a lack of language knowledge or to insufficient automatisation of this knowledge. To generate a more detailed profile of the language user it is necessary to use on-line tests in addition to off-line tests. On-line tests can be defined as time-critical tests that require the candidate to respond while the ongoing stimulus is still being processed. Performances on on-line tests, such as the Lexical Decision test, reflect the status of the processes underlying a language skill (Rietveld & Van Heuven 2001). In our investigation, a second pilot study was set up to assess the added value of on-line tests as a complement to off-line tests in obtaining 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. We conducted Pilot Study II in order to answer the following two questions:

Research questions in Pilot Study II
1. Do on-line tests give additional information about L2 users in comparison to an off-line knowledge test, and do on-line tests thereby help to obtain a more detailed language proficiency profile?

2. To what extent is it possible to predict L2 listening comprehension performance on the basis of language knowledge tests and a measurement of the status of the word recognition process?

Several tests were used to determine the language profile of the second language learners participating in Pilot Study II. There were tests that measured listening comprehension, language knowledge (grammar), the status of visual and auditory word recognition process, and memory capacity. Pilot Study II is described in Chapter 4.

Pilot studies II and I were conducted in order to do some preparatory work for the main study. The main goal of the research described in this thesis was to compare a training method that focuses on lower-order recognition skills with a more traditional higher-order training focusing on
global understanding, in their effect on listening comprehension. In this respect, our investigation follows the idea propounded by Segalowitz and Segalowitz (1993) that automatisation of aural word recognition is conditional for successful listening comprehension in a second language. The research question of the main experiment of our study can be formulated as:

**Research question in the Training Study**

*What is the most effective training method to improve listening comprehension performance of intermediate L2 learners,*

(i) *a method focusing on improvement of lower-order word recognition skills or*

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

To answer this question a training study was set up, involving two experimental training conditions and a non-training control condition, in a between-group design with adult learners of Dutch as a second language at intermediate levels of proficiency. The first experimental group, the so-called *Recognition group*, performed listening tasks focused on the identification of typically Dutch speech sounds and on the recognition of spoken words in concatenated speech. The second experimental group, named the *Comprehension group*, trained listening comprehension in a ‘traditional way’ by performing tasks focussing on understanding the global meaning of aural input. Before the actual training started there was a test session to assess participants’ pre-treatment level of proficiency. After training, there was a second test session to test their post-treatment level of proficiency in order to investigate the effect of training. A third group was formed as the Control group, which was excluded from the training programs but participated in the pre and post-tests.

The main research question of the Training Study is concerned with the differential effect of training lower-order processing skills and higher-order comprehension strategies. In addition to this main question we also wanted to investigate two subsidiary issues. The first one pertains to individual differences in participants’ capacity to keep verbal information in relation to their listening comprehension skills. To investigate this issue we assessed participants’ working memory capacity in addition to their listening comprehension. The second issue is concerned with automatisation. Before and after training participants performed a word recognition test. We investigated whether word recognition had become more automated after training. The training study is described in Chapter 5.
1.5. The outline of the thesis in summary

In this first, introductory chapter, some background was given about the scientific and social relevance for the present study. Then the three empirical studies (Pilot Study I, Pilot Study II, and the Training Study) were briefly introduced.

In Chapter 2 the listening process is described in various respects relevant to our empirical studies. Chapter 3 describes Pilot Study I, aimed at developing an empirical criterion that seeks to distinguish automatised word recognition of native speakers from non-automatised word recognition of non-native speakers. This criterion is also used in Pilot Study II, which is described in Chapter 4. In that chapter we investigate the relationship between language knowledge, speed and accuracy of the spoken word recognition process, and general listening comprehension. To what extent can the first two variables successfully predict listening comprehension performance? In Chapter 5, the training study that we conducted to answer the main research question of this thesis is presented, including a detailed description of the method and the results. A general discussion as well as some general conclusions and suggestions for future research are given in Chapter 6, which concludes this thesis.
Chapter 2

Background

2.1. Introduction

In the field of second language learning four basic language skills are distinguished. These are ordered along two dimensions:

(i) **modality**, that is the difference between the auditory language mode versus the visual mode, and

(ii) **processing activity**, that is the process of either encoding or decoding.

Together these two binary dimensions define the so-called four skills, as indicated in Table 2.1.

*Table 2.1: The four language skills.*

<table>
<thead>
<tr>
<th>Modality</th>
<th>Processing activity</th>
<th>encoding</th>
<th>decoding</th>
</tr>
</thead>
<tbody>
<tr>
<td>auditory</td>
<td>speaking</td>
<td>listening</td>
<td></td>
</tr>
<tr>
<td>visual</td>
<td>writing</td>
<td>reading</td>
<td></td>
</tr>
</tbody>
</table>

The listening skill is defined, then, as the process by which the language user decodes auditory input, i.e., speech. This use of the term ‘listening’ deviates from its ordinary use, in which listening simply refers to the act of intentional hearing, in much the same way as looking refers to the act of intentional seeing.

In other disciplines the listening skill is often called listening comprehension, which is the complex of processes that transform an auditory stimulus to a mental reconstruction on the part of the listener of the speaker’s intention. In this thesis we will use the terms ‘listening (skill)’ and ‘listening comprehension’ indiscriminately.
The aim of this chapter is to provide the reader such theoretical background information on listening and language learning as is needed to follow the exposition in the later chapters.

### 2.2. The listening process

To communicate adequately it is important for a listener to understand what the speaker says. Understanding spoken language can be described as an inferential process based on the perception of several cues rather than a simple match between sounds and meaning. The listening comprehension process is a combination of (roughly) four sub-processes or modules:

(i) **hearing**: the auditory reception of an acoustic signal (perception), as is also performed on non-speech sounds,

(ii) **categorisation of sounds**: categorising incoming sounds in terms of the sound categories of the language

(iii) **word recognition**: breaking up the stream of sounds into linguistic units (morphemes, words) and retrieving their meaning from long term memory, and

(iv) **comprehension**: integrating the meanings of the words in their sequence into an interpretation of the entire utterance, i.e., a reconstruction of the speaker’s communicative intention (= message).

Obviously, the third and fourth modules draw heavily on linguistic knowledge. Both lexical knowledge, needed to recognize words, and knowledge of the rule system of the language, needed to decode the grammatical relationships among the words that make up the sentence, are part of the listener’s linguistic competence (Ellis 2000). However, especially at the third stage, knowledge of the world, i.e., non-linguistic knowledge, plays an often-indispensable role in the listening process.

Figure 2.1 shows the different sub-processes of the listening comprehension process. Each of the sub-processes displayed in Figure 2.1 can be a source of listening comprehension problems (see section 2.3.).
Before embarking on a discussion of the modules included in Figure 2.1, it is expedient to introduce a parameter that is often used with reference to psycholinguistic processing models, viz. the direction of the information flow in terms of bottom-up versus top-down. In Figure 2.1 sensory information enters the system at the periphery, which is popularly called the bottom-end of the system. In successive stages the lower-order information is transformed to more abstract, higher-order, representations. Auditory details are lost in the transformation process, each time when a higher-order representation is derived. The loss of detail is probably a prerequisite if the brain has to keep a longer stretch of speech in working memory. The most abstract representation, generated by the most central module, is the overall interpretation of the discourse, the gist of the message, in which details of individual sounds, words and grammatical structures are absent.

In addition to the bottom-up flow of information, i.e., from sensory periphery to central processing by the brain, there is a top-down information flow from the brain down to the periphery. Top-down information flow may occur between modules. For instance, when a (predictable) sound in a spoken word is (electronically) replaced by a noise burst (or cough), the listener will recognize the word as if no sound was missing; in fact, the listener will not be able to tell the experimenter which sound exactly was missing from the word. The word was recognized from partial sensory input, and the missing phoneme was restored (hence: ‘phoneme restoration effect’, cf. Obusek & Warren 1973; Samuel 1990) on the basis of the complete lexical specification. Information available at a higher, more central, representation is fed back, then, to the lower level. Top-down information is typically used to limit the number of choices that has to be made by a lower-level module, which serves to free processing capacity in working memory. When a process, or module contained by it, crucially depends on both bottom-up and top-down information streams, it is called ‘interactive’; when the process uses bottom-up information only, it is called ‘autonomous’.
As is indicated by the bi-directional arrows in Figure 2.1, we take the view that the modules in the listening process are typically interactive; only the peripheral hearing process is considered to be autonomous.

The first sub-process of the listening skill is **hearing**. We define hearing as the relatively peripheral process of filtering the relevant speech signal from the multitude of acoustic events that simultaneously strike the ear. Even under optimal listening conditions the speaker’s voice must be isolated from background noise. In more complicated listening situations, such as may occur when several speakers talk at the same time, in a noisy environment, the process by which a listener is able to attend to just one speaker’s voice is ill understood. For the purpose of the present study we will assume that the sub-process of filtering the relevant signal from the complex auditory input is language independent, and never poses a specific problem for the foreign language learner.

The linguistic processing of the output of the hearing stage is the focus of this thesis. As a first language-specific transformation listeners **categorise** incoming sounds in terms of the sound categories of their language, this is the second sub-process of listening. Speech sounds differ along a limited number of phonetic dimensions, each of which may depend on several auditory cues. For instance, whether a consonant is perceived as an intervocalic voiced /b/ or voiceless /p/ – as in the minimal pair ruby ~ rupee – depends on a multiplicity of auditory cues, such as the duration and intensity of a noise burst, the duration of a (nearly) silent interval separating the plosive from the preceding vowel, the duration of the preceding vowel, the abruptness of the amplitude and formant frequency changes at the edges between the consonant and the abutting vowels, and the course of the fundamental frequency at the onset of the post-consonantal vowel (Hayward 2000: 117-121, Rietveld & van Heuven 2001: 62-64). It is commonly held

---

1 There is at least one aspect of hearing that demonstrably involves the use of top-down information, i.e., speaker separation on the basis of prosodic continuity. It has been shown (Darwin, 1975 in Nooteboom, 1996: 669-670) that a listener is able to attend to one speaker’s voice in the midst of competing noises and competing (even louder) speakers. This is the so-called cocktail-party effect described by Cherry (1953). One very important cue the listener uses to track one speaker’s voice is the predictability of the pitch pattern; the listener will attend to the speaker whose voice continues on the pitch level that the listener expects from the preceding spoken context. Intonation patterns differ between languages. Therefore, generating predictions of the pitch level of the next syllable involves language-specific, top-down information. For expository reasons, we will nevertheless assume that hearing proceeds in a strictly bottom-up fashion.
that a dozen binary phonetic dimensions (theoretically yielding ca. $2^{12} = 4096$ different sound categories or ‘phonemes’, 15-75 in each language) suffice to capture the variety of vowels and consonants in the languages of the world. Each of these binary dimensions is coded in turn by a limited number of primitive acoustic cues of the type exemplified above. However, languages differ widely in the number and type of phonemes, and in the precise division of the multi-dimensional phonetic space over the sound categories that exist in the language.

Apart from the segmental categories that define the various vowels and consonants in the language, spoken utterances are characterised by so-called suprasegmental (or: prosodic) properties. These are properties that do not belong to specific, individual speech sounds but subsume larger units of at least the size of a syllable. Prosodic features serve to break up the continuous utterance into smaller chunks (clauses and phrases, sometimes inaptly referred to as thought units) by the presence of pauses and boundary-marking pitch changes, and to mark one syllable or word as the focus of the speaker’s attention within the chunk (accentuation) (Nooteboom 1997, and references given therein). Typically the suprasegmental properties guide the interpretation process whilst the segmental cues serve to access words in the mental lexicon (Cutler 2001).

The third sub-process of the listening skill is **word recognition**. The process of word recognition has been the object of phonetic and psycholinguistic research for years; the extensive body of research has resulted in several word recognition models. A few well-known models are (in chronological order) the Logogen model (Morton 1969), the Cohort model (Marslen-Wilson & Welsh 1978), the TRACE model (McClelland & Elman 1986), the Shortlist model (Norris 1994) and the Merge model (Norris et al 2000). Parameters in which the models differ are, firstly, the assumptions about the nature of the elements that make contact with the mental lexicon (e.g. phonemes in the Logogen model but sound features in the TRACE model) and, secondly, the presence or absence of a feedback possibility (top-down flow); the interactive TRACE model allows both bottom-up and top-down information streams whereas the Shortlist model is of the autonomous type and allows bottom-up information only.

The models mentioned here are based on the word recognition process in the native language. The word recognition process in a non-native language has to deal with the incongruence between native and non-native language elements. It is only recently that researchers have begun to work on an explicit non-native word recognition model. So far this research resulted in the BIA (Bilingual Interactive Activation) model for visual non-native word recognition (Dijkstra & van Heuven 1998; Dijkstra, van Heuven, &
Grainger 1998) and the BIMOLA (Bilingual Model of Lexical access) model, which is still under development, for non-native spoken word recognition (Grosjean 1988, 1997).

The fourth sub-process of the listening skill is the **comprehension** (and/or interpretation) of the stream of sounds that is uttered by the speaker. When the incoming words are recognized, they are assigned to grammatical categories (various classes of content words and function words) and the structural and semantic relationships between the words are established: the input gets parsed. Because of the redundancy in communication it is not necessary for the listener to make a precise and exhaustive grammatical analysis; moreover, the rate of delivery and the limited processing capacities of the listener make it rather impossible for language users to construct detailed grammatical analyses. The parsing process critically depends on the learner’s knowledge of the language.

That **knowledge of the language** influences the comprehension process seems obvious; it is only when language users know words and grammatical structures of a language that they can recognise them.

The comprehension sub-process does not only depend on lexical and grammatical, but also pragmatic knowledge and ‘knowledge of the world’.

As far as **pragmatics** is concerned, there are three key notions that play a role here (Rost 2002):

(i) **Deixis.** Deictic elements are often employed by a speaker, and interpreted by the listener, as elements that do not directly – but only indirectly – refer to objects in the ‘real world’. Consider the following discourse: *Beatrix is the queen of The Netherlands. She lives in The Hague.* Here the proper noun *Beatrix* holds a direct and constant correspondence to a person in the real world. The interpretation of the deictic element *she* in the second sentence is indirect and variable. To resolve the referent of *she*, the listener first has to establish that *she* is co-referential with *Beatrix*. Deictic elements are commonly used for making variable, context-dependent reference to, for example, moments in time (e.g., *yesterday, now*), locations in space (e.g., *here, there*), and objects in the time-space continuum (through pronouns).

(ii) **Intention.** The notion of intention refers to the speakers’ goals when they produce an utterance; they want to have some kind of influence on the listener, and act in a certain way to accomplish the desired influence. Generally the speaker’s intention is completely determining the choice of words, the grammatical structures used
and sometimes also the gestures. There are situations, however, where an inferential leap is required on the part of the listener in order to correctly interpret the speaker’s intention. Irony or sarcasm typically has to be inferred from the context. Or, to give one more example of inferential work that may have to be done by the listener, a speaker saying ‘*It is rather cold in here*’ should actually be interpreted as issuing a subtle command for the addressee to get up and close the window (Searle 1975).

(iii) **Strategic use.** Speakers generally have multiple options at their disposal to accomplish their goal. It is an integral part of the comprehension sub-process that the listener does not only recognize what it is that the speaker wishes to accomplish (i.e., his intention) but that he is also aware of the way the speaker uses to reach this objective, for instance by being authoritarian, by expressing anger, by being friendly or polite, etc. The so-called indirect speech act illustrated under (ii) above is a clear example of a speaker-strategy that requires a lot of inferential work on the part of the listener.

**Knowledge of the world** does also play a role in the comprehension and interpretation of speech. For a good interpretation of direct speech, and certainly for the interpretation of indirect speech, it is important that the listener and the speaker share background knowledge. Effective recovery of background knowledge and use of the pragmatic keys described above will optimise the listening process. Notions that are linked with the activation of the listeners’ background knowledge are ‘schema’ and ‘script’.

Schemata are general cognitive constructs of knowledge that listeners use to model current events and situations and thereby bring them into alignment with familiar patterns of experience (Widdowson 1983). Listeners use schemata to order and categorise information that makes it possible to interpret and remember this information. A more detailed cognitive framework than a schema is called a script. Scripts, then, can be conceived of as specific modules within a larger schema. A script is typically procedural and comprises the time-course of an event. A well-known example illustrating the notions schema and script is the act of having dinner in a restaurant. Most listeners have some notion of events that typically happen in a restaurant; the knowledge of the environment (tables, napkins, food, check) can be marked as a schema, the knowledge about the order of an event (waiter asks what one wants to eat, after the meal he or she will give the check) is a script.

Every time a language user is engaged in reading or listening, existing schemata and scripts are updated or new ones are made. The many
schemata and scripts adults have are interrelated and their level of activation is influenced by, for example, the frequency and the recency of their use.

Bransford and colleagues explored the importance of prior knowledge and the processes of remembering and comprehension. Bransford and Johnson (1973, in Harley 1998), for example, showed that ideas that were presented in a read-aloud text, were better recalled when context was given before the reading, in other words when the correct schema and scripts had already been activated. They also showed that remembering the ideas of a text is facilitated when the context is very familiar to the listeners. Here we assume that better recall was caused by better comprehension. More direct evidence for the importance of activating scripts and schemata was provided by Sulin and Dooling (1974, in Harley 1998), who showed that activated background knowledge may interfere with the comprehension of an utterance; it can make listeners remember ideas that were not presented. Schemata and scripts enable listeners to comprehend references to people and events within the cognitive framework that were not explicitly mentioned by the speaker. This way the communication between speaker and listener is made more efficient, accelerating and facilitating the comprehension process.

2.3. The L2 listening process

There are obvious differences between acquiring listening comprehension skills in the first and in the second language. The acquisition of the L1 listening comprehension skill happens largely at the same time that the child develops its general cognitive abilities. The development of the L2 listening processes occurs after the learner’s cognitive development is more or less completed. When acquiring a foreign language, learners have to learn mainly what the L2-words and grammatical forms refer to (Gathercole & Thorn 1998); they have to acquire the differences between their L1 and the L2. In this light the idea of ‘conceptual redeployment’ has to be mentioned (Churchland 1999).

Conceptual redeployment can be defined as the process in which a fully developed conceptual framework that is in regular use in some domain of experience or comprehension, for example L1 comprehension, comes to be used for the first time in a new domain, for example in L2 comprehension. In the case of L2 learning, adults must reorganize semantic knowledge of words and concepts into a new domain of language use,  

---

2 This of course does not hold for bilingual children. The processes described here are based on L2 learning in a monolingual environment.
namely in their L2. The L2 learners already have cognitive frameworks but these have to be filled out with new (L2) language forms. The fact that new references (of known concepts) have to be learned and the change of conceptual frameworks while the cognitive development of the learner has already been developed makes learning a second language a more conscious process than L1 acquisition.

For successful second language learning it is important that there is enough correct language input. The kind of input that second language learners receive has been the object of research for several years. The speech input directed by L1 speakers towards L2 learners, which is often referred to as ‘foreigner talk’ shows some similarities with the modified input of child-directed speech; adjustments are noted at several linguistic levels. For example the phonology and the phonetics of foreigner talk differ from regular L1 talk in that there is more use of pauses, that there are fewer contractions and that there is a wider pitch range.

The discussion on the effect of this modified input and the way L2-learners deal with non-modified input has not been ended yet. In L2-research four different approaches of dealing with input can be distinguished: (i) studies that focus on the frequency of linguistic structures in the input, (ii) studies that focus on discourse and the way discourse construction is linked to acquisition, (iii) studies that relate the input to the output of L2-learners, and (iv) studies that focus on the comprehensibility of the input (see Ellis 2000 for a detailed description of these kind of studies).

A theory about the relation between input and improving comprehension is the ‘Input Hypothesis’ (Krashen 1982, 1985). The Input Hypothesis says that learners naturally develop their understanding and comprehension of the language by understanding only input that is slightly above their current language proficiency level. If the current level of the language learner is denoted as $i$; then the next-higher level $i+1$ can be reached by training with understanding and comprehension input at level $i+1$. Input materials at the current level $i$ contribute no new information, whereas input at level $i+2$ would be too difficult to understand at all. Lack of

---

3 Ellis (2000) gives an overview of similarities and differences between child-directed speech and foreigner talk.
4 The Input Hypothesis further claims that comprehensible input is not the only requirement for language learning: learners need to be willing to ‘let the input in’ they comprehend. The hypothesis claims that fine-tuning is not necessary; simplification and the help of extra-linguistic information and context make the input comprehensible. A last point of the theory is that learner production does not contribute to acquisition. (see Ellis 1994 for a discussion of studies for and against the Input Hypothesis)
empirical support (and even of testability) of this theory is a source of criticism.

As was made clear in the previous paragraphs, acquiring listening comprehension in an L2 is a different cognitive process than the acquisition of listening skills in one’s mother tongue. When learning a second language, learners already know how to listen in their L1; listeners have (unconscious) knowledge of (L1) listening strategies and their effects on the listening comprehension process. They can use some of these strategies, schemata and scripts while listening to the L2. In fact, for beginning language learners this higher-order knowledge is the most important source they have. By leaning on these mainly higher-order strategies, there will be less focus on the exact linguistic features of the utterance, i.e., less focus on the input itself. It is thus not the case that all input contributes to the learner’s knowledge and understanding of the language; not all input becomes ‘intake’. Sharwood Smith (1986) makes a distinction between input that helps the learner interpret the utterance and input that learners use to expand their language knowledge. Only the latter kind of input may become intake. Ellis (2000) defines intake as ‘information stored in temporary memory which may or may not be subsequently accommodated in the interlanguage system’. Chaudron (1985, in Ellis 2000) describes intake as a process that mediates between target language input and the learner’s internalised set of rules.

As was made clear, language learners may lean on higher-order processes as well as on lower-order processes while processing language input. What now is the importance of these two kinds of processing in language learning? Research has been done to establish the relative importance of lower-order versus higher-order processes, and to determine the relationship between these processes and overall listening proficiency. Tsui and Fullilove (1998, p. 433) state that: ‘While some studies have found that less-skilled readers/listeners are deficient in top-down processing skills, others have contradicted this, citing evidence that, in fact, less-skilled readers/listeners lack bottom-up processing skills.’ The authors looked at the results on (part of) The Hong Kong Certificate of Education Examination (HKCEE) English Language, in order to establish the importance of bottom-up versus top-down processes as a discriminator of listening performance, and this way found out what kind of processes are typically used by skilled listeners. The material had the design of the paradigm often used in listening research and education, namely a listening text with multiple-choice questions. There were furthermore, two kinds of schemata: the first type of schema activated by the initial input was consistent with the following input; in the second type, the schema activated in the beginning was not consistent with the subsequent input. In this latter case, participants had to be able to revise the schema in order to give a correct answer; here the bottom-up
processes are important. In the case of the ‘matching’ schemata, participants can rely largely on top-down processes to give a correct answer. The authors concluded that bottom-up processing was more important than top-down processing in discriminating the listening performance. They state that ‘... on the one hand because less skilled L2 listeners are weak in bottom-up processing, initially they need plenty of contextual support to compensate for the lack of automatised linguistic decoding skills. On the other hand, they need to learn to become less reliant on guessing from contextual or prior knowledge and more reliant on rapid and accurate decoding of the linguistic input.’ (p. 449). In other words, as the decoding (or recognition) process becomes more automatised, the reliance on the higher-order processes may become weaker. This indicates that skilled learners make use of both top-down processes and (automatised) bottom-up processes, less skilled learners lack the automatisation of the lower-order processes.

In conclusion we can state that both bottom-up and top-down processes are important for successful listening comprehension. In second language learning, the difference between skilled and less-skilled learners lies in the status of the lower-order processes (ranging from more controlled to fully automatised). This implies that to improve the listening comprehension skills of second language learners, the focus has to be on lower-order processes.

2.4. Problems in listening

Listening problems (that result in miscommunication) may have several different sources; the problem can be pure physical, for example a damage of the inner ear, the problem can be linguistic, for example a problem with the recognition of speech sounds, or the problem can be of a more abstract cognitive level, for example with the interpretation of the message.

Let us now describe in more detail the problems L2-learners may have in executing some specific sub-processes of listening comprehension. A first sub-process that can be a source of problems is the identification of L2-phonemes. One popular view is that native speakers identify speech sounds due to the so-called perceptual magnet effect, which means that they recognize sounds by comparing them to prototypical phonemes (Rost 2002). These prototypes lie within a range of allowable variation; as long as the characteristics of the input sound lie within this range, the sound can be categorized correctly. Lack of knowledge of the L2 prototypes that leads to wrong categorisation, could be a first problem for second language learners in processing spoken language; if they do not have enough knowledge of the L2-specific phonemes, they will not be able to categorise the input sounds
correctly. Clear examples are the difficulties Chinese listeners have with the categorization of (English or Dutch) /l/ versus /r/. These two sound categories are not contrastive in Chinese as they are in English (e.g. lead ~ read). Since Chinese people do not observe the contrast in their mother tongue (they hear an unspecified liquid), it is only possible for them to recognize these sounds after intensive practice.

Of course, not all L2 sounds will cause problems. A distinction can be made between three different classes of L2 speech sounds, namely (Flege 1987):

(i) **Identical sounds.** These sounds of the L2 are the same as the sounds of the L1; L2 learners have nothing to learn. An example of an identical sound in Dutch and English is the lax high-mid front vowel /u/.

(ii) **Similar sounds.** These sounds are not the same in L1 and L2; in terms of IPA transcription they share the same base symbol and differ in a diacritic only. An example is the /s/ in Dutch and English, which has intensity above 5 KHz in English (having a sharp timbre) but not in Dutch (leading to a duller timbre causing confusion with English /ʃ/). Subtly incorrect pronunciation of similar sounds will always give the L2 learner away as a non-native; it is believed that even long exposure and intensive training is of no avail here.

(iii) **New sounds.** These sounds show a substantial difference between the languages. At first, learners do not know that they have to pronounce a sound that they do not have in their native language. After a period of time (and intensive practice), the sound will become established as an extra category. An example is the English /æ/ like in hat ‘hoed’ which is a new category that has to be established by Dutch learners in between Dutch /ɛ/ as in pret ‘fun’ and Dutch /ɑ/ as is praat ‘talk’. New sounds are a source of difficulties that L2-learners experience, which may be overcome by setting up new, or redefining existing, prototypes and ranges of allowable variation for them. This may come about only after prolonged exposure to the L2 and/or explicit feedback.

---

5 The split of the undifferentiated liquid category into /l/ versus /r/ as is required of Chinese learners of English, would be yet another example of setting up ‘new’ sound categories.
Flege (1987) states that language users learn to categorize speech sounds in terms of phonetic categories of their L1 before the age of seven (in fact, they begin at birth – or even earlier). When the categorization is completed, new sounds (for example L2 sounds) will preferably be categorized in one of the existing and relevant L1 categories. Young children are still able to ‘create’ new categories for the L2 sounds whereas older children and adults are not.

A second sub-process that makes listening comprehension so difficult is the recognition of the spoken words. A characteristic of continuous speech that makes listening comprehension so difficult is that speech contains no clear auditory equivalent to the inter-word white spaces that we find in written text. The lack of clear word boundary markers in languages such as Dutch makes it hard to determine word beginnings and endings (the segmentation problem). It is important to realize that sounds in continuous speech are fluently coarticulated and assimilated, not only within words but also across word boundaries. These so-called Sandhi processes are only blocked at higher-order linguistic boundaries, at the level of the Intonation Phrase boundary (and higher) in Dutch (Berendsen 1986). The lack of explicit word boundary markers and the distortions of word shapes due to assimilation and reduction are the reasons why words that are known by the language user when presented visually, are often not recognized when they are part of continuous speech. Although phenomena like Sandhi and co-articulation are efficient for the speaker, since they facilitate pronunciation, the lack of word-boundary markers can cause misunderstanding or even the complete breakdown of word recognition.

Problems can also be caused by a form-meaning mismatch. If a listener identifies the form of a word correctly but has not enough knowledge of its meaning, the recognition process (and ultimately the communication process) will fail accordingly. Word-recognition problems, whether caused by faulty word-boundary identification or by insufficient lexical knowledge, are major sources of miscommunication in listening comprehension; this is, as is discussed, particularly the case in the second language listening comprehension process.

On the comprehension level researchers point out the effort for ‘complete understanding’ as a characteristic of L2 learners that causes problems

---

6 Implicit word-boundary markings of a subtle nature are sometimes available in spoken language. For instance, there are languages that use stress to mark the beginning or end of a word. Hungarian, for example, has main stress on the first syllable of the word whereas Dari stresses the final syllable.

7 In this context form can be defined as an auditory presentation of the concept.
(Fearch & Kasper 1986, Ur 1984). L2 learners often try to get a complete understanding of the utterance, anxious to miss any aspect of the message. A characteristic of a good L2-listener is the ability to distinguish between important and less important aspects of the utterance, as is also done in L1-listening. In languages such as Dutch and English prosody is used to signal the position of important parts of the sentence. Typically important words are pronounced more carefully and more slowly (some 10 percent, Nooteboom & Eefting, 1991; van Heuven 1998). Moreover, the prosodic head (stressed syllable) of important words is associated with a perceptually conspicuous change in vocal pitch (accentuation). Native listeners use these changes in speaking rate and pitch to predict the positions in the utterance where important words and syllables will be spoken. Experiments have revealed that (English) listeners attend more forcefully to bottom-up information at points in time where accented syllables are expected (Cutler, 1976; Cutler & Darwin, 1981). When the L2 listener is not familiar with the prosodic system of the target language the attention-focussing mechanism will not be employed. As a result the L2 listener will continually direct all his resources to bottom-up signal information, leaving less capacity for the recoding of the recognized units to a more abstract (semantic) representation. The importance of the division of work over the various modules in memory will be addressed in the next section. For now, we conclude that problems in L2-listening can be due to lower-order skills (recognition) as well as higher-order skills (comprehension).

In the two remaining sections of the present chapter we survey the literature on working memory and automatisation, respectively, in so far as these concepts are relevant to our study.

### 2.5. Memory

This section first briefly considers working memory (WM) as a component of memory in general and then focuses on the components that make up working memory, giving special attention to the so called phonological loop and the measurement of working memory capacity.

According to Carroll (1994), human memory can be divided into three units: the sensory stores, working memory, and long-term memory. Figure 2.2 illustrates the relations between the units (based on Carroll 1994).
The sensory stores take in the colours, tastes and tones we experience and retain these for a short period of time in an unanalysed form. The information of the sensory stores is refreshed and analysed in working memory. Long-term memory (LTM) is the unit that contains, amongst others, knowledge of the world. The functioning of working memory and the relation between WM and LTM is discussed below.

As Figure 2.2 shows, the input goes via the sensory stores and working memory to long-term memory. That WM is not just a gateway for information towards LTM, is indicated by the arrows from LTM to the other two units, they point out that knowledge that is stored in LTM may influence the processes of the sensory stores and WM. Evidence for LTM influence on the other units is given by for example Hulme, Maughan and Brown (1991). They showed that working memory span is better for words than for nonwords as it is easier to remember words than nonwords. Furthermore, Gathercole (1995) showed that memory span will be influenced positively not only by concrete concepts of words but also by knowledge about the structure of the language. She found that nonword recall was better for more wordlike nonwords than for less wordlike nonwords. Some theories therefore regard WM not as a unit separate from LTM but as that part of LTM which is currently in a state of activation beyond a certain threshold (Miyake & Shah 1999).

2.5.1. Working memory

The term working memory is used in several different ways in different areas of cognitive science (Miyake & Shah 1999 provide an in-depth
presentation and discussion of eleven WM models). In this present thesis the
definition of Baddeley is adopted: Working memory refers to a limited
capacity system of temporary storage and manipulation of input that is
necessary for complex tasks such as comprehension and reasoning (Miller et
al. 1960; Baddeley & Hitch 1974). The concept of WM is based on the
concept of Short-term memory (STM), the biggest difference between them
being a matter of dynamics. STM has mostly been considered as a passive
repository while WM has been attributed both a storage and a processing
function (Carroll 1994). Baddeley and Hitch (1974) proposed a WM model
that has been developed through the years to the most recent model shown in
Figure 2.3 (Baddeley 1986; Baddeley 2000).8 The three components of the
model presented are the visuo-spatial sketchpad, the episodic buffer, and the
phonological loop.

![Diagram of Working Memory Model](image)

**Figure 2.3: The working memory model (Baddeley 2000).**

As can be seen in Figure 2.3, highest in the hierarchy of working memory is
the central executive, which is a supervisory controlling system. The central
executive is assumed to be modality free; it acts as a link between three
peripheral slave systems. The visuo-spatial sketchpad and the phonological
loop are modality specific; the visuo-spatial sketchpad deals with visual

---

8 The largest difference between the revised model and the original memory model
is the focus on the processes of integrating information. The original model focused
on the isolation of the subsystems.
The episodic buffer is a new component in comparison with the Baddeley WM model of 1986. The need for a new storage system that is capable of dealing with all kinds of input was felt, amongst other things, by the lack of explanation for the recall of prose effect. This effect indicates that humans are able to recall 15-20 idea units, if these units are related. However, for the recall of unrelated words, failures occur once the number of words exceeds five or six. An explanation for this contrast is that in the case of related words, knowledge from LTM is involved. This indicates that information from various sources is used during the processing of related words: LTM, and, depending on the presentation mode of the items, phonological loop or visuo-spatial sketchpad. The combined information has to be stored in some kind of system. In the new WM model, the episodic buffer performs this back-up storage. Baddeley (2000) assumes the episodic buffer to be: ‘... a limited-capacity temporary storage system that is capable of integrating information from a variety of sources. It is assumed to be controlled by the central executive, which is capable of retrieving information from the store in the form of conscious awareness, of reflecting on that information and, where necessary, manipulating and modifying it. The buffer is episodic in the sense that it holds episodes whereby information is integrated across space and potentially extended across time. (...) it is assumed to be a temporary store (...) It is (...) assumed to play an important role in feeding information into and retrieving information from episodic long term memory (LTM).’ (p. 421). Figure 2.3 shows that the central executive and the episodic LTM may indeed interact. The latter is able to exchange information with the phonological loop and the visuo-spatial sketchpad via the language and visual semantics respectively.

As stated above, an important aspect of WM in (second) language listening is the phonological loop, since this is the component that deals with speech input. This component will therefore be described in detail in the following section.

**Phonological loop**

In every day life we often have to hold in mind non-visual and non-spatial information, like for instance a telephone number or a person’s name, for a short time. We try to facilitate this task by constructing a phonological representation of the input and by rehearsing this construct (out loud or in silence). The rehearsing process takes place in what is called the Phonological Loop. The Phonological Loop consists of two components:
(i) **a storage system**, i.e., the system for the representation of verbal information (in phonological or phonetic form) and

(ii) **a rehearsal mechanism**. The incoming material is encoded in the phonological store according to its sound-based characteristics. The auditory memory traces that are comprised in the temporary store decays within two seconds unless they are re-activated by the rehearsal mechanism.

Figure 2.4 shows a model of this Phonological Loop (Baddeley & Hitch, 1974, Baddeley 1986).

![Figure 2.4: Baddeley's (1986) model of the phonological loop.](image)

In essence, the phonological loop is knowledge-free as phonological information is kept in the store, regardless of its correspondence to a familiar or an unfamiliar sound pattern. However, experimental work shows that the loop (and indeed the WM as a unit) is linked to language knowledge, and, as described in the previous subsection, the bridge between knowledge and the input is formed by the episodic buffer.

Newport’s ‘less-is-more’ hypothesis (1990) states that the benefit language learners have of a larger WM, forms, at the end, an obstacle for reaching high levels of achievement. This hypothesis may offer an explanation for the so-called critical period phenomenon, pertaining to an age-related decline in ability to acquire a foreign language (or even a first language). The argument Newport gives in favour of her hypothesis is that their large WM makes adult language learners focus on the input as a whole. As a result they do not make a detailed analysis of the encoded elements. The limited WM of children, however, makes it impossible for them to
process the input as a whole. They are forced to process the input in detail. According to the less-is-more hypothesis this eventually leads to a higher ultimate attainment, despite the fact that the acquisition process will initially be slower. Children’s detailed input analysis causes a better understanding of the internal structure of the language to be learned. However, as Miyake and Friedman (1998) state, the less-is-more hypothesis does not mean that a large WM is no longer important. The less-is-more principle compares the WM of young children with that of adults, while the theories discussed earlier in this section compare the WM capacities among adults. Research shows that some adult second-language learners are able to achieve a very high level of performance (for example in pronunciation), even if they started to learn the language at a later age (Bongaerts et al. 1997; Bialystok & Hakuta 1994). A larger WM might be the distinctive feature between the adults that do reach this near-native level and those that do not.

Experimental research indicates that the phonological loop has an important role in (i.e., highly correlates with) L1 vocabulary acquisition (Gathercole & Baddeley 1989) as well as in L2 vocabulary acquisition (Service 1992). Results of research with bilingual and monolingual children show that familiarity with a language influences the functioning of the phonological loop in a positive way (Gathercole & Thorn, 1998). Taken together, these studies indicate that the capacity of the loop seems to be dependent on the familiarity with the sound of the item and on (L2) language proficiency (Miyake & Friedman 1998). Therefore, the best way of testing the loop in order to establish its basic capacity is to use a test based on a language unfamiliar to the test taker. The absence of long-term representations of the items will make the outcome a more pure measure of the phonological loop.

2.5.2. Measuring working memory capacities

In the previous section it was said that both functions of WM, storage and analysis, play an important role in language processing. Since WM capacity is assumed to be limited, storage and processing have to compete with each other and trade-off effects might be expected (Daneman & Carpenter 1980). Since both storage and processing are important in WM it is important that

---

9 A study conducted by Geva & Ryan (1993) shows that the correlation between L2 WM and L2 proficiency is not just a reflection of general intelligence. The results of this study also suggest that the role of WM in L2 is stronger than in L1.

10 In our Training Study, we have chosen for a measure using unfamiliar words (nonwords) but familiar phonemes and phoneme combinations. See Chapter 5 for details.
both aspects are involved in measures of WM capacity. The two components that are always present in WM tests are:

(i) a measure to test whether the input information has been processed to a relevant degree, and

(ii) a measure of the information that has to be stored and remembered over the course of a test trial.

In most cases the so-called WM span is then operationalised as the maximum amount of information that the test taker can remember. In other words, the span is based on a measure of the second component. In this section several tests will be described that have been reported in the literature. The descriptions will be limited to tests of the WM as a whole, and tests of the phonological loop.\textsuperscript{11}

A test that is typically used as a measure of Working Memory is the Reading Span test, originally developed by Daneman and Carpenter (1980). In the test both components of working memory are involved: the processing component for dealing with the current sentence and the storage component for storing words to be remembered. In each test trial, participants are given some sentences that are presented one by one. Participants are asked to read them out loud and to try to remember the last word of each sentence. Reading involves the processing component of WM, while rehearsal of the sentence final words involves its storage component. Participants are told that the number of sentences per trial increases over the course of the test. Their reading span is defined as the maximum number of sentences that they can read while correctly recalling the last word. It is scored in a simple way by calculating the number of words correctly recalled.

There are several adapted versions of the reading span test. Differences between the derived and the original tests can be found in the modality of input presentation (e.g., written vs. oral) or in the sentence processing assignment (e.g., reading aloud vs. making a semantic judgment on sentence content). In the Training Study of this thesis a listening span test was constructed and used as a measure of working memory capacity differing from Daneman and Carpenter’s original reading span test in terms of modality. A detailed description of our listening span test is given in Chapter 5.

Examples of tests that are used for measuring the capacity of the phonological loop are the nonword repetition test, and the serial recognition test (Gathercole 1999; Gathercole et al. 2001). These tests rely on the storage

\textsuperscript{11} See Gathercole (1999) for tests to measure the visuo-spatial component.
and rehearsal mechanism of the phonological loop and are based on the same principle as the reading span test:

(i) participants have to process the information to be able to give a correct answer, and
(ii) the amount of information that has to be processed increases over the course of the test.

In a digit span test, series of digits are presented which the participants have to repeat while in a nonword repetition test participants repeat series of nonwords. In a serial recognition test, a series of words or nonwords is presented twice. The order of the words (or nonwords) in the second presentation, however, may or may not remain the same as the order in the first presentation. Participants have to indicate whether they think the order is the same or different. An advantage of tests that makes use of nonwords is that they rely on LTM to a much lesser degree than when existing words are used. An additional advantage of the serial recognition test over a repetition test is that performance is not influenced by articulatory and phonological output abilities. Because of these assets a serial recognition test was used in the Training Study of this thesis. A detailed description of this serial recognition test is given in Chapter 5.

2.6. Automatisation

In this section the concept of automatisation is described. Section 2.6.1. gives some theoretical background and addresses problems that arise in defining the concept of automatisation. Section 2.6.2. presents the coefficient of variability, a possible gauge for measuring automatisation.

2.6.1. Theoretical background

Every time we try to learn a new skill (e.g., learning a new language, learning to play an instrument) we have to deal with the fact that, in the beginning, it is a time consuming and effortful process. However, we know that when we practice a lot, the process will become easier. After a certain period of (intensive) practice, the process runs off without effort, without awareness, and much faster than in the beginning. The process has then become automatised.

Notwithstanding an extensive literature on automaticity of cognitive processes, it is hard to find a clear definition of what automatisation exactly
is. One of the earliest definitions of an automatic process is given by Posner & Snyder (1975): ‘... the process occurs without intention, without giving rise to any conscious awareness, and without producing interference with other ongoing mental activity.’ (p. 56). Several characteristics have been proposed to distinguish between automatic and controlled processes. The typical criteria were: the speed of the processes, the error rate, the effort required, the capacity used, and the level of possible control. Later, researchers rejected several of these criteria as unnecessary or insufficient to make a distinction between the two kinds of processes. The last twenty years have witnessed several developments in automatisation research. First, the idea emerged that controlled and automatised processes are not mutually exclusive. Most researchers now posit a continuum of automatic processes at one extreme, controlled processes at another, and many processes of a mixed nature in between. Second, a shift occurred from theories that reflect automatisation in relation with how much attention was required to theories that present automatisation as a reflection of how memory is used. Third, views changed with regard to when a process has become automatised. Some researchers ascribe automatisation to the simple speed-up of component processes (Carlson et al. 1989; Logan 1988) while others define automatisation not as a simple speed-up (a quantitative change) but as a qualitative change in the composition of component process (Segalowitz & Segalowitz 1993). The emphasis that Segalowitz lays on the qualitative change makes his position on automatisation rather unique (Segalowitz & Segalowitz 1993; Segalowitz & Gatbonton 1995; Segalowitz, Segalowitz & Wood 1998; Segalowitz 2000). Segalowitz, nevertheless, subscribes to the practical importance of improved speed and accuracy for achieving fluency in language use. He makes a distinction between ‘performance fluency’, pertaining to the observable speed and accuracy of component processes, and ‘cognitive fluency’, pertaining to the efficiency with which various component processes are blended and balanced (Segalowitz et al 1998; Segalowitz 2000). After intensive practice the blend will change in favour of automatised processes. Performance will then become faster and more stable because the influence of or the reliance on the relatively slow and variable controlled processes is eliminated or reduced. However, when necessary, an otherwise automatic process, such as the pronunciation of planned utterances

---

12 There are some positions that lie between the two approaches mentioned here. Cheng (1985) for example describes automatisation as a reorganization of components, whereas both Shallice (1982) and Stanovich (1990) characterize automatisation as a shift from reliance on higher-order to a more independent processing mechanism, taking place in what Fodor (1983) calls ‘informationally encapsulated’ processing units. In such units, information processing finds place without involvement of higher- or lower-level information.
in a language that we command fluently, can be held under conscious control, e.g., when we want to pronounce certain words or sounds in an uncommon way. Or, in the case of receptive language use, we can, if the situation demands, pay special attention to the way someone else pronounces certain words or sounds, which would otherwise go by unnoticed.

To determine the place of automatisation in skill development, two theories are often cited in the literature on second language acquisition, namely the rule-based ACT-R theory of Anderson (1993) and the item-based Instance theory of Logan (1988).\(^\text{13}\) Important concepts of the ACT-R theory are declarative knowledge (e.g., the knowledge that ‘work’ is a verb and that ‘Petra’ is a proper noun) and procedural knowledge. Procedural knowledge has the form of production rules of the if \(x\) then do \(y\) type (e.g., if the subject of a sentence is taken by a proper noun with the features third person and singular, and if the predicate has the feature present tense, the verb takes the suffix -s, leading to the computation of sentences like ‘Petra works’). The idea is that in early skill acquisition, performances are based on production rules to retrieve the relevant knowledge out of declarative memory and putting them together in working memory. Due to practice, components of the skill that are repeated can become routinised, forming chunks. These chunks can be retrieved fast and efficiently; they are used relatively independently of conscious control. Thus, procedural knowledge can be said to become increasingly automatic.

In Logan’s (1988) Instance theory, automatisation means a shift from algorithm-based performance to memory-based performance. In this theory ‘performance is automatic when it is based on single-step direct-access retrieval of past solutions from memory’ (p. 493). Each time the algorithm is used, a new memory trace corresponding to the rule is formed. The more the rule is used, the more memory traces there will be. Each time an action has to be performed, there will be a race between the original algorithm and the memory traces. This race will eventually be won by direct access to memory traces. The raised activation levels of the memory traces, being the result of increased experience, will make it more efficient to retrieve the traces than to execute the algorithm itself.

2.6.2. Measuring automatisation

How now to measure automatisation? All theories of automatisation regard speed of processing as an important indicator of automatisation. This can be

\(^{13}\) The ACT-R theory was preceded by the ACT* theory (1983) that was preceded by the original ACTE theory (Anderson, 1976). ACT stands for Adaptive Control of Thought.
reflected in the speed with which a behavioural response is given to a stimulus. However, Segalowitz and Segalowitz (1993) claim, as stated above, that simple speed-up is not a true sign of automatisation since it does not necessarily involve a restructuring of sub-processes. In their 1993 paper they propose the ‘coefficient of variability’ (CV_{RT}) as a measure for automatisation. This ‘coefficient of variability’ is defined as the standard deviation (SD) of reaction times divided by the mean reaction time (RT) of responses in, for example, a lexical decision task. It is a measure of variability that is corrected for the latency of responding. In case of a mere speed-up, the standard deviation (SD_{RT}) is reduced proportionally to the reduction of the RT. Thus, there will be no change in the relative variability of RT (CV_{RT}). In the case of restructuring, however, some time-consuming controlled processes will be eliminated or modularised. For example, skilled readers do not pay attention anymore to the form of letters; they can quickly recognize words like bad, dab, dad, and baby, without paying attention to the difference in form of the letters ‘b’ and ‘d’. Children learning to read, however, have to invest considerable time and effort to master the automatic recognition of these letters, paying special attention to the components of their forms. Skilled readers will omit these initial component processes. Consequently, not only will the overall RT of word recognition be reduced but also so will the SD_{RT} be reduced (more than proportionally). As a result, the CV_{RT} will reduce as well. The reduction of the CV_{RT} can be regarded as a first criterion for determining automatisation. In addition to this CV_{RT} reduction, another criterion has to be met before the conclusion that a process became automatised can be drawn. There has to exist a positive correlation between RT and CV_{RT}. Because in the case of restructuring (automatisation) every reduction in RT will be accompanied by a more than proportional reduction in SD_{RT}, CV_{RT} should be positively correlated with overall RT. This means that a reduced CV_{RT} combined with a positive correlation between CV_{RT} and RT can be seen as an index of automatisation.\footnote{Before Segalowitz started to use CV_{RT}, this index was not generally used in cognitive psychology research. It had, however, been used in other areas such as the study of action patterns in animal behaviour.}

Let us take as an example a spoken word recognition task that is performed by a number of individuals a number of times in a period of several days. If automatisation were solely a matter of speed-up, then we would expect a shortening of reaction times during the course of the experiment. If, however, automatisation were a matter of quantitative as well as qualitative change, we would expect that certain component processes,
such as the recognition of the sounds and phonemes, would be reduced, compiled or even omitted altogether. This would result in faster reaction times but also in a more than proportionally reduced SD. As a result there would be a decline of the $CV_{RT}$. In this thesis we will use the $CV_{RT}$ to make a distinction between simple speed-up and automatisation (see Chapter 5).
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.  Fig. 3.1b Non-overlapping populations.
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 automatized 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 automatized and non-automatized 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 automatized are now included in the automatized 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 automatization 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 automatized word recognition process. In summary, the research question of the present study can be formulated as follows: is there a criterion to distinguish between automatized and non-automatized 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\(^1\); 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

\(^1\) 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 disproportionately 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:

*Ik heb de sleutel van het HUIS nodig, niet van de GARAGE.*
‘I need the key of the HOUSE, not of the GARAGE.’

*Is dat de pagel van JAN of van PIET?*
‘Is she the pagel of JAN or of 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 ∆ values), i.e., reactions on words minus reactions on nonwords.

<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>∆</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 (\(\Delta\)) that are significant are marked by ‘*’.

<table>
<thead>
<tr>
<th></th>
<th>Overarticulated</th>
<th>Underarticulated</th>
<th>(\Delta)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RT % cor.</td>
<td>RT % cor.</td>
<td>RT %</td>
</tr>
<tr>
<td>Words</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L1</td>
<td>906 (127)</td>
<td>99 (3)</td>
<td>954 (190)</td>
</tr>
<tr>
<td></td>
<td>L2 1047 (183)</td>
<td>97 (6)</td>
<td>1134 (181)</td>
</tr>
<tr>
<td>(\Delta)</td>
<td>-141*</td>
<td>2</td>
<td>-180*</td>
</tr>
<tr>
<td>Nonwords</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L1</td>
<td>1136 (195)</td>
<td>89 (16)</td>
<td>1088 (218)</td>
</tr>
<tr>
<td></td>
<td>L2 1701 (369)</td>
<td>63 (21)</td>
<td>1453 (364)</td>
</tr>
<tr>
<td>(\Delta)</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 $CV_{RT}$ measure for automatisation. 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>$CV_{RT}$</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 $CV_{RT}$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 $CV_{RT}$ 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 $CV_{RT}$ 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), $CV_{RT} = .28$). The result now shows the expected correlation between $CV_{RT}$ and mean RT.\(^4\) 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).
Table 3.4: Results of the application of the CV\textsubscript{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\textsubscript{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\textsubscript{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\textsubscript{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.
CHAPTER 3

Figure 3.2: Scatter plot of listeners in the two-dimensional plane defined by Mean RT and CV<sub>RT</sub>

Since we had to exclude a certain part of the data in order to get a significant correlation between CV<sub>RT</sub> and mean RT, and using the CV<sub>RT</sub> 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 automatization. Since we believe that besides speed, accuracy is also an important indicator of automatization, 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 pre-given 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
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 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 very few 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.
Chapter 4

Pilot study II:
The predictability of
listening comprehension

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>Poor</td>
<td>Poor (category #2)</td>
</tr>
<tr>
<td>Good</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.

![Test battery diagram]

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.

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

**English**
- Would you like to go to the movies tonight?
- No, I’ll be .......................at home.
  A) starting
  B) staying
  C) talking
  D) searching

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?

Dutch               English
A) hebt             A) weared
B) hebben           B) wears
C) heeft            C) wore
D) heb              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 intern 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 timeout. 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 administer 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 to 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>Overlapping errors</td>
<td>141 (35%)</td>
<td>141 (35%)</td>
</tr>
<tr>
<td>Mode Exclusive errors</td>
<td>209 (53%)</td>
<td>66 (17%)</td>
</tr>
<tr>
<td>Total number of errors</td>
<td>350 (88%)</td>
<td>207 (52%)</td>
</tr>
<tr>
<td>Total numbers of items x subjects</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 (200)</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, $t(19) = -0.13$, $p = 0.89$, nor for the native speakers, $t(19) = 2.09$, $p = 0.05$. The difference between the percentages correct, however, are significant: non-native speakers $t(19) = -4.61$, $p < 0.01$, native speakers $t(19) = -4.39$, $p < 0.01$. 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>df</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Auditory LD</td>
<td>19</td>
<td>-2.80</td>
<td>&lt; .05</td>
<td>19</td>
<td>10.37</td>
<td>&lt; .01</td>
</tr>
<tr>
<td>Visual LD</td>
<td>19</td>
<td>-4.54</td>
<td>&lt; .01</td>
<td>19</td>
<td>4.72</td>
<td>&lt; .01</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>Offset RT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.02</td>
<td>-.11</td>
</tr>
<tr>
<td></td>
<td>.22</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>% correct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>df</td>
<td>t</td>
</tr>
<tr>
<td>Words</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>19</td>
<td>.004</td>
</tr>
<tr>
<td>Overarticulated words</td>
<td></td>
<td>.31</td>
</tr>
<tr>
<td>Underarticulated word</td>
<td></td>
<td>-.40</td>
</tr>
<tr>
<td>Nonwords</td>
<td></td>
<td>5.88</td>
</tr>
<tr>
<td>Overarticulated nonwords</td>
<td></td>
<td>5.75</td>
</tr>
<tr>
<td>Underarticulated nonword</td>
<td></td>
<td>4.67</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: Distribution of participants by Percentage correct and Mean RT on visual presented nonwords.

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></th>
<th>Words</th>
<th>Nonwords</th>
<th>Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RT</td>
<td>% correct</td>
<td>RT</td>
</tr>
<tr>
<td>L1</td>
<td>1002 (299)</td>
<td>99 (2)</td>
<td>1208 (364)</td>
</tr>
<tr>
<td>L2</td>
<td>1130 (206)</td>
<td>97 (3)</td>
<td>1900 (360)</td>
</tr>
<tr>
<td>Δ</td>
<td>125</td>
<td>2</td>
<td>686 *</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>
<thead>
<tr>
<th>Mode</th>
<th>Word rec. RT</th>
<th>Word rec. %</th>
<th>Comp. %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visual</td>
<td>1436 (246)</td>
<td>85 (11)</td>
<td>71 (14)</td>
</tr>
<tr>
<td>Auditory</td>
<td>988 (204)</td>
<td>74 (8)</td>
<td>58 (13)</td>
</tr>
</tbody>
</table>

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 % correct</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>Audit. LD RT</td>
<td>−.03</td>
<td>−.10</td>
<td>.05</td>
<td>−.26</td>
<td>.29</td>
<td>.08</td>
<td>−.09</td>
<td></td>
</tr>
<tr>
<td>Visual LD % 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>.08</td>
<td>−.23</td>
<td>.08</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reading span</td>
<td>−.24</td>
<td>−.17</td>
<td>−.20</td>
<td>−.36</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reading comp.</td>
<td></td>
<td>.36</td>
<td></td>
<td>.56*</td>
<td>.62*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Listening comp.</td>
<td></td>
<td></td>
<td></td>
<td>.53*</td>
<td>.50*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Basic Grammar</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.81*</td>
<td></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>

* p ≤ 0.05

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>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Good</td>
</tr>
<tr>
<td>Good</td>
<td>Listening good 4</td>
</tr>
<tr>
<td>Poor</td>
<td>Listening good 4 b</td>
</tr>
<tr>
<td></td>
<td>Listening good 0</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.
Chapter 5  

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.
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.

Figure 5.2: Overview of the tests used in Test stage 1.
In the training study, 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 /\ep/ /\ip/ /\yp/ /\op/ /\ep/
    geep gip goop guup gep /\ep/ /\ip/ /\op/ /\yp/ /\ep/

(2) huf hul hur hug hus /\of/ /\ol/ /\or/ /\oy/ /\os/
    huf hul hur hug hus /\of/ /\ol/ /\or/ /\oy/ /\os/

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>X</td>
<td></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>X</td>
<td></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></td>
<td>X</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*
(1996). 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 5.4: Overview of overlapping parts between pre and posttests (number of items).

<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\(^3\); 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 choose 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*  
1) *Hij stopt voor een auto langs de weg*  
2) *Hij is onder een auto gekomen*  
3) *Hij verzamelt geld om een auto te kopen*  
4) *Hij verzamelt foto’s van auto’s*  
5) *Ik weet het echt niet*

<table>
<thead>
<tr>
<th>Dutch Description</th>
<th>English Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Hij stopt voor een auto langs de weg</em></td>
<td>1) He stops for a car along the road</td>
</tr>
<tr>
<td><em>Hij is onder een auto gekomen</em></td>
<td>2) He was run over by a car</td>
</tr>
<tr>
<td><em>Hij verzamelt geld om een auto te kopen</em></td>
<td>3) He collects money to buy a car</td>
</tr>
<tr>
<td><em>Hij verzamelt foto’s van auto’s</em></td>
<td>4) He collects pictures of cars</td>
</tr>
<tr>
<td><em>Ik weet het echt niet</em></td>
<td>5) I really do not know</td>
</tr>
</tbody>
</table>

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:
_an overtime hour 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 investeringsbegroting 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 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._

<table>
<thead>
<tr>
<th>Waar</th>
<th>Niet Waar</th>
</tr>
</thead>
</table>

‘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 /ø/-/y/ and /ø/-/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 VERSCHILLENDR
   ‘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 automatisation 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.
2 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. 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>Total</td>
<td>83</td>
<td>15</td>
<td>4</td>
<td>64</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></th>
<th>Basic Grammar</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mean (max = 60)</td>
<td>Percent correct</td>
<td>Mean (max = 53)</td>
<td>Percent correct</td>
</tr>
<tr>
<td>Recognition</td>
<td>21</td>
<td>48 (6)</td>
<td>81 (10)</td>
<td>44 (4)</td>
<td>84 (7)</td>
</tr>
<tr>
<td>Comprehension</td>
<td>31</td>
<td>47 (6)</td>
<td>78 (9)</td>
<td>44 (4)</td>
<td>83 (8)</td>
</tr>
<tr>
<td>Control</td>
<td>12</td>
<td>47 (5)</td>
<td>78 (8)</td>
<td>45 (4)</td>
<td>85 (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>Percentage correct</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>Percentage correct</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>Percentage correct</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 $\Delta$ 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 Mean (max = 27)</th>
<th>Percent correct</th>
<th>Posttest Mean (max = 27)</th>
<th>Percent correct</th>
<th>$\Delta$ Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recognition</td>
<td>21</td>
<td>17 (4)</td>
<td>63 (14)</td>
<td>19 (5)</td>
<td>69 (17)</td>
<td>2</td>
</tr>
<tr>
<td>Comprehension</td>
<td>31</td>
<td>16 (4)</td>
<td>59 (14)</td>
<td>17 (5)</td>
<td>63 (17)</td>
<td>1</td>
</tr>
<tr>
<td>Control</td>
<td>12</td>
<td>18 (3)</td>
<td>68 (11)</td>
<td>19 (4)</td>
<td>72 (16)</td>
<td>1</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 $r = .67, p < .001 \ (N = 64)$.

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>Posttest</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Pretest</td>
<td>Percent correct</td>
</tr>
<tr>
<td>Recognition</td>
<td>21</td>
<td>1300 (499)</td>
<td>86 (5)</td>
</tr>
<tr>
<td>Comprehension</td>
<td>31</td>
<td>1249 (430)</td>
<td>85 (6)</td>
</tr>
<tr>
<td>Control</td>
<td>12</td>
<td>1103 (274)</td>
<td>84 (8)</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 RT</th>
<th>Percent correct</th>
<th>Posttest RT</th>
<th>Percent correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recognition</td>
<td>21</td>
<td>950 (248)</td>
<td>74 (21)</td>
<td>801 (132)</td>
<td>77 (12)</td>
</tr>
<tr>
<td>Comprehension</td>
<td>31</td>
<td>907 (216)</td>
<td>75 (8)</td>
<td>779 (146)</td>
<td>77 (8)</td>
</tr>
<tr>
<td>Control</td>
<td>12</td>
<td>820 (168)</td>
<td>81 (9)</td>
<td>672 (89)</td>
<td>82 (10)</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: \(\text{RT} F(2, 61) = 1.4, p = .26, \eta^2 = .04\), 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 \(\text{RT} F(2, 60) = 2.5, p = .09, \eta^2 = .08\), 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: \(\text{RT} F(2, 60) = 6.2, p = .003, \eta^2 = .17\). 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 5.11: Performance on overlapping and new items: reaction times (in ms), and percentage correct by condition. Standard deviations are given between parentheses.

<table>
<thead>
<tr>
<th>Words</th>
<th>Condition</th>
<th>N</th>
<th>Overlapping items</th>
<th>New items</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></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>
</tr>
<tr>
<td></td>
<td>Comprehension</td>
<td>31</td>
<td>593 (149)</td>
<td>94 (10)</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>12</td>
<td>540 (71)</td>
<td>93 (7)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Nonwords</th>
<th>Condition</th>
<th>N</th>
<th>Overlapping items</th>
<th>New items</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></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>
</tr>
<tr>
<td></td>
<td>Comprehension</td>
<td>31</td>
<td>1062 (292)</td>
<td>60 (18)</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>12</td>
<td>919 (253)</td>
<td>72 (20)</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>
<thead>
<tr>
<th>Overarticulated items</th>
<th>New items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Words</td>
<td></td>
</tr>
<tr>
<td>Condition</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Recognition</td>
<td>21</td>
</tr>
<tr>
<td>Comprehension</td>
<td>31</td>
</tr>
<tr>
<td>Control</td>
<td>12</td>
</tr>
<tr>
<td>Nonwords</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Recognition</td>
<td>21</td>
</tr>
<tr>
<td>Comprehension</td>
<td>31</td>
</tr>
<tr>
<td>Control</td>
<td>12</td>
</tr>
</tbody>
</table>
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></td>
<td>RT</td>
<td>% correct</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>

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 post</th>
<th>Reading comp.</th>
<th>Voc. size</th>
<th>Listening span</th>
<th>Serial rec.</th>
</tr>
</thead>
<tbody>
<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>.046</td>
<td>.005</td>
<td></td>
</tr>
<tr>
<td>Reading Comp.</td>
<td></td>
<td>.478*</td>
<td>.077</td>
<td>−.023</td>
<td></td>
</tr>
<tr>
<td>Vocabulary Size</td>
<td></td>
<td></td>
<td>−.134</td>
<td>.183</td>
<td></td>
</tr>
<tr>
<td>Listening Span</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.234</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>%</th>
<th>RT</th>
<th></th>
<th>%</th>
<th>RT</th>
<th></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>
<td></td>
<td></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>
<td></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>
<td></td>
<td></td>
</tr>
<tr>
<td>% post</td>
<td></td>
<td></td>
<td>.246*</td>
<td>.254*</td>
<td>.265*</td>
<td>.160</td>
<td></td>
<td></td>
<td></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>% 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>-.094</td>
<td>.703*</td>
<td>-.374*</td>
<td>-.011</td>
<td>-.403*</td>
<td>-.147</td>
<td>-.203</td>
<td></td>
<td></td>
</tr>
<tr>
<td>% pre</td>
<td>-.112</td>
<td>.149</td>
<td>-.055</td>
<td>.181</td>
<td>.233</td>
<td>.111</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RT post</td>
<td>-.203</td>
<td>-.079</td>
<td>-.402*</td>
<td>.038</td>
<td>-.002</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% post</td>
<td>.106</td>
<td>.261*</td>
<td>.016</td>
<td>.347*</td>
<td></td>
<td></td>
<td></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></td>
<td></td>
<td>−.279*</td>
<td>.584*</td>
<td>−.114</td>
</tr>
<tr>
<td>LD % pre</td>
<td></td>
<td></td>
<td>−.277*</td>
<td>.093</td>
<td></td>
</tr>
<tr>
<td>RT SV pre</td>
<td></td>
<td></td>
<td></td>
<td>−.094</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></td>
<td>.412*</td>
<td>−.189</td>
<td></td>
</tr>
<tr>
<td>LD % post</td>
<td></td>
<td></td>
<td>−.226</td>
<td>.340*</td>
<td></td>
</tr>
<tr>
<td>RT SV post</td>
<td></td>
<td></td>
<td></td>
<td>−.203</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\textsubscript{RT}, computed by dividing SD\textsubscript{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\textsubscript{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\textsubscript{RT} for the results on the Sentence Verification test, while section 5.3.4.2 presents CV\textsubscript{RT} for the Lexical Decision test.

5.3.4.1. Coefficient of variation for the Sentence Verification test

Table 5.15 shows the CV\textsubscript{RT} of the Sentence Verification test. The correlation coefficient between CV\textsubscript{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: $CV_{RT}$ of Sentence Verification by group, for both test moments, with correlation coefficient between $CV_{RT}$ and mean RT in parentheses.

<table>
<thead>
<tr>
<th>$CV_{RT}$</th>
<th>Comprehension</th>
<th>Recognition</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>$CV_{RT}$ pretest</td>
<td>.34 (.57*)</td>
<td>.40 (.52*)</td>
<td>.24 (.28)</td>
</tr>
<tr>
<td>$CV_{RT}$ posttest</td>
<td>.36 (.55*)</td>
<td>.33 (.37)</td>
<td>.24 (.39)</td>
</tr>
<tr>
<td>$CV_{RT}$ overlapping items</td>
<td>.45 (.35)</td>
<td>.29 (.42)</td>
<td>.25 (.04)</td>
</tr>
</tbody>
</table>

$p = < .001$

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

5.3.4.2. Coefficient of variation for the Lexical Decision test

Table 5.16 shows the $CV_{RT}$ 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: $CV_{RT}$ of Lexical Decision by group, for both test moments, with correlation coefficient in parentheses.

<table>
<thead>
<tr>
<th>$CV_{RT}$</th>
<th>Comprehension</th>
<th>Recognition</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>$CV_{RT}$ pretest</td>
<td>.21 (.22)</td>
<td>.30 (.68*)</td>
<td>.21 (.32)</td>
</tr>
<tr>
<td>$CV_{RT}$ posttest</td>
<td>.20 (.50*)</td>
<td>.20 (.43)</td>
<td>.13 (.16)</td>
</tr>
<tr>
<td>$CV_{RT}$ overlapping items</td>
<td>.22 (.30)</td>
<td>.20 (.30)</td>
<td>.13 (.04)</td>
</tr>
</tbody>
</table>

$p = < .001$

Within all groups the $CV_{RT}$ 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 automated. For that a positive correlation with the mean RT is needed. As Table 5.16 shows, there are no positive significant correlations between the $CV_{RT}$ 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 automated due to the training, despite a decline in $CV_{RT}$. There is a significant correlation between the $CV_{RT}$ and the mean RT in the posttest for the Comprehension group. However, since the difference between the $CV_{RT}$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.
Chapter 6  Conclusions and suggestions for further research

6.1. Recapitulation of research questions and main findings

The study described in this thesis was mainly set up to compare the effect on listening comprehension of a training of lower-order skills with a training of higher-order strategies. Closely linked to this topic is the claim of Segalowitz and Segalowitz (1993) arguing that automatisation of the recognition of (most of) the words spoken is conditional for successful listening comprehension in a second language. The study described in this thesis can therefore also (partly) be seen as an empirical study to investigate the validity of that claim.

A first step that had to be taken in order to find an answer to the main question described above was the development, in Pilot Study 1 (Chapter 3), of a criterion to distinguish between automatised and non-automatised processes. The criterion that we developed was based on the assumption that the process of word recognition is fully automatised in one’s mother tongue but not in a second language. It was discussed in Chapter 2 and Chapter 3 that intensive training could cause the word recognition processes to become more automatised. In that case 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. To answer the research question an auditory lexical decision experiment was set up. The experiment in which 25 native speakers and 25 non-native speakers of Dutch participated, included despite the distinction between words and nonwords, also overarticulated versus underarticulated items. As was described in Chapter 3, the main difference between overarticulated and underarticulated speech is the degree of carefulness in the pronuciation whereby overarticulated items are pronounced very carefully whereas underarticulated speech can be defined as sloppy speech. In every condition (i.e., overarticulated words, overarticulated nonwords, underarticulated words, and underarticulated nonwords) there were 1, 2, and 3-syllable items, for example huis ‘house’, ogant*, omgeving ‘environment’. First we tried to make a distinction between the (automatised) processes of the native speakers and the (non-automatised) processes of the non-native speakers, by using the measure of automatisation that Segalowitz and Segalowitz (1993)
suggest, namely the Coefficient of variability and the correlation between this coefficient and the mean RTs. The results of using this measure showed only a perfect distinction between native and non-native speakers when the performances of a number of participants were excluded from further analyses. Since we preferred using all available data and since we think that besides speed, accuracy is also an important indicator of automatisation we decided to develop our own criterion to measure automatisation. Discriminant analyses based on the results of the lexical decision test showed that it is possible to make an almost perfect distinction between the performances of native and non-native speakers in the recognition of one and two-syllable overarticulated nonwords. Therefore we could conclude that the criterion to distinguish between automatised (L1) and non-automatised (L2) processes can be found in the speed and accuracy of participants’ reactions on one and two-syllable nonwords. The criterion found was used in the further experiments of this thesis.

A second step in the way to answer the main research question was the investigation, in Pilot Study II (described in Chapter 4), of the relation between language knowledge, status of the word recognition process (in terms of speed and accuracy) and general listening comprehension skills. More concretely, the question was whether it is possible to categorise language learners into fixed categories based on the performances on tests assessing the language knowledge and the skills described above. We distinguished three possible categories. Participants in category 1 were assumed to be learners with a good knowledge of the language that are able to use this knowledge under time pressure, they were assumed to have good listening comprehension skill. Participants in category 3 were assumed to be learners with poor results on the knowledge tests, by definition, automatisation of knowledge is therefore not possible, we predicted that they had a poor listening comprehension skill. Finally, participants in category 2 were assumed to be learners with sufficient language knowledge that lack the aural word recognition skills necessary to successfully use the knowledge under time pressure, they were therefore predicted to have a poor listening comprehension skill. In addition to the investigation of the relation between language and skills, Pilot Study II also investigated the complementary value of on-line tests, which are defined as time-critical, to off-line tests, in the compilation of a detailed language proficiency profile of a language learner. A test-battery that included knowledge tests, a Listening comprehension test, an on-line Lexical decision test and a memory test, was administered. The performances on these tests of the 20 second language learners that participated in this pilot study indicated that it is not simple to categorize language learners into fixed categories. The relation between language knowledge, word recognition and listening comprehension was not
as straightforward as we expected; it was necessary to determine sub-categories in order to categorise the participating L2 learners. Six of the learners had sufficient language knowledge and the status of their word recognition process did also meet the criterion we determined, but they had a bad listening comprehension skill while 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. The necessity to use sub-categories, however, indicated the usefulness of on-line tests for compiling language profiles; without the use of these tests wrong conclusions can easily be drawn regarding the profile of language learners. Without the use of the Auditory Lexical Decision test, for example, one could easily draw the wrong conclusion that the participants that did not pass the (off-line) Listening Comprehension test were poor second language listeners. The results of the (on-line) Lexical decision test, however, made clear that these participants that performed poorly on the general listening test, had good performances on the online decision test. In other words, reaction times and percentages correct on the on-line tasks made it possible to make rather subtle differentiations between the language skills of L2-learners resulting in a detailed language proficiency profile. Furthermore, the results on the auditory lexical decision test of this second pilot study were in line with the results of the test used in the first pilot study. The parallel in the results of Discriminant analyses between the two pilot studies strengthens the validity of the criterion that was the result of Pilot Study I (Chapter 3).

The information gained in the two pilot studies described above was used in the training study (described in Chapter 5). It was already stated at the beginning of this chapter that the training study was set-up to investigate the differential effect of training lower-order processing skills versus higher order comprehension strategies. The study was set-up in three stages: Test stage 1 – Training – Test stage 2. During the test stages, knowledge tests (Grammar and Vocabulary) to establish participants’ knowledge of Dutch were administered. There were also tests included (e.g. memory tests) the results of which were expected to indicate a relation between these tests and general listening comprehension skills. The most important tests, however, 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). In this study, two experimental groups were compiled, a so-called Recognition group, which received a training in the processing of lower-order skills like word recognition, and a so-called Comprehension group, which received a training in higher-order comprehension strategies. A Control group was also included the participants of which were excluded
from the training; they participated, however, in the two test stages. Since at Test stage 1, no significant differences were found between the groups, it was expected that a conclusion about the training effect could be drawn based on the results obtained in Test stage 2.

In the training study, the notion of automatisation and its relation to listening comprehension was highlighted. The link for this study between automatisation of the word recognition process and listening comprehension lies in the claim of Segalowitz and Segalowitz (1993), which was also stated above. We reformulated their claim applied to our main research question in two hypotheses. The first hypothesis stated that the students that participated in the Recognition group will perform better on the post-training general listening comprehension test than would the participants in the Comprehension group. The second hypothesis stated that, even if no positive evidence could be found to support the first hypothesis, a significant difference will be found in the performances on the post-training Sentence Verification test or at least in the performance on the post-training Lexical Decision test since these tests measure processes that form the focus of the method of the Recognition group. However, the study did not show any of these effects. The results of the study did not give a clear answer to the main research question: no significant differences were found on the listening comprehension test between the two experimental groups. The data did also give no clear additional information about the relation of individual characteristics of the learners (e.g., the memory span) and listening comprehension.

Based on the results obtained in the present study we can not conclude that training lower-order skills is more beneficial for second language learners than training listening comprehension in the more traditional (comprehension) way. However, neither do the results allow the conclusion that the traditional method is better. It is reiterated here that the training study should not be regarded as a comparison of two methods competing for educational relevance but as a comparison of training effects of two theoretically relevant component processes. As we did not find support for our hypotheses, there is even less reason to draw conclusions from this study for L2 instruction. As we stated in chapter 5, there are theoretical reasons to claim that listening comprehension can best be acquired through a combination of training in processing information at lower levels of information (up to the sentence level) and a training in applying strategies in dealing with information at higher levels of information (beyond the sentence level). However, as long as solid empirical evidence for such a claim is lacking, it is prudent for researchers to refrain from giving advice to educationalists. Since the results on the Staatsexamen
CONCLUSIONS

NT2 (as was described in Chapter 1) do indicate that the traditional training method is not optimal, finding a more satisfying method is still necessary.

6.2. Suggestions for application of the findings and for further research

The overall results of the present study do not allow us to recommend a change in the training of general listening comprehension skills towards training in lower-order skills such as word recognition. However, we would like to make a few recommendations based on the findings described in this thesis.

A first recommendation concerns the use of on-line tests in compiling a detailed proficiency profile of a language learner. The usefulness, or even the necessity, of using these time-critical tests providing information about the underlying processes of a skill, was discussed in Chapter 4. It was shown that when the results on the on-line tests were not included in the assessment of the learners’ language proficiency, erroneous conclusions could be drawn concerning the skill tested. We would like to suggest that the use of on-line tests might uncover some of the weaknesses of a language learner’s language command; once this is known, the language teacher can develop a specific training programme to foster the learner’s language skill. In the study described in this thesis, the listening comprehension skill was the target of the investigation; we, therefore, mainly used auditory tests. Future research could concentrate on, for example, the visual mode or on production. If the findings of such research would be in line with the results presented here, it would be possible to develop a test battery and a method of data analysis and categorisation such that a general (language proficiency) profile of a language learner could be formed.

Concerning the development of the automatisation theory, we think that our results contribute to the idea that automatised and controlled processes are not a dichotomy but that there is a continuum between the two extremes. The results of the LDA-analyses described both in Chapter 3 and in Chapter 4 show a continuum between the L1 and L2-speakers’ performances (expressed by mean RT and percent correct) such that performance of the L1-speakers were interpreted as automatised and performance of the L2-speakers as more controlled. The distance between native speakers’ and non-native speakers’ performance supports the assumption that the process of spoken word recognition is automatised in one’s mother tongue but not in a second language. We therefore think that using LDA-analyses based on the mean results (RT and percent correct) of a
lexical decision test of L2-speakers and L1-speakers produces an index of the status of the spoken word recognition process.

In the discussion and conclusion section of Chapter 5 (section 5.4.4), the chapter in which the training study was described, a few reasons for the lack of clear training effects were given. It was said that the most likely reason why neither of the two training methods had a positive impact on listening comprehension was lack of overlap between the materials used in the training and the materials used in the tests. This was described as a lack of transfer appropriate learning. It was stated that we may have been too optimistic in assuming that the similarity between the cognitive processes at the time of learning and processes at the time of testing would be sufficient to allow successful transfer of training. There was, of course, some overlap, as some of the words used in the training also occurred in the texts used in the Listening Comprehension test. This overlap was, however, not at the focus of our attention. In further research this focus could be established by manipulating the overlap between training materials and test materials. Manipulating the overlap in a way that the amount of overlap is in focus might result in clearer results concerning the effect of training. If this would be the case, the results would subscribe the transfer appropriate learning theory.

In the discussion section of Chapter 5, the remark was also made that it is rather impossible to control the actual amount of input that participants receive. The participants of the experimental groups in the present study received a lot of input during the intensive training and all participants followed a Dutch course in which they received a lot of input. The uncontrollable aspect concerning the input is the amount of input the participants received in contacts outside the classroom and the test situation in which they were confronted with Dutch (active or passive). A possibility to control the amount of input to a larger degree and that way get more clear results is by conducting the training in another language environment.

One last suggestion concerns the fact that the results of the training study do not indicate the importance of individual characteristics of language learners (i.e. memory span) in relation to listening comprehension skills. This might be due to the fact that the tests we used were either completely in Dutch (Listening span test) or they followed the grammatical and phonological rules of Dutch (Serial recognition test). It might be useful, in further, research to use also memory tests in the participants L1, the results on these tests might give insight in the relation between listening comprehension skills and the individuals’ memory span. Using L1-memory tests and comparing the results on these tests with results on L2-memory tests, might give information about the language dependency of the memory
capacity, a very interesting research topic the results of which might give insight in the importance of memory span in language learning.
References


REFERENCES


REFERENCES


REFERENCES


REFERENCES


REFERENCES


Appendix A     Item list of the auditory LD test

If the participants heard the items of the first set in the underarticulated version, than the items of the second set were in the overarticulated version and vice versa.

<table>
<thead>
<tr>
<th>Set 1</th>
<th>One syllable items</th>
<th>Two syllable items</th>
<th>Three syllable items</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bril</td>
<td>Rug</td>
<td>Knie</td>
</tr>
<tr>
<td></td>
<td>bril</td>
<td>reχ</td>
<td>kni</td>
</tr>
<tr>
<td></td>
<td>Wagen</td>
<td>Aantal</td>
<td>Nadruk</td>
</tr>
<tr>
<td></td>
<td>wayɔn</td>
<td>antɔl</td>
<td>nadrɔk</td>
</tr>
<tr>
<td></td>
<td>Afdeling</td>
<td>ðafelŋ</td>
<td>ðadrɔkŋ</td>
</tr>
</tbody>
</table>

<p>| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |
|       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |</p>
<table>
<thead>
<tr>
<th>Set 2</th>
<th>One syllable items</th>
<th>Two syllable items</th>
<th>Three syllable items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brief</td>
<td>brief</td>
<td>Verkeer</td>
<td>Maatregel</td>
</tr>
<tr>
<td>Reis</td>
<td>reis</td>
<td>Bakker</td>
<td>Onderwijs</td>
</tr>
<tr>
<td>Kaart</td>
<td>kart</td>
<td>Avond</td>
<td>Omgeving</td>
</tr>
<tr>
<td>Huis</td>
<td>hëys</td>
<td>Indruk</td>
<td>Gemeente</td>
</tr>
<tr>
<td>Plaat</td>
<td>plats</td>
<td>Kamer</td>
<td></td>
</tr>
<tr>
<td>Sla</td>
<td>sla</td>
<td>Waarheid</td>
<td></td>
</tr>
<tr>
<td>Tijd</td>
<td>teit</td>
<td>Teken</td>
<td></td>
</tr>
<tr>
<td>Gang</td>
<td>γαη</td>
<td>Uitgang</td>
<td></td>
</tr>
<tr>
<td>Lucht</td>
<td>λοη</td>
<td>Eiland</td>
<td></td>
</tr>
<tr>
<td>Vraag</td>
<td>vraη</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fiets</td>
<td>fits</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jas</td>
<td>jōs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dap</td>
<td>dop</td>
<td>Binkel</td>
<td>Ofiestedel</td>
</tr>
<tr>
<td>Flaat</td>
<td>flat</td>
<td>Giebes</td>
<td>kneitreeper</td>
</tr>
<tr>
<td>Kag</td>
<td>καη</td>
<td>Kimgaad</td>
<td>Oeregien</td>
</tr>
<tr>
<td>Meip</td>
<td>meip</td>
<td>Koemer</td>
<td>Desmeike</td>
</tr>
<tr>
<td>Rig</td>
<td>ρη</td>
<td>Sloger</td>
<td></td>
</tr>
<tr>
<td>Sud</td>
<td>σοη</td>
<td>Verluip</td>
<td></td>
</tr>
<tr>
<td>Tup</td>
<td>τοη</td>
<td>Bronter</td>
<td></td>
</tr>
<tr>
<td>Mip</td>
<td>μιπ</td>
<td>Driner</td>
<td></td>
</tr>
<tr>
<td>Tag</td>
<td>τοη</td>
<td>Libor</td>
<td></td>
</tr>
<tr>
<td>Glut</td>
<td>γλοη</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kluim</td>
<td>klεym</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zof</td>
<td>zοη</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Appendix B  
### Mean duration of LD items

Table B.1: Mean duration of experimental items in ms, broken down by lexicality, context and length (the number of items is given between parentheses).

<table>
<thead>
<tr>
<th></th>
<th>1-syllable items</th>
<th>2-syllable items</th>
<th>3-syllable items</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Overarticulated items</td>
<td>Underarticulated items</td>
<td>Overarticulated items</td>
</tr>
<tr>
<td>Words</td>
<td>513 (23)</td>
<td>225 (23)</td>
<td>602 (20)</td>
</tr>
<tr>
<td>Nonwords</td>
<td>461 (23)</td>
<td>216 (23)</td>
<td>616 (20)</td>
</tr>
</tbody>
</table>
Appendix C  
Mean RT for L1 and L2 speakers

Table C.1: Mean reactions (reactions times (RT) and percentage correct (%)) for L1 and L2 speakers, broken down by lexicality, context, and length (SD). Differences (Δ) that are significant are marked by ‘*’.

<table>
<thead>
<tr>
<th></th>
<th>Language</th>
<th>Overarticulated</th>
<th>Underarticulated</th>
<th>Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>RT</td>
<td>%</td>
<td>RT</td>
</tr>
<tr>
<td>1-syllable Words</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L1</td>
<td>885 (139)</td>
<td>98 (6)</td>
<td>971 (211)</td>
<td>78 (19)</td>
</tr>
<tr>
<td>L2</td>
<td>1069 (204)</td>
<td>95 (12)</td>
<td>1242 (233)</td>
<td>65 (23)</td>
</tr>
<tr>
<td>Δ</td>
<td>–184*</td>
<td>3</td>
<td>–271*</td>
<td>13*</td>
</tr>
<tr>
<td>1-syllable Nonwords</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L1</td>
<td>1044 (160)</td>
<td>87 (17)</td>
<td>1031 (226)</td>
<td>82 (14)</td>
</tr>
<tr>
<td>L2</td>
<td>1641 (437)</td>
<td>58 (25)</td>
<td>1442 (454)</td>
<td>71 (26)</td>
</tr>
<tr>
<td>Δ</td>
<td>–597*</td>
<td>29*</td>
<td>–411*</td>
<td>11</td>
</tr>
<tr>
<td>2-syllable Words</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L1</td>
<td>893 (127)</td>
<td>100 (2)</td>
<td>966 (231)</td>
<td>84 (11)</td>
</tr>
<tr>
<td>L2</td>
<td>1016 (205)</td>
<td>100 (2)</td>
<td>1073 (231)</td>
<td>78 (15)</td>
</tr>
<tr>
<td>Δ</td>
<td>–123</td>
<td>0</td>
<td>–77</td>
<td>6</td>
</tr>
<tr>
<td>2-syllable Nonwords</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L1</td>
<td>1219 (281)</td>
<td>89 (17)</td>
<td>1112 (227)</td>
<td>86 (12)</td>
</tr>
<tr>
<td>L2</td>
<td>1808 (346)</td>
<td>60 (26)</td>
<td>1440 (360)</td>
<td>73 (21)</td>
</tr>
<tr>
<td>Δ</td>
<td>–589*</td>
<td>29*</td>
<td>–328*</td>
<td>13*</td>
</tr>
</tbody>
</table>
### 3-syllable Words

<table>
<thead>
<tr>
<th>Language</th>
<th>Overarticulated</th>
<th>Underarticulated</th>
<th>Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RT</td>
<td>%</td>
<td>RT</td>
</tr>
<tr>
<td>L1</td>
<td>993 (147)</td>
<td>100 (0)</td>
<td>918 (164)</td>
</tr>
<tr>
<td>L2</td>
<td>1037 (170)</td>
<td>99 (7)</td>
<td>1038 (200)</td>
</tr>
<tr>
<td>Δ</td>
<td>–44</td>
<td>1</td>
<td>–120*</td>
</tr>
</tbody>
</table>

### 3-syllable Nonwords

<table>
<thead>
<tr>
<th>Language</th>
<th>Overarticulated</th>
<th>Underarticulated</th>
<th>Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RT</td>
<td>%</td>
<td>RT</td>
</tr>
<tr>
<td>L1</td>
<td>1208 (206)</td>
<td>96 (14)</td>
<td>1178 (238)</td>
</tr>
<tr>
<td>L2</td>
<td>1726 (505)</td>
<td>87 (19)</td>
<td>1532 (353)</td>
</tr>
<tr>
<td>Δ</td>
<td>–518*</td>
<td>9</td>
<td>–354</td>
</tr>
</tbody>
</table>
## Appendix D

### Item list of the visual LD test

<table>
<thead>
<tr>
<th>Words:</th>
<th>Nonwords:</th>
</tr>
</thead>
<tbody>
<tr>
<td>auto</td>
<td>aaf</td>
</tr>
<tr>
<td>citroen</td>
<td>alfaar</td>
</tr>
<tr>
<td>boterham</td>
<td>davelo</td>
</tr>
<tr>
<td>bal</td>
<td>beel</td>
</tr>
<tr>
<td>dochter</td>
<td>bakkel</td>
</tr>
<tr>
<td>diploma</td>
<td>lopelmoook</td>
</tr>
<tr>
<td>berg</td>
<td>bijk</td>
</tr>
<tr>
<td>eigenaar</td>
<td>beulaar</td>
</tr>
<tr>
<td>horloge</td>
<td>makantie</td>
</tr>
<tr>
<td>blad</td>
<td>blank</td>
</tr>
<tr>
<td>emmer</td>
<td>brimmel</td>
</tr>
<tr>
<td>medicijn</td>
<td>mekering</td>
</tr>
<tr>
<td>dak</td>
<td>det</td>
</tr>
<tr>
<td>fabriek</td>
<td>buffrouw</td>
</tr>
<tr>
<td>officier</td>
<td>meraniek</td>
</tr>
<tr>
<td>dier</td>
<td>drak</td>
</tr>
<tr>
<td>koffer</td>
<td>detel</td>
</tr>
<tr>
<td>regering</td>
<td>tarasiet</td>
</tr>
<tr>
<td>doos</td>
<td>dreed</td>
</tr>
<tr>
<td>maaltijd</td>
<td>dolfien</td>
</tr>
<tr>
<td>geest</td>
<td>virecteur</td>
</tr>
<tr>
<td>mijnheer</td>
<td>lagi</td>
</tr>
<tr>
<td>geld</td>
<td>ficht</td>
</tr>
<tr>
<td>moeder</td>
<td>galper</td>
</tr>
<tr>
<td>grens</td>
<td>frood</td>
</tr>
<tr>
<td>ontbijt</td>
<td>gruipel</td>
</tr>
<tr>
<td>ham</td>
<td>fus</td>
</tr>
<tr>
<td>paleis</td>
<td>hintel</td>
</tr>
<tr>
<td>kerk</td>
<td>gol</td>
</tr>
<tr>
<td>regen</td>
<td>coldijn</td>
</tr>
<tr>
<td>kind</td>
<td>gork</td>
</tr>
<tr>
<td>schouder</td>
<td>krempel</td>
</tr>
<tr>
<td>knoop</td>
<td>kang</td>
</tr>
<tr>
<td>sleutel</td>
<td>kroesdie</td>
</tr>
<tr>
<td>laars</td>
<td>lorst</td>
</tr>
<tr>
<td>suiker</td>
<td>leptisch</td>
</tr>
<tr>
<td>lijf</td>
<td>oop</td>
</tr>
<tr>
<td>tegel</td>
<td>nenster</td>
</tr>
<tr>
<td>lip</td>
<td>rolk</td>
</tr>
<tr>
<td>toren</td>
<td>pochtel</td>
</tr>
<tr>
<td>mens</td>
<td>streul</td>
</tr>
<tr>
<td>wereld</td>
<td>sunder</td>
</tr>
<tr>
<td>oma</td>
<td>suif</td>
</tr>
<tr>
<td>winkel</td>
<td>tiekel</td>
</tr>
<tr>
<td>oor</td>
<td>suk</td>
</tr>
<tr>
<td>rok</td>
<td>wapel</td>
</tr>
<tr>
<td>snoep</td>
<td>tan</td>
</tr>
<tr>
<td>touw</td>
<td>tasp</td>
</tr>
<tr>
<td>zak</td>
<td>wor</td>
</tr>
<tr>
<td>zee</td>
<td></td>
</tr>
</tbody>
</table>
## Appendix E

Mean performances of L1 and L2 speakers by condition

Table E.1: Mean reactions (reaction times (RT) and percentage correct (%)) for L1 and L2 speakers, broken down by lexicality, context, and length (SD). Differences (Δ) that are significant are marked by ‘*’.

### Words

<table>
<thead>
<tr>
<th>Language</th>
<th>Overarticulated</th>
<th>Underarticulated</th>
<th>Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RT</td>
<td>%</td>
<td>RT</td>
</tr>
<tr>
<td>L1</td>
<td>523 (381)</td>
<td>98 (5)</td>
<td>813 (393)</td>
</tr>
<tr>
<td>L2</td>
<td>548 (162)</td>
<td>96 (5)</td>
<td>779 (219)</td>
</tr>
<tr>
<td>Δ</td>
<td>-25</td>
<td>2</td>
<td>34</td>
</tr>
</tbody>
</table>

### Nonwords

<table>
<thead>
<tr>
<th>Language</th>
<th>Overarticulated</th>
<th>Underarticulated</th>
<th>Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RT</td>
<td>%</td>
<td>RT</td>
</tr>
<tr>
<td>L1</td>
<td>773 (415)</td>
<td>94 (5)</td>
<td>1009 (404)</td>
</tr>
<tr>
<td>L2</td>
<td>1577 (435)</td>
<td>57 (21)</td>
<td>1637 (506)</td>
</tr>
<tr>
<td>Δ</td>
<td>-804*</td>
<td>37*</td>
<td>-628*</td>
</tr>
</tbody>
</table>

As can be seen in Table E.1, the differences between the mean reaction times to overarticulated and underarticulated words are significant for both groups. The differences between percentages correct to these items is only significant for the native speakers, for the non-native speakers, reacting to underarticulated words is not significantly more difficult than reacting to overarticulated words. Among the differences within the groups on reacting to non-words, the only two significantly differences that are found are the difference in percentage correct and RT between overarticulated and underarticulated nonwords for the native speakers.
<table>
<thead>
<tr>
<th>Item-list of the sentence verification test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre</td>
</tr>
<tr>
<td>------</td>
</tr>
<tr>
<td>Koken doe je bijna altijd in de keuken.</td>
</tr>
<tr>
<td>Een geheim mag je aan iedereen vertellen.</td>
</tr>
<tr>
<td>Vlees kan je het best bij de kapper kopen.</td>
</tr>
<tr>
<td>Als je het koud hebt, kun je het best een trui aantrekken.</td>
</tr>
<tr>
<td>De maand oktober komt na de maand september.</td>
</tr>
<tr>
<td>Een schilderij hangt men meestal aan de muur.</td>
</tr>
<tr>
<td>Meisjes spelen meer met poppen dan jongens.</td>
</tr>
<tr>
<td>Een mens kan wel honderd kilometer per uur lopen.</td>
</tr>
<tr>
<td>In de lente krijgen bomen nieuwe bladeren.</td>
</tr>
<tr>
<td>Als je een huis huurt moet je huur betalen.</td>
</tr>
<tr>
<td>Op een horloge kan je zien wat voor weer het is.</td>
</tr>
<tr>
<td>In Nederland heeft helemaal niemand een fiets.</td>
</tr>
<tr>
<td>Als je niet goed ziet, kan je het best een bril dragen.</td>
</tr>
<tr>
<td>Als je honger hebt, kan je het best wat eten.</td>
</tr>
<tr>
<td>Brood wordt gemaakt van water, meel en een snuifje zand.</td>
</tr>
<tr>
<td>In Nederland kan je wandelen in hoge bergen.</td>
</tr>
<tr>
<td>Wanneer een mes bot is, snijdt het niet goed.</td>
</tr>
<tr>
<td>In een restaurant kan je alleen brood krijgen.</td>
</tr>
<tr>
<td>Tijdens de vakantie moet je elke dag naar school.</td>
</tr>
<tr>
<td>Een bed gebruik je om koekjes te bakken.</td>
</tr>
<tr>
<td>Een bril moet je dragen als je goed kan zien.</td>
</tr>
<tr>
<td>In Nederland wordt er geen kaas gemaakt.</td>
</tr>
<tr>
<td>Een gebouw heeft minimaal 4 muren.</td>
</tr>
<tr>
<td>In sommige landen draagt men baby’s op de rug.</td>
</tr>
<tr>
<td>Een computer gebruik je om de vloer te vegen.</td>
</tr>
<tr>
<td>Je moet niet naar de dokter als je ziek bent.</td>
</tr>
<tr>
<td>In een disco kan je wat drinken en dansen.</td>
</tr>
<tr>
<td>Zij werken in dezelfde fabriek dus zijn ze collega’s.</td>
</tr>
<tr>
<td>Eén meter is even lang als zeventig centimeter.</td>
</tr>
<tr>
<td>Een vaas gebruik je om bloemen in te zetten.</td>
</tr>
<tr>
<td>Aan zee is er meestal meer wind dan in het binnenland.</td>
</tr>
<tr>
<td>Om te kunnen voetballen heb je een stoel nodig.</td>
</tr>
<tr>
<td>Als je het warm hebt moet je de verwarming aanzetten.</td>
</tr>
<tr>
<td>In Engeland wordt er veel thee gedronken.</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>--------------------------------</td>
</tr>
<tr>
<td>Aan sport doen is niet goed voor je gezondheid.</td>
</tr>
<tr>
<td>Je kan jezelf zien als je in de spiegel kijkt.</td>
</tr>
<tr>
<td>Iedereen in de wereld heeft een naam.</td>
</tr>
<tr>
<td>Om de pap zoeter te maken kan je suiker toevoegen.</td>
</tr>
<tr>
<td>Als je gezond bent moet je vaak medicijnen slikken.</td>
</tr>
<tr>
<td>In een café moet je nooit de rekening betalen.</td>
</tr>
<tr>
<td>Stenen blijven drijven als je ze in het water gooit.</td>
</tr>
<tr>
<td>Als je wil lezen als het donker is, moet je licht aandoen.</td>
</tr>
<tr>
<td>In een woordenboek kan je geen woorden opzoeken.</td>
</tr>
<tr>
<td>Iemand die brood bakt is een journalist.</td>
</tr>
<tr>
<td>Om een slot open te maken, heb je een sleutel nodig.</td>
</tr>
<tr>
<td>Een bus is een voertuig voor slechts één persoon.</td>
</tr>
<tr>
<td>De meeste kamers hebben 7 hoeken.</td>
</tr>
<tr>
<td>’s Nachts wordt een auto vaak in een garage gezet.</td>
</tr>
<tr>
<td>Als je een huisdier hebt, moet je het elke dag voeren.</td>
</tr>
<tr>
<td>Om te weten hoe laat het is, kan je op de klok kijken.</td>
</tr>
</tbody>
</table>
### Appendix G  
**Item-list of Vocabulary Size test (Hazenberg 1994)**

<table>
<thead>
<tr>
<th>Words and INL-frequency used in our study</th>
<th>Item number Hazenberg test (1994 Appendix IV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 aspect</td>
<td>3869</td>
</tr>
<tr>
<td>2 godsdienst</td>
<td>2187</td>
</tr>
<tr>
<td>3 koesteren</td>
<td>1575</td>
</tr>
<tr>
<td>4 klank</td>
<td>1391</td>
</tr>
<tr>
<td>5 sparen</td>
<td>1141</td>
</tr>
<tr>
<td>6 prijzen</td>
<td>1049</td>
</tr>
<tr>
<td>7 atmosfeer</td>
<td>914</td>
</tr>
<tr>
<td>8 combineren</td>
<td>754</td>
</tr>
<tr>
<td>9 evaluatie</td>
<td>633</td>
</tr>
<tr>
<td>10 hiel</td>
<td>533</td>
</tr>
<tr>
<td>11 mentaliteit</td>
<td>462</td>
</tr>
<tr>
<td>12 dromerig</td>
<td>403</td>
</tr>
<tr>
<td>13 onzeggen</td>
<td>357</td>
</tr>
<tr>
<td>14 energiek</td>
<td>316</td>
</tr>
<tr>
<td>15 onroerend</td>
<td>282</td>
</tr>
<tr>
<td>16 afwegen</td>
<td>249</td>
</tr>
<tr>
<td>17 afbakenen</td>
<td>226</td>
</tr>
<tr>
<td>18 manifest</td>
<td>204</td>
</tr>
<tr>
<td>19 professie</td>
<td>187</td>
</tr>
<tr>
<td>20 spoeden</td>
<td>170</td>
</tr>
<tr>
<td>21 desgewenst</td>
<td>155</td>
</tr>
<tr>
<td>22 weerbarstig</td>
<td>141</td>
</tr>
<tr>
<td>23 schipbreuk</td>
<td>128</td>
</tr>
<tr>
<td>24 fanfare</td>
<td>116</td>
</tr>
<tr>
<td>25 toelage</td>
<td>107</td>
</tr>
<tr>
<td>26 rooskleurig</td>
<td>97</td>
</tr>
<tr>
<td>27 sterfte</td>
<td>89</td>
</tr>
<tr>
<td>28 asgrauw</td>
<td>81</td>
</tr>
<tr>
<td>29 spitsuur</td>
<td>75</td>
</tr>
<tr>
<td>30 baldakijn</td>
<td>68</td>
</tr>
<tr>
<td>31 rakker</td>
<td>63</td>
</tr>
<tr>
<td>32 verhelen</td>
<td>58</td>
</tr>
<tr>
<td>Index</td>
<td>Term</td>
</tr>
<tr>
<td>-------</td>
<td>----------------</td>
</tr>
<tr>
<td>33</td>
<td>glazuur</td>
</tr>
<tr>
<td>34</td>
<td>sokkel</td>
</tr>
<tr>
<td>35</td>
<td>want</td>
</tr>
<tr>
<td>36</td>
<td>vla</td>
</tr>
<tr>
<td>37</td>
<td>tiranniseren</td>
</tr>
<tr>
<td>38</td>
<td>overvoeren</td>
</tr>
<tr>
<td>39</td>
<td>wrochten</td>
</tr>
<tr>
<td>40</td>
<td>zwikken</td>
</tr>
<tr>
<td>41</td>
<td>pennen</td>
</tr>
<tr>
<td>42</td>
<td>kletskous</td>
</tr>
<tr>
<td>43</td>
<td>gala</td>
</tr>
<tr>
<td>44</td>
<td>toverkol</td>
</tr>
<tr>
<td>45</td>
<td>slakkenhuis</td>
</tr>
<tr>
<td>46</td>
<td>betreffend</td>
</tr>
<tr>
<td>47</td>
<td>doodgooien</td>
</tr>
<tr>
<td>48</td>
<td>ceramiek</td>
</tr>
<tr>
<td>49</td>
<td>zegepralen</td>
</tr>
<tr>
<td>50</td>
<td>diabolo</td>
</tr>
</tbody>
</table>
Beste student,

Vaak wordt het onderdeel ‘Luisteren’ van de cursus Nederlands als tweede taal een van de moeilijkste onderdelen gevonden. Luisteren is dan ook een erg complex proces.

Om goed te kunnen luisteren zijn twee dingen belangrijk.

Je moet de woorden van de spreker verstaan en Je moet begrijpen wat er door de spreker gezegd wordt.

Met verstaan bedoelen we het herkennen van de woorden in een zin. Om goed te kunnen luisteren, moet je weten welke woorden de spreker uitgesproken heeft. Met begrijpen bedoelen we het geven van betekenis aan de herkende woorden en aan de zinnen. Je moet dus eigenlijk eerst verstaan wat de spreker zegt om het daarna te kunnen begrijpen. Omdat spreken snel gaat, is het vaak moeilijk om als luisteraar de spreker te blijven volgen, zeker als er niet in je moedertaal gesproken wordt.

In deze lesbrief zullen enkele tips gegeven worden die je kunnen helpen bij het luisteren naar gesproken Nederlands.

**Verstaan**

De belangrijkste tip die we je kunnen geven, is dat wat je hoort vaak niet hetzelfde is als wat je leest!

**Enkele voorbeelden:**

Je hoort: Daar staat ie dan. Je ziet: Daar staat hij dan.

Blijf ’ns even staan. Blijf eens even staan.
Iedereen moedat doen. Iedereen moet dat doen

Uitgesproken woorden, worden vaak vervormd onder invloed van andere woorden, of onder invloed van de snelheid waarmee ze uitgesproken worden. De verschillen tussen de geschreven en de gesproken vormen
kunnen door verschillende processen veroorzaakt worden. De meeste processen hebben als consequentie dat de woorden makkelijker en sneller uitgesproken kunnen worden maar ze veranderen niets aan de betekenis van de woorden. Vaak merken moedertaalsprekers de verschillen niet eens op. Hieronder worden enkele van die processen beschreven.

Assimilatie
Met assimilatie wordt het proces bedoeld dat ervoor zorgt dat klanken gedeeltelijk gelijk worden aan de volgende of voorafgaande klanken.

**Enkele voorbeelden:**

- klabdeur → i.p.v. klapdeur
- mizdaad → i.p.v. misdaad
- aampakken → i.p.v. aanpakken
- wassak → i.p.v. waszak

Deletie
Met deletie wordt het proces bedoeld dat sprekers soms letters weglaten.

**Enkele voorbeelden:**

- Ik heb ’m gezien → i.p.v. Ik heb hem gezien
- We gaan een stukje lope → i.p.v. We gaan een stukje lopen
- Wis je dat → i.p.v. Wist je dat
- wandle → i.p.v. wandelen

Toevoeging
Met toevoeging wordt het proces bedoeld dat sprekers soms klanken toevoegen.

**Enkele voorbeelden:**

- hij kompt → i.p.v. hij komt
- erreg → i.p.v. erg
- zeejarend → i.p.v. zee-arend

Reductie
Met reductie wordt het proces bedoeld dat mensen een klinker soms vervangen door een sjwa (Dat is de klank /ə/ van de).
Enkele voorbeelden:

- menuut i.p.v. minuut
- benaan i.p.v. banaan
- moter i.p.v. motor

Als je een van deze processen opmerkt in gesproken Nederlands, moet je er niet te lang bij blijven stilstaan. Je weet nu dat deze processen bestaan en dat ze niets aan de betekenis van de woorden veranderen. Probeer je er niet door te laten afleiden en probeer gewoon verder te gaan met het luisteren naar de spreker.

Begrijpen

Het doel van met elkaar te praten en naar elkaar te luisteren is het overbrengen van een boodschap. Het is dan ook erg belangrijk dat de luisteraar begrijpt wat de spreker wil zeggen. Hieronder staan enkele tips die je kunnen helpen de boodschap (dat wat de spreker zegt) beter te begrijpen.

activeren van schema’s

Als je naar een gesproken tekst gaat luisteren is het nuttig even over het onderwerp na te denken. Zo bouw je een verwachting op van wat er gezegd kan worden en is het geheel beter te volgen. Als je bijvoorbeeld luistert naar een tekst over twee mensen die uit gaan eten in een restaurant dan kan je je al een beetje voorstellen wat je kan verwachten. Zo weet je bijvoorbeeld dat er een kelner zal zijn en dat de mensen iets zullen bestellen aan de hand van een menukaart.

signaalwoorden

In een gesproken tekst komen soms woorden voor die een specifieke functie hebben wat de structuur/opbouw van de informatie betreft. Deze woorden noemen we signaalwoorden. Het doel van deze woorden is je alert maken op hetgeen dat zal volgen.

Enkele voorbeelden:

* Ten eerste ..., ten tweede ... Deze woorden worden altijd bij een opsomming gebruikt. Als je iemand ten eerste hoort zeggen kan je er zeker van zijn dat er ook ten tweede gezegd zal worden.
* Omdat, doordat Na omdat en doordat wordt er een reden gegeven.
* Bovendien Dit woord wordt meestal gebruikt nadat de spreker zijn eigen mening van iets gegeven heeft. Na bovendien geeft de spreker een argument voor zijn mening.

luisterhoudingen
Je kan op verschillende manieren naar een gesproken tekst luisteren. Je kan luisteren zonder echt op te letten, je kan luisteren om iets te leren, je kan luisteren omdat je iets specifiek wilt horen, je kan luisteren naar een tekst omdat je weet dat er vragen over gesteld zullen worden. Deze verschillende houdingen zorgen ervoor dat je anders naar de spreker luistert. Als je bijvoorbeeld iets wilt leren zal je aandachtiger naar de spreker luisteren dan wanneer je niet echt hoeft op te letten. Als je voor het luisteren even bedenkt wat je doel is (bijvoorbeeld het kunnen oplossen van vragen over de tekst) zal dat je helpen je doel sneller te bereiken.

steunen op context
Soms komt het voor dat je niet alles wat gezegd wordt, begrijpt. Dat is niet erg. Het kan best zijn dat je de tekst als geheel toch begrijpt ondanks dat je een paar zinnen of een aantal woorden niet begrepen hebt. Wanneer je een fragment niet helemaal begrijpt, kan je best gewoon door luisteren en proberen aan de hand van de context te achterhalen wat er met het fragment bedoeld werd. Je moet dan dus proberen om de nodige informatie uit de context te halen.
Hetzelfde geldt voor woorden die typisch zijn voor een bepaald vak, vakjargon. Vaak maakt de context duidelijk wat het woord betekent. Het is dus niet altijd nodig dat je de term kent.

tekstsoort
Er zijn verschillende soorten gesproken tekst, bijvoorbeeld interviews, dialogen, monologen, instructies, ... Elke tekstsoort heeft zijn eigen eigenschappen. Wanneer je weet naar wat voor een tekst je luistert weet je dus ook wat je kan verwachten. Als je bijvoorbeeld naar een interview luistert weet je dat er vragen gesteld zullen worden en dat deze vragen beantwoord zullen worden.

Het toepassen van deze tips kan je oefenen tijdens de luistertraining.

We willen je alvast bedanken voor je medewerking en wensen je veel succes.
Appendix I  Off-line test performances broken down by university

Table I.1 shows the overall results on the so-called Selection tests and the Listening Comprehension test, broken down by university and experimental group. The results indicate that the assumption that the students of the control group are comparable with the other UvA students is appropriate.

Table I.1: Mean results on off-line tests (Basic Vocabulary (voc), Basic Grammar (gram) and Listening comprehension (listn)) broken down by university and experimental group. Standard deviation is given between parentheses.

<table>
<thead>
<tr>
<th>University</th>
<th>Experimental group</th>
<th>N</th>
<th>Voc max=60</th>
<th>Gram max=53</th>
<th>Listn max=27</th>
</tr>
</thead>
<tbody>
<tr>
<td>University of Amsterdam</td>
<td>Recognition</td>
<td>10</td>
<td>50 (10)</td>
<td>46 (4)</td>
<td>19 (4)</td>
</tr>
<tr>
<td></td>
<td>Comprehension</td>
<td>17</td>
<td>50 (6)</td>
<td>45 (4)</td>
<td>18 (3)</td>
</tr>
<tr>
<td>Free University</td>
<td>Recognition</td>
<td>8</td>
<td>48 (5)</td>
<td>42 (3)</td>
<td>17 (4)</td>
</tr>
<tr>
<td>Leiden University</td>
<td>Comprehension</td>
<td>9</td>
<td>43 (4)</td>
<td>41 (3)</td>
<td>13 (3)</td>
</tr>
<tr>
<td></td>
<td>Recognition</td>
<td>3</td>
<td>44 (6)</td>
<td>46 (3)</td>
<td>13 (2)</td>
</tr>
<tr>
<td>University of Amsterdam</td>
<td>Comprehension</td>
<td>5</td>
<td>45 (7)</td>
<td>46 (4)</td>
<td>14 (4)</td>
</tr>
</tbody>
</table>

As can be seen in Table I.1, the average scores of students of UvA are higher than the mean scores of the other students. A one-way ANOVA analysis shows that there is indeed a university-effect for the comprehension group: Basic Vocabulary, $F(2,28) = 5.2$, $p < .05$, Basic Grammar, $F(2,28) = 3.4$, (ns.), and Listening Comprehension $F(2,28) = 7.0$, $p < .05$. These significant results were, however, not confirmed by the Student-Newman-Keuls post-hoc test. The two-way interaction between university and condition was also not significant. These results and the more or less equal spreading of the UvA-students over the experimental conditions allows the conclusion that there is no real difference between the experimental groups in case of the Selection tests and the Listening Comprehension test.
1. Inleiding

Iedereen die ooit een vreemde taal heeft geleerd kent waarschijnlijk het frustrerende gevoel dat ondanks het gevolgde onderwijs het communiceren met moedertaalsprekers van die taal uiterst moeizaam verloopt. De anderstalige gesprekspartner praat (volgens de tweede-taalverwerver) te snel en te onzorgvuldig, waardoor woorden die eigenlijk wel gekend zijn, niet snel genoeg herkend worden in de gesproken vorm. Met andere woorden, T2-gebruikers kennen vaak wel de woorden maar zijn niet in staat ze te herkennen, met name niet wanneer de tijdsdruk te hoog is om de benodigde taallelementen snel genoeg uit het geheugen op te halen. Dat dit niet enkel geldt voor zogenaamde receptieve vaardigheden, zoals luisteren, maar ook voor productieve processen, zoals spreken, blijkt uit het gegeven dat taalgebruikers voor of na een conversatie vaak wél precies weten wat ze hoe moeten zeggen in de vreemde taal maar tijdens het gesprek zelf niet verder komen dan wat gestamel.

De situaties zoals hierboven geschetst duiden op een verschil tussen het gebruiken van taal met en zonder tijdsdruk. Het lijkt erop dat T2-gebruikers niet altijd in staat zijn de kennis die ze hebben, op een afdoende wijze te gebruiken. Dit geeft aan dat naast het leren van de feitelijke kennis van een taal, bijvoorbeeld woordenschat en grammatica, het vloeiend leren gebruiken van die kennis een voorwaarde is om succesvol te kunnen communiceren. In het talenonderwijs tegenwoordig wordt echter met name aandacht besteed aan het verwerven van kennis en niet zozeer aan het op adequate wijze kunnen ophalen (automatiseren) van die kennis. Natuurlijk is

---

1 De studie zoals beschreven in dit proefschrift is gericht op het verwerven van een tweede taal (T2). Het verschil tussen het leren van een tweede taal en een vreemde taal ligt in de context waarin de taal geleerd wordt. Men spreekt van tweede-taalverwerving wanneer de taal geleerd wordt in een omgeving waarin doeltaal gesproken wordt (bijvoorbeeld het leren, door een Fransman, van het Nederlands in Nederland of Vlaanderen); men spreekt van vreemde-taalverwerving wanneer de taal geleerd wordt in een omgeving waarin de doeltaal niet de omgevingstaal is (bijvoorbeeld het leren van Nederlands, door een Fransman, in Frankrijk).
kennisverwerving een onontbeerlijke stap in de taalverwerving maar wanneer dit het enige onderwerp is van het onderwijs zullen T2-leerders nauwelijks vloeiend gebruik kunnen maken van de taal, laat staan dat ze ooit voor een moedertaalspreker van die taal kunnen doorgaan.

In het huidige politiek-economische klimaat, waarin internationale communicatie steeds belangrijker wordt en immigranten verplicht worden de taal van het gastland te leren, is de vraag naar efficiënte methodes om een T2 te leren belangrijker dan ooit. Daarvoor is onder andere onderzoek nodig naar de verbetering van de luistervaardigheid. Het in deze dissertatie beschreven onderzoek beoogt daaraan een bijdrage te leveren.

2. Het luisterproces

In de beginperiode van het onderzoek naar het luisterproces in een tweede taal werd luisteren beschouwd als de tegenhanger van lezen. Tegenwoordig heeft men steeds meer oor voor de fundamentele verschillen tussen de twee vaardigheden lezen en luisteren. Een van de verschillen is het feit dat luisteraars getuige zijn van het totstandkomingproces van de uiting. Luisteraars nemen de aarzelingen, herhalingen, en verbeteringen van de spreker waar terwijl de lezer enkel met het ‘perfecte’ eindproduct wordt geconfronteerd. Een ander groot en duidelijk waarneembaar verschil is het verschil in tijdsdruk waarin de input, zijnde spraak of tekst, verwerkt moet worden. In de meeste lees situaties kunnen lezers de tekst in hun eigen tempo verwerken, zonder enige vorm van tijdsdruk. Als zij een moeilijke zin moeten lezen, kunnen ze deze naar believen enige malen herhalen. Luisteraars daarentegen hebben wat wordt genoemd een on-line contact met de spreker, zij staan in rechtstreeks contact en zijn afhankelijk van de snelheid waarmee de spreker zijn boodschap weergeeft. Het is zeer waarschijnlijk dat net dit verschil in gegeven verwerkingstijd luisteren voor veel vreemde-taalleerders zoveel moeilijker maakt dan lezen.

Dat het leren luisteren soms moeizaam gaat, is niet zo verwonderlijk. Het luisterproces is het complexe resultaat van verschillende subprocessen. Ruwweg kunnen we drie deelprocessen onderscheiden:

(i) **horen**: uit al het omgevingsgeluid moet, in het geval van luisteren naar talige informatie, de spraak van het omgevingsgeluid gescheiden worden

(ii) **woordherkenning**: de stroom van spraakklanken wordt in linguïstische eenheden verdeeld (bijvoorbeeld woorden) en de betekenis van deze eenheden wordt vanuit het geheugen geactiveerd
(iii) **begrijpen**: het integreren van de woordbetekenissen in de gehele uiting en het interpreteren van het geheel.

Het eerste subproces wordt aangeduid als een lagere-orde proces terwijl begrijpen wordt beschouwd als een hogere-orde proces. Het woordherkenningsproces ligt min of meer op de grens; in deze samenvatting beschouwen we het primair als een lagere-orde proces.

Elk van deze subprocessen kan problemen opleveren. Dat geldt voor moedertaalsprekers van een taal maar zeker ook voor tweede-taalverwervers. Een groot verschil tussen eerste en tweede-taalverwerving is het verschil in cognitieve ontwikkeling op het moment dat de taal geleerd wordt. Het leren van een eerste taal gebeurt gelijktijdig met de algemene cognitieve ontwikkeling terwijl het leren van een tweede taal meestal gebeurt wanneer deze ontwikkeling reeds voltooid is.\(^2\)

Een probleem dat (zeker beginnende) tweede-taalverwervers ondervinden is het foutief, of geheel niet, identificeren van de klanken uit de nieuwe taal. Het kan zo zijn dat er in de nieuw te leren taal klanken zijn opgenomen die niet als zodanig voorkomen in de moedertaal van de taalleerder. Dit is bijvoorbeeld het geval bij de fonemen /l/ en /t/ die in het Nederlands betekenisonderscheidend zijn (zoals in het minimale paar rat – lat) maar niet in het Chinees. Chineestalige leerders van het Nederlands horen in eerste instantie het verschil tussen deze twee klanken niet: ze kennen het contrastieve onderscheid niet in hun eigen taal en herkennen het dus ook niet in het Nederlands. Enkel een intensieve training op het onderscheiden van de klanken kan leiden tot een succesvol herkenningsproces.

Bovenstaand voorbeeld geeft al het belang weer van aandacht voor dergelijke processen in het vreemde talenonderwijs. Zoals reeds eerder gesteld wordt in de meeste taalcursussen luisteren onderwezen met een sterke nadruk op begrijpend luisteren. Minder aandacht wordt er geschonken aan het herkennen van individuele gesproken woorden of klanken terwijl deze herkenningsprocessen erg belangrijke onderdelen zijn van het luisterproces. Resultaten van bijvoorbeeld het Staatsexamen Nederlands als tweede taal niveau II, laten zien dat de traditionele methode met de nadruk op begrijpen niet optimaal is (Kerkhoff 1997); vele studenten behalen geen voldoende op het luisteronderdeel terwijl ze wel slagen voor de overige vaardigheden (bijvoorbeeld schrijven).

\(^2\) Een uitzondering hierop vormen kinderen die vanaf de geboorte volledig tweetalig opgevoed worden.
3. Het onderzoek

Het idee dat het benadrukken van lagere-orde processen belangrijk is om goed te leren luisteren in een tweede of vreemde taal, is nooit eerder door middel van een gecontroleerd trainingsexperiment onderzocht. Wij hebben geen kennis van een studie die al eerder het effect van een lagere-orde training heeft vergeleken met het effect van de traditionele luistertraining waarin juist het leren begrijpen wordt benadrukt. Het ontbreken van dergelijk onderzoek vormt de motivatie voor de huidige studie. Een bijkomend interessant aspect van onze studie is dat op een empirische wijze de automatiseringsclaim van Segalowitz en Segalowitz (1993) wordt onderzocht. Deze onderzoekers beweren dat het automatiseren van de gesproken woordherkenning een noodzakelijke voorwaarde is om te komen tot goede luistervaardigheid. Het eerder geopperde idee dat het nuttig kan zijn om op lagere-orde processen als bijvoorbeeld woordherkenning, te focussen is duidelijk sterk gerelateerd aan deze claim. Zoals reeds eerder werd gezegd, is het noodzakelijk dat taalgebruikers in staat zijn de kennis automatisch toe te passen om goed te kunnen communiceren. In die zin zal het automatiseren van die kennis het luisterproces verbeteren.

Ons onderzoek is dus opgezet om het effect te onderzoeken van het trainen van lagere-orde processen en het trainen van hogere-orde processen op de verwerving van algemene luistervaardigheid in een vreemde of tweede taal. De studie is echter niet zozeer een vergelijking tussen twee methodes als wel een onderzoek waarin het relatieve effect van verschillende manieren om luistervaardigheid te trainen onderzocht wordt.

Een eerste stap in het onderzoek was een manier te vinden om een onderscheid te kunnen maken tussen de volledig geautomatiseerde processen van moedertaalsprekers van het Nederlands en de in mindere mate geautomatiseerde of meer gecontroleerde processen van verwervers van het Nederlands als tweede taal (NT2). Hiertoe werd in hoofdstuk 3 een computergestuurd zogenaamd lexicale-decisie experiment afgenomen bij moedertaalsprekers als NT2-leerders. Tijdens het experiment kregen de proefpersonen klankreeksen te horen waarvan ze zo snel mogelijk moesten beslissen of het wel of niet om een bestaand woord van het Nederlands ging. In het experiment kwamen bestaande woorden van het Nederlands voor, bijvoorbeeld hemel, als ook zogenaamde nonwoorden, bijvoorbeeld pagel. Deze nonwoorden zijn gevormd volgens de regels van het Nederlands en zouden dus bestaande woorden kunnen zijn, ze hebben echter geen betekenis. Zowel de woorden als de nonwoorden kwamen voor in een zorgvuldig uitgesproken versie als in een minder goed geartikelde en sneller uitgesproken versie. De resultaten laten zien dat het beste criterium om de processen van moedertaalsprekers en de anderstaligen te
scheiden, gevonden kan worden in een combinatie van snelheid en accuraatheid in de reacties op de netjes uitgesproken nonwoorden. Het gevonden criterium werd gedurende het verdere onderzoek gebruikt als maat om vast te stellen of bij een luisteraar het woordherkenningsproces al dan niet volledig geautomatiseerd is.

Hoofdstuk 4 beschrijft de tweede stap in het onderzoek. In deze studie keken we of de status van de algemene luistervaardigheid van NT2-leerders voorspeld kan worden aan de hand van enerzijds taalkennis en anderzijds de status van woordherkenning (volledig geautomatiseerd of niet). We onderzochten of het mogelijk is om taalleerders onder te verdelen in drie categorieën:

(i) Categorie 1: Leeriders hebben een goede kennis van de taal en de status van hun woordherkenningsproces haalt het door ons vastgestelde criterium. Wij voorspelden voor deze groep dan ook een goede luistervaardigheid.


(iii) Categorie 3: Leeriders hebben een goede kennis van de taal maar omdat ze niet in staat zijn die kennis automatisch te gebruiken voorspelden we een slechte luistervaardigheid voor deze groep.

Er werd in deze studie gebruik gemaakt van tijdkritische tests, de zogenaamde on-line toetsen, waarbij de proefpersonen moeten reageren terwijl de input nog verwerkt wordt (bijvoorbeeld een lexicale-decisietoets), en van off-line taken, waarbij de proefpersonen tijd krijgen om na te denken alvorens te antwoorden (bijvoorbeeld een schriftelijke grammaticatoets). De resultaten laten zien dat de door ons vooropgestelde categorieën niet afdoende waren voor het categoriseren van de proefpersonen; de relatie tussen de factoren was niet zo eenduidig als werd verondersteld. Zo waren er proefpersonen die ondanks een goede taalkennis (vocabulairekennis) en een hoge geautomatiseerdheid van het woordherkenningsproces toch onvoldoende scoorden op de begrijpend luistertoets en anderen die bij een goede taalkennis maar een slechts zwak geautomatiseerde woordherkenning toch goed scoorden op begrijpend luisteren.

De resultaten van deze studie geven duidelijk het belang van de toevoeging van on-line toetsen aan, ter verkrijging van een gedetailleerd en correct profiel van de taalleerder. Zonder deze tijdkritische toetsen kan snel
een verkeerde conclusie getrokken worden aangaande de luistervaardigheid. De resultaten op on-line toetsen geven subtiele verschillen tussen taalleerders weer en maken het mogelijk om gedetailleerde taalvaardigheidsprofielen op te stellen. Taaldocenten zouden deze profielen kunnen gebruiken om de zwakke deelprocessen bij hun leerlingen op te sporen en dan via gerichte training het proces van begrijpend luisteren te optimaliseren.

Hoofdstuk 5 beschrijft in detail het materiaal, de procedure en de resultaten van de trainingsstudie (de hoofdstudie) waarin de effecten van het trainen van lagere- en hogere-orde processen op algemene luistervaardigheid worden onderzocht. De onderzoeksvraag werd als volgt geformuleerd:

\[
\text{Wat is het relatieve effect van twee training methodes om de luistervaardigheid van tweede taalverwervers te verbeteren:}
\]

(i) een methode die gericht is op het verbeteren (automatiseren?) van lagere-orde woordherkenningsvaardigheden,

(ii) een methode die zich richt op het verbeteren van hogere-orde vaardigheden?

Voorafgaand aan, en volgend op, de feitelijke training waren er toetsperioden. Tijdens deze perioden werden de deelnemende studenten onderworpen aan allerlei toetsen, gaande van woordenschat- en grammaticatoetsen tot computergestuurde geheugentoetsen. Een aantal toetsen werd zowel voor als na de training afgenomen omdat op die manier een eventueel effect van de training op de deelvaardigheden waargenomen kon worden. Aan de studie namen twee experimentele groepen deel. De zogenaamde Herkenningsgroep, kreeg een training gericht op het automatiseren van lagere-orde processen zoals het herkennen van woorden in verbonden spraak. De andere experimentele groep, de Begrijpgroep, volgde een training die voornamelijk bestond uit traditionele luisteroefeningen waarbij geluisterd wordt naar een tekst terwijl tekstbegriffvragen opgelost moeten worden. Naast de experimentele groepen was er ook een Controlegroep, de studenten uit deze groep deden wel mee aan de toetsen die rondom de training werden georganiseerd maar zij kregen verder geen vorm van luistertraining. Het opnemen in het design van een controlegroep bestaande uit mensen van een zelfde proefpersonenpopulatie maakt het mogelijk een onderscheid te maken tussen echte trainingseffecten en effecten die toe te schrijven zijn aan overige zaken.
Er werden geen grote verschillen in de prestaties gevonden tussen de Herkennings- en Begrijpsgroep. Met andere woorden, de resultaten laten geen duidelijk verschil zien tussen de effecten die de trainingsmethodes hebben op algemene luistervaardigheid. In het slot van hoofdstuk 5 worden enkele mogelijke redenen genoemd voor de afwezigheid van het verwachte trainingseffect. Een daarvan is dat, in het geval van de Herkenningsgroep, er een te geringe overlap was tussen de trainingstof en de stof van de luistervaardigheidstoets waardoor een trainingseffect niet zichtbaar kon worden. Wij konden aan de hand van de gevonden resultaten dan ook niet concluderen dat het trainen van lagere-orde processen een positiever resultaat oplevert dan de hogere-orde begripstraining. Er werd echter ook niet vastgesteld dat de klassieke manier van trainen (begripstraining) de betere is.

4. Algemene conclusies

Ook al zijn duidelijke resultaten bij de hoofdstudie achterwege gebleven, ons onderzoek levert niettemin een relevante bijdrage aan het onderzoeksveld. Een eerste punt van aandacht betreft de ontwikkeling van de automatiseringstheorie. De resultaten zoals beschreven in hoofdstuk 3 en bevestigd in hoofdstuk 4 tonen aan dat het onderscheid tussen gecontroleerde en geautomatiseerde processen geen dichotomie is. Er is eerder sprake van een continuüm. De resultaten geven ook steun voor de aanname dat het proces van gesproken woordherkenning geautomatiseerd is bij volwassen moedertaalsprekers terwijl dat niet volledig het geval is bij de NT2-leerders die aan het experiment deelnamen. Het criterium, percentage correct en reactiesnelheid van de responsies op zorgvuldig uitgesproken nonwoorden, is een doeltreffende maat om de graad van geautomatiseerdheid van het woordherkenningsproces te becijferen.

Een tweede opbrengst van de studie betreft het gebruik van tijdkritische on-line toetsen om een gedetailleerd profiel van een taalleerder te verkrijgen. Het tweede experiment toonde voor luistervaardigheid aan hoe belangrijk het is niet alleen naar taalkennis te kijken (kennis van woorden en grammatica) maar ook naar de snelheid waarmee die kennis kan worden toegepast. Kijkend naar de resultaten van het hoofdexperiment kunnen we niet concluderen dat het trainen van lagere-orde vaardigheden gunstiger is dan het trainen van begrijpend luisteren zoals in dit experiment onderzocht, d.w.z. (1) met tweede-taalverwervers die aan de in hoofdstuk 5 genoemde selectievoorwaarden voldeden, (2) met de door ons gekozen operationalisatie van lagere-orde en hogere-orde training en (3) met de door ons gekozen luistervaardigheidstoets. Het is belangrijk te bedenken dat deze studie niet
werd opgezet om twee kant-en-klare onderwijsmethodes met elkaar te vergelijken. Immers, zoals in de inleiding van deze samenvatting gezegd is, er zijn theoretische redenen om aan te nemen dat een optimale luistervaardigheidstraining bestaat uit componenten die zowel lagere-orde als hogere-orde processen bestrijken. In het onderwijs moeten dus ook beide typen processen geoefend worden. Maar voor de vraag hoe dat precies het beste gebeuren kan, rekening houdend met de drie hierboven genoemde parameters, is verder onderzoek noodzakelijk.
Curriculum vitae
