

KATHOLIEKE UNIVERSITEIT LEUVEN
Faculteit Psychologie en Pedagogische Wetenschappen
Centrum voor Gezins- en Orthopedagogiek

Elements of Speech Technology Based Reading Assessment and Intervention

Development and Analysis of a Children's Oral Reading Corpus and Evaluation of a Computerized Reading Tutor

Proefschrift aangeboden
tot het behalen van de graad van
Doctor in de Pedagogische Wetenschappen
door **Leen Cleuren**

Promotor: Prof. Dr. Pol Ghesquière
Copromotor: Prof. Dr. Ir. Hugo Van hamme

2009

Leen Cleuren, Elements of Speech Technology Based Reading Assessment and Intervention. Dissertation submitted to obtain the degree of Doctor in Educational Sciences, 2009. Supervision: Prof. Dr. Pol Ghesquière and Prof. Dr. Ir. Hugo Van hamme.

The four studies presented in the present dissertation can be situated within a multidisciplinary research project that wanted to implement automatic speech recognition (ASR) and automatic speech synthesis (i.e., text-to-speech, TTS) tools in reading assessment and intervention. The implementation of ASR in reading assessment enables this process to become automated and objective while still enabling the child to read aloud; it also provides reading speed measures that cannot (easily) be obtained by means of paper-and-pencil reading tests. Within the context of reading intervention, ASR enables a computer to become a real reading tutor that can present reading material to a child and is able to really listen to a child's reading performance: it is able to track a child's progress and follow as the child reads along, and to detect reading errors so that immediate feedback on these errors can be given. Naturally sounding, highly intelligible, and phonetically correct TTS enables this tutor to give oral instructions, to encourage a child, to read to or read along with a child, and to provide different kinds of individualized speech feedback and decoding assistance when an error is made or help is asked for. It also ensures a program to be 'open' so that the reading tutor's users are not restricted to the reading material chosen and implemented by the authors of the program.

In a first study, the development of the CHildren's Oral REading Corpus (CHOREC) was discussed, and the quality of the annotations forming part of that corpus was evaluated. CHOREC was developed to statistically characterize children's reading behaviour, and to supply a speech recognizer with a model that contains information on the nature and prevalence of common oral reading errors. It contains recorded and annotated read speech (42 GB or 130 hours) of 400 Dutch speaking elementary school children with or without reading difficulties. Analyses of inter- and intra-annotator agreement were carried out in order to investigate the consistency with which reading errors were detected, orthographic and phonetic transcriptions were made, and reading errors and reading strategies were labelled. Percentage agreement scores and kappa values both showed that agreement between annotations, and therefore the quality of the annotations, was high, so that the corpus could be reliably used for research purposes.

In a second and third study, to contribute to the literature on the reading process in orthographically transparent languages, and to obtain information on children's oral reading behaviour that could be of interest to be implemented into the language model of the speech recognizer, both a quantitative and qualitative analysis of oral reading errors annotated in the CHOREC corpus was carried out. In general, the results of the quantitative analyses replicated the results that were found in the orthographically opaque English language. More specifically, pseudowords were harder to read (i.e., with lower accuracy) than unrelated real words, and the latter were harder to read than real words embedded in a story context. Longer unrelated real words or pseudowords were read with lower accuracy than words or pseudowords with fewer syllables, and stories at a child's frustrational level induced relatively more reading errors than stories at a child's mastery or instructional level. Although differences were not always statistically significant, results also indicated that boys read less accurately than girls, that reading accuracy increased with increasing grade, that good word decoders obtained higher accuracy scores than average word decoders, who in turn were more accurate than poor word decoders, and that readers with a higher text reading level read with higher accuracy than readers with a lower text reading level.

The results of the qualitative analyses substantially strengthened the orthographic depth hypothesis with regard to the nature of reading in Dutch. Although not being an all-or-nothing phenomenon, children seemed to be mainly relying on sublexical strategies to read unrelated real words, and on lexical strategies to read real words embedded in a story. Children also had greater confidence in trying to read difficult words when these were embedded in a story context. Results also showed that when real words became longer, children increasingly relied on sublexical strategies to read these words. At the same time, when real words became longer or the relative difficulty of stories increased, children became less confident in trying to read difficult words. There also was a developmental shift in the use of sublexical and lexical strategies to read words: the older children became, the more they relied on lexical strategies to read both unrelated real words and real words in a story. Also, the lower a child's decoding skills, the more he/she relied on lexical strategies to decode an unknown word during story reading. Finally, the results showed that girls were less likely than boys to use a sublexical strategy to read an unfamiliar word during unrelated real word reading.

The last study of this dissertation describes an intervention study that used a randomized pretest-posttest matched-group design to investigate whether improvements in reading accuracy and speed could be obtained in elementary school children with reading difficulties that used a prototype of a computerized reading tutor with TTS feedback for approximately 4.5 hours during four weeks. The intervention comprised instructional-level non-repetitive text reading with synthesized phoneme-by-phoneme feedback. Each child of a matched pair was randomly assigned to an intervention ($N = 10$) or control group ($N = 10$), and was pre- and post-tested (one week and four weeks after intervention) on both a paper-and-pencil and a computerized real word, pseudoword, and story reading test. Mixed model analyses of covariance demonstrated that the intervention group outperformed the control group with regard to both reading accuracy and speed, mainly with regard to real word and pseudoword reading. Additionally, children's pre-intervention real word and pseudoword reading performance level seemed to have a differential effect on intervention efficacy.

Leen Cleuren, Bijdragen aan een op Spraaktechnologie Gebaseerde Diagnostiek en Behandeling van Leesproblemen. Proefschrift aangeboden tot het verkrijgen van de graad van Doctor in de Pedagogische Wetenschappen, 2009. Promotoren: Prof. Dr. Pol Ghesquière en Prof. Dr. Ir. Hugo Van hamme.

De vier studies die in dit proefschrift worden voorgesteld, kunnen worden gesitueerd binnen een multidisciplinair onderzoekproject dat automatische spraakherkenning (i.e., automatic speech recognition - ASR) en spraaksynthese (i.e., text-to-speech - TTS) wil gebruiken bij de leesdiagnostiek en -behandeling. De implementatie van ASR in de leesdiagnostiek maakt dit proces geautomatiseerd en objectief terwijl het kind toch luidop kan blijven lezen; het verschaft ook gedetailleerde informatie over de leessnelheid van een kind die niet (gemakkelijk) door middel van een pen-en-papier leestest kan verkregen worden. Binnen de context van leesbehandeling stelt ASR een computer in staat om een echte leestrainer te worden die leesoefeningen aan het kind kan presenteren en werkelijk naar een kind kan luisteren. De computer is in staat om op elk moment te weten waar in de tekst een kind precies aan het lezen is, en leesfouten te ontdekken zodat hierop meteen feedback kan worden gegeven. Natuurlijk klinkende, duidelijk verstaanbare en fonetisch correcte TTS laat deze leestrainer toe om mondelinge instructies te geven, om een kind aan te moedigen, om voor te lezen aan of mee te lezen met een kind, en om verschillende soorten geïndividualiseerde spraakfeedback te verlenen wanneer een fout wordt gemaakt of om hulp wordt gevraagd. TTS maakt het programma ook 'open' zodat de gebruikers van de leestrainer niet beperkt worden tot het leesmateriaal dat door de auteurs van het programma werd uitgekozen en geïmplementeerd.

In een eerste studie werd de ontwikkeling van het CHildren's Oral REading Corpus (CHOREC) besproken, en werd de kwaliteit van de annotaties die deel uitmaken van dat corpus geëvalueerd. CHOREC werd ontwikkeld om het leesgedrag van kinderen op een statistische manier te kunnen karakteriseren, en om een spraakherkenner te kunnen voorzien van een model dat informatie bevat over de aard en de prevalentie van regelmatig voorkomende leesfouten. Het bevat opgenomen en geannoteerde gelezen spraak (42 GB of 130 uur) van 400 Nederlandstalige basisschoolkinderen met of zonder leesmoeilijkheden. De inter- en intra-beoordelaarsbetrouwbaarheid werd geanalyseerd om de consistentie na te gaan waarmee leesfouten werden gedetecteerd, orthografische en fonetische transcripties werden gemaakt, en leesfouten en leesstrategieën werden gelabeld. Zowel de percentage overeenkomst scores als de kappawaarden waren groot en gaven aldus aan dat de annotaties van hoge kwaliteit waren en dat het corpus op een betrouwbare manier voor onderzoeksdoeleinden kon gebruikt worden.

In een tweede en derde studie werd een kwantitatieve respectievelijk kwalitatieve analyse van leesfouten (geannoteerd in het CHOREC corpus) uitgevoerd. Hiermee werd een bijdrage geleverd aan de literatuur over het leesproces in orthografisch transparante talen, en werd interessante informatie over het leesgedrag van kinderen verkregen die vervolgens kan gemodelleerd worden in een taalmodel voor een spraakherkenner. De resultaten van de kwantitatieve analyses repliceerden de resultaten die werden gevonden bij Engelstalige kinderen. Pseudowoorden waren moeilijker om te lezen (i.e., met lagere nauwkeurigheid) dan echte woorden, en de laatstgenoemden waren moeilijker te lezen dan echte woorden ingebed in een verhaalcontext. Langere echte woorden of pseudowoorden werden minder nauwkeurig gelezen dan echte woorden of pseudowoorden met minder lettergrepen, en verhalen op het kind zijn/haar frustratieniveau lokten meer fouten uit dan verhalen op instructie- of beheersingsniveau. Hoewel de volgende verschillen niet altijd statistisch significant waren, gaven de resultaten aan dat jongens minder accuraat lazen dan meisjes, dat de leesaccuratesse toenam naarmate kinderen ouder werden, dat goede lezers minder fouten maakten dan gemiddelde lezers, die op hun beurt minder fouten maakten dan slechte lezers, en dat lezers met een hoger tekstleesniveau minder fouten maakten dan lezers met een lager tekstleesniveau.

De resultaten van de kwalitatieve analyses bevestigden de 'orthographic depth hypothesis' m.b.t. de aard van het lezen in het Nederlands. Hoewel dit nooit een alles-of-niets fenomeen was, toonden de resultaten aan dat kinderen zich hoofdzakelijk op sublexicale strategieën baseerden om losse echte woorden te lezen, en op lexicale strategieën om echte woorden ingebed in een verhaal te lezen. Kinderen hadden ook een groter vertrouwen in het proberen te lezen van moeilijke woorden wanneer deze in een verhaalcontext ingebed waren. De resultaten toonden ook aan dat wanneer woorden langer werden, kinderen zich meer en meer op sublexicale strategieën baseerden om deze woorden te lezen. Bovendien, wanneer woorden langer werden of de relatieve moeilijkheid van een verhaal groter werd, durfden kinderen minder vaak proberen om moeilijke woorden te lezen. Er was ook een leeftijdsgebonden verschuiving in het gebruik van sublexicale en lexicale strategieën om woorden te lezen: hoe ouder kinderen werden, hoe meer ze zich baseerden op lexicale strategieën om zowel losse echte woorden als echte woorden ingebed in een verhaal te lezen. Ook hoe lager de decodeervaardigheden van een kind, hoe meer dit kind zich baseerde op lexicale strategieën om een onbekend woord ingebed in een verhaal te lezen. Tot slot toonden de resultaten aan dat meisjes minder vaak dan jongens een sublexicale strategie gebruiken om onbekende losse woorden te lezen.

De laatste studie van dit proefschrift beschrijft een interventiestudie waarin een gerandomiseerd pretest-posttest gematchte-groepen-design werd gebruikt om te onderzoeken of de leesaccuratesse en leessnelheid van basisschoolkinderen met leesproblemen kon worden verbeterd door gedurende 4 weken met een prototype van een computergestuurde leestrainer met TTS-feedback te oefenen voor een totaal van ongeveer 4.5 uur. De interventie bestond uit het lezen van verhalen op instructieniveau waarbij gesynthetiseerde foneem-per-foneem feedback werd gegeven. Elk kind uit een gematcht paar werd willekeurig toegewezen aan een interventie- ($N = 10$) of controlegroep ($N = 10$), en werd vooraf en achteraf (één week en vier weken na interventie) getest door middel van zowel een pen-en-papier als een computergestuurde woordleestest, pseudowoordleestest, en verhalenleestest. Mixed model covariantieanalyses toonden aan dat de interventiegroep beter presteerde dan de controlegroep tijdens het lezen van losse echte woorden en pseudowoorden m.b.t. zowel leesaccuratesse als leessnelheid. Bovendien bleken verschillen in het pre-interventie-leesniveau van kinderen een differentieel effect op de interventiedoeltreffendheid te hebben.

Contents

Acknowledgements **ix**

1 General Introduction **1**

2 CHildren's Oral REading Corpus (CHOREC): Description and Assessment of Annotator Agreement **11**

3 A Quantitative Oral Reading Error Analysis in Elementary School Children: Evidence from Dutch **37**

4 Oral Reading Errors of Dutch Speaking Elementary School Children: A Qualitative Analysis at the Word Level **57**

5 Evaluation of a Reading Tutor with Synthesized Speech Feedback for Dutch Speaking Elementary School Children with Reading Difficulties **83**

6 General Conclusions **105**

A Test Material **121**

B Reading Strategy Labels **143**

C Reading Error Labels **145**

D Phonetic Symbols **151**

References **153**

Acknowledgements

Many people have contributed to this dissertation in many different ways. First, I would like to acknowledge my promotor Prof. Dr. Pol Ghesquière and copromotor Prof. Dr. Ir. Hugo Van hamme for giving me the opportunity to participate in an inspiring interdisciplinary research project, and for their advice and suggestions during the development of this doctoral research. Likewise, I would like to thank my colleagues at the SPACE project and in particular Dr. Jacques Duchateau for his insightful comments and constructive criticisms. I would also like to thank Dr. Bart Boets, Dr. Jacques Duchateau, Prof. Dr. Rianne Janssen, Prof. Dr. Lieven Verschaffel, and Prof. Dr. Inge Zink for being members of the supervision committee, and Prof. Dr. Marc De Bodt, Prof. Dr. Rianne Janssen, and Prof. Dr. Lieven Verschaffel for being members of the doctoral jury.

My greatest gratitude also goes to my colleagues and former colleagues at the Centre for Parenting, Child Welfare, and Disabilities for their friendship and support, and for the warm atmosphere that always characterized the 'fifth floor' as well as the 'other side'. In particular, I would like to thank Dr. Sofie Kuppens for her generous help and advice on statistical analyses, but above all for being a really great friend. Finally, I would like to address a very special thank you to Kevin who has always been able to make me laugh when I was worried, to reassure me when I was feeling insecure, and to simply be my very best friend. *Jongen*, thank you.

For the financial support to complete this research, my acknowledgements are due to the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT-Vlaanderen, Belgium) that financed the SPACE project (sbo/040102): SPeech Algorithms for Clinical and Educational applications, home page: http://www.esat.kuleuven.be/psi/spraak/projects/SPACE.

Chapter 1

General Introduction

1.1 Using Automatic Speech Recognition in the Assessment of Word Decoding Skills

1.1.1 Rationale

In Flanders, elementary school children's progress is regularly assessed in order to detect early learning difficulties such as reading disabilities. Because adequate reading (with good comprehension) can be seen as the resulting product of decoding (or 'technical' reading) and linguistic comprehension, both skills should be assessed ('simple view of reading'; Georgiou, Das, & Hayward, 2008; Gough & Tunmer, 1986; Hoover & Gough, 1990). Nevertheless, although there is much more to adequate reading than word decoding, there is never reading without decoding. Moreover, being able to accurately and fluently decode printed words is an important prerequisite to come to reading comprehension (Samuels, 2002). Therefore, the *assessment of word decoding skills* was the first main focus of the research presented in the present dissertation[1].

An adequate battery to assess word decoding skills typically contains instruments that ask a child to read aloud unrelated real words, pseudowords (i.e., words that do not exist but are well pronounceable in a particular language), or stories (of graded difficulty level) (cf., Jenkins, Fuchs, van den Broek, Espin, & Deno, 2003; Protopapas & Skaloumbakas, 2007; Protopapas, Skaloumbakas, & Bali, 2008). According to the dual-route theory of reading[2] (cf., Coltheart, Curtis, Atkins, & Haller, 1993), when reading real words, a

direct, lexical route is most often used for well-known and exception words, whereas an indirect, nonlexical, phonological route is used for low-frequency and new regular words. However, the assessment of a child's ability to read pseudowords provides specific information on the efficiency with which the indirect, nonlexical, phonological route is followed since pseudowords are presumed to have no lexical representation. Assessment of connected text reading (i.e., stories) is important to obtain a good indication of the actual difficulties experienced in academic situations in which the reading of texts usually plays an important role. Although the presence of context (defining syntactical and semantic restrictions) has been associated with fewer oral reading errors compared to reading series of unrelated (pseudo)words (e.g., Allington & McGill-Franzen, 1980; Archer & Bryant, 2001; Goodman, 1965), it has also been regarded as an additional burden potentially accentuating the word-level reading problem via cognitive resource straining (Protopapas et al., 2008).

Examples of well-known and commonly used standardized real word reading tests in Flanders are the 'One-Minute-Test' (Brus & Voeten, 1995), the 'Three-Minutes-Test' (Verhoeven, 1995), and the 'Tests of LVS-CSBO' (Bilhaert & Grysolle, 1998). An example of a commonly used standardized pseudoword reading test is 'de Klepel' (van den Bos, Spelberg, Scheepstra, & de Vries, 1994); whereas at the sentence level, a well appreciated standardized test in Flanders is the 'AVI test' (Visser, van Laarhoven, & ter Beek, 1998). Although the quality of these paper-and-pencil instruments is not questioned, it is clear that the administration of these tests is very time-consuming: a teacher or clinician needs to sit besides the child and scrupulously listen to his/her reading in order to detect reading errors or to monitor timings, or has to record the reading session and needs to perform this time-consuming assessment process afterwards. Another disadvantage of this way of assessment is that, although usually performed by experienced examiners who, normally, used a same assessment protocol, the evaluation suffers from examiner bias. This is, for instance, due to differences in familiarity with the child under assessment, or due to differences in the choice to do the assessment 'on line' or afterwards based on the recorded reading session. Third, in a language such as Dutch which has fairly consistent grapheme-to-phoneme correspondence rules, children suffering from a phonological deficit (lying at the basis of their reading difficulties) will usually make up arrears regarding reading accuracy when adequate instruction is given. Their reading speed, however,

will remain a stumbling-block (e.g., de Jong & van der Leij, 2003). Therefore, apart from an accurate detection of oral reading errors, a precise registration of the overall time needed to complete a particular test and the time needed to read a particular word in that test is equally important. However, the registration of a detailed timing measure at the word level is practically unfeasible in paper-and-pencil reading tests.

Recently, a change in assessment procedure has been proposed as a solution for the first problem associated with traditional word decoding assessment at the text level. More specifically, to save time in the assessment of a child's story reading performance by means of the AVI test, a recent update of this test has been accompanied with a different assessment strategy: first, children are assessed by means of a collectively administered 'silent' reading test; after that, only the poorest achievers or children who stagnated or obtained a dubious result on this test are additionally assessed by means of the individually administered and 'oral' AVI test (cf., Cleuren & Ghesquière, 2008; de Wijs, 2008). Another alternative to address the disadvantages associated with traditional paper-and-pencil assessment, and one which has received large attention internationally, is the development of computerized tools to identify reading disabilities (Protopapas et al., 2008). However, to develop entirely automated tools that need no intervention or supervision, no tasks with spoken responses can be included. Instead, a child needs to use a mouse, keyboard, or touch screen response.

A way to overcome all disadvantages of traditional word decoding assessment at once ánd to enable a child to really read aloud, is by implementing *automatic speech recognition* in the reading assessment process. By means of a speech recognizer, oral reading assessment can become automated and objective, and accurate measures of a child's reading performance can be obtained. An increasing amount of people are getting used to automatic speech recognition in their daily lives. Using speech to control your stereo or mobile phone, or to dictate texts to your computer does no longer sound futuristic. In these types of applications, the speech recognizer's goal is to detect what the speaker intended to say and not to know what the speaker exactly said in case of hesitations or mispronunciations. In contrast, automatic speech recognition within the context of reading assessment (and instruction, cf., infra) not only needs to be able to track the child's reading (i.e., to know at each moment in time which word exactly a child is reading) but also to accurately detect reading errors and, in case of hesitations or mispronunci-

ations, to know what exactly happened so that a qualitative diagnosis can be made and appropriate feedback can be given (Duchateau et al., 2009). Given the fact that acoustical differences between two different phonemes (e.g., /A/ in 'man' [*man*] versus /a/ in 'maan' [*moon*]) can be very small, it is a very difficult task for a speech recognizer to reliably decide whether a child pronounced a particular phoneme 'well enough', or whether he/she clearly read a different phoneme than the one that was written down. Moreover, regional influences can result in acoustical differences that are equally small and which make an accurate recognition even more difficult to obtain. As such, in spite of the availability of the text being read, automatic speech recognition within the context of reading assessment (and instruction) is a very challenging task.

1.1.2 Development of a Speech Recognition Based Assessment Tool

For the development of a speech recognition based reading assessment tool, there were two options, namely: (a) to develop a computerized version of the already existing standardized paper-and-pencil Three-Minutes-Test, Klepel, and AVI test (cf., supra), or (b) to develop a completely new computerized reading assessment battery. However, since the aim was to develop a *computerized* reading assessment tool, the fact that children were asked to read computer presented text instead of traditional paper presented text had to be taken into account; in other words, the so-called 'test mode effect' had to be dealt with. This effect, on which past research outcomes have been inconsistent, refers to the fact that identical paper-and-pencil and computerized tests do not always result in equivalent (reading) performance with respect to speed, accuracy, and comprehension. Both reader characteristics (e.g., familiarity with computers) and technological issues (e.g., screen size and resolution) can lie at the basis of differences in outcome between the two modes (cf., Bodmann & Robinson, 2004; Clariana & Wallace, 2002; Hadwin, Winne, & Nesbit, 2005; Hardré, Crowson, Xie, & Ly, 2007; Kim & Huynh, 2008; Leeson, 2006; McDonald, 2002; Sim & Horton, 2005; Wang, Jiao, Young, Brooks, & Olson, 2008).

As a consequence of the test mode effect, a computerized version of already available standardized paper-and-pencil reading tests would have asked for new reference group standards anyway. Therefore, it was decided to de-

velop a whole new computerized test battery to assess children's unrelated real word reading, pseudoword reading, and story reading performance. At the same time, some objections often raised with respect to the 'Klepel' (van den Bos et al., 1994) and the 'AVI test' (Visser et al., 1998) could be taken into account. These objections included the fact that the pronunciation of particular pseudowords of the 'Klepel' caused confusion, and second, the fact that some stories of the 'AVI test' contained vocabulary that sounded very northern Dutch and were therefore not suitable for the Flemish population. In imitation of the procedure used in the Three-Minutes-Test, it was also decided to develop a real word and a pseudoword reading test which both consisted of three lists of (pseudo)words of different length (instead of one list containing words of increasing difficulty as is the case in the One-Minute-Test). As such, a better view could be provided on a child's reading performance at a particular level of word difficulty (as expressed by the number of syllables). Finally, it was decided to construct a real word and pseudoword reading test in which each child read the same number of words so that the time a child needed to read the list varied.

The resulting Real Word Reading Test (RWRT) and PseudoWord Reading Test (PWRT) contained three lists of respectively 40 1-syllable, 40 2-syllable and 40 3- or 4-syllable real words or pseudowords. The Story Reading Test (SRT) consisted of 9 graded text stories (vocabulary of 538 distinct word forms), ranging from AVI 1 to AVI 9[3]. Also, a reading tutor platform (on which more detailed information can be found in paragraph 1.2.2) was developed so that the computerized presentation of words and pseudowords could be done one-by-one, whereas stories could be presented paragraph-by-paragraph so that no scrolling through a page was needed (cf., Kim & Huynh, 2008). At the same time, the platform provided the possibility to make high quality recordings of a child's readings which were necessary to ensure good automatic speech recognition. In Appendix A, the exact content of each test can be found as well as information on the selection process of RWRT words and PWRT pseudowords, and on the construction of the different SRT stories.

1.2 Using Automatic Speech Recognition and Synthesis in the Remediation of Reading Difficulties

1.2.1 Rationale

Not all children arrive at fluent, effortless reading which is so crucial for reading comprehension. Whenever reading problems are detected, early, regular (preferably daily), and adequate intervention is highly needed in order to prevent the child from dropping further and further behind his or her classmates, not only with respect to his or her reading development but also with respect to appropriate functioning in other school subjects since text usually plays an important role in all of them (Strickland, 2002). Therefore, the *remediation of reading difficulties*, and more specific word decoding difficulties, was the second main focus of the research presented in the present dissertation.

Daily reading intervention can often not be provided due to practical classroom issues (e.g., often only one teacher is in charge of a group of 20 to 30 children) and governmental health insurance constraints (e.g., only a limited amount of individual reading therapy is covered by the national health care system). Therefore, the use of a computerized reading intervention program could offer a welcome solution. As correctly pointed out by Irausquin, Drent, & Verhoeven (2005), the use of a computer has different advantages:

> Computers allow children to work independently, at their own pace, and with as much repetition as needed. Moreover, with the help of computer-assisted instruction, the difficulty level of the exercises can be increased step by step, in an adaptive way, to prevent frustration. The computer also offers possibilities beyond those of written exercises such as precise setting and adjustment of presentation times for words or letters (e.g., flash exercises), simultaneous presentation of visual and auditory information that enables online assisted reading with the computer, and direct feedback. Last but not least, children are often very motivated to work with computers. (p. 248)

Moreover, a computer equipped with automatic speech recognition and speech synthesis can do even more: it can offer individualized, pedagogically sound intervention sessions without the need for a clinician or teacher to be

around. More specifically, *automatic speech recognition* can be implemented in reading instruction in a similar way as has been discussed with respect to a speech recognition based reading assessment tool (cf., supra). Automatic speech recognition enables a computer to act as a real reading tutor which can present reading material to a child and is able to really listen to a child's reading performance: it can track a child's progress and therefore follow as he/she reads along, and can detect reading errors so that immediate feedback on these errors can be given. Additionally, computers are well suited to provide different kinds of speech feedback and decoding assistance for difficult words by means of either digitized (recorded) or *synthesized speech*[4] and the ability to highlight words or word parts (MacArthur, Ferretti, Okolo, & Cavalier, 2001). The offering of feedback has been shown to be of great importance for the development of reading for both children with and without reading difficulties (e.g., McCoy & Pany, 1986; Pany & McCoy, 1988; Perkins, 1988; Spaai, Ellermann, & Reitsma, 1991; Wise & Olson, 1998). Since a beginning or struggling reader who makes a reading mistake has great difficulties correcting him- or herself, a constantly available intervening computerized reading tutor equipped with automatic speech synthesis to provide educationally sound feedback could be very helpful.

As was the case for automatic speech recognition, people are also increasingly getting familiar with computer speech in their daily lives. For instance, digitized speech is increasingly being used by public transport companies to broadcast information on times of arrival, platform numbers, delays etc., as well as by GPS car navigation systems to inform their users which exit to take, which road to continue on etc. More advanced uses of computer speech (i.e., synthesized speech) can be found in, for instance, websites that can automatically be read to people with a visual disability, or in applications for people who are not able to use their own voice. Within the context of reading instruction, more sophisticated synthesized speech (as opposed to digitized speech) is needed to ensure that programs can be 'open' so that its user is not restricted to the learning material chosen and implemented by the authors of the program. More specifically, automatic speech synthesis in reading instruction can be used to give instructions, to encourage the child, to read to or read along with the child, to give feedback when an error was made, or to provide help when this was asked for. As such, whereas the goal of most speech synthesis applications is only to provide fluent speech, automatic speech synthesis to be used in reading instruction must also be able to

synthesize separate phonemes or syllables as these synthesis modes are necessary to give educationally sound feedback to the learning child (Duchateau et al., 2009).

1.2.2 Development of a Speech Recognition and Speech Synthesis Based Reading Intervention Tool

The motive to develop a speech recognition and speech synthesis based reading tutor was not to let it completely replace any form of human support, but rather to let it be of use in situations where no human help and tutoring was available. Four main modules could be distinguished: (1) a module for the construction of new reading exercises which were presented word-by-word or paragraph-by-paragraph on several consecutive screens (in case of unrelated word reading respectively connected text reading), (2) a module for the management of students and the assignment of particular exercises to a particular student, (3) a module to actually exercise with the reading tutor while recordings were made of every reading session, and (4) a module for the management and assessment of the recorded reading sessions (Duchateau et al., 2009).

Automatic speech recognition was used to automatically track where the child was exactly reading so that the progress through the reading task was known. Such accurate tracking was necessary to enable the reading tutor to automatically advance from one screen to the next as the child had read the last word of the previous screen. However, the information provided by the tracker was also used to generate feedback for the child (e.g., by highlighting the word he/she should read), or was used in exercises in which only the words already read and the very next word to be read were visible (while the rest of the text remained masked). Information provided by the tracker was also used to generate feedback for the speech synthesizer that could use this information to slow down, speed up, or restart the speech synthesis while reading along with the child. Finally, automatic speech recognition was used to assess a child's reading speed as well as reading accuracy at the word level so that information on the total time needed to complete the exercise as well as on the number of correctly read words, the number of correctly read words at a first try, and on the number of omitted words could be provided. Moreover, this information could also be used to provide additional exercising to the child for only those words that were not read correctly (at a first try).

Synthesized speech was used to enable the reading tutor to read to or read along with the child as well as to provide three different kinds of oral corrective feedback or help whenever an error was made against that word, help was asked for, or the child remained silent: whole-word-feedback, syllable-by-syllable feedback, and phoneme-by-phoneme feedback (cf., McCoy & Pany, 1986; Perkins, 1988; Spaai et al., 1991; Wise & Olson, 1998). In case of the first type of feedback, the computerised reading tutor supplied the child with the whole word (e.g., mother); the second and third types of feedback concerned the provision of segmentation cues by segmenting the erroneously read word into its constituent syllables (e.g., mo-ther) or phonemes (e.g., m-o-th-e-r). Additionally, a synthesiser was built that was able to emphasise particular phonemes during phoneme-by-phoneme feedback since human teachers or clinicians seem to emphasise the wrongly read phoneme in a word during feedback (which was noticed during some own clinical experience in the field of reading intervention). For example, if a child has to read the word '**w**ood' but reads the word '**m**ood' instead, the human tutor tends to overemphasise the phoneme /w/ in the segmental (phoneme-by-phoneme) feedback presented to the child.

1.3 Aim and Outline of This Dissertation

First, to enable a speech recognizer to track which (pseudo)word or where in the text a child is reading, and to accurately detect oral reading errors, the CHildren's Oral REading Corpus (CHOREC), a Dutch database of recorded and annotated children's readings, was developed so that the speech recognizer could be supplied with detailed information on children's oral reading behaviour. Also, to ensure that the annotation quality was high, an assessment of annotator consistency was performed (Chapter two). However, this database was not only of use for the development and testing of the speech recognizer, but also offered the unique possibility to reading researchers to accurately characterize the phenotypic reading performance pattern of children with and without reading disabilities with respect to oral reading errors. As a result, both a quantitative and qualitative analysis of oral reading errors was carried out (Chapters three and four). Then, an intervention study was carried out to investigate whether improvements in reading accuracy and speed could be obtained in elementary school children with reading difficulties who exercised with a preliminary version of the reading tutor (Chapter five). To

conclude, Chapter six summarizes the main findings of this dissertation and additionally highlights some important methodological findings and practical implications. It also discusses some limitations of the presented research, and provides some ideas for future research.

Notes

[1]The present dissertation is the result of research that has been conducted within the scope of a research project called 'SPeech Algorithms for Clinical and Educational applications' (SPACE). The main goals of this project were: (1) to improve speech recognition and speech synthesis technologies to make them suitable as supporting technologies for educational and clinical applications, and (2) to demonstrate the benefits of speech technology based tools for these applications, namely for the development of an automated reading tutor (used for assessment and remediation) and for the development of a tool to measure intelligibility of pathological speech (e.g., dysarthria). More information can be found on http://www.esat.kuleuven.be/psi/spraak/projects/space/.

[2]Dual-route theories have been challenged (Coltheart et al., 1993). For example, the reading of pseudowords has been shown to be influenced by their similarity to real words, which indicates that they were not read only by grapheme-to-phoneme correspondence rules (i.e., an indirect, nonlexical, phonological route) (Humphreys & Evett, 1985). An alternative model that does not have a dual-route architecture is the parallel-distributed-processing (PDP) connectionist model of Seidenberg & McClelland (1989) in which both routes are assumed to interactively operate in parallel.

[3]In the Netherlands and Flanders, the AVI index is used to distinguish between texts of different technical difficulty level. The index is largely based on the reading index A (Brouwer, 1963) which takes into account word, sentence and text features. However, only very recently, CITO (i.e., Centraal Instituut voor Toetsontwikkeling [*Central Institute for Test Development*]) proposed a new reading index to base the AVI index on, namely the CILT index. This index is based on the percentage of high-frequency words and average word length (in letters) (cf., Cleuren & Ghesquière, 2008; de Wijs, 2008)

[4]Digitized speech sounds like recorded speech; it is speech that is converted into a digitized format and then reconverted back into speech. However, synthesized speech is computer generated speech that is created by concatenating pieces of recorded speech that are stored in a database, based on various acoustic/phonetic algorithms (Wilkinson & Hennig, 2007).

Chapter 2

CHildren's Oral REading Corpus (CHOREC): Description and Assessment of Annotator Agreement

Abstract[1]

Within the scope of the SPACE project, the CHildren's Oral REading Corpus (CHOREC) was developed in order to statistically characterize children's reading behaviour, and to supply a speech recognizer with a model that contains information on the nature and prevalence of common oral reading errors. This database contains recorded and annotated read speech (42 GB or 130 hours) of 400 Dutch speaking elementary school children with or without reading difficulties. Analyses of inter- and intra-annotator agreement were carried out in order to investigate the consistency with which reading errors were detected, orthographic and phonetic transcriptions were made, and reading errors and reading strategies were labelled. Percentage agreement scores and kappa values both showed that agreement between annotations, and thus the quality of the annotations, was high. Taken all double or triple annotations (for 10% resp. 30% of the corpus) together, % agreement varied between 86.4% and 98.6%, whereas kappa varied between 0.72 and 0.97 depending on the annotation tier that was assessed. Degree of annotation detail, school type and reading task type seemed to account for systematic differences in % agreement, but these differences disappeared when kappa values were calculated. Finally, annotation differences with respect to phoneme labels, reading strategy labels and reading error labels were analyzed.

2.1 Introduction

In Flanders (Belgium), elementary school children's school progress is regularly assessed in order to detect early learning difficulties such as reading disabilities (RD). With respect to a child's reading development, much attention is paid to the assessment of word decoding skills as these are crucial for adequate reading. Whenever word decoding problems arise, early, regular and adequate intervention is highly needed in order to prevent a child from dropping further and further behind his or her classmates, not only with respect to his or her reading development but also with respect to appropriate functioning in other school subjects since text usually plays an important role in all of them.

To detect RD in children, various valuable paper-and-pencil diagnostic instruments are being used at the moment (e.g., Visser et al., 1998; Verhoeven, 1995; van den Bos et al., 1994). Unfortunately, although the quality of these instruments is not questioned, this traditional paper-and-pencil reading assessment process is a very time-consuming one, taking away valuable time that could much better be invested in the actual guidance and treatment of those children that experience reading difficulties. Moreover, another disadvantage of this way of reading skill assessment is that its evaluation suffers from examiner bias. As a result, an automated, more reliable and objective reading assessment tool is in great demand.

A way to address these disadvantages is by implementing automatic speech recognition in the reading assessment process. Recently, different research projects have made a great effort in implementing speech technology in the assessment (and intervention) of reading difficulties, such as the LISTEN project (Carnegie Mellon University) (Mostow & Aist, 2001), the Foundations to Literacy Reading Tutor project (Colorado University) (Wise et al., 2005), and the TBALL project (UCLA) (Kazemzadeh et al., 2005). In Flanders, the SPACE project (SPeech Algorithms for Clinical and Educational applications) aimed to automate the reading assessment process and to make it more objective by using speech recognition technology. Additionally, the project wanted to develop interactive, speech synthesis supported tools that enable a virtual reading tutor to act as a fluent reader model and to provide immediate feedback to the learning child by means of natural, highly intelligible, and phonetically correct speech. More information on the SPACE project can be found here: http://www.esat.kuleuven.be/psi/spraak/projects/SPACE.

In spite of the availability of the text being read, the use of automatic

speech recognition within the context of reading assessment and instruction is a very challenging task due to child-related and reading-related developmental processes. For example, due to a child's variable maturation, articulatory competencies can differ and due to variable word decoding skills, oral reading of novice readers or readers with RD can be fraught with oral reading errors. Within the scope of the SPACE project - in order to improve the speech recognizer's ability to track a child's reading and to accurately detect oral reading errors - CHOREC (CHildren's Oral REading Corpus) was developed, a Dutch database of recorded and annotated children's oral reading. CHOREC provides sufficient data to statistically characterize children's reading behaviour, and to supply the speech recognizer with a model that contains information on the nature and prevalence of common oral reading errors (cf., Mostow, Beck, Winter, Wang, & Tobin, 2002). However, this database is not only of use for the development and testing of speech technology applications. It also offers the unique possibility to researchers in the field of learning disabilities to accurately characterize the phenotypic reading performance pattern of children with and without reading disabilities with respect to reading strategy use and reading error occurrence. As such, this contribution adds to the existing children's speech databases for both speech technology and educational research.

Before this corpus can be used for automated analysis, preferably all recorded reading sessions are manually segmented, transcribed and labelled. The quality of these human annotations relies heavily on various annotator characteristics, such as familiarity with the recorded material, amount of annotation training, motivation etc.; and on various external influences such as time pressure, changes in annotation protocol etc. (Bayerl & Paul, 2007). It is clear then that a quality estimation of these annotations, namely an analysis of annotator agreement and consistency in segmentation, transcription, and labelling, is recommended (Cucchiarini, 1996). To do so, different methods to assess the quality of corpus annotations have been proposed in the literature, such as pairwise percentage agreement and the kappa statistic (e.g., Dilley, Breen, Bolivar, Kraemer, & Gibson, 2006; Kazemzadeh et al., 2005; Pitt, Johnson, Hume, Kiesling, & Raymond, 2005; Yoon, Chavarria, Cole, & Hasegawa-Johnson, 2004).

In this chapter, the CHOREC database and its annotation protocol are described. Then, two annotation tiers are highlighted: the oral reading strategy tier and the oral reading error tier. Finally, an analysis of inter- and

intra-annotator agreement is presented with respect to the transcription and labelling of recorded children's oral reading. More specifically, attention goes out to the consistency with which orthographic and phonetic transcriptions were made, reading errors were detected, and reading errors and reading strategies were labelled.

2.2 Method

2.2.1 Corpus Creation and Annotation

Participants

The speakers/readers in the CHOREC corpus were 400 Dutch speaking elementary school children (6- to 12-years-old) enrolled in a Flemish regular elementary school ($N = 274$) or in a school for children with specific learning disabilities ($N = 126$). For all children, parental consent was obtained, as well as relevant information on grade, age, gender, mother tongue, and school curriculum (see Table 2.1), and on place of birth, place of residence, reading level, reading method used to learn how to read, and presence of reading disabilities.

Reading Material

For every child, a newly developed computerized reading test battery was administered which contained a real word reading test (RWRT), a pseudoword reading test (PWRT), and a story reading test (SRT). Both the RWRT and the PWRT contained three lists of respectively 40 1-syllable, 40 2-syllable and 40 3- or 4-syllable real words or pseudowords. The SRT consisted of nine graded text stories (vocabulary of 538 distinct word forms), ranging from AVI 1 to AVI 9 in difficulty, and 103 to 223 words in length. Real words and pseudowords were presented individually; text stories were presented paragraph-by-paragraph on the computer screen. Each child read minimally one and maximally three real word lists and pseudoword lists, and minimally one and maximally four text stories depending on that particular child's reading level. Recordings were done in three sessions of maximally 20 minutes each.

	Regular Schools (N = 274)				Special Schools (N = 126)
	Grade 1	Grade 2	Grade 3	Grade 4	
Age Range (years)	5.93-7.15	7.09-9.04	8.09-9.59	8.92-10.71	7.03-13.23
Gender					
N Boys	25	40	34	39	95
N Girls	30	38	34	34	31
Mother Tongue					
N Dutch	55	78	61	71	117
N Other			1	2	9
School Curriculum[a]					
N Normal	51	74	62	64	
N Behind	1	4	6	7	
N Ahead	3			2	

Note. [a] Regular school children could be on schedule with respect to their school curriculum, or could have doubled or skipped a grade.

Table 2.1: *Several characteristics of the speakers/readers in CHOREC*

Recording Material and Setup

Recordings were made in any room available at the participating school. People were restricted from entering the room during recordings, but environmental noise such as school bells, children entering the play ground etc. could not be avoided. For the making of the recordings, a reading tutor platform was developed, a tool - installed on a laptop computer - that enabled us to record each list of words and each story separately. Children's speech was recorded at 22050 Hz by means of two microphones: a close-talking microphone and a desk microphone, both connected with the laptop through a preamplifier. In total, 42 GB or 130 hours of speech was recorded.

Children were instructed to try to read the words, pseudowords or paragraphs presented on the screen, as accurately and as fluently as possible.

Most children were very motivated and eager to participate in the reading recordings. Some children however, often having severe reading difficulties, needed to be motivated by the promise that they could listen to their own recordings afterwards (cf., Kazemzadeh et al., 2005).

Annotation Procedure

The recordings were segmented, transcribed and labelled manually by means of a customized version of 'Praat' (http://www.Praat.org/), a free-source tool for speech analysis and synthesis. This tool provided the possibility to attach a time-aligned text-grid (containing different tiers) to the speech sound. As such, each tier provided another layer of descriptive information about the speech sound that particular tier was attached to. The eight annotation tiers used in the CHOREC database included both information on utterances directly resulting from the child's efforts to read what was presented on the computer screen, as well as information on background noise and reading task-related and task-unrelated unforeseen utterances made by both the child and the examiner. Each tier originally contained segment boundaries corresponding to the timing of new stimulus presentation. In case of the RWRT and PWRT, boundaries indicated the presentation of a new real word or pseudoword; in case of the SRT, they indicated the presentation of a new paragraph.

Initially, when a recording was loaded in Praat, the first two tiers, that contained the original reading task (with original resp. adjusted boundaries), were automatically loaded along with it. Then, two passes were made through each speech file during annotation. Through a first pass (resulting in a 'partial' (p01) annotation file containing six tiers), the following four tiers were annotated and added: (1) orthographic transcription tier, (2) a tier for the broad-phonetic transcription of what was actually read, (3) a tier for the annotation of utterances made by the examiner, and (4) a tier for the annotation of background noise. Through a second pass, only for those words the child hesitated or made a reading error, labels of oral reading strategies and errors were added to the p01-annotation file (resulting in a 'full' (f01) annotation file containing eight tiers). During annotation, the annotator could rely on the audio signal and on visual speech wave representations of the recorded speech. Each annotation tier can be briefly described as follows:

Tier 1 In order to carefully describe and annotate oral reading recordings, it was often advisable to move or remove existing boundaries or insert new ones whenever the child hesitated or made a reading error. This implicated, however, that information about the exact timing of reading stimulus presentation was lost. Because latency and production times could be of great importance for situating readers within a developmental model of reading (cf., Ehri, 1991) or for discriminating fast direct word recognition strategies from slow internalized or externalized unit-by-unit decoding strategies, tier 1 was used for preserving the original boundaries (and the original reading task) corresponding to the exact timings of word and paragraph presentation.

Tier 2 As in tier 1, tier 2 did not allow for making changes with respect to the orthographic transcription of the original reading task. Segment boundaries, however, could be manually moved or removed and newly inserted when necessary for annotation. This meant that this tier allowed for the insertion of new segments whenever needed. Segment boundaries in tiers 3, 4, 5, and 6, were linked to those in tier 2.

Tier 3 and 4 As was the case in tier 1 and 2, tier 3 initially contained the orthographic transcription of the original reading task. Tier 4 initially consisted of a concatenation of lexicalized phonetic transcriptions, including all possible correct pronunciations but ignoring crossword assimilation. If the reading task was read fluently without any errors or hesitations, tier 3 and 4 remained unchanged. However, whenever the child struggled, made reading errors (also including word omissions and real word substitutions) in a particular word or inserted words or interjections unrelated to the printed reading task, manual adjustments to this word had to be made in both tier 3 and 4 in order to capture each of these reading attempts and errors. Appearance of assimilation between words or dialect influences within a word was only captured whenever changes had to be made to tier 3 and/or tier 4.

Tier 5 Tier 5 provided the possibility to describe oral reading behaviour used to arrive at the correct or incorrect reading of words. More details about the labels used in the reading strategy tier can be found in Appendix B.

Strategies for arriving at the correct reading of words, besides correct decoding within the first trial, were: (a) repeating a correctly decoded word

once or more, (b) partially or completely, correctly or incorrectly spelling out a word before correctly synthesizing it, and (c) incorrectly decoding a word in the first trial but reading it correctly in the final trial. Strategies that did not result in the correct reading of words, besides incorrect decoding within the first trial, were: (a) partially or completely correctly or incorrectly spelling out and incorrectly synthesizing a word or not synthesizing it at all, (b) correctly decoding a word within the first trial but then 'correcting' it wrongly in a second trial, (c) mistakenly omitting or inserting a word, and (d) asking for a partial or complete prompt of a word before carrying on reading it.

Tier 6 Tier 6 was used for the classification of the type of oral reading errors occurring at a particular word during reading. To do so, an oral reading error classification based on surface level features was used, on which more details can be found in Appendix C.

Reading errors could occur at four different levels: at the level of the paragraph, the sentence, the word, or the grapheme. At the paragraph level, errors included omissions or repetitions of a whole line or sentence. At the sentence level, a part of a sentence could be omitted or repeated; a word could be mistakenly inserted; a word could change places with another word in the sentence; or a word could be substituted for a word being a synonym of or semantically related to that word. At the level of the word, reading errors were generally characterized by errors directly resulting from an incorrect compensatory decoding strategy. A spelling strategy could result in letter-by-letter or syllable-by-syllable reading without (correctly) synthesizing; and a guessing strategy could result in the substitution of a word or pseudoword by another word or pseudoword respectively, or the substitution of a pseudoword by a real word. Finally, four main categories of reading errors could be distinguished at the grapheme level: sequential errors, substitution errors, deletion errors, and insertion errors. Sequential errors included changing letter order of adjacent and nonadjacent letters. Substitution errors included vowel substitutions, consonant or consonant cluster substitutions, and vowel-consonant substitutions. Deletion and insertion errors were characterized by vowel deletions respectively insertions, consonant or consonant cluster deletions respectively insertions, and syllable deletions respectively insertions.

Tier 7 and 8 Tier 7 was used to orthographically transcribe possible extra utterances by the examiner. Segment boundaries were placed so as to surround the utterance as closely as possible. In tier 8, boundaries were put to catch background noise.

A detailed annotation protocol made sure that annotation was done uniformly by all annotators. The basis for this protocol was obtained by adopting the protocol used in the project CGN (Corpus Gesproken Nederlands, Spoken Dutch Corpus, see http://lands.let.kun.nl/cgn): for example, use of Praat, protocol on orthographic transcriptions, and protocol on broad phonetic transcriptions (same phone set and symbols). However, at some points, deviations from those protocols were needed. For instance, for CHOREC, some CGN annotation layers did not apply (e.g., syntactic annotation, lexicon link-up), while other annotation layers needed to be added (e.g., text to be read, reading strategy annotation, reading error annotation). Furthermore, the orthographic protocol could be simplified, and it now addressed the topic of disfluencies: since phonetic transcriptions were always available in CHOREC (in CGN only for 10%), a detailed orthographic transcription at the grapheme level of what had been said (during disfluencies) was not needed; as such avoiding a non-unique mapping between the grapheme and phoneme string.

A total of 10 annotators (1 doctoral researcher and 9 students), all native speakers of Dutch and none of them having prior experience in annotating audio files, worked in the project. The doctoral researcher supervised the project and trained each of the other annotators by letting them independently work through the annotation protocol and by giving them personal feedback after manually correcting their first two days of annotation work. After that, constant supervision was given so that questions regarding the annotation protocol and particular annotation problems could immediately be answered. After the annotations were finished, these were formally checked automatically and manually corrected if necessary.

2.2.2 Analysis of Inter- and Intra-Annotator Agreement

Re-annotations

To be able to assess inter-annotator agreement with respect to the detection of reading errors, and orthographic and phonetic transcriptions, p-annotation files were independently provided by 3 different annotators (A1, A2, and A3) for 30% of the corpus (p01, p02, p03). To assess intra-annotator agreement with respect to reading strategy and reading error labelling, f-annotation files were annotated twice by one annotator for 10% of the corpus, with a period of at least one month (up to one year) between the two annotations (f01, f02/f01b). For 50% of the double f-annotations, the reading strategy and reading error tiers were stripped from the f01-file which was then re-annotated for these tiers (resulting in an f01b-file). For the other 50% of double f-annotations, p02-files were used as the basis annotation to add the reading strategy and reading error tiers to (resulting in an f02-file). The audio files chosen for triple p-annotations (p01, p02, p03) and double f-annotations (f01, f01b/f02) were selected in such a way that they formed a representative sample of all reading tasks used in the making of CHOREC (RWRT, PWRT, SRT), and both school types (regular schools, special schools). The resulting database consisted of individual p- and f-annotation files for each list of (pseudo)words or each story read by the child, and for each version of re-annotation.

Tiers under Analysis

For the analysis of inter-annotator agreement, the focus was on the orthographic transcription tier (tier 3) and the phonetic transcription tier (tier 4); for the analysis of intra-annotator agreement, the reading strategy tier (tier 5) and the reading error tier (tier 6) were the tiers of interest. Figure 2.1 illustrates these 4 tiers by giving an example of an annotated erroneously read sentence.

In the orthography tier, a '*' denoted an erroneous attempt to read a word whereas a correct attempt was orthographically written down. If the erroneous attempt resulted in another existing Dutch word (e.g., 'als' means 'if'), this word was added between brackets. A '*s' visualized the fact that an annotator heard 'something' but could not write it out because she was not

Expected	Els zoekt haar schoen onder het bed. [Els looks for her shoe under the bed.]						
Observed	Als *(says 'something')* zoekt haar sch…schoen onder bed. [Als *(says 'something')* looks for her sh…shoe under bed.]						
	Els	zoekt	haar	schoen	onder	het	bed.
Orthography	*(als)	*s zoekt	haar	* schoen	onder		bed
Phonetics	Als	*s zukt	har	sx sxun	Ond@r		bEt
Strategy	f	*sg	g	agg	g	O	g
Error	e/4						

Figure 2.1: *Example of an annotated sentence*

able to distinguish the different reading attempts and to register how these attempts were exactly pronounced. Whenever the last attempt for a word was a '*' or '*s', this word was assessed as being read incorrectly (even if the child first read the word correctly but made an error or hesitated after that); when the last attempt was correctly read (and therefore was fully written down orthographically), this word was assessed as being read correctly (even if the child made one or more preceding errors or hesitations before this last correct attempt).

Second, the phonetics tier provided a broad phonetic transcription of what was actually read by the child. Before any manual annotation started, this tier automatically provided a concatenation of all correct phonetic transcriptions for a particular word. If all words in a segment were directly read correctly, these canonical transcriptions remained; else, all words in that segment were transcribed phonetically. '*s' was just copied from the orthography tier.

Third, the strategy used by a child to read a particular word was annotated in the reading strategy tier (e.g., 'f' = wrong within the first trial, '*sg' = correct after some nearly inaudible reading attempts, 'g' = correct within the first trial, 'agg' = repetition of an initial part of the word; 'O' = omission of a word). Fourth, for those words for which the last attempt was erroneous, the (combination of) reading errors made were labelled in the reading error tier. An error classification system containing 40 reading error categories (each category represented by a letter or a number) lied at the basis of this tier. In case of a correct last attempt, a '0' label was attributed to the word in order to make the total amount of entities in this tier correspond to the

amount of words the child had to read.

Thus, for each (erroneous or correct) attempt of the child to read a particular word, a different string (made of letters and/or tags) was annotated in the orthography and phonetics tier, such that the total amount of entities (strings) in the two tiers corresponded. In the reading strategy tier, one string of letters represented all reading attempts for a particular word; whereas in the reading error tier, a string of numbers (joined by a separator symbol) represented the errors made in the last reading attempt of a word. Therefore, the total amount of entities in these two tiers corresponded to the amount of words the child had to read.

Agreement Metrics

The two most popular measures to express annotator agreement were used in the present study: (1) percentage agreement, and (2) the kappa statistic. Also, for each of these measures, a 95% confidence interval (CI) was calculated in order to analyze whether differences in agreement were statistically significant.

Percentage agreement Pairwise inter-annotator agreement scores (expressed as % agreement) were calculated for each of the three pairs (A1-A2, A2-A3, A1-A3) of annotators; and this with respect to both orthographic transcription and phonetic transcription agreement. Intra-annotator agreement scores (expressed as % agreement) were calculated with respect to reading error and strategy labelling agreement. Here, the pairwise percentage agreement results represented only a single comparison pair (Yoon et al., 2004). Percentage agreement was obtained by dividing the total number of pairwise agreements between annotators for each word, by the total number of possible pairwise agreements (sum of total number of disagreements and total number of agreements), and then multiplying this by 100% (Cucchiarini, 1996). In the present study, percentage agreement scores were calculated at the word level.

Cohen's kappa When one wants to correct for chance agreement, the unweighted kappa statistic (which varies between 0 and 1) is commonly used to evaluate annotator agreement at the label level (for the exact formula see Cohen, 1960). A kappa statistic of 0.6 or higher indicated a substantial

agreement; a kappa of 0.8 or higher indicated an almost perfect agreement (Landis & Koch, 1977). In this study, kappa values were calculated at the label level.

Sometimes two annotations did not contain the same number of symbols so that the two strings of symbols were not equally long. As a result, before two strings could be compared on the level of their individual labels (i.e., in case of orthographic symbols, phonetic symbols, and oral reading strategy labels), they had to be aligned. To do this automatically, a dynamic programming algorithm was applied to obtain the best correspondence between symbols in the two strings (cf., Cucchiarini, 1996). In case of oral reading error labels, no alignment was needed since the order in which errors were labelled did not play a role. Instead, it was assessed whether two annotations agreed on the number and the types of error labels added to the oral reading error tier.

2.3 Results

Table 2.2 gives an overview of all % agreement scores and kappa values that were calculated to assess annotator agreement for the different annotation levels.

2.3.1 Description of Agreement Between Annotators

Detection of Reading Errors

The first analyses concerned the annotation of a word assessed as being read correctly (last attempt is the orthographic transcription of the word under assessment) versus a word assessed as being read incorrectly (last attempt is a '*' or a '*s'). Two annotators were said to agree if they both annotated the last attempt as a '*' or '*s', or both annotated the last attempt orthographically. For these analyses, added real words between brackets (cf., supra) were not taken into account.

The overall inter-annotator agreement for the detection of reading errors was 95.96% with a 95% confidence interval (CI) of 95.87-96.04 ($\kappa = 0.796$; 95% CI 0.792-0.800). When looking at agreement between annotators for the two different school types, the % agreement score for regular schools was higher than for special schools: 96.32% (96.22-96.42) versus 95.21% (95.04-

95.37); whereas the kappa value for regular schools was lower than for special schools: 0.779 (0.773-0.785) versus 0.816 (0.809-0.822) (both differences were statistically significant). When comparing agreement scores for the different reading tasks, there were more clear and statistically significant differences with respect to the % agreement scores: 95.20% (RWRT; 95.00-95.39) versus 90.59% (PWRT; 90.32-90.86) versus 98.37% (SRT; 98.30-98.44); as well as significant differences with respect to the corresponding kappa values: $\kappa = 0.735$ (RWRT; 0.725-0.746) versus $\kappa = 0.776$ (PWRT; 0.770-0.783) versus $\kappa = 0.794$ (SRT; 0.785-0.804).

Orthographic Transcriptions

The second group of analyses concerned the exact orthographic transcription. Two annotators were said to agree if both annotated a word exactly alike, namely with the same number of '*', and the same number of orthographic transcriptions of the word under assessment; and all that in the same order. For these analyses, '*s' was not taken into account but words between brackets were. Kappa values at the orthographic label level were not calculated because of the fact that there were exactly 2716 different labels in the orthography tier. According to Cicchetti (1972), a total of 2 x 2716^2 observations would have been necessary to obtain meaningful kappa values; a condition that was not fulfilled at all (number of observations = 259804).

The overall inter-annotator agreement for the making of orthographic transcriptions was 90.79% (95% CI 90.66-90.91). Transcriptions for regular schools showed significantly better agreement between annotators than for special schools: 92.13% (91.99-92.27) versus 87.93% (87.68-88.18). The agreement score for the annotation of SRTs (95.56%; 95.44-95.68) was significantly higher than for RWRTs (88.92%; 88.63-89.21), which was again significantly higher than for PWRTs (80.50%; 80.13-80.86).

Phonetic Transcriptions

A third cluster of analyses was carried out to assess agreement between annotators with respect to the broad phonetic transcription of the word under assessment. Two annotators were said to agree if their phonetic transcriptions of a particular word were exactly alike. For these analyses, only those words were assessed for which the canonical phonetic transcription, often containing several allowed pronunciations for that word, was reduced to only

		All Data	School Type		Task Type		
			Regular Schools	Special Schools	RWRT	PWRT	SRT
Reading Error Detection							
	(1)	95.96 %	96.32 %	95.21 %	95.20 %	90.59 %	98.37 %
	(2)	0.796	0.779	0.816	0.735	0.776	0.794
Orthographic Transcriptions							
	(1)	90.79 %	92.13 %	87.93 %	88.92 %	80.50 %	95.56 %
Phonetic Transcriptions							
	(1)	86.37 %	88.51 %	82.18 %	78.87 %	68.45 %	94.34 %
	(2)	0.930	0.937	0.917	0.907	0.888	0.964
Reading Strategy Labeling							
f01-f01b	(1)	98.64 %	98.72 %	98.45 %	98.35 %	96.75 %	99.26 %
	(2)	0.966	0.961	0.971	0.960	0.966	0.956
f01-f02	(1)	91.50 %	93.09 %	88.38 %	91.25 %	77.79 %	95.96 %
	(2)[a]	0.779	0.802	0.744	0.774	0.733	0.711
Reading Error Labeling							
f01-f01b	(1)	97.77 %	98.01 %	97.22 %	97.55 %	92.55 %	99.32 %
	(2)	0.911	0.899	0.921	0.896	0.575	0.933
f01-f02	(1)	94.14 %	95.39 %	91.71 %	94.57 %	80.88 %	98.24 %
	(2)	0.717	0.722	0.706	0.709	0.660	0.848

Note. [a] For the calculation of these kappa values, strings containing a '*s' strategy label were excluded.

Table 2.2: *% Agreement scores (1) and kappa values (2) for the different analyses*

one transcription option (otherwise, it was not known which transcription was the 'chosen' one). % agreement scores were calculated at the word level, whereas kappa values were calculated at the phoneme label level.

The overall inter-annotator agreement for the making of phonetic transcriptions was 86.37% (95% CI 86.20-86.54) with a corresponding kappa value of 0.930 (95% CI 0.929-0.930). Audio files from regular schools were annotated with significantly higher agreement than files coming from special schools: percentages agreement were 88.51% (88.31-88.71) versus 82.18% (81.85-82.51); the corresponding kappa values were 0.937 (0.936-0.937) versus 0.917 (0.915-0.918). The agreement percentage for phonetic transcriptions of SRTs (94.34%; 94.19-94.48) was much higher than those for RWRTs (78.87%; 78.35-79.38) and PWRTs (68.45%; 67.94-68.95); a tendency that was also shown by the kappa values for the different reading tasks ($\kappa = 0.964$; 0.963-0.965 vs. 0.907; 0.906-0.909 vs. 0.888; 0.886-0.889). All differences were statistically significant.

Reading Strategy Labelling

The fourth series of analyses concerned the intra-annotator agreement with respect to the labelling of reading strategies. Two annotations were said to agree if exactly the same labels in the same order were attributed to the word under assessment. Results for the f01-f01b and f01-f02 comparison are discussed separately. % agreement scores were calculated at the word level, whereas kappa values were calculated at the reading strategy label level.

f01-f01b comparison The overall intra-annotator agreement for the labelling of reading strategies was 98.64% (95% CI 98.41-98.83) whereas the kappa value was 0.966 (95% CI 0.961-0.970). When looking at agreement between annotations for the two different school types, % agreement scores were almost equal: 98.72% (regular schools; 98.45-98.94) versus 98.45% (special schools; 97.99-98.81); the kappa values were 0.961 (regular schools; 0.955-0.968) versus 0.971 (special schools; 0.964-0.978). However, the % agreement score for the annotation of SRTs (99.26%; 99.04-99.43) was significantly higher than for RWRTs (98.35%; 97.71-98.81), which was again significantly higher than for PWRTs (96.75%; 95.91-97.42). On the contrary, kappa values for the three different reading tasks did not show this tendency: $\kappa = 0.956$ (0.945-0.966) versus $\kappa = 0.960$ (0.949-0.971) versus $\kappa = 0.966$ (0.958-0.973)

(the differences were not statistically significant).

f01-f02 comparison The overall intra-annotator agreement for the labelling of reading strategies was 91.50% with a 95% confidence interval of 90.95-92.02 ($\kappa = 0.779$; 0.767-0.791). Reading strategy labelling for regular schools showed significantly better agreement between annotations than for special schools: 93.09% (92.47-93.66) versus 88.38% (87.28-89.39), $\kappa = 0.802$ (0.786-0.817) versus $\kappa = 0.744$ (0.725-0.763). The agreement score for the annotation of SRTs (95.96%; 95.45-96.42) was significantly higher than for RWRTs (91.25%; 89.96-92.39), which was again significantly higher than for PWRTs (77.79%; 75.93-79.54). However, when comparing the corresponding kappa values, it seemed that agreement was lower for SRTs ($\kappa = 0.711$; 0.686-0.736) than for RWRTs ($\kappa = 0.774$; 0.748-0.800), and that there were no significant differences in agreement between annotations for SRTs, RWRTs and PWRTs ($\kappa = 0.733$; 0.713-0.752). It must be noted here that, because of automatic alignment difficulties, annotation pairs containing a '*s' strategy label were not used in the kappa value calculations. In these cases, '*s' could be substituted for an enormous amount of different strategy label combinations, as such making it impossible to obtain interpretable aligned strings.

Reading Error Labelling

The last series of analyses were carried out to assess the intra-annotator agreement with respect to the labelling of reading errors. Two annotations were said to agree if exactly the same labels (not necessarily in the same order) were attributed to the word under assessment. Again, results for the f01-f01b and f01-f02 comparison are discussed separately. % agreement scores were calculated at the word level, whereas kappa values were calculated at the reading error label level.

f01-f01b comparison The overall intra-annotator agreement for the labelling of reading errors was 97.77% (95% CI 97.49-98.02); the corresponding kappa was 0.911 (95% CI 0.903-0.919). Audio files from regular schools were annotated with comparable agreement as files coming from special schools: 98.01% (97.68-98.29) versus 97.22% (96.63-97.71), $\kappa = 0.899$ (0.887-0.911) versus $\kappa = 0.921$ (0.909-0.933). The % agreement for phonetic transcriptions of SRTs (99.32%; 99.11-99.48) was significantly higher than that for RWRTs

(97.55%; 96.80-98.13) and PWRTs (92.55%; 91.35-93.59); the corresponding kappa's were 0.933 (SRT; 0.919-0.948) versus 0.896 (RWRT; 0.875-0.917) versus 0.575 (PWRT; 0.557-0.593).

f01-f02 comparison The overall intra-annotator agreement for the labelling of reading errors was 94.14% with a 95% CI of 93.67-94.57 ($\kappa = 0.717$; 95% CI 0.702-0.731). Reading error labelling for regular schools showed significantly better agreement between annotations than for special schools with respect to % agreement: 95.39% (94.87-95.86) versus 91.71% (90.76-92.57). However, kappa values did not differ significantly: $\kappa = 0.722$ (0.701-0.742) versus $\kappa = 0.706$ (0.686-0.726). The % agreement score for the annotation of SRTs (98.24%; 97.89-98.53) was significantly higher than for RWRTs (94.57%; 93.51-95.46), which was again significantly higher than for PWRTs (80.88%; 79.12-82.53). The kappa value for SRTs ($\kappa = 0.848$; 0.791-0.906) seemed to be significantly higher than the value for RWRTs ($\kappa = 0.709$; 0.673-0.744) and PWRTs ($\kappa = 0.660$; 0.639-0.681). However, there were no significant differences between the kappa values for RWRTs and PWRTs.

2.3.2 Interpretation of Differences in Agreement

All Data

Taken all double and triple annotations together, kappa values ranged from 0.717 to 0.966 and % agreement scores ranged from 86.37 to 98.64%. This indicated that agreement between annotators, and therefore the quality of annotations, was high.

When comparing inter-annotator agreement across the different annotation tiers, it could be seen that percentage agreement diminished when the annotation became more detailed: scores were always higher for the detection of reading errors than for orthographic transcriptions, and % agreement scores for the latter were always higher than for phonetic transcriptions. However, kappa values were always smaller for the detection of reading errors than for phonetic transcriptions. This is exactly the opposite as seen for the % agreement scores, and can be explained by the fact that, when calculating kappa values, there is a correction for agreement by chance. The smaller the number of categories to choose from, the easier the classification problem, and the larger the correction.

When % agreement scores and kappa values for the f01-f01b and f01-f02 comparison for both reading strategy and reading error labelling were compared, agreement was almost always higher for the f01-f01b comparison. This is easily explained by the influence of differences in annotations at the phonetic transcription level for the f01-f02 comparison. However, no explanation was found yet for the inverse relationship in kappa values with respect to reading error labelling in PWRTs.

Influence of School Type

When comparing kappa's for regular schools versus special schools, values ranged from 0.706 to 0.971; % agreement scores ranged from 82.18 to 98.72%. This indicated that, for both school types separately, agreement was high.

With respect to the % agreement scores, inter-annotator agreement depended on school type: annotations for regular schools were always better than those for special schools with respect to reading error detection, orthographic transcriptions, and phonetic transcriptions. These differences went hand in hand with differences in the amount of errors made by these two groups. Children enrolled in a special school (19.3% errors) made more reading errors than those enrolled in a regular school (12.2% errors) (which is not counterintuitive). The more errors were made, the more opportunities there were for annotators to possibly disagree. However, intra-annotator % agreement did not depend on school type unless with respect to f01-f02 reading strategy and error labelling (but this was probably just a side-effect of the above mentioned inter-annotator disagreement at the phonetic transcription level).

In contrast, no systematic significant differences in inter-annotator kappa values were detected: the picture was different for the two annotation tiers. More precisely, annotations for regular schools agreed worse than for special schools with respect to the detection of reading errors, but better than for special schools with respect to the phonetic transcriptions. With respect to reading strategy and reading error labelling, there were again no significant differences between the two school types with respect to intra-annotator agreement unless with respect to f01-f02 reading strategy labelling (which was again probably just a side-effect of inter-annotator disagreement with respect to phonetic transcriptions).

Influence of Reading Task Type

Kappa values for the different reading tasks ranged from 0.575 to 0.966, whereas % agreement scores ranged from 68.45 to 99.32%. This indicated that, for the different reading tasks separately, agreement between annotators was substantial to almost perfect.

With respect to both inter- and intra-annotator agreement, % agreement scores for SRTs were always better than those for RWRTs which were again better than those for PWRTs. Again, these differences can be explained by the differences in the amount of errors made by these three groups. For all children it seemed that most errors were made in the pseudoword reading test (37.4% errors), and the least in the story reading test (5.5% errors); in the real word reading test children made 14.3% errors. This, again, comes up to our expectations. Pseudoword reading causes more reading errors because for those words, children cannot use a direct or lexical route and need to rely more on phonological processes. Story reading causes less reading errors because in that case, extra contextual information defines syntactical and semantic restrictions for a particular word. The more errors were made, the more possible disagreements.

Again, with respect to the kappa values, the picture was not the same for all annotation tiers. Regarding the detection of reading errors, agreement for SRTs was higher than for PWRTs, and agreement for PWRTs was higher than for RWRTs. With respect to the phonetic transcriptions, the agreement for SRTs was again the highest, but now agreement for PWRTs was higher than for RWRTs. Reading strategy labelling agreement was the same for all three reading tasks; there only seemed to be a significant difference in the f01-f02 comparison that showed better agreement for RWRTs than for SRTs. However, reading error labelling agreement seemed to be different for the different reading tasks: SRTs were annotated with higher agreement than RWRTs, and annotations for RWRTs agreed better than for PWRTs (except for the f01-f02 comparison, in which agreement was the same for RWRTs and PWRTs).

2.3.3 Confusions

Phoneme Labelling

To know which phonemes were more likely to get confused by the different annotators, a confusion matrix was set up and, taken all data together, agreement scores were calculated for the different phonemes used for the making of the phonetic transcriptions. Figure 2.2 indicates for which phonemes the percentage agreement score was less than 90% (agreement scores for all other phonemes ranged from 90.3% to 96.2%). For these phonemes, the matrix reveals that the most common confusions happened between /i/ and /I/, /o/ and /O/, /@/ and /E/, /y/ and /Y/, /2/ and /y/, /x/ and /G/, /n/ and /N/, /S/ and /s/, /g/ and /G/, and /z/ and /Z/ (see Appendix D for more details on the phonetic symbols). These confusions were largely attributable to the fact that the different annotators came from different Flemish regions. Because of this, their dialects differed and so did their judgments of vowel length and type (explaining confusions between the long and short 'i', 'o', 'u', and between /2/ and /y/), and voicing (explaining confusions between /x/ and /G/). Confusions between /@/ and /E/, and /n/ and /N/ were not unexpected because of the fact that those phonemes show large acoustical resemblance. The /g/-/G/ and /z/-/Z/ confusions were probably just the result of erroneous typing (and did not occur very frequently); whereas the /S/-/s/ confusion was caused by the alignment program that aligns these two phonemes in words like for example /mE+[sj/S]@/ (where there is the need to choose between /sj/ and /S/). Additionally, vowels seemed to be deleted/inserted often; this was especially seen for the /@/, /I/, /O/, /E/, /i/, /O/, and the /Y/.

Reading Strategy Labelling

With respect to reading strategy labelling, to know which reading strategy labels were more likely to get confused, a confusion matrix was set up for the f01-f01b comparison. Figure 2.3 pictures the % agreement scores for the different strategy labels (see Appendix B for more details on reading strategy labels). Agreement ranged from 32.4% up to 99.8%. Most confusions happened between 'lsf' (i.e., wrong letter spelling) and 'af' (i.e., wrong start of a word), 'ssf' (i.e., wrong syllable spelling) and 'alf' (i.e., wrong end of a word), 'af' and 'f' (i.e., wrong decoding within one trial), 'ssg' (i.e., correct

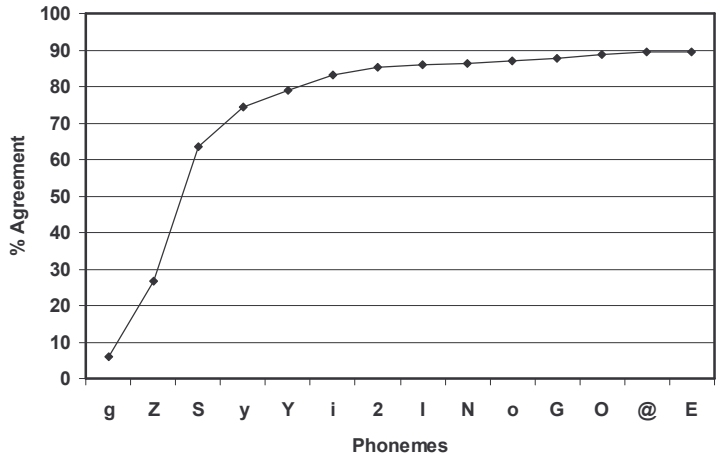

Figure 2.2: *Percentage agreement per phoneme*

syllable spelling) and 'ssf', and 'lsg' (i.e., correct letter spelling) and 'lsf'. The low score for '(W)' (i.e., change of word order) is due to its low occurrence; the % agreement score for 'O' (i.e., word omission) did not reach the 100% because of deletions/insertions of that label. Other labels had % agreement scores ranging from 92.9 to 99.8%.

Did You Hear That? The Case of '*s'

As described before, a '*s' in the orthographic transcription tier visualized the fact that an annotator heard 'something' but could not write it out because she was not able to distinguish the different reading attempts (which would otherwise be annotated as a '*' for each attempt), and to register how these attempts were exactly pronounced. Of course, the more effort an annotator put into listening to a child's reading, the less '*s' was used and the more difficult the annotation task became. As such it is clear that annotator characteristics and external influences came into play here.

A '*s' in the orthography tier was simply copied to the reading strategy tier; whereas every '*' received a separate reading strategy label in that tier. To have a better picture of the importance of these '*s' disagreements, a confusion matrix for the reading strategy f01-f02 comparison was set up. This revealed that 15.3% of all disagreements were caused by a '*s' disagreement, namely disagreements where annotator 1 was able to annotate what was said

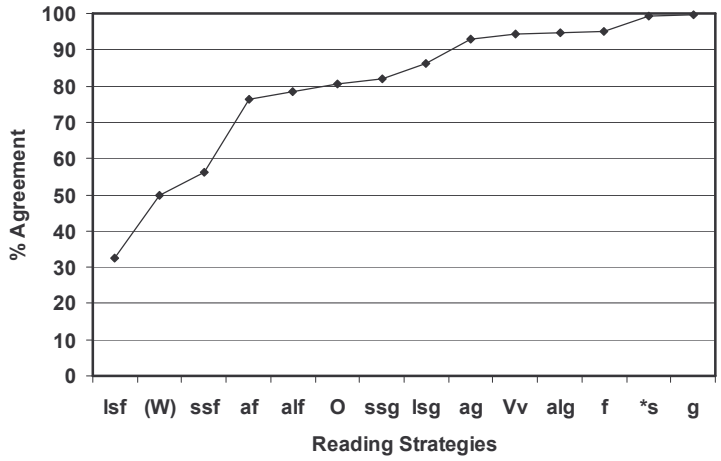

Figure 2.3: *Percentage agreement per strategy label*[2]

by the child, whereas the other annotator decided she was not able to do that and put a '*s' instead.

Reading Error Labelling

Again, a confusion matrix was set up to investigate % agreement scores for each reading error label (see Appendix C for more details on reading error labels). For the 10 most frequently annotated reading error labels (which accounted for 92.3% of all reading error label annotations), agreement ranged from 44.1 up to 99.9%. Figure 2.4 pictures the % agreement scores for these 10 reading error labels. No systematic confusions were detected, except for consistent disagreement with respect to the annotation of short-long vowel substitutions.

2.4 Conclusion

A speech recognizer is only able to accurately track a child's reading and accurately recognize oral reading errors if it is provided with sufficient data to statistically characterize children's reading behaviour and with a model that contains information on the nature and prevalence of common oral reading errors. In order to obtain this information, CHOREC, a Dutch database of recorded and carefully annotated children's reading and reading errors, was

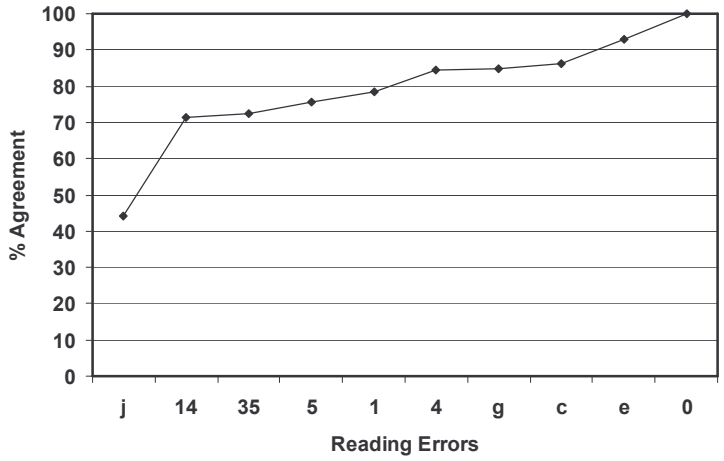

Figure 2.4: *Percentage agreement per error label*

developed.

To assess the quality of the annotations, inter- and intra-annotator agreement was analyzed with respect to the detection of reading errors, orthographic and phonetic transcriptions, and reading strategy and reading error labelling. Percentage agreement scores and kappa values both showed that agreement between annotations, and therefore the quality of the annotations, was high. Taken all double or triple annotations together, % agreement varied between 86.4% and 98.6%, whereas kappa varied between 0.72 and 0.97 depending on the annotation tier that was assessed. Degree of annotation detail, school type and reading task type seemed to account for systematic differences in % agreement, but these differences disappeared when kappa values were calculated that corrected for chance agreement.

When having a closer look at which phonemes were more likely to get confused during phonetic transcriptions, it seemed that, for the phonemes with the lowest inter-annotator agreement scores, confusions could be explained by acoustical resemblance between phonemes and dialect differences between annotators. Additionally, vowels seemed to be easily inserted or deleted. With respect to reading strategy and reading error label disagreement, systematic confusions were detected for the former but not for the latter. Furthermore, '*s' disagreements seemed to account for 15.3% of all disagreements in the f01-f02 strategy labelling comparison.

Notes

[1]Reformatted from: (1) Cleuren, L., Duchateau, J., Sips, A., Ghesquière, P., & Van hamme H. (2006). Developing an automated assessment tool for children's oral reading. In *Proceedings of the 9th International Conference on Spoken Language Processing* (pp. 817-820). Pittsburgh, USA; and (2) Cleuren L., Duchateau J., Ghesquière P., & Van hamme H. (2008). Children's oral reading corpus (CHOREC): Description and assessment of annotator agreement. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*. Marrakech, Morocco. 8 pages, ISBN 2-9517408-4-0, EAN 9782951740846.

[2]The strategy labels 'vkg', 'vkf' and 'Vg' were almost never used and were therefore left out of the figure.

Chapter 3

A Quantitative Oral Reading Error Analysis in Elementary School Children: Evidence from Dutch

Abstract[1]

In order to contribute to the literature on reading error analyses in orthographically transparent languages, the present study examined whether and how various child (i.e., gender, age, word decoding level, text reading level) and reading task (i.e., relative text difficulty, lexical status of the reading material, presence of context, word length) variables were associated with the occurrence of oral reading errors. The subjects in this study were 248 regular elementary school children who read lists of real words and pseudowords of different length, and stories of different (relative) difficulty level. Results indicated that the nature of the associations between the various child and reading task variables and general reading accuracy was in line with previous research findings regarding an orthographically less transparent language such as English. However, the use of an accuracy measure that is more accurate than a simple count of the number of correctly read words was promoted. A measure that would take into account the total number of different errors made within a particular word would probably result in more significant differences between subgroups of children. Also, in order to obtain a full understanding of oral reading error behaviour in an orthographically transparent language such as Dutch, future oral reading error research should include an investigation of the accompanying reading speed data.

3.1 Introduction

In Flanders (the Dutch speaking part of Belgium), elementary school children's reading development is regularly assessed, which provides the opportunity to detect reading difficulties early on. Unfortunately, important drawbacks of current reading assessment pertains to the time-consuming process and often serious examiner bias related to traditional paper-and-pencil reading assessment. As a result, an automated, more reliable and objective reading assessment tool is in great demand.

The SPACE project (SPeech Algorithms for Clinical and Educational applications) aimed to address these drawbacks by implementing automatic speech recognition in the reading assessment process. However, the use of automatic speech recognition within the context of reading assessment is a very challenging task because oral reading of novice readers or readers with reading difficulties can be fraught with oral reading errors. To improve the speech recognizer's ability to accurately detect oral reading errors, an error analysis is needed in order to identify a set of common reading errors that eventually can be modelled in the speech recognizer (cf., Mostow et al., 2002). Therefore, a Dutch CHildren's Oral REading Corpus (CHOREC) was developed in order to statistically characterize children's reading behaviour, and to supply the speech recognizer with a model that contains information on the nature and prevalence of common oral reading errors (Cleuren, Duchateau, Sips, Ghesquière, & Van hamme, 2006; Cleuren, Duchateau, Ghesquière, & Van hamme, 2008).

In addition to being useful for the development and testing of speech technology applications, this database also offers the unique possibility to accurately characterize the phenotypic reading performance pattern of Dutch elementary school children of various reading ability with respect to oral reading errors. Moreover, oral reading error analysis increases our understanding of the reading process because errors reveal what the reader chose as the basis for his response in a particular instance and thereby reveal the strategies readers use to identify words (Danielsson, 2002; Goikoetxea, 2006; Goodman, 1969; Leu, 1982; Weber, 1968).

Importantly, the impact of a specific language's orthographic transparency has increasingly been taken into account in research on both normal and problematic reading development. Studies have shown that the difficulties experienced by beginning or struggling readers can differ between languages

depending on the consistency of that language's grapheme-to-phoneme correspondence rules. Young (normally developing) readers in (fairly) transparent languages such as for example Greek, Finnish, German, Italian, Spanish, Welsh, and Dutch, experience less decoding problems than their English speaking peers. Also, children with reading difficulties in these languages will make up arrears regarding reading accuracy when adequate instruction is given; their reading speed, however, will remain a stumbling-block (e.g., de Jong & van der Leij, 2003; Ellis et al., 2004; Landerl & Wimmer, 2008; Landerl, Wimmer, & Frith, 1997; Patel, Snowling, & de Jong, 2004; Protopapas & Skaloumbakas, 2007; Seymour, Aro, & Erskine, 2003; Spencer & Hanley, 2003; Wimmer, 1993, 1996; Ziegler & Goswami, 2006). As such, the Dutch language is regarded as being an orthographically fairly transparent language (cf., Seymour et al., 2003), in which reading can almost entirely be mastered on the basis of consistent grapheme-to-phoneme correspondence rules. This consistency stands in harsh contrast to the extreme inconsistencies found in the English orthography. As most reading research (inclusive of research on oral reading errors) has been conducted in English, a concern has emerged among reading researchers that findings may not be applicable to other languages with more consistent mappings between orthography and phonology (Protopapas & Skaloumbakas, 2007). As a result from these insights, although still scarce, a more recent and renewed focus on oral reading error analysis in orthographically more transparent languages has emerged (e.g., Danielsson, 2001, 2002, 2003; Ellis et al., 2004; Goikoetxea, 2006; Gupta, 2004; Gupta & Jamal, 2006, 2007; Nergård-Nilssen, 2006).

In the specific English literature on oral reading error analyses, several child and reading task variables have been associated with oral reading error occurrence. Age is a first important variable in this respect because older and therefore more fully developed and more experienced readers read more accurately; the correspondence between the target word and what is actually read, becomes larger with age (Francis, 1984; Harding, Beech, & Sneddon, 1985; Moscicki & Tallal, 1981; Weber, 1970). For example, research has shown that during the first three years of school, the similarity between the target story and what was actually read increased for English-speaking children (Francis, 1984; Harding et al., 1985). Others (Moscicki & Tallal, 1981) found that older English readers (8;1 to 9;3 years) were significantly more accurate than younger readers (6;3 to 7;4 years) when reading nonsense syllables. Weber (1970) also reported that both good and weak first grade English readers

improved their story reading with respect to graphic similarity throughout the school year.

Another child variable that has been associated with oral reading error occurrence, is whether a child is developing normally or is confronted with reading difficulties inappropriate for his/her reading age. Young children with reading difficulties make more errors than a chronological age matched normal reading control group. Allington & McGill-Franzen (1980), for example, found that good fourth grade English readers made significantly fewer errors than poor fourth grade English readers when reading connected text or words without context.

Various reading task variables may also associate with oral reading error occurrence, such as the effect of relative text difficulty on the number of oral reading errors. The importance of this variable lies in the fact that some apparent (quantitative or qualitative) differences between different ability groups in many studies may have been generated by their different responses to the same text (e.g., Allington & McGill-Franzen, 1980; Biemiller, 1979; Blaxall & Willows, 1984; Leu, 1982). With respect to quantitative differences in oral reading error occurrence, the research of Allington & McGill-Franzen (1980) showed that when fourth grade good English readers were presented with material of different relative difficulty level, more errors were made when the text was more difficult, which confirmed the importance of the fact that oral reading errors of children of different ability made during text reading should only be compared on reading material of the same relative difficulty level.

The lexical status of the reading material (real words versus pseudowords) has also been associated with oral reading error occurrence. To read pseudowords, one needs to rely more on phonological processes (use of grapheme-to-phoneme correspondence rules) than to read real words; an indirect, phonological route needs to be followed to a greater extent than a more direct, lexical route (Siegel, 2003). As a result, because English grapheme-to-phoneme correspondences are highly inconsistent and because English speaking children are generally not taught by means of a systematic phonics approach (cf., Landerl, 2000), more errors occur when pseudowords are read instead of real words (e.g., Baddeley, Logie, & Ellis, 1988; Reynolds & Besner, 2008; Szezsulski & Manis, 1987). For instance, the research of Baddeley et al. (1988) showed that 12-year-old dyslexic readers of English (and their reading age controls) made significantly more errors when reading orthographically

regular pseudowords compared to equivalent real words. Additionally, others (Szezsulski & Manis, 1987) demonstrated that both normal and dyslexic English readers made more errors when reading pseudowords compared to regular or irregular real words.

It has also been demonstrated by various researchers that the presence of context (defining syntactical and semantic restrictions) is associated with the occurrence of oral reading errors. This influence appears from the fact that considerably fewer errors are made when reading connected text compared to reading series of unrelated (pseudo)words (e.g., Allington & McGill-Franzen, 1980; Archer & Bryant, 2001; Goodman, 1965; Jenkins et al., 2003). Although criticised, Goodman was the first to demonstrate that children are able to read words in context that they are unable to read in isolation (Goodman, 1965). This contextual facilitation effect was still supported when the experimental design adequately controlled for word exposure (Archer & Bryant, 2001). Allington & McGill-Franzen (1980) also showed that both good and poor fourth grade English readers more accurately read connected text compared to the same words in a random order.

Finally, (pseudo)word features such as (pseudo)word length (in terms of number of letters or syllables), and complexity (e.g., with respect to consonant combinations) have been associated with the occurrence of oral reading errors. Children make more reading errors when reading longer or more complex (pseudo)words (e.g., Baddeley et al., 1988; Moscicki & Tallal, 1981). Baddeley et al. (1988) concluded that a length effect in oral real word reading was shown by 12-year-old dyslexic readers of English (and their reading age controls). Moscicki & Tallal (1981) found that young (6;3 to 9;3 years) average to good English readers made proportionally more errors on digraphs than on graphs when reading nonsense syllables.

The objective of the present study was to contribute to the literature on reading error analyses in orthographically transparent languages by examining associations between important child and reading task variables and the occurrence of oral reading errors. Such information would reveal a more accurate characterization of the phenotypic performance pattern of children of various reading ability in Dutch (i.e., an orthographically transparent language, cf., Seymour et al., 2003), and would provide the speech recognizer with a better model of oral reading errors. Two research questions were addressed.

Research on children learning to read in English (i.e., an orthographically opaque language) already demonstrated that the child's reading age and reading level, the text's relative difficulty, the lexical status of the reading material, the presence of context, and the length and complexity of (pseudo)words to be read are associated with the occurrence of oral reading errors. Therefore, the main research question in this study was: "How can these associations be characterized in an orthographically transparent language such as Dutch, in which speed instead of accuracy often remains the largest stumbling-block?". Additionally, although no evidence for differences between boys and girls was found in the specific reading error literature, it has been shown by various researchers that girls outperform boys with respect to literacy throughout the school years (Horne, 2007). As a result, an additional research question in this study was: "Is there a gender difference as regards oral reading error occurrence in an orthographically fairly transparent language such as Dutch?".

To answer the above formulated research questions, 248 regular elementary school children read lists of 1-syllable, 2-syllable, and 3- or 4-syllable real words and pseudowords, and stories of different (relative) difficulty level. Children's reading was digitally recorded so that reading errors could be manually annotated and quantitatively analyzed afterwards.

3.2 Method

3.2.1 Participants

The subset of native Dutch speaking readers in the CHOREC (Children's Oral REading) corpus that was investigated in the present study comprised 248 children (121 girls and 127 boys) enrolled in Flemish regular elementary schools ($N = 5$). Because of pragmatic considerations, the participating schools were selected to be within easy reach of the researchers involved in the making of the database. The participating classes were selected by the school. Children attended grade 1 to 4 and met the additional requirements of not having skipped or doubled a grade (age range 6.07 to 10.06 years) so that all children of the same grade had the same birth year, and of not having any articulation problems[2]. For all children, parental consent was obtained. Table 3.1 provides an overview of several sample characteristics.

	Grade 1	Grade 2	Grade 3	Grade 4
CA (years)	6.07-7.09	7.09-8.05	8.09-9.09	9.09-10.06
N boys	24	38	31	34
N girls	27	35	31	28

Table 3.1: *Chronological age range (CA) and number of boys and girls within each grade*

3.2.2 Materials

Reading Level Assessment

To gather information on the children's reading abilities, two aspects of reading were assessed by means of two different Dutch standardized paper-and-pencil reading tests, namely pseudoword decoding and connected text reading.

Klepel (van den Bos et al., 1994) The child was asked to accurately and fluently read a list that contains 1-, 2-, 3-, and 4-syllable pseudowords (gradually rising in difficulty) for 2 minutes. This test resulted in a raw score that expressed the number of correctly read pseudowords (i.e., total number of pseudowords read minus the number of erroneously read pseudowords), which was then converted into a standard score ($M = 10$, $SD = 3$). Children with a standard score of less than 7 (i.e., 1 *SD* below the mean) were labelled as poor decoders, whereas children with a standard score of more than 13 (i.e., 1 *SD* above the mean), were labelled as good decoders. The children falling between these two categories (standard score between 7 and 13) were labelled as average decoders.

AVI test (Visser et al., 1998) The child was asked to accurately and fluently read 1 or more out of 9 stories of graded difficulty level (AVI 1 - 9) in order to detect the child's mastery level of text reading. For a child to obtain a particular AVI level, the corresponding AVI story had to be read with a maximum number of errors and within a maximum amount of time. A child

that did not fulfil the requirements to obtain a text reading level of AVI 1, obtained an AVI 0 score.

Quantitative Analysis of Oral Reading Errors

For the quantitative analysis of oral reading errors, children were presented with three different computerized reading tasks that were developed within the SPACE project: a real word reading test, a pseudoword reading test and a story reading test. For all three tests, high-quality digital audio recordings were made to ensure an accurate count of wrongly read (pseudo)words afterwards.

Real Word Reading Test (RWRT) and PseudoWord Reading Test (PWRT) The child was asked to read three different lists of 40 1-syllable, 40 2-syllable, and 40 3- and 4-syllable words (RWRT) or pseudowords (PWRT). For each (pseudo)word list (six in total), (pseudo)words were presented one-by-one on a computer screen.

Story Reading Test (SRT) The child was asked to read 1, 2 or 3 stories out of 9 SRT stories of graded difficulty level (AVI 1 - 9, 103 to 223 words in length). The choice of SRT stories was determined by the child's mastery AVI level as assessed by the AVI test (Visser et al., 1998), so that three different stories corresponding to the child's personal mastery, instructional and frustrational level were read. For example, a child with an AVI 3 mastery level had to read the AVI 3 (mastery), AVI 4 (instructional) and AVI 5 (frustrational) SRT stories. If children's AVI level was below 1 or above 7, only 1 or 2 stories were read. All stories were presented paragraph-by-paragraph on a computer screen.

3.2.3 Assessment Procedure

All assessments took place in a quiet room at the child's own school in three individual 30-minute sessions on three separate days (with 1-week intervals). The order of tests and reading tasks was counterbalanced across these three days. Most children were very motivated and eager to participate in the tests and reading tasks. Some children however, often those with reading difficulties (which made up 7% of all participating regular elementary school

children), needed to be motivated by the promise that they could listen to their own recordings afterwards (cf., Kazemzadeh et al., 2005).

3.2.4 Analysis Strategy

A global accuracy score for each subtask (list or story) within a reading task (RWRT, PWRT, SRT) was calculated dividing the number of (pseudo)words that were erroneously read in a child's final trial to read that (pseudo)word by the total number of (pseudo)words read within that list or story. The resulting proportions reflected a multilevel structure because subtask scores were nested in children, and children were nested in schools. A three-level logistic multilevel model was fitted via the proc glimmix of the SAS computer program to appropriately account for the nature of the data (i.e., proportions) and the nested data structure.

To answer the research questions, six independent variables were included in the model, namely: gender (boys, girls), grade (1st, 2nd, 3rd, 4th), word decoding level (poor, average, good), text reading level (AVI 1-9), and subtask (individual (pseudo)word list or story within the RWRT, PWRT and SRT). The independent variable subtask (i.e., effect of each list or story within a particular reading task) was entered as a within-subjects factor into the model. As such, each subject was tested at each level (i.e., nine different subtests) of the independent variable subtask. The remaining independent variables, namely gender, grade, word decoding level, and text reading level were additionally entered as interaction effects (between each of these four variables and the variable subtask) into the fixed part of the model. An alpha-level of .01 was used for the evaluation of each test.

3.3 Results

3.3.1 Task Predictors of Oral Reading Errors

The results from the logistic multilevel analysis showed that the effect of 'subtask' was statistically significant ($F(8,1458) = 1707.99$, $p < .0001$). Inspection of the least squares means estimated values at the subtask level (i.e., 1-syllable, 2-syllable, 3+4-syllable (pseudo)words, stories at mastery, instructional, frustrational level) clearly showed that most reading errors were made in the pseudoword reading subtasks, and the least in the story reading sub-

tasks (see Table 3.2). The results on the real word reading subtasks lied between the two other subtask types; all differences between two subtasks of a different task were statistically significant (ps $< .0001$). Inspection of the least squares means estimated values also revealed that word length had a significant effect on both real word and pseudoword reading: 1-syllable (pseudo)words were read more accurately than 2-syllable (pseudo)words, and 2-syllable (pseudo)words were read more accurately than 3+4-syllable pseudo(words) (ps $< .001$). Relative text difficulty also had a significant effect on reading accuracy: stories at both the mastery and instructional level were read with higher accuracy than stories at the frustrational level (ps $< .001$). However, between stories at the mastery and instructional level, there was no significant difference.

		Adjusted mean	Standard error	Antilogit [a]
RWRT	1-syllable	2.500	0.080	0.924
	2-syllable	2.273	0.080	0.907
	3+4-syllable	2.080	0.081	0.889
PWRT	1-syllable	1.810	0.076	0.859
	2-syllable	0.852	0.075	0.701
	3+4-syllable	-0.127	0.075	0.468
SRT	Mastery	3.513	0.078	0.971
	Instructional	3.560	0.079	0.972
	Frustrational	3.334	0.078	0.966

Note. [a] Because the adjusted means were predicted by a logistic multilevel model, an 'antilogit' transformation was needed in order to obtain accuracy scores that could again be interpreted as the number of correct words divided by the total number of read words.

Table 3.2: *Accuracy scores for all data: adjusted means, standard errors and antilogit of adjusted means for each subtask of the Real Word Reading Test (RWRT), PseudoWord Reading Test (PWRT), and Story Reading Test (SRT)*

3.3.2 Child Predictors of Oral Reading Errors

The interaction effects between subtask and child independent variables were entered into the fixed part of the model and the following statistically significant effects emerged (see Tables 3.3 and 3.4): gender ($F(9,1343) = 3.51$, $p < .001$), grade ($F(27,1343) = 3.03$, $p < .0001$), word decoding level ($F(18,1343) = 4.31$, $p < .0001$), and text reading level ($F(75,1343) = 3.68$, $p < .0001$).

	RWRT			PWRT			SRT		
	1syl	2syl	3+4syl	1syl	2syl	3+4syl	Mast.	Instr.	Frustr.
					Gender				
Boys	0.907	0.884	0.841	0.823	0.578	0.388	0.970	0.967	0.946
Girls	0.918	0.904	0.878	0.834	0.590	0.405	0.980	0.970	0.946
					Grade				
1st	0.846	0.815	0.791	0.823	0.540	0.357	0.982	0.965	0.946
2nd	0.897	0.859	0.819	0.794	0.560	0.370	0.969	0.954	0.926
3rd	0.934	0.914	0.895	0.849	0.611	0.429	0.974	0.974	0.960
4th	0.946	0.947	0.909	0.846	0.623	0.433	0.973	0.977	0.947

Note. [a] Because the adjusted means were predicted by a logistic multilevel model, an 'antilogit' transformation was needed in order to obtain accuracy scores that could again be interpreted as the number of correct words divided by the total number of read words.

Table 3.3: *Influence of gender and grade on the accuracy scores: antilogit of adjusted means for each subtask of the Real Word Reading Test (RWRT), PseudoWord Reading Test (PWRT), and Story Reading Test (SRT)*

First, although an investigation of the least squares means estimates revealed that boys tended to perform less accurately than girls on all subtests except for stories at the frustrational level; tests of simple effects only indicated a significant association between gender and the reading of 3+4-syllable real words ($F(1,1343) = 8.75$, $p < .001$), and between gender and stories at the mastery level ($F(1,1343) = 17.25$, $p < .0001$).

Second, tests of simple effects revealed a significant association between

	RWRT			PWRT			SRT		
	1syl	2syl	3+4syl	1syl	2syl	3+4syl	Mast.	Instr.	Frustr.
				Word decoding level					
Poor	0.877	0.845	0.776	0.721	0.381	0.250	0.969	0.948	0.923
Average	0.928	0.904	0.873	0.835	0.648	0.410	0.977	0.969	0.954
Good	0.925	0.922	0.908	0.897	0.709	0.550	0.978	0.982	0.955
				Text reading level					
AVI 0	0.784	0.620	[b]	0.619	0.262	0.152	[c]	0.949	0.929
AVI 1	0.834	0.774	0.621	0.713	0.451	[b]	0.960	0.970	0.909
AVI 2	0.878	0.866	0.781	0.788	0.490	0.336	0.983	0.963	0.938
AVI 3	0.920	0.928	0.845	0.858	0.556	0.453	0.965	0.977	0.964
AVI 4	0.904	0.896	0.867	0.811	0.583	0.364	0.979	0.983	0.938
AVI 5	0.939	0.917	0.887	0.845	0.610	0.374	0.977	0.969	0.965
AVI 6	0.921	0.922	0.885	0.840	0.660	0.414	0.965	0.963	0.947
AVI 7	0.933	0.928	0.900	0.873	0.699	0.489	0.979	0.971	0.957
AVI 8	0.943	0.934	0.908	0.892	0.703	0.478	0.981	0.962	[c]
AVI 9	0.963	0.951	0.922	0.919	0.783	0.589	0.977	[c]	[c]

Note. [a] Because the adjusted means were predicted by a logistic multilevel model, an 'antilogit' transformation was needed in order to obtain accuracy scores that could again be interpreted as the number of correct words divided by the total number of read words. [b c] Data are missing because the task was too difficult ([b]) or did not exist ([c]).

Table 3.4: *Influence of word decoding and text reading level on the accuracy scores: antilogit of adjusted means for each subtask of the Real Word Reading Test (RWRT), PseudoWord Reading Test (PWRT), and Story Reading Test (SRT)*

grade and the reading of 1-syllable ($F(3,1343) = 4.52$, $p < .01$), 2-syllable ($F(3,1343) = 8.11$, $p < .0001$), and 3+4-syllable real words ($F(3,1343) = 6.59$, $p < .0001$), and between grade and the reading of stories at the instructional ($F(3,1343) = 5.76$, $p < .001$) and at the frustrational level ($F(3,1343) = 6.88$, $p < .001$). For each of the RWRT tasks, the least squares means estimates showed that children in grade 4 read more accurately than children in grade 3, that children in grade 3 read more accurately than children in grade 2, and that the last ones read more accurately than children in grade 1. However, the least squares means estimates for stories at the instructional and frustrational level did not show such a clear picture of accuracy improvement. Moreover, the differences between subsequent grades were only statistically significant with respect to the grade 2-grade 3 transition (for all 5 tests), and the grade 3-grade 4 transition (for the reading of 2-syllable real words). With respect to the reading of pseudowords and stories at the mastery level, an investigation of the least squares means clearly showed a tendency of higher accuracy with increasing grade for the 2-syllable and 3+4-syllable pseudoword reading tasks, but a less clear picture for the 1-syllable pseudoword reading task and stories at the mastery level.

Third, significant associations between word decoding level and reading subtask were detected with respect to the reading of 1-syllable real words ($F(2,1343) = 5.42$, $p < .01$), 3+4-syllable real words ($F(2,1343) = 6.39$, $p < .01$), 1-syllable pseudowords ($F(2,1343) = 12.06$, $p < .0001$), 2-syllable pseudowords ($F(2,1343) = 17.26$, $p < .0001$), 3+4-syllable pseudowords ($F(2,1343) = 13.23$, $p < .0001$), stories at the instructional level ($F(2,1343) = 7.66$, $p < .001$), and stories at the frustrational level ($F(2,1343) = 5.10$, $p < .01$). In general, good decoders obtained higher accuracy scores than average decoders, and the latter obtained higher accuracy scores than poor decoders. However, differences between good and average decoders were only significant for the reading of 1-syllable and 3+4-syllable pseudowords and the reading of stories at the instructional level; whereas differences between average and poor decoders were significant for the reading of 3+4-syllable real words, 1-syllable, 2-syllable, and 3+4-syllable pseudowords, and stories at the instructional level.

Fourth, there was a significant association between text reading level and the reading of real words (1syl: $F(9,1343) = 3.32$, $p < .001$; 2syl: $F(9,1343) = 5.52$, $p < .0001$; 3+4syl: $F(8,1343) = 4.15$, $p < .0001$) and pseudowords (1syl: $F(9,1343 = 4.69$, $p < .0001$, 2syl: $F(9,1343) = 6.21$, $p < .0001$,

3+4syl: $F(8,1343) = 5.03$, $p < .0001$), and between text reading level and the reading of stories at the mastery ($F(8,1343) = 3.02$, $p < .01$), instructional ($F(8,1343) = 2.94$, $p < .01$), and frustrational level ($F(8,1343) = 2.58$, $p = .0122$). Inspection of the unrelated real word and pseudoword least squares means for each group revealed that, in general, children with lower AVI levels made more reading errors than children with higher AVI levels. With respect to the stories at mastery, instructional and frustrational level, this effect was not expected (because the difficulty of the story material was adapted to the child's personal AVI mastery level), and also not found.

3.4 Discussion

The aim of the present study was to examine whether and how various child and reading task variables were associated with the occurrence of oral reading errors. The present results are important for two reasons: (1) a speech recognizer that is used for automatic reading assessment can be supplied with a model that contains information on the prevalence of oral reading errors so that a recognition of these errors can be done with higher accuracy, and (2) a more accurate characterization of the phenotypic performance pattern of children learning to read in an orthographically fairly transparent language such as Dutch can be obtained.

First, the results of the quantitative analyses with respect to the task predictors were clearly in line with previous research findings on oral reading errors in an orthographically much more opaque language (i.e., English). More specifically, results indicated that, independently of word length, unrelated pseudowords were read with lower accuracy than unrelated real words (cf., Baddeley et al., 1988; Reynolds & Besner, 2008; Szezsulski & Manis, 1987). Interesting though, the difference in reading accuracy became more substantial as real words and pseudowords became longer. So, although research has shown that children who are reading in an orthographically transparent language experience substantially less difficulties in a pseudoword reading task than their English speaking peers due to the higher grapheme-phoneme regularity in their language (cf., Aro & Wimmer, 2003), these results indicate that, even for readers in the orthographically fairly transparent language of Dutch, pseudoword reading is clearly more difficult than real word reading, especially when (pseudo)words consist of two or more syllables. Additionally, results indicated that, independently of relative text difficulty, unrelated real

words were read with lower accuracy than real words in a text (cf., Allington & McGill-Franzen, 1980; Archer & Bryant, 2001; Goodman, 1965). Of course, in the present study, no attempt was made to compare exactly the same words in versus out of an embedded context. Still, the extremely high average accuracy scores of 97% made clear that, even though reading accuracy for unrelated real words was already very high for children learning to read in the orthographically transparent language of Dutch, the presence of context resulted in even fewer reading errors, even for stories at children's individual frustrational level.

The data also showed that, also for Dutch speaking readers, longer real words or pseudowords were less accurately read than shorter (pseudo)words (cf., Baddeley et al., 1988; Siegel, 2003). Because of the fact that Dutch has fairly transparent grapheme-to-phoneme correspondence rules, this result was not at all unexpected. Because of this, children can strongly rely on sublexical processes to read unrelated real words and pseudowords. Knowledge of the regularities in the orthography-phonology relationship can be successfully used to convert a string of graphemes into a synthesized sequence of phonemes by means of a left-to-right parse. As a result, the longer a (pseudo)word, the more possibilities to make an error somewhere during this conversion process (cf., Spencer & Hanley, 2003). Moreover, because in case of pseudowords the reliance on sublexical strategies to read the word is even larger, the effect of word length on reading accuracy was expected and found to be larger too. Additionally, results indicated that more errors occurred when a child read texts at his/her frustrational level compared to his/her mastery or instructional level (cf., Allington & McGill-Franzen, 1980). Since the nine texts used in this study were constructed in such a way that they represented a particular technical difficulty level (i.e., AVI level) and since the technical difficulty level of a text was based on - among other features - the average number of syllables, the effect of relative text difficulty level can to a large extent be put on a par with the effect of word length. However, no significant accuracy differences emerged between stories at the mastery and instructional level. This may be explained by the way a child's mastery, instructional and frustrational level of text reading was operationalised, namely the last two levels being one respectively two AVI levels higher than the child's mastery level of text reading as assessed by the AVI test (Visser et al., 1998). However, it can be doubted whether the relation between the relative difficulty of a text for a particular child is that straightforward. For

example, a child with a mastery text reading level of AVI 3 was assumed to have AVI 4 as his/her instructional level and AVI 5 as his/her frustrational level, which was probably not always the case. Moreover, as children in the present study learned to read in the orthographically fairly transparent language of Dutch, it could have been reading speed instead of accuracy that was differentiating between story reading at the mastery and instructional level.

Second, with respect to the child predictors, the results were generally in line with previous research findings on oral reading errors in an orthographically less transparent language such as English. More specifically, results indicated that, in general, boys read less accurately than girls (cf., Horne, 2007), that reading accuracy increased with increasing grade (cf., Francis, 1984; Harding et al., 1985; Moscicki & Tallal, 1981; Weber, 1970), and that better word decoders or readers with a higher mastery text reading level obtained higher accuracy scores than poorer word decoders or readers with a lower mastery text reading level (cf., Allington & McGill-Franzen, 1980). However, although the least squares means values showed that boys performed less accurate than girls on all subtasks (except for stories at the frustrational level), the differences between boys and girls only reached statistical significance with respect to 3- and 4-syllable real words and stories at the mastery level. This first result is in line with the research of Seymour et al. (2003) who did not find any statistical differences either between boys and girls with respect to a 1-syllable real word and a 1- and 2-syllable pseudoword reading test. As such it seems that word length plays a role: girls and boys do not differ largely with respect to reading accuracy when 1-syllable or 2-syllable real words are read, but differ significantly with respect to the reading of 3- and 4-syllable real words. However, no such argument could be used for the reading of 3- and 4-syllable pseudowords as girls did not differ significantly from boys with respect to this reading task. Perhaps the absence of this effect is due to the fact that, following the literature on English oral reading errors, this study used a rather rough accuracy measure, namely a simple count of the number of (pseudo)words that were read correctly, divided by the total number of (pseudo)words read. As children mainly rely on sublexical processing strategies to read pseudowords and a left-to-right-parse is used to convert a string of graphemes into phonemes, more than one error is likely to be made within a particular pseudoword. Therefore, a more fine-grained accuracy measure would probably have been useful. A possible

strategy to obtain a more sensitive accuracy measure may be to count the total number of different errors within a wrongly read (pseudo)word (and not only count a word as correct or incorrect), and divide this total by the total number of (pseudo)words read. This more sensitive accuracy measure would possibly have resulted in a significant effect of gender with respect to 3- and 4-syllable pseudoword reading. Finally, to interpret the fact that girls were significantly more accurate than boys during the reading of a story at a child's mastery level, accompanying reading speed would have been needed so that a possible difference in speed-accuracy trade-off between boys and girls could have been investigated.

Similarly, although the least squares means values showed that reading accuracy increased with increasing grade for all but one real word reading and pseudoword reading tasks (i.e., not for 1-syllable pseudowords), the effect of grade was only statistically significant with respect to 1-syllable, 2-syllable and 3+4-syllable real words, whereas the effect with respect to pseudoword reading did not reach the level of significance. The same argument for the lack of significance with respect to pseudoword reading can be used as before: a more sensitive accuracy measure would possibly have resulted in significant differences with respect to pseudoword reading too. Results also showed that, only after four to five months of reading instruction, children learning to read in the orthographically fairly transparent language of Dutch already obtained very high reading accuracy scores with respect to the reading of short and long unrelated real words (i.e., 1-, 2-, and 3- and 4-syllables), and with respect to the reading of short (i.e., 1-syllable) pseudowords. Although accuracy improved significantly over grades with respect to unrelated real word reading, there was no dramatic decrease in reading errors in second grade as was demonstrated in other orthographically transparent languages (e.g., Goikoetxea, 2006; Seymour et al., 2003); in the present study, the largest increase in reading accuracy was seen between grades two and three for all real word reading subtasks, and between grades three and four in case of 2-syllable real words. The lack of even a trend for increasing accuracy with increasing grade with respect to 1-syllable pseudoword reading can possibly be tackled by the aforementioned 'lack of sensitivity' hypothesis: as accuracy for 1-syllable pseudowords is already high in first grade, a possible improvement over grades could lie in the number of errors made within a particular pseudoword and not in the total number of pseudowords read correctly. In contrast, it came as a surprise that a significant accuracy difference was found

between grade two and grade three children reading stories at their personal instructional or frustrational level. A story at a child's instructional or frustrational level should be of the same relative difficulty level for all children independently of their grade or of their personal reading abilities. Moreover, accuracy for these stories was high, which makes this finding even more unexpected. A possible explanation may pertain to a shift in the speed-accuracy trade-off between grade 2 and grade 3. The least squares means data showed that children in grade 2 tended to have lower accuracy scores than children in grade 1 and in grade 3. Probably, these grade 2 children made a different speed-accuracy trade-off and combined lower accuracy with faster reading (cf., Dennis & Evans, 1996). However, a concern that was mentioned before pertains to the fact that the operationalization of a child's mastery, instructional, and frustrational level of text reading can be criticized and could have created an artefact.

Finally, the least squares means results demonstrated that, in general, good decoders were more accurate readers than average decoders, who in turn were more accurate readers than poor decoders. Only with respect to 1-syllable real words, no real accuracy difference could be seen between good and average decoders, which was probably just the result of a ceiling effect with respect to this fairly easy reading task. The data further demonstrated that differences between children of different word decoding ability were the largest for 3+4-syllable real words, for pseudowords of all word length and for stories at the instructional level. The effect of word decoding ability with respect to instructional level story reading will not be interpreted because of the possible doubtful operationalization of relative text difficulty. However, the results with respect to pseudoword reading can be interpreted as a confirmation of the nonword reading deficit (cf., Herrmann, Matyas, & Pratt, 2006) as differences between good, average and poor decoders were most clearly there when reading 1-syllable, 2-syllable, or 3- and 4-syllable pseudowords. The results also demonstrated that, in general, children with a higher mastery text reading level were more accurate readers than children with a lower mastery text reading level with respect to real word and pseudoword reading. As a child's mastery text reading level increases with reading experience, this effect can largely be put on a par with the effect of grade. The lack of this effect with respect to story reading is as expected since the difficulty level of the stories children had to read was adjusted to their mastery level of text reading.

3.5 Limitation and Conclusion

A major limitation of the present study has to do with the procedure used for the sampling of the subjects that did possibly not result in a representative subgroup of the whole population. Therefore, the results of the present study should be interpreted with caution, and cannot simply be generalized to all Flemish elementary school children. As has been mentioned in the discussion section, a second limitation pertains to the way a child's mastery, instructional and frustrational level of text reading was operationalised, namely the last two levels being one respectively two AVI levels higher than the child's mastery level of text reading as assessed by the AVI test (Visser et al., 1998). It can be doubted whether the relation between the relative difficulty of a text for a particular child and the AVI level of a text (as determined by several word and text features), is that straightforward. However, despite these limitations, an important contribution to the oral reading error literature has been made.

More specifically, the present study revealed that a number of reading task and child independent variables were associated with oral reading error occurrence in Dutch elementary school children. The present results indicated that the nature of these associations with respect to general reading accuracy was in line with previous research findings on an orthographically less transparent language such as English. However, the present study also suggested that, in an orthographically transparent language such as Dutch, a more fine-grained accuracy measure is probably needed in order to obtain a significant accuracy effect of the different reader characteristics. Moreover, to have a better grasp of the speed-accuracy trade-off that was used by a particular reader, an analysis of reading speed should complement an analysis of oral reading errors. The former often remains a larger stumbling-block for Dutch speaking novice readers or readers with reading difficulties than reading accuracy. Therefore, future research on this topic needs to use a more sensitive accuracy measure as well as a speed measure in order to obtain a better understanding of oral reading error behaviour in Dutch.

Notes

[1]Cleuren, L., Kuppens, S., Duchateau, J., Van hamme, H., & Ghesquière, P. (2009). A quantitative oral reading error analysis in elementary school children: Evidence from

Dutch. Manuscript in preparation for submission.

[2]CHOREC consisted of recordings of 274 regular elementary school children. Because of the requirements of being native Dutch speaking, not having skipped or doubled a grade, and of not having any articulation problems, only a subset of 248 children was used in the present study.

Chapter 4

Oral Reading Errors of Dutch Speaking Elementary School Children: A Qualitative Analysis at the Word Level

Abstract[1]

In order to find further evidence for the orthographic depth hypothesis with respect to the occurrence of lexical and sublexical reading strategies in Dutch, and in order to further increase the understanding of the Dutch reading process in general, oral reading errors at the word level of 248 first to fourth grade elementary school children were investigated in both unrelated real word reading and story reading. Additionally, the association between errors at the word level and a child's (reading) age, gender, and word decoding level was examined within each reading task type. First, the results of the present study substantially strengthen the orthographic depth hypothesis with respect to the nature of reading in Dutch, and consolidate the effect that the presence of context has been known to have on strategies used to read real words. Moreover, the results demonstrate that the nature of this association between the different types of reading material and error distribution at the word level does not depend on the (relative) difficulty of the text material, nor on a child's word decoding level, (reading) age, or gender. Second, the present study makes clear that word length and relative text difficulty, as well as (reading) age, word decoding level, and gender have an influence too on which reading strategies (i.e., sublexical and/or lexical) are favoured during unrelated real word and/or story reading.

4.1 Introduction

Although recent research on oral reading errors is rather scarce, the attractive power of this type of research still lies in the fact that oral reading error analysis offers a unique way to characterize the phenotypic reading performance pattern of readers with and without reading difficulties. At the same time, it helps to increase the understanding of the reading process since errors may reveal what strategies readers used to identify words (cf., Goodman, 1969; Leu, 1982; Weber, 1968).

Interestingly, the research on oral reading error analysis has known a large evolution: for a long time, errors (then often called 'miscues') were seen as a result of the different linguistic sources available to the reader (e.g., Biemiller, 1970, 1979; Blaxall & Willows, 1984; A. S. Cohen, 1974; D'Angelo & Mahlios, 1983; Goodman, 1965; Harding et al., 1985; Leslie, 1980; Miller & Isakson, 1978; Thomson, 1978; Weber, 1970). Errors were evaluated on their syntactical or semantic acceptability, and on their graphical similarity with what should have been read (for an overview, see Goodman, 1969; Leu, 1982; Weber, 1968). As a result, according to this psycholinguistic approach, the main difficulties experienced by a beginning reader were situated at the sentence level (Moscicki & Tallal, 1981).

However, as Savage and colleagues rightly mentioned, the importance of word level factors (e.g., phonological awareness and word decoding skills) became increasingly clear so that researchers started to get more focused on the development of more detailed accounts of single word reading errors (Savage, Stuart, & Hill, 2001). From this perspective, oral reading errors that are phonologically similar to the target word possibly suggest the use of grapheme-to-phoneme correspondence rules, whereas phonologically dissimilar errors possibly suggest that other processes are used to read the word (e.g., Chiappe & Siegel, 1999; Francis, 1984; Laing, 2002; Moscicki & Tallal, 1981; Savage et al., 2001; Werker, Bryson, & Wassenberg, 1989).

As for most reading research areas, the research on oral reading errors had always been predominantly focused on English readers. More recently, however, the impact of a language's orthographic transparency (i.e., the consistency in representing phonemes by means of graphemes) on the difficulties experienced at the beginning of reading development, has increasingly been taken into account. As a result from these insights, a renewed focus on oral reading error analysis with respect to orthographically more transparent lan-

guages has emerged too (e.g., Danielsson, 2001, 2002, 2003; Ellis et al., 2004; Goikoetxea, 2006; Gupta, 2004; Gupta & Jamal, 2006, 2007; Hoxhallari, van Daal, & Ellis, 2004; Nergård-Nilssen, 2006; Seymour et al., 2003; Spencer & Hanley, 2003).

More specifically, the 'orthographic depth hypothesis' has linked the ease of acquiring reading in a particular language to the nature of the orthography of that language (Katz & Frost, 1992). Young (normally developing) readers of rather transparent languages such as for example Albanian, Greek, Finnish, German, Italian, Spanish, Welsh and Dutch, experience less decoding problems than their English or French speaking peers (e.g., Ellis et al., 2004; Hoxhallari et al., 2004; Landerl & Wimmer, 2008; Patel et al., 2004; Seymour et al., 2003; Spencer & Hanley, 2003; Thorstad, 1991; Ziegler & Goswami, 2006).

Additionally, the orthographic depth hypothesis has also postulated that the nature of reading is different for languages of different orthographic transparency: readers adapt their processing strategies to the demands of the orthography they are reading (Goswami, Gombert, & de Barrera, 1998). This theory, based on the dual route model of reading (Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001), postulates that transparent orthographies are easier to read using sublexical reading strategies that involve phonology, whereas readers of opaque orthographies are more dependent on the lexical processing route which operates at the whole-word level (cf., Siegel, 2003). As a result, the majority of errors made by children (both normally developing and dyslexic) that learn to read in such a language are non-word errors, whereas the majority of errors made by children learning to read in an orthographically more opaque langue are real words or non-responses (e.g., Danielsson, 2001; Ellis & Hooper, 2001; Ellis et al., 2004; Gupta & Jamal, 2006, 2007; Hoxhallari et al., 2004; Laing, 2002; Seymour et al., 2003; Spencer & Hanley, 2003; Stuart & Coltheart, 1988; Wimmer & Goswami, 1994; Wimmer & Hummer, 1990)

With respect to non-response errors, Seymour et al. (2003), for instance, found that normal first grade readers of the fairly transparent orthography of Dutch did not make any non-response errors when reading unrelated real words. Similarly, Ellis et al. (2004) who investigated unrelated real word reading in normal 6- to 15-year-old readers of several languages of different orthographic transparency, found that the number of non-response errors was very small (2%) for readers of the transparent orthography of Albanian, but

was much larger (22%) for English readers. Hoxhallari et al. (2004) made the same conclusion with respect to unrelated real word reading in Albanian: non-response errors in first grade normal readers were very rare (4.4%). Finally, Ellis & Hooper (2001) investigated the oral reading errors made by normal second grade Welsh (i.e., orthographically transparent language) and English readers when reading unrelated real words. Although they did find a substantial proportion of non-response errors in readers of Welsh (14%), this type of errors was still much more present in readers of English (31%).

With respect to real word and non-word errors, Gupta & Jamal (2007), for instance, found that, when reading unrelated real words, normal third grade bilingual (Hindi and English) readers made only non-word errors when reading in the orthographically transparent language of Hindi, whereas they made more real word than non-word errors when reading in English. Similarly, Ellis et al. (2004) who investigated unrelated real word reading in normal 6- to 15-year-old readers, found that Albanian children made relatively more non-word errors than real word errors, whereas the proportion of real word errors was much higher (but equalled the proportion of non-word errors) in English readers. Similar findings with respect to unrelated real word reading in Albanian were also obtained by Hoxhallari et al. (2004). These authors found that much more non-word errors were made than real word errors (64.9% vs. 30.7%). Similarly, the results of a study by Ellis and Hooper (2001) showed that Welsh second grade normal readers made more non-word errors than real word errors when reading unrelated real words (73% vs. 24%), whereas the opposite was true for English readers (25% vs. 45%).

In spite of a language's orthographic transparency, when words are presented in a context, additionally available syntactical and semantic clues can be used during reading so that one possibly relies more on lexical strategies to read words (Archer & Bryant, 2001; Goodman, 1965). As a result, no difference in reading strategy use is expected for readers of English when reading words in context as opposed to unrelated real words, whereas readers of orthographically more transparent languages will possibly make a shift towards a larger use of lexical word recognition strategies. This hypothesis was consolidated by the research of Laing (2002) and Danielsson (2001). With respect to normal third grade English children reading connected texts, Laing (2002) found a larger proportion of real word errors (61.0%) than non-word errors (26.0%), which is in line with what had been demonstrated for

unrelated real word reading in English. Danielsson (2001), however, found that when preschool or first grade normal readers of the fairly transparent orthography of Swedish were reading connected texts, they also made relatively more real word errors than non-word errors, which is exactly the opposite of what had been demonstrated for unrelated real word reading in orthographically transparent languages.

Finally, no large attention has been paid to the association between a child's word reading skills and oral reading error occurrence at the word level. A notable exception is the study of Gupta & Jamal (2006, 2007) who analyzed oral reading errors made by dyslexic third grade bilingual (Hindi and English) readers during unrelated real word reading. The authors found that, as was also shown for normal readers of an orthographically transparent language, far more non-word errors were made than real word errors when reading in Hindi. However, at the same time, the proportion of real word errors was larger than for normal readers. Interestingly, the authors also found that more non-word errors than real word errors were made when reading in English, which was not in line with what could be expected about reading in an orthographically opaque language, but was explained by the authors as the result of the child's first language influence (i.e., Hindi) that was stronger for dyslexic readers than normal readers. Additionally, Laing (2002) investigated the reading errors of third grade English children with below-average reading and below-average general language performance when reading connected texts. The results showed that, equal to the average performing third grade readers from the same study, the proportion of real word errors (68.3%) was larger than the proportion of non-word errors (15.0%).

The present study wanted to contribute to the literature on orthographically more transparent languages by performing a qualitative analysis of Dutch oral reading errors at the word level. As such, the study aimed at: (1) finding further evidence for the orthographic depth hypothesis with respect to the nature of reading in Dutch, and (2) further increasing the understanding of the Dutch reading process by accurately characterizing the phenotypic reading performance pattern with respect to oral reading errors of Dutch speaking elementary school children of different age, gender, and word decoding level.

In this study, five research questions were pursued. First, we aimed to investigate the proportions of non-response errors, real word errors and non-word errors in two different kinds of reading material: unrelated real words

and real words presented in a text (i.e., stories). On the basis of the literature, it was hypothesized that Dutch, a language with a fairly transparent orthography, independently of differences in word decoding level, would promote non-word errors to a larger extent than real word errors or non-response errors when reading unrelated real words. Furthermore, it was hypothesized that, independently of differences in word decoding level, the proportion of real word errors would be larger and exceed the proportion of non-word errors when reading story material as opposed to unrelated real words, and that the proportion of non-response errors would be smaller when reading story material than when reading unrelated real words.

Second, to our knowledge, no explicit attention has been paid to the association between the (relative) difficulty of the reading material and oral reading error occurrence at the word level. Therefore, a second research question set out to investigate the proportions of non-response errors, real word errors and non-word errors made when reading unrelated real words of different word length, and stories of different relative text difficulty. It was hypothesized that longer words would more strongly promote the use of sublexical reading strategies, and therefore elicit a larger proportion of non-word errors (cf., Spencer & Hanley, 2003). With respect to story reading, it was hypothesized that relatively easier stories would lead to a larger use of lexical strategies, and therefore evoke a larger proportion of real word errors.

Third, to explore the developmental course of oral reading errors at the word level, we aimed to investigate the proportions of non-response errors, real word errors and non-word errors in children of different (reading) age when reading unrelated real words of different word length or stories of different relative difficulty. It was hypothesized that the proportion of non-response and non-word errors would decrease with increasing age, whereas the proportion of real word errors would increase with increasing age (i.e., because of a larger use of lexical strategies, cf., Wimmer & Goswami, 1994).

Fourth, we aimed to investigate the proportions of non-response errors, real word errors and non-word errors in children with good, average or poor word decoding skills when reading unrelated real words of different word length or stories of different relative difficulty. On the basis of the literature, it was hypothesized that the lower a child's word decoding skills, the larger the proportion of real word errors that would be made.

Fifth, we wondered whether boys and girls would differ in their use of reading strategies to read different types of reading material (i.e., unrelated

real words of different length versus stories of different relative difficulty) as expressed by their oral reading errors at the word level. Although no evidence for quantitative differences between boys and girls was found in the specific oral reading error literature on orthographically transparent languages (Seymour et al., 2003), it seemed worthwhile to explore the association between gender and oral reading errors at the word level.

To test these hypotheses, oral reading errors of 248 first to fourth grade elementary school children were examined with respect to the reading of unrelated real words of different word length, and the reading of connected texts of different relative difficulty (i.e., task variables). Additionally, the association between errors at the word level and grade, gender, and word decoding level (i.e., child variables) was examined within each reading task type.

4.2 Method

4.2.1 Participants

Within the scope of the SPACE project (i.e., SPeech Algorithms for Clinical and Educational applications), a Dutch CHildren's Oral REading Corpus (CHOREC) was developed so that children's oral reading errors could be statistically characterized (Cleuren et al., 2006; Cleuren, Duchateau, et al., 2008). The subset of native Dutch speaking readers in CHOREC that was investigated in the present study comprised 248 children (121 girls and 127 boys) enrolled in Flemish regular elementary schools ($N = 5$). Because of pragmatic considerations, the participating schools were selected to be within easy reach of the researchers involved in the making of the database. The participating classes were selected by the school. Children attended grade 1 to 4 and met the additional requirements of not having skipped or doubled a grade (age range 6.07 to 10.06 years) so that all children of the same grade had the same birth year, and of not having any articulation problems[2]. For all children, parental consent was obtained. Table 4.1 provides an overview of several sample characteristics.

		Grade 1	Grade 2	Grade 3	Grade 4
CA	(years)	6.07-7.09	7.09-8.05	8.09-9.09	9.09-10.06
Gender	*N* boys	24	38	31	34
	N girls	27	35	31	28
WDL	Good	1	29	8	8
	Average	42	41	49	52
	Poor	8	3	5	2

Table 4.1: *Chronological age range (CA), gender distribution, and distribution of word decoding level (WDL) within each grade*

4.2.2 Materials

Reading Level Assessment

To gather information on the children's reading abilities at the middle of each grade (i.e., February), two aspects of reading were assessed by means of two different Dutch standardized paper-and-pencil reading tests: pseudoword decoding and connected text reading.

De Klepel (van den Bos et al., 1994) The child was asked to accurately and fluently read a list that contains 1-, 2-, 3-, and 4-syllable pseudowords (gradually rising in difficulty) for 2 minutes. This test resulted in a raw score that expressed the number of correctly read pseudowords (i.e., total number of pseudowords read minus the number of erroneously read pseudowords), which was then converted into a standard score ($M = 10$, $SD = 3$). Children with a standard score of less than 7 (i.e., 1 *SD* below the mean) were labelled as poor decoders, whereas children with a standard score of more than 13 (i.e., 1 *SD* above the mean), were labelled as good decoders. The children falling between these two categories (standard score between 7 and 13) were labelled as average decoders.

AVI test (Visser et al., 1998) The child was asked to accurately and fluently read 1 or more out of 9 stories of graded difficulty level (AVI 1 - 9) in order

to detect the child's mastery level of text reading. For a child to obtain a particular AVI level, the corresponding AVI story had to be read with a maximum number of errors and within a maximum amount of time. A child that did not fulfil the requirements to obtain a text reading level of AVI 1, obtained an AVI 0 score.

Oral Reading Error Analysis

For the analysis of reading errors, children were presented with two different computerized reading tasks that were developed within the SPACE project: a real word reading test and a story reading test. For both tests, high-quality digital audio recordings were made to ensure an accurate analysis of oral reading errors afterwards.

Real Word Reading Test (RWRT) The child was asked to read three different lists of 40 1-syllable, 40 2-syllable, and 40 3- and 4-syllable words, which were selected so as to be highly familiar to young children. For each of the three word lists, words were presented one-by-one on a computer screen.

Story Reading Test (SRT) The child was asked to read 1, 2 or 3 stories out of 9 SRT stories of graded difficulty level (AVI 1 - 9, 103 to 223 words in length). The choice of SRT stories was determined by the child's mastery AVI level as assessed by the AVI test (Visser et al., 1998), so that three different stories corresponding to the child's personal mastery, instructional and frustrational level were read. For example, a child with an AVI 3 mastery level had to read the AVI 3 (mastery), AVI 4 (instructional) and AVI 5 (frustrational) SRT stories. If children's AVI level was below 1 or above 7, only 1 or 2 stories were read. All stories were presented paragraph-by-paragraph on a computer screen.

Table 4.2 provides an overview of the distribution of tasks over children. There was a substantial amount of missing data. Data could be missing because of four reasons: (1) because of a bad recording, a particular test could not be annotated afterwards, (2) because of time restrictions, a particular test was not assessed, (3) because of a child's limited reading ability, a particular test was not assessed to prevent he/she from getting frustrated, and (4) because of the fact that a child's mastery and/or instructional story reading

level was below AVI 1, or because his/her instructional and/or frustrational level was above AVI 9, a particular test could not be assessed.

4.2.3 Procedure

Assessment Procedure

All assessments took place in a quiet room at the child's own school in three individual 30-minute sessions on three separate days (with 1-week intervals). The order of tests and reading tasks was counterbalanced across these three days. Most children were very motivated and eager to participate in the tests and reading tasks. Some children however, often those with reading difficulties, needed to be motivated by the promise that they could listen to their own recordings afterwards (cf., Kazemzadeh et al., 2005).

Oral Reading Error Categorization Procedure

To categorize oral reading errors, all children's RWRT and SRT recordings were carefully analyzed by means of 'Praat', a free speech analysis and synthesis program (http://www.Praat.org/). As such, not only the audio signal but also the visual speech wave representation of the recorded speech could be used during categorization; and - only for those words for which the last try to read the word was incorrect - error labels were added to an annotation layer that was time-aligned to the speech sound. Errors were categorized at four different levels: at the level of the paragraph (only for SRT), the sentence (only for SRT), the word, and the grapheme (cf., Cleuren et al., 2006; Cleuren, Duchateau, et al., 2008). Different error categories could simultaneously be attributed to a single word.

For the study presented in this paper, only the errors at the word level (not those at the paragraph, sentence or grapheme level) were taken into account. At this level, three error categories could be distinguished: non-response errors, non-word errors, and real word errors. *Non-response errors* were words for which the child remained silent, or that were partially or completely prompted by the examiner because of a non-response by the child. *Non-word errors* consisted of letter/syllable soundings and pseudoword errors. *Letter/syllable soundings* were correct or erroneous letter-by-letter and/or syllable-by-syllable reading without a subsequent try to synthesize the word. *Pseudoword errors* were errors that were substitutions of a real

	Grade 1			Grade 2			Grade 3			Grade 4		
	G	A	P	G	A	P	G	A	P	G	A	P
N children	*N* = 1	*N* = 42	*N* = 8	*N* = 29	*N* = 41	*N* = 3	*N* = 8	*N* = 49	*N* = 5	*N* = 8	*N* = 52	*N* = 2
RWRT												
1-syllable	1	38	6	27	40	3	7	47	5	8	46	2
2-syllable	1	22	2	27	40	3	8	47	5	8	48	2
3+4-syllable	1	2	0	27	39	2	7	47	5	8	48	2
SRT												
Mastery	1	13	0	27	38	3	6	44	5	7	41	1
Instructional	1	25	2	20	37	3	0	38	5	0	16	1
Frustrational	1	32	6	12	36	3	0	25	5	0	8	1

Table 4.2: *Number of times a particular task was assessed within a particular grade and word decoding level (G = good, A = average, P = poor)*

word by a pseudoword. Finally, *real word errors* were errors that were substitutions of a real word by another (semantically or graphemically alike) real word. Table 4.3 provides an example for each of the different error types.

	Target Word		Actually read	
Error Type	Dutch	English	Dutch	English
Non-response errors	kat[a]	cat	Child remains silent	
Real word errors				
Semantically alike	kat[a]	man	poes[a]	guy
Graphemically alike	kat[a]	cat	kar[b]	cut
Non-word errors				
Letter soundings	kat[a]	cat	k…a…t	c…a…t
Syllable soundings	voetbal[c]	football	voet…bal	foot…ball
Pseudoword errors	kat[a]	cat	kal[d]	cal[d]

Note. [a] Means 'cat'. [b] Means 'jug'. [c] Means 'football'. [d] Pseudoword.

Table 4.3: *Examples for the different error categories*

Categorization Agreement

To assess oral reading error categorization agreement, 10% of the recordings were investigated twice by one rater (i.e., first author) with a period of at least one month (up to one year) between the two investigations. Then, percentage agreement scores and kappa values (Cohen, 1960) were calculated to obtain an indication of intra-rater agreement. As such, it was shown that agreement between the two error categorizations was high. Overall kappa values varied between 0.911 and 0.966, and overall percentage agreement scores varied between 97.77% and 98.64% (for more details, see Cleuren, Duchateau, et al., 2008).

4.2.4 Analysis Strategy

To obtain the data for the present study, all labelled errors were automatically extracted from the annotated recordings together with information on the child that made that particular error, and on the reading task in which the error was made. The above formulated research questions were answered by constructing contingency tables (for the whole data set ánd for each child characteristic separately) on which chi-square tests of independence or Fisher's exact tests (in the case of small expected frequencies) were performed. In addition, descriptive statistics were used to characterize the data. All analyses were carried out by use of the SAS system, version 9.1, and hypotheses were tested at an .01 alpha level to correct for multiple testing.

4.3 Results

In total, 4075 errors were annotated at the word level: 1744 errors were made in the unrelated real word reading tasks (RWRT), and 2331 errors were made in the story reading tasks (SRT). The error distributions over tasks and children are presented in Tables 4.4 and 4.5.

4.3.1 Task Variables

Main Effects

Reading task type Taking all RWRT and SRT data together, a chi-square test of independence indicated that there was a significant association between reading task type and the distribution of errors at the word level ($\chi^2(2, N = 4075) = 715.04$, $p < .0001$). When reading unrelated real words, children made more non-word errors (59.12%) than real word errors (26.20%). At the same time, there was a considerable proportion of non-response errors (14.68%). When reading stories, the proportion of real word errors substantially increased (66.92%), as such becoming larger than the decreased proportion of non-word errors (30.54%). Also, the proportion of non-response errors (2.53%) was considerably lower than when reading unrelated real words.

Word length A clear association between word length and error proportions at the word level could only be obtained when the data coming from first grade children was left out of the analyses. This can be explained by the fact

		Non-response errors		Real word errors		Non-word errors			
						Letter/syllable soundings		Pseudoword errors	
		N	%	*N*	%	*N*	%	*N*	%
RWRT		**256**	**14.68**	**457**	**26.20**	**272**	**15.60**	**759**	**43.52**
	1-syllable[a]	4	1.99	98	48.76	4	1.99	95	47.26
	2-syllable[a]	29	10.39	81	29.03	24	8.60	145	51.97
	3+4-syllable[a]	63	12.96	112	23.05	134	27.57	177	36.42
SRT		**59**	**2.53**	**1560**	**66.92**	**231**	**9.91**	**481**	**20.63**
	Mastery	5	0.72	465	67.39	59	8.55	161	23.33
	Instructional	20	2.79	482	67.32	83	11.59	131	18.30
	Frustrational	34	3.68	613	66.27	89	9.62	189	20.43

Note. [a] Only second, third, and fourth grade data.

Table 4.4: *Error frequencies and row percentages for each task and subtask*

		Non-response errors				Real word errors				Non-word errors							
										Letter/syllable soundings				Pseudoword errors			
		RWRT		SRT		RWRT		SRT		RWRT		SRT		RWRT		SRT	
		N	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%
Grade	1	160	20.57	35	8.86	166	21.34	240	60.76	110	14.14	53	13.42	342	43.96	67	16.96
	2	84	14.43	13	1.73	143	24.57	493	65.56	101	17.35	75	9.97	254	43.64	171	22.74
	3	12	4.49	4	1.04	93	34.83	271	70.57	46	17.23	39	10.16	116	43.45	70	18.23
	4	0	0.00	2	1.82	55	47.01	91	82.73	15	12.82	5	4.55	47	40.17	12	10.91
WDL	Good	8	6.67	4	0.94	41	34.17	262	61.36	18	15.00	43	10.07	53	44.17	118	27.63
	Average	206	15.74	49	2.97	336	25.67	1110	67.27	188	14.36	166	10.06	579	44.23	325	19.70
	Poor	42	13.33	6	2.36	80	25.40	188	74.02	66	20.95	22	8.66	127	40.32	38	14.96
Gender	Boys	122	12.15	21	1.63	253	25.20	863	77.11	206	20.52	126	9.80	423	42.13	276	21.46
	Girls	134	18.11	38	3.64	204	27.57	697	66.70	66	8.92	105	10.05	336	45.41	205	19.62

Table 4.5: *Error frequencies and row percentages for each grade, word decoding level (WDL), and gender*

that, in first grade, there was a lot of missing data, especially with respect to the reading of 3+4-syllable unrelated real words (because this task was too difficult for these children). Taking all the second, third, and fourth grade RWRT data together, a chi-square test of independence indicated that there was a significant association between word length and error proportions at the word level ($\chi^2(4, N = 966) = 53.85$, $p < .0001$). When reading unrelated real words, the proportion of non-word errors increased with increasing word length (49.25 vs. 60.57 vs. 63.99%), as well as the proportion of non-response errors (1.99 vs. 10.39 vs. 12.96%). In contrast, the proportion of real word errors decreased with increasing word length (48.76 vs. 29.03 vs. 23.05%). For all three error types, the largest increase/decrease occurred between 1- and 2-syllable unrelated real words.

Relative text difficulty Taking all SRT data together, a chi-square test of independence indicated that there was a significant association between relative text difficulty and error proportions at the word level ($\chi^2(4, N = 2331) = 14.56$, $p = .0057$). When reading real words in a text, the proportion of non-response errors increased with increasing relative text difficulty (0.72 vs. 2.79 vs. 3.68%). With respect to the proportions of non-word errors (31.88 vs. 29.89 vs. 30.05) and real word errors (67.39 vs. 67.32 vs. 66.27), there was no clear association with relative text difficulty.

Interaction Effects with Child Variables

Grade Additional contingency table analyses for each grade showed that the significance and nature of the association between reading task type and the distribution of errors at the word level did not depend on grade. However, the significance and nature of the association between word length and error proportions at the word level did depend on grade. More specifically, a significant association was only found in second and third graders. Also, the proportion of non-response errors made by third graders did not further increase between 2- and 3+4-syllable unrelated real words. Finally, the significance of the association between relative text difficulty and error proportions at the word level disappeared when investigated for each grade separately.

Word decoding level Additional contingency table analyses for each level of word decoding showed that the significance and nature of the association

between reading task type and the distribution of errors at the word level did not depend on word decoding level. However, the significance and nature of the association between word length and error proportions at the word level did depend on word decoding level. More specifically, a significant association (of the same nature) was only found in average word decoders. Finally, the significance of the association between relative text difficulty and error proportions at the word level disappeared when investigated for each level of word decoding separately.

Gender Additional contingency table analyses for each gender showed that the significance and nature of the association between reading task type and the distribution of errors at the word level did not depend on gender. However, the significance and nature of the association between word length and error proportions at the word level did depend on gender. More specifically, although a significant association was found in both boys and girls, there was a difference between both groups with respect to the proportions of real word errors and non-response errors: whereas boys showed the expected decrease respectively increase with increasing word length, girls' proportion of real word errors did not decrease between 2- and 3+4-syllable unrelated real words and their proportion of non-response errors slightly decreased again between 2- and 3+4-syllable unrelated real words. Finally, the significance of the association between relative text difficulty and error proportions at the word level disappeared when investigated for each gender separately.

4.3.2 Child Variables

Main Effects

The association between the different child variables (i.e., grade, word decoding level, and gender) and error proportions at the word level was investigated separately for the reading of unrelated real words and real words in a text (i.e., stories). For these analyses, the (relative) difficulty of the reading material was not taken into account.

Grade A chi-square test of independence showed that there was a significant association between grade and error proportions at the word level when reading unrelated real words ($\chi^2(6,\ N = 1744) = 90.30,\ p < .0001$). More specifically, the proportion of non-response errors decreased with increasing

grade (20.57 vs. 14.43 vs. 4.49 vs. 0.00%), whereas the proportion of real word errors increased with increasing grade (21.34 vs. 24.57 vs. 34.83 vs. 47.01%). With increasing grade, the proportion of non-word errors only decreased between third and fourth grade (58.10 vs. 60.99 vs. 60.68 vs. 52.99%). As a result, fourth grade children made only a slightly larger proportion of non-word errors (52.99%) than real word errors (47.01%), and did not make any non-response errors (0.00%) when reading unrelated real words. When reading stories, their proportion of non-response errors did not decrease (0.75%) because of a floor effect with respect to this type of error that was present during unrelated real word reading.

A chi-square test of independence also showed that there was a significant association between grade and error proportions at the word level when reading real words in a text ($\chi^2(6,\ N = 2331) = 75.34,\ p < .0001$), and that the nature of this association was the same as for unrelated real word reading. More specifically, the proportion of non-response errors decreased with increasing grade (8.05 vs. 1.62 vs. 0.86 vs. 0.75%), whereas the proportion of real word errors increased with increasing grade (62.30 vs. 66.09 vs. 67.92 vs. 75.56%). With increasing grade, the proportion of non-word errors only decreased between third and fourth grade (29.66 vs. 32.28 vs. 31.22 vs. 23.68%).

Word decoding level A chi-square test of independence showed that there was no significant association between word decoding level and error proportions at the word level when reading unrelated real words ($\chi^2(4,\ N = 1744) = 10.07,\ p = .0393$). In contrast, there was a significant association when reading real words in a text (i.e., stories, $\chi^2(4,\ N = 2331) = 20.97,\ p = .0003$). More specifically, the proportion of non-word errors decreased with decreasing word decoding level (37.70 vs. 29.76 vs. 23.62%), whereas the proportion of real word errors increased with decreasing word decoding level (61.36 vs. 67.27 vs. 74.02%). The proportion of non-response errors was the lowest for good word decoders (0.94 vs. 2.97 vs. 2.36%).

Gender A chi-square test of independence showed that there was a significant association between gender and error proportions at the word level when reading unrelated real words ($\chi^2(2,\ N = 1744) = 16.20,\ p = .0003$). More specifically, girls made a larger proportion of non-response errors (18.11 vs. 12.15%) and a smaller proportion of non-word errors (54.32 vs. 62.65%) than

boys. The proportion of real word errors was only slightly larger for girls than for boys (27.57 vs. 25.20%).

A chi-square test of independence also demonstrated a significant association between gender and error proportions at the word level when reading real words in a text (i.e., stories, $\chi^2(2, N = 2331) = 9.64$, $p = .0081$). As for unrelated real words, girls made a slightly larger proportion of non-response errors (3.64 vs. 1.63%) and a slightly smaller proportion of non-word errors (29.67 vs. 31.26%) than boys. The proportion of real word errors did almost not differ between girls and boys (66.70 vs. 67.11%).

Interaction Effects with Other Child Variables

Grade Additional contingency table analyses for each grade showed that there was still no significant association between word decoding level and error proportions at the word level when reading unrelated real words. When reading real words in a text, the association was only significant in second graders (i.e., subgroup with the smallest amount of missing data), but of the same nature as that of the association that was found for the whole data set.

Additional contingency table analyses for each grade also showed that, only for the subgroup of first graders, there was a significant association (that was of the same nature as for the whole data set) between gender and error proportions at the word level when reading unrelated real words. When reading real words in a text, this association was only marginally significant in the subgroup of first graders (and of the same nature as for the whole data set).

Word decoding level Additional contingency table analyses for each level of word decoding showed that, only in the subgroups of average and poor word decoders, there was a significant association between grade and error proportions at the word level when reading both unrelated real words and real words in a text. When reading unrelated real words, the nature of the association for the subgroup of average word decoders was the same as for the whole data set; however, it was different for the subgroup of poor word decoders with respect to the proportions of real word errors and non-word errors. More specifically, with increasing grade, the proportion of non-word errors also showed a substantial increase between first and second grade, and a substantial decrease between second and third grade. The proportion

of real word errors first showed a decrease between first and second grade before increasing from second grade on. When reading real words in a text, the nature of the association for the subgroup of average word decoders was different from that for the whole data set with respect to the proportion of real word errors: more specifically, there was no relative increase in real word errors between grade two and grade three. For the subgroup of poor readers, the nature of the association was different from that for the whole data set with respect to the proportions of both non-response errors and real word errors. More specifically, although the proportion of non-response errors gradually became smaller from first to third grade, it became larger again in fourth grade. With respect to the proportion of real word errors, there was a large increase between first and second grade, but not anymore after grade two.

Additional contingency table analyses for each level of word decoding also showed that, only for the subgroup of poor word decoders, there was a significant association (that was of the same nature as for the whole data set) between gender and error proportions at the word level when reading unrelated real words. When reading real words in a text, this association was no more significant in none of the subgroups of different word decoding level.

Gender Additional contingency table analyses for each gender showed that the significance and nature of the association between grade and error proportions at the word level did not depend on gender when reading unrelated real words. When reading real words in a text, the significance of the association did also not depend on gender but its nature did: more specifically, in the subgroup of girls, the proportion of non-response errors did not decrease anymore between third and fourth grade. Moreover, the proportion of real word errors did only slightly increase between second and third grade, whereas there was no decrease at all between third and fourth grade with respect to non-word errors.

Additional contingency table analyses for each gender also showed that there was still no significant association between word decoding level and error proportions at the word level when reading unrelated real words. When reading real words in a text, the association was only significant in boys, but of the same nature as that of the association that was found for the whole data set.

4.4 Discussion

The aim of the present study was to: (1) find further evidence for the orthographic depth hypothesis with respect to the nature of reading in Dutch, and (2) further increase the understanding of the Dutch reading process by accurately characterizing the phenotypic reading performance pattern with respect to oral reading errors of Dutch speaking elementary school children of different age, gender, and word decoding level. To do so, oral reading errors of 248 first to fourth grade elementary school children were examined with respect to the reading of unrelated real words of different word length, and the reading of connected texts of different relative difficulty (i.e., task variables). Additionally, the association between errors at the word level and grade, gender, and word decoding level (i.e., child variables) were examined within each reading task type.

First, the results of this study substantially strengthen the orthographic depth hypothesis with respect to the nature of reading in Dutch. More specifically, independently of differences in grade, word decoding level, or gender, much more non-word errors than real word or non-response errors were made when reading unrelated 1-syllable, 2-syllable, or 3+4-syllable real words. This result indicates that in Dutch, children seem to be mainly relying on sublexical processing strategies to read unrelated real words, namely on the assemblage of phonemes by means of grapheme-to-phoneme correspondence rules. Because of the fairly orthographic transparency of Dutch, knowledge of the regularities in the orthography-phonology relationship can be successfully used to convert a string of graphemes into a synthesized sequence of phonemes. As a consequence, non-word errors are the result of a left-to-right parse of a grapheme sequence which has failed because of an incomplete knowledge of grapheme-to-phoneme correspondences, or because of difficulties with pronunciation assemblage. This result is in line with the research of Gupta & Jamal (2006, 2007) who also reported a preponderance of non-word errors in both dyslexic and normal readers when reading unrelated real words in the orthographically transparent Hindi language. It is also in line with the research of Ellis et al. (2004) and Hoxhallari et al. (2004) who found similar results in normal readers when reading unrelated real words in the orthographically transparent Albanian language, and with the research of Seymour et al. (2003) and Ellis & Hooper (2001) who found similar results with respect to unrelated real word reading in Dutch (among other

languages) respectively Welsh. However, an interesting contrast is provided by Spencer & Hanley (2003), who also found a larger proportion of non-word errors than real word errors in readers of the orthographically transparent orthography of Welsh at the age of 7 (i.e., 3rd year of formal reading instruction). However, at the same time, they found the opposite relation between non-word and real word errors at the age of 6 (i.e., 2nd year of formal reading instruction), which leaves open the possibility that these children were mainly using lexical reading strategies during the early stages of reading.

In contrast, the present study showed that when reading real words in a text (i.e., stories of different relative difficulty), independently of differences in grade, word decoding level, or gender, much more real word errors than non-word errors were made, and that the proportion of non-response errors had become very small when compared to unrelated real word reading. This result indicates that to read stories in Dutch, children seem to rely more strongly on lexical strategies: because of the presence of context, words can be fairly successfully read on the basis of only a partial visual analysis of the reading stimulus. As a consequence, when this whole-word lexical access on the basis of partial visual analysis goes wrong, the likely result is another real word. This result also indicates that children have greater confidence in trying to read the words when reading real words in a story so that they make fewer non-response errors than when reading unrelated real words. This is in line with the research of Danielsson (2001) who demonstrated a preponderance of real word errors in both fluent and less-fluent young readers when reading unfamiliar stories in the orthographically fairly transparent Swedish language.

Of course, this never is an all-or-nothing phenomenon. Although the type of reading material can be associated with a particular predominant reading strategy, children combine both lexical and sublexical strategies when reading unrelated real words or real words in a text. As a result, a considerable proportion of real word errors and non-response errors was also found when children were reading unrelated real words, whereas a considerable proportion of non-word errors was also found when they were reading real words in a text. Especially because children were not questioned about the content of the stories they had read, decoding was probably favoured at the expense of understanding so that non-word errors were still very likely to occur (cf., Danielsson, 2001). These findings are in line with, for instance, the research of Ellis et al. (2004) who also found a considerable proportion

of non-response errors in readers of the orthographically fairly transparent Greek language, and a considerable proportion of real word errors in both Greek and Albanian readers when reading unrelated real words. Likewise, Spencer & Hanley (2003) and Ellis & Hooper (2001) also found a substantial proportion of non-response errors in Welsh unrelated real word reading; and Hoxhallari et al. (2004), Spencer & Hanley (2003), ánd Ellis & Hooper (2001) found a substantial proportion of real word errors in Albanian respectively Welsh unrelated real word reading. With respect to story reading, Danielsson (2001) also found a substantial proportion of non-word errors in Swedish beginning readers. Gupta & Jamal (2006, 2007), however, only demonstrated this occurrence of real word errors during unrelated real word reading in dyslexic third grade Hindi readers (as opposed to normal readers), whereas in the present study, a considerable proportion of real word errors was found in good, average and poor word decoders. Likewise, Ellis et al. (2004) and Hoxhallari et al. (2004) only found a very small proportion of non-response errors in Albanian readers that read unrelated real words. However, this interesting contrast can be explained by the fact that, for normal readers, the greater the orthographic transparency of a language, the more likely a sublexical, phonological reading strategy is consistently used when reading unrelated real words and the more confident a child is in trying to read a new word (Hoxhallari et al., 2004). Interestingly, Gupta & Jamal (2007) also associated the proportion of real word errors with the spoken frequency of the unrelated real words to be read: the higher the frequency, the larger the proportion of real word errors. Since the unrelated real words in the present study were chosen to be highly familiar, the results are in line with those of Gupta & Jamal (2006, 2007).

Second, the present study also provides new insights on the association between word length, relative text difficulty, grade, word decoding level, and gender, and the nature of reading in Dutch, an orthographically fairly transparent language. A first conclusion is that the proportion of non-word errors and non-response errors became larger with increasing word length, whereas the proportion of real word errors became smaller with increasing word length. These results indicate that, when longer words are read, the difference between the proportions of non-word errors and real word errors becomes larger: they rely even more on sublexical processing strategies instead of lexical strategies when reading longer unrelated real words. At the same time, the increase of the proportion of non-response errors indicates that

children feel less confident to try to read words when these become longer. A second conclusion is that only the proportion of non-response errors became larger with increasing relative text difficulty. This result indicates that the hypothesis that children rely more on lexical strategies when stories become easier to read, cannot be confirmed.

From the data it could also be concluded that, although children predominantly used a sublexical strategy to read unrelated real words and a lexical strategy to read real words in a text, the proportion of real word errors increased with increasing grade when reading both unrelated real words and real words in a text. At the same time, the proportion of non-response errors substantially decreased with increasing grade, and the proportion of non-word errors substantially decreased between grade three and four. These results indicate that there is a developmental shift in the use of sublexical and lexical strategies to read real words. The older children become, the more they rely on lexical strategies to read a word. This explains why, in grade four, children's proportion of non-word errors was only slightly larger than their proportion of real word errors. As such, this result is not in line with the research of Spencer & Hanley (2003), who found that Welsh children at the very beginning of reading instruction (at the age of 6) mainly used lexical reading strategies and only made a shift towards the use of sublexical reading strategies one year later. However, it is in line with the research of Wimmer & Goswami (1994) who found that German children moved into reading by relying heavily on the use of a sublexical reading strategy; however, to achieve fast and fluent reading, they gradually more often used a lexical strategy as they became older.

The data also showed that the proportion of non-word errors decreased and the proportion of real word errors increased with decreasing word decoding level when reading real words in a text. The proportion of non-response errors was the lowest for the good word decoders. This result indicates that, when reading a story, the lower a child's word decoding skills, the more he/she tends to guess the word by means of a lexical strategy instead of trying to really use grapheme-to-phoneme correspondence rules to decode an unknown word.

Finally, during unrelated real word reading, girls made a larger proportion of non-response errors and real word errors, and a smaller proportion of non-word errors than boys. This result indicates that, when reading unrelated real words, girls are less likely than boys to use a sublexical strategy to read

an unfamiliar word. Instead, they are more likely to guess or to not respond at all.

4.5 Limitations and Conclusion

This study was limited in several ways. A first limitation of the present study has to do with the procedure used for the sampling of the subjects that did possibly not result in a representative subgroup of the whole population. Therefore, the results of the present study should be interpreted with caution, and cannot simply be generalized to all Flemish elementary school children.

The second limitation pertains to the fact that there was a substantial amount of missing data, mainly in first grade with respect to unrelated real word reading, and in first, third and fourth grade with respect to stories of each relative difficulty level (see Table 4.2). As a result, interaction effects between task and child variables or between child variables only - although described in the results section - were not discussed, as it was difficult to disentangle the effect of missing data within a particular subgroup from the effect of the characteristic on which the subgroup classification was based. As such, the present study clearly shows the important influence of several child characteristics on the relative use of lexical or sublexical reading strategies; however, future research should use a balanced design and should avoid missing data in order to obtain a full understanding of this association.

A third limitation pertains to the way a child's mastery, instructional and frustrational level of text reading was operationalised, namely the last two levels being one respectively two AVI levels higher than the child's mastery level of text reading as assessed by the AVI test (Visser et al., 1998). However, it can be doubted whether the relation between the relative difficulty of a text for a particular child and the AVI level of a text (as determined by several word and text features), is that straightforward. Moreover, children could not obtain a mastery level of text reading above AVI 9, since this was the highest text reading level that was assessed. This probably prevented us from finding any association between relative text difficulty and non-word or real word error proportions at the word level.

However, despite these limitations, the present study substantially strengthens the orthographic depth hypothesis with respect to the nature of reading in Dutch which is characterized by a predominant use of sublexical strategies to read real words. It also consolidates the effect that the presence of context

has on the strategies used to read real words, namely a shift towards a larger use of lexical strategies. Moreover, this study confirms that the nature of the association that is found between the different types of reading material and error distribution at the word level does not depend on the (relative) difficulty of the reading material nor on a child's (reading) age, word decoding level, or gender. Second, the present study makes clear that word length and relative text difficulty, as well as a child's (reading) age, word decoding level, and gender have an influence too on the emphasis that is put on sublexical and/or lexical reading strategies during unrelated real word and/or story reading.

Notes

[1]Cleuren, L., Kuppens, S., Duchateau, J., Van hamme, H., & Ghesquière, P. (2009). Oral reading errors of Dutch speaking elementary school children: A qualitative analysis at the word level. Manuscript submitted for publication.

[2]CHOREC consisted of recordings of 274 regular elementary school children. Because of the requirements of being native Dutch speaking, not having skipped or doubled a grade, and of not having any articulation problems, only a subset of 248 children was used in the present study.

Chapter 5

Evaluation of a Reading Tutor with Synthesized Speech Feedback for Dutch Speaking Elementary School Children with Reading Difficulties

Abstract[1]

Using a randomized pretest-posttest matched-group design, the present intervention study investigated whether improvements in reading accuracy and speed could be obtained in elementary school children with reading difficulties that used a reading tutor prototype for approximately 4.5 hours during four weeks. The intervention comprised instructional-level non-repetitive text reading with synthesized phoneme-by-phoneme feedback. Each child of a matched pair was randomly assigned to an intervention ($N = 10$) or control group ($N = 10$), and were pre- and post-tested (one week and four weeks after intervention) on various paper-and-pencil and computerized real word, pseudoword, and story reading tests. Mixed model analyses of covariance demonstrated that the intervention group outperformed the control group with respect to both reading accuracy and speed, mainly with respect to real word and pseudoword reading. Additionally, children's pre-intervention real word and pseudoword reading performance level seemed to have a differential effect on intervention efficacy.

5.1 Introduction

The consistency in representing phonemes by means of graphemes can differ largely across languages. This orthographic transparency influences the difficulties experienced by both young, normally developing readers (e.g., Landerl & Wimmer, 2008; Patel et al., 2004; Ziegler & Goswami, 2006), and children who laboriously or not at all arrive at fluent reading. In a language with rather consistent grapheme-phoneme relations (such as for example Greek, Finnish, German, Italian, Spanish and Dutch), children suffering from a phonological deficit will make up arrears regarding reading accuracy when adequate instruction is given; their reading speed, however, will remain a stumbling-block. Therefore, in a language such as Dutch, an older weak or dyslexic reader is able to decode rather accurately but will not succeed in doing this at a reasonable speed; despite good decoding skills, this reader does not arrive at the development of a direct word recognition strategy (e.g., de Jong & van der Leij, 2003; Landerl et al., 1997; Protopapas & Skaloumbakas, 2007; Wimmer, 1996).

Whenever reading problems arise, early, regular and adequate intervention is highly needed in order to prevent a child from dropping further and further behind his or her classmates, not only with respect to his or her reading development but also with respect to appropriate functioning in other school subjects since text usually plays an important role in all of them. Research regarding treatment of reading difficulties has often focused on intervention aiming at improving a child's phonological skills (e.g., Macaruso, Hook, & McCabe, 2006). This research has indicated that accuracy, and consequently reading comprehension, improved significantly by means of that kind of intervention programs; however, it did not foster improvements in reading speed. Because reading difficulties in orthographically transparent languages are not permanently characterized by accuracy but rather by speed problems, and because lack of adequate reading speed negatively influences reading comprehension, it is important that researchers as well as practitioners call attention to reading speed and adequate intervention to improve it (Thaler, Ebner, Wimmer, & Landerl, 2004).

So what should adequate intervention comprise in order to build up fluency in struggling readers that learn to read in an orthographically transparent language such as Dutch? To start, an important condition to become an automated reader is to read a lot (NICHHD, 2000; Samuels, 2002). Unfortu-

nately, those that need it the most often read the least: Children experiencing problems in word recognition exercise less in reading because they read less and slower, a phenomenon called the 'Matthew effect' (Stanovich, 1986). As such, it is clear that reading itself needs to be central in an adequate intervention program for children facing reading difficulties.

However, the most effective way to use reading as the basis of reading intervention to build up fluency (accuracy as well as speed) is still subject of discussion. Although many other intervention strategies exist (see Kuhn & Stahl, 2003, for a review), one general approach is to let children read aloud under supervision and to give them feedback on what they read. Intervention strategies that fall within this approach can be roughly divided into two categories: repetitive versus non-repetitive strategies. The first technique, namely the technique of repeated reading, is based on the theory of automatic information processing in reading (LaBerge & Samuels, 1974), and has, according to the National Reading Panel, a clear and positive effect on reading fluency (NICHHD, 2000). However, results from studies by Rashotte & Torgesen (1985) and Homan, Klesius, & Hite (1993) suggested that when the same amount of practice was done, and the same type of feedback was given, repeated reading was not more effective than non-repetitive strategies to speed up reading.

Additionally, to improve a child's reading fluency, a teacher or clinician usually provides a variety of activities and strategies, occurring before, during or after reading. These activities and strategies include providing feedback and help while the child is reading aloud and struggles or makes a mistake. Frequently studied feedback techniques are those of general feedback and corrective feedback. The first type simply indicates whether a particular performance is right or wrong, whereas corrective feedback goes further than only identifying what is good and what is wrong and provides the reader with help to correct the error (Pany & McCoy, 1988; Perkins, 1988; Spaai et al., 1991). Corrective feedback typically focuses on decoding written symbols by using part-to-whole strategies and thus, on the grapho-phonemic aspects of reading in order to help the child decode the words and read the passage more fluently and accurately (Crowe, 2003). Much research has been conducted on this topic (for an overview, see Crowe, 2003; MacArthur et al., 2001; Wise & Olson, 1998), and has generally shown that corrective feedback strategies (such as for example phoneme-by-phoneme feedback, cf., infra) yield better results than general feedback strategies.

Irrespectively of which intervention strategy is favoured, the enhancement of word reading fluency requires daily reading practice. Given the fact that daily practice can often not be provided due to practical classroom issues and governmental health insurance constraints, the use of a computerized reading intervention program could offer a welcome solution (Irausquin et al., 2005; Lynch, Fawcett, & Nicolson, 2000; MacArthur et al., 2001; NICHHD, 2000). Moreover, as stated by MacArthur et al. (2001), computers are well suited to provide different kinds of speech feedback and decoding assistance for difficult words by means of either digitized (recorded) or synthesized speech and the ability to highlight words or word parts. It has been demonstrated by various researchers that computerized reading intervention with speech feedback fosters word recognition skills and text reading fluency of novice readers and readers with reading difficulties (e.g., Elbro, Rasmussen, & Spelling, 1996; Jiménez et al., 2007; Spaai et al., 1991; van Daal & Reitsma, 2000; van Daal & van der Leij, 1992; Wise & Olson, 1998; Wise, Ring, & Olson, 2000).

Additionally, computers also become better and better suited to listen to a child reading aloud by means of automatic speech recognition. In general, there are two ways to use speech recognition in reading instruction. A first way is to use speech recognition as an assistive technology; namely, as a way for the child to 'write' texts without really writing or typing; the computer listens to the child and transcribes the child's words into text (e.g., Higgins & Raskind, 2004). A second way to implement automatic speech recognition in reading instruction is by using it in a computer that can act as a real reading tutor. Such a tutor can present reading material to the child and is able to really listen to the child's reading performance: it can track the child's progress and therefore follow as the child reads along, and can detect reading errors so that feedback on these errors can be given.

In Flanders, the SPACE project (SPeech Algorithms for Clinical and Educational applications) wants to make oral reading assessment automated and objective by improving speech technology in order to make it suitable as a supporting technology for an automated reading tutor (Cleuren et al., 2006; Duchateau, Cleuren, Van hamme, & Ghesquière, 2007). Moreover, the project not only aims to develop a reading tutor that can really listen to a child reading aloud, but also a tutor that is able to synthesize natural, highly intelligible and phonetically correct speech. More precisely, the newly developed speech synthesizer developed within this project makes it possible to synthesize various feedback modes such as whole-word (i.e., fluently read

sentences or words; e.g., 'mother'), phoneme-by-phoneme (e.g., 'm-o-th-e-r'), and syllable-by-syllable (e.g., 'mo-ther') feedback, with the two last feedback modes being synthesis modes that were not yet available in most speech synthesizers (Latacz, Kong, & Verhelst, 2007; Latacz, Kong, Mattheyses, & Verhelst, 2006). As a result, interactive tools combining both speech recognition and synthesis are being developed so that the tutor can act as a fluent reader model and is able to provide immediate feedback to the learning child.

The present study aimed to evaluate the Dutch reading tutor prototype[2] with synthesized speech feedback developed within the SPACE project. More specifically, this research was designed to investigate whether substantial improvements in reading accuracy and reading speed could be obtained in elementary school children experiencing reading difficulties who used a preliminary version of the reading tutor. It was hypothesized that instructional-level non-repetitive text reading with corrective phoneme-by-phoneme feedback would have a beneficial influence on both reading accuracy and speed of elementary school children with reading difficulties. However, because it directly became clear that not all children's pre-intervention word and pseudoword reading skills were below average (as was originally intended), an additional and exploratory research question to be answered by this study was whether children's initial performance level had an influence on the reading tutor's intervention efficacy.

5.2 Method

5.2.1 Participants

The study was conducted in two Flemish neighbouring regular elementary schools during the first trimester. At the end of the previous school year, remedial teaching coordinators were asked to nominate children with severe reading difficulties who met the following requirements: normal intelligence, Dutch speaking (mother tongue), and absence of sensory difficulties (normal vision and hearing). As such, 28 2nd to 6th grade elementary school children were selected; for all children parental consent was obtained.

Out of the total sample of 28 children, 10 pairs of children were selected (8 girls and 12 boys) who were individually matched with respect to chronological age, IQ (as measured by Raven's Progressive Matrices; Raven, Court,

& Raven, 1992), and mastery text reading level (as measured by the AVI test; Visser et al., 1998). This resulted in four pairs attending grade 2, two pairs attending grade 3, two pairs attending grade 4, and two pairs attending grade 5; the age range for the whole group was 6.75 to 10.75 years. Both children of a grade 2 pair attended the same class in order to minimize the possible effects of different reading instruction methods. In Table 5.1, subject characteristics are presented for both the experimental and control group separately. No significant group differences were demonstrated with respect to none of the matching variables ($ps > .10$).

	Intervention children ($N = 10$; 6 girls)		Controls ($N = 10$; 2 girls)	
	M	*SD*	*M*	*SD*
Chronological age (months)	103.30	18.32	100.60	15.06
Estimated IQ	100.80	13.43	100.20	11.26
AVI level	3.30	2.31	3.10	2.18
Educational level mother[a]	4.10	1.20	4.40	0.84
Educational level father[b]	4.30	1.06	4.67	0.71

Note. [a b] Both parents were asked to indicate their educational level as an index of socio-economic status. These data were used to classify the parents into six categories as described by the international Standard Classification of Education (ISCED) developed by the UNESCO (OECD, 1999).

Table 5.1: *Characteristics of the matched pairs (N = 20)*

5.2.2 Materials

Pre- and Posttest Materials

One-Minute-Test (OMT) (Brus & Voeten, 1995) and Klepel (van den Bos et al., 1994) The child was asked to read a list that contains 1-syllable, 2-syllable, and 3- and 4-syllable words (OMT) or pseudowords (Klepel) (gradually rising in difficulty) for 1 (OMT) or 2 minutes (Klepel). The result on this test was a raw score that expressed the number of correctly read

(pseudo)words (i.e., = total number of (pseudo)words read - number of erroneously read (pseudo)words), which was then converted into a standard score ($M = 10$, $SD = 3$).

AVI test (Visser et al., 1998) The child was asked to accurately and fluently read 1 or more out of 9 stories of graded difficulty level (AVI 1 - 9) in order to detect the child's mastery level of text reading. For a child to obtain a particular AVI level, the corresponding AVI story had to be read with a maximum number of errors and within a maximum amount of time. A child that did not fulfil the requirements to obtain a text reading level of AVI 1, obtained an AVI 0 score.

Real Word Reading Test (RWRT) and PseudoWord Reading Test (PWRT) The child was asked to read three different lists of 40 1-syllable, 2-syllable, and 3- and 4-syllable words (RWRT) or pseudowords (PWRT). (Pseudo)words were presented one-by-one on a computer screen. Reading speed and the total number of errors were registered. The added value of the RWRT and PWRT over the OMT and Klepel was that speed and accuracy were independently registered. Moreover, because 1-syllable, 2-syllable, and 3- and 4-syllable (pseudo)words were assessed separately, it allowed for a more differentiated analysis of reading performance.

Story Reading Test (SRT) The child was asked to read 2 or 3 stories out of 9 SRT stories of graded difficulty level (AVI 1 - 9). The choice of SRT stories was determined by the child's mastery AVI level as assessed by the AVI test (Visser et al., 1998). For example, a child with an AVI 3 mastery level had to read the AVI 3 (mastery), AVI 4 (instructional) and AVI 5 (frustrational) SRT stories. If children did not yet master the story at level AVI 1, or already mastered the story at level AVI 8, only 2 stories needed to be read. All stories were presented paragraph-by-paragraph on a computer screen. Reading speed and the total number of errors were registered for each story.

Intervention Materials

During each intervention session, the child read (parts of) an instructional-level story (i.e., a story of a difficulty level above the child's mastery level

as assessed by the AVI test). Children's storybooks were carefully selected from commercial reading materials by the first author, and were converted to an appropriate computer version that could be presented on a computer screen by means of the reading tutor that offered the one-to-one tutoring. In each session, new reading material was presented so that no re-reading of the same text occurred.

Reading Tutor Setup and Speech Synthesis

Setup During the pre- and posttest, children sat in front of a laptop computer on which the computerized reading tests (RWRT, PWRT, and SRT) were shown by means of the reading tutor interface that simultaneously recorded each test administration. In case of the paper-and-pencil OMT, Klepel and AVI test, the reading tutor interface was only used to obtain high-quality audio recordings. During the intervention sessions, children sat in front of a large touch screen that was connected to the same laptop computer. Now, the reading material was presented on the touch screen and two speakers were connected to the laptop to give speech feedback. Again, high-quality recordings were made.

Speech synthesis By means of the so-called 'Wizard-of-Oz' method (Dahlbäck, Jönsson, & Ahrenberg, 1993), the human researcher replaced the speech recognition module: Although no real speech recognition was used, children really thought that the computer was able to listen to them. However, whenever the child made an error that was not spontaneously corrected immediately afterwards, the reading tutor highlighted the erroneously read word and gave phoneme-by-phoneme feedback to the child (with a maximum of three times for each word). The child was then asked by the tutor to blend the phonemes together in order to read the word correctly before carrying on with the reading of the story. When the child encountered a difficult word, phoneme-by-phoneme feedback could also be asked for by touching the word.

5.2.3 Procedure

Design

A randomized pretest-posttest matched-group design served as the framework for this intervention study. Both children within a matched pair were

randomly assigned to an experimental (intervention) or control group (no extra intervention) following a pretest assessment of reading (decoding) skills. Following intervention, a posttest and follow-up reading skill assessment (one week respectively four weeks after intervention) was administered to each child.

General Procedure

All assessments and interventions took place in a quiet room at the child's own school. As is depicted in Figure 5.1, one week before the start of the intervention period, pretest assessments were individually administered to all children ($N = 20$) in two 30-minute sessions on two consecutive days.

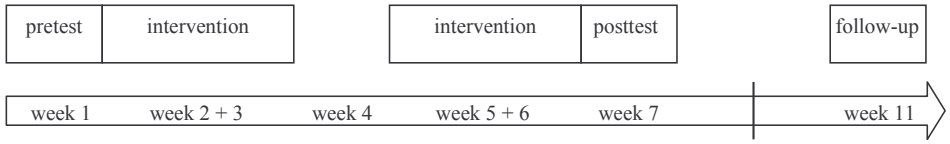

Figure 5.1: General procedure

Following the pretest, the children in the intervention group ($N = 10$) individually received two 2-week intervention blocks that contained fourteen 20-minute reading sessions: seven reading sessions in the first 2-week block, another seven sessions in the second 2-week block. This resulted in a total of approximately 4.5 hours of training for each child ($M = 4.61$, $SD = 0.22$). Between the two intervention blocks, there was one week of holiday in which no intervention was given. In order to familiarize children and examiners with the routine of the intervention sessions, one pre-intervention 15-minute reading session was organized in the pretest-week. The group of control children ($N = 10$) did not receive any form of extra intervention.

Afterwards, posttest (after one week) and follow-up (after four weeks) assessments were individually administered to all children in two 30-minute sessions on two consecutive days. For the tests consisting of two parallel versions (i.c., OMT, Klepel, AVI test), the different versions were alternately administered; for the others (i.c., RWRT, PWRT, SRT), the same three lists of words or pseudowords, or the same 3 stories (or 2 stories) were read at the three different assessment moments.

Computerized Reading Tutor Intervention Sessions

The intervention featured individual 20-minute computerized reading sessions in which the child read instructional-level story material. The reading tutor registered the child's progress in each story so that for each new reading session, the child could carry on reading the story from the point that was reached the session before. Every time an error was made or help was asked for, phoneme-by-phoneme feedback was provided by the reading tutor.

Intervention children were motivated via an improvement graph that was shown after every session, together with a verbal positive comment by the reading tutor ("Well done", "Congratulations" etc.). At the end of each week, they received a sticker that could be exchanged for a present at the end of the intervention study.

5.2.4 Analysis Strategy

Effects of Intervention

To determine the effects of intervention, separate analyses were run for each variable of each reading task at both posttest and follow-up. All hypotheses were tested at the .05 level of significance. To account for the dependent nature of the data (i.e., matched pairs), mixed model analyses of covariance (ANCOVA) were conducted, using proc mixed of the SAS computer program (Littell, Milliken, Stroup, Wolfinger, & Schabenberger, 2006). The pretest value of each variable was treated as a fixed covariate; and dependency of the data was modelled by entering 'matched pair' as a random factor in all analyses. The effect of 'group' (intervention versus control) was investigated by entering this variable as a between-subjects factor.

Since the sample sizes in this intervention study were low, it was disputable to draw inferences after a failure to reject the null hypothesis. Therefore, R^2 was calculated as a measure of effect size in order to detect differences of clinical interest. As described by Roberts & Monaco (2006), R^2 can be obtained by (1) subtracting the variance estimate from the model where both the pretest value and 'group' (intervention versus control) are entered as predictors, from the variance estimate from the model where only the pretest value is entered as a predictor (i.e., baseline model); and (2) dividing this subtraction by the variance estimate of the baseline model. As such, R^2 represents the fraction of variation that is explained by entering 'group'

(intervention versus control) into the model (Singer, 1998).

Influence of Performance Level on Intervention Effects

Although both school's remedial teaching coordinator was asked to only select children with severe reading difficulties, it could be seen that, with respect to the One-Minute-Test, 7 (of which 4 controls) out of 20 children obtained a pre-intervention 'OMT-Standard score' of at least 8. With respect to the 'Klepel-Standard score', the same held for 8 (of which 4 controls) out of 20 children. This dichotomy between the remedial teacher's opinion and some children's performance on the OMT and Klepel was likely due to a spontaneous catch-up during the summer holiday. Six out of seven (in case of the OMT), and seven out of eight children (in case of the Klepel) attended 2nd grade when participating in this study. As a result, their diagnosis of 'having severe reading difficulties' was made very early (at the end of first grade), which has likely led to false positives (cf., Wanzek & Vaughn, 2007).

Because we wondered if lower performing children's response to intervention was different from that of average performing children, the total group of 20 children was divided into a 'clinical' group ($N = 14$; 7 pairs) and an 'average' group ($N = 6$; 3 pairs). Children belonging to the average group obtained a standard score of at least 8 on both the OMT and the Klepel; children belonging to the clinical group obtained a standard score of 7 or less on at least one of both tests. Important to keep in mind was that the average group solely consisted of 2nd grade children (mean age = 85.83, $SD = 4.12$), whereas the clinical group consisted of 2nd to 5th grade children (mean age = 108.86, $SD = 14.64$).

By allowing the effect of the 'group' variable (intervention versus control) to vary at the pair level, and by entering the 'group' x 'performance' (average versus clinical) interaction into the fixed part of the original model, the differential effect of intervention according to children's performance level was investigated.

5.3 Results

5.3.1 Effects of Intervention

For both the intervention and control children, means (pretest) or adjusted means (posttest, follow-up), standard deviations or errors, and R^2 for all accuracy and fluency variables of all six reading tests (OMT, Klepel, AVI test, RWRT, PWRT, and SRT) are reported in Tables 5.2, 5.3, 5.4, and 5.5. There was 3.1% of missing data resulting from lost recordings due to technical difficulties and could therefore be regarded as missing at random. Another 0.2% of missing data was due to the fact that one particular child's story reading mastery level was below AVI 1 so that no mastery level story was available for that child.

Real Word Reading

Reading accuracy The mixed model analyses significantly showed that, at posttest, 1-syllable ($F(1,20) = 11.15$, $p = .003$, $R^2 = .36$) as well as 2-syllable real words ($F(1,20) = 4.77$, $p = .04$, $R^2 = .19$) were more accurately read by the intervention group than by the control group. At follow-up, the difference between the two groups stayed marginally significant with respect to the reading of 1-syllable real words ($F(1,9.41) = 3.49$, $p = .09$, $R^2 = .17$), but became insignificant with respect to the reading of 2-syllable real words ($F(1,9.9) = 2.44$, $p = .15$, $R^2 = .20$). However, looking at R^2, indicating the amount of variability that could be explained by the 'group' variable (intervention versus control), a clear intervention effect on accuracy was still obvious with respect to the 2-syllable real words at follow-up ($R^2 = .20$). Furthermore, at follow-up, there seemed to be a small positive (though insignificant) effect of intervention with respect to the variable 'OMT-Standard score' ($F(1,20) = 2.25$, $p = .15$, $R^2 = .10$): the intervention group obtained a higher mean standard score than the control group with respect to the standardized One-Minute-Test. With respect to the reading of 3+4-syllable real words, there were no significant differences between the two groups nor meaningful R^2 values at both posttest and follow-up.

Reading speed No significant differences were observed between groups at none of the assessment moments by means of mixed model analyses. However, R^2 statistics demonstrated that a substantial amount of variance in real

word reading speed was attributed to the intervention. Namely, it seemed that 1-syllable and 2-syllable real words were read faster by the intervention group than by the control group at both posttest (1-syllable words: $F(1,20) = 2.15$, $p = .16$, $R^2 = .43$; 2-syllable words: $F(1,7.21) = 3.30$, $p = .11$, $R^2 = .29$) and follow-up (1-syllable words: $F(1,7.24) = 1.12$, $p = .32$, $R^2 = .58$; 2-syllable words: $F(1,8.85) = 2.42$, $p = .15$, $R^2 = .20$). Again, these group differences were not detected with respect to 3+4-syllable real words.

		Pretest		Posttest[a]			Follow-up[a]		
		M	*SD*	*M*	*SEM*	R^2	*M*	*SEM*	R^2
					One-Minute-Test				
Standard score	C	6.30	3.83	7.14	0.32	0.02	7.35	0.29	0.10
	I	6.20	3.61	6.96	0.32		7.95	0.29	
					Klepel				
Standard score	C	6.40	3.20	7.20	0.42	0.01	6.90	0.58	0.00
	I	6.40	3.03	7.00	0.42		7.10	0.58	
					AVI test				
Mastery level	C	3.10	2.18	3.48	0.29	0.01	3.79	0.20	0.10
	I	3.30	2.31	3.32	0.29		4.21	0.20	

Note. C = control group ($N = 10$); I = intervention group ($N = 10$).
[a] Means adjusted for pre-intervention values.

Table 5.2: *(Adjusted) means, standard deviations or errors, and R^2 for the One-Minute-Test, Klepel and AVI test*

Pseudoword Reading

Reading accuracy Mixed model analyses showed no significant differences between groups nor meaningful R^2 values at posttest. At follow-up, although no differences between groups could be seen with respect to the standard score obtained on the Klepel, it seemed that an intervention effect could be seen for PWRT pseudowords of each length: intervention children made

fewer errors than control children. Moreover, the differences between groups were marginally significant with respect to 1-syllable pseudowords ($F(1,9.03) = 4.29$, $p = .07$, $R^2 = .37$), and statistically significant with respect to 3+4-syllable pseudowords ($F(1,8.47) = 5.41$, $p = .05$, $R^2 = .40$). With respect to 2-syllable pseudowords, the difference was insignificant but still very clear ($F(1,9.67) = 2.32$, $p = .16$, $R^2 = .23$).

Reading speed At posttest, mixed model analyses showed no significant differences between the intervention and control group, nor meaningful R^2 values. However, at follow-up, the differences in speed between the two groups became statistically significant. Surprisingly, children forming part of the intervention group needed more time than the control children to read the three different lists of 40 1-syllable ($F(1,8.99) = 10.44$, $p = .01$, $R^2 = .57$), 2-syllable ($F(1,10.2) = 9.42$, $p = .01$, $R^2 = .55$), and 3+4-syllable pseudowords ($F(1,8.38) = 6.60$, $p = .03$, $R^2 = .46$).

Story Reading

Reading accuracy At posttest, mixed model analyses showed a small though insignificant effect of group (intervention versus control) with respect to the reading of stories at the instructional level: children in the intervention group made fewer errors than control children ($F(1,10.4) = 2.97$, $p = .11$, $R^2 = .19$). With respect to the reading of stories at the mastery level, a R^2 value of .14 showed that there was a small though insignificant effect of group: unexpectedly, the intervention group made more errors than the control group ($F(1,8.33) = 1.21$, $p = .30$, $R^2 = .14$). Furthermore, at posttest, no significant accuracy differences or meaningful R^2 values were detected with respect to the reading of stories at the frustrational level.

At follow-up, differences between groups with respect to stories at the instructional level had disappeared, and were still not present with respect to stories at the frustrational level. However, the unexpected difference with respect to stories at the mastery level had increased and had even become statistically significant ($F(1,7.79) = 11.33$, $p = .01$, $R^2 = .61$).

Reading speed Mixed model analyses showed no significant differences between groups at none of the assessment moments with respect to none of the story speed variables. However, an R^2 value of .20 indicated that at posttest

		Pretest		Posttest[a]			Follow-up[a]		
		M	*SD*	*M*	*SEM*	R^2	*M*	*SEM*	R^2
					1 syllable				
Errors	C	7.40	7.43	5.09[b]	0.50	0.36	5.15[d]	0.76	0.17
	I	5.00	8.62	2.71[b]	0.50		3.15[d]	0.76	
Time	C	84.54	33.19	76.14	3.27	0.43	68.11	3.62	0.58
	I	81.77	32.51	69.37	3.27		64.36	3.62	
					2 syllables				
Errors	C	9.80	14.90	6.48[c]	1.02	0.19	5.99	1.10	0.20
	I	9.10	14.08	3.32[c]	1.02		3.81	1.10	
Time	C	118.00	60.43	107.90	8.32	0.29	87.14	6.32	0.20
	I	111.40	59.80	93.70	8.32		82.26	6.32	
					3+4 syllables				
Errors	C	16.11[e]	23.96	9.23[e]	1.93	0.02	8.70[e]	1.70	0.04
	I	10.67[e]	8.17	7.44	1.93		6.63	1.70	
Time	C	152.89[e]	65.00	132.96[e]	8.19	0.00	111.58[e]	6.74	0.06
	I	157.00[e]	70.84	128.41	8.19		115.58	6.75	

Note. C = control group ($N = 10$); I = intervention group ($N = 10$).
[a] Means adjusted for pre-intervention values. [b c] The differences between the control and intervention group are found to be statistically significant at $p = .05$. [d] The difference between the control and intervention group is found to be marginally significant at $p = .10$. [e] $N = 9$.

Table 5.3: *(Adjusted) means, standard deviations or errors, and R^2 for the Real Word Reading Test*

the intervention group tended to read stories at the frustrational level faster than the control group ($F(1,9.91) = 1.61$, $p = .23$, $R^2 = .20$); but this difference in reading speed had disappeared again at follow-up.

5.3.2 Influence of Performance Level on Intervention Effects

It was detected that intervention did have an effect on the accuracy with which 3+4-syllable real words were read, but that this effect could only be seen in the average performing, younger group. More specifically, at posttest, the interaction effect between group and performance was significant ($F(1,8.24) = 9.90$, $p = .01$), and a test of simple effects showed that 3+4-syllable words were more accurately read by the average children from the intervention group than those from the control group ($F(1,7.86) = 9.28$, $p = .02$). With respect to the clinical children, there was no difference between groups. At follow-up, the interaction effect stayed marginally significant ($F(1,9.73) = 4.31$, $p = .07$), and the positive impact of intervention could still be seen in the average group ($F(1,9.11) = 4.72$, $p = .06$) whereas no group difference could be found in the clinical group.

Another differential effect of intervention could be seen with respect to the time that was needed to read the 3+4-syllable pseudowords at follow-up. Then, the 'group' x 'performance' interaction effect was marginally significant ($F(1,7.88) = 4.66$, $p = .06$), and a test of simple effects showed that the 3+4-syllable pseudowords were more slowly read by the clinical experimental children than by the clinical control children ($F(1,8.1) = 14.54$, $p = 0.01$), whereas no group differences could be seen within the younger, average group. In other words, the previously detected 3+4-syllable pseudowords speed difference between the intervention and control group was only due to differences within the clinical group.

A third differential effect of intervention could be seen with respect to the accuracy of stories read at the frustrational level at follow-up. At that assessment moment, the interaction between group and performance was significant ($F(1,8.46) = 7.67$, $p = .02$), and a test of simple effects showed a significant effect of intervention in the younger, average group ($F(1,8.93) = 6.76$, $p = .03$). In other words, although no significant effect of intervention was detected for the total group, it was detected in the younger, average group: in this group, intervention children made significantly fewer errors

		Pretest		Posttest[a]			Follow-up[a]		
		M	*SD*	*M*	*SEM*	R^2	*M*	*SEM*	R^2
					1 syllable				
Errors	C	12.00	7.96	10.25	1.76	0.05	13.71[f g]	1.79	0.37
	I	14.10	8.60	9.65	1.76		10.39[f]	1.72	
Time	C	109.60	37.38	101.11	6.14	0.00	85.81[b g]	7.51	0.57
	I	110.90	30.31	101.29	6.14		97.54[b]	7.46	
					2 syllables				
Errors	C	28.70	18.15	25.09[g]	3.13	0.01	28.77[g]	3.15	0.23
	I	34.10	20.44	23.62	2.96		22.81	2.96	
Time	C	154.90	63.92	135.05[g]	9.75	0.03	114.41[c g]	11.35	0.55
	I	160.70	41.80	141.12	9.58		135.89[c]	11.25	
					3+4 syllables				
Errors	C	57.22[g]	30.02	53.09[g]	4.31	0.02	60.49[d g]	5.38	0.40
	I	67.20	36.07	49.91[g]	4.33		46.50[d g]	5.42	
Time	C	200.89[g]	62.27	191.00[g]	14.10	0.04	171.26[e g]	16.81	0.46
	I	234.00	69.52	200.42[g]	14.13		203.14[e g]	16.85	

Note. C = control group ($N = 10$); I = intervention group ($N = 10$).
[a] Means adjusted for pre-intervention values. [b c d e] The differences between the control and intervention group are found to be statistically significant at $p = .05$. [f] The difference between the control and intervention group is found to be marginally significant at $p = .10$. [g] $N = 9$.

Table 5.4: *(Adjusted) means, standard deviations or errors, and* R^2 *for the PseudoWord Reading Test*

than their matched controls.

A last differential effect of intervention was detected with respect to the speed with which stories at the frustrational level were read at posttest. It seemed that the effect that was described before was in fact due to the younger, average performing group: the intervention children of this group needed significantly less time to read the story than the individually matched control group ($F(1,8.56) = 10.09$, $p = .01$). Consequently, the 'group' x 'performance' interaction was significant ($F(1,8.63) = 7.65$, $p = .02$) for 3+4-syllable pseudowords at posttest.

5.4 Discussion

The aim of this intervention study was to investigate whether improvements in reading accuracy and reading speed could be obtained in Dutch speaking elementary school children with reading difficulties that used the reading tutor prototype with synthesized speech feedback developed within the SPACE project. To do so, a randomized pretest-posttest matched-group design was used in which each child of a matched pair was randomly assigned to an intervention or control group. By means of the reading tutor, the intervention group received individualised intervention sessions that focussed on non-repetitive instructional-level text reading with phoneme-by-phoneme feedback when an error was made or help was asked for.

The reading tutor feedback consisted of synthesized speech (as opposed to digitized speech) that made it possible to develop an 'open' program so that new text material for each individual child could be imported into the reading tutor at any moment. Based on the opinions of the three examiners, and based on the lack of negative reaction with respect to this matter from the children that exercised with the reading tutor, it could be concluded that the speech synthesis quality was good: phoneme-by-phoneme feedback sounded naturally, was clearly audible and well understandable. Children got quickly used to the reading tutor's voice and were very understanding with respect to the reading tutor's occasional mistakes (e.g., accidentally prompting feedback although the word was decoded correctly).

To evaluate the reading tutor's effectiveness, children's reading skills were assessed before (pretest), immediately after (posttest), and four weeks after (follow-up) the intervention period by means of several word reading, pseudoword reading and text reading paper-and-pencil and computerized

		Pretest		Posttest[a]			Follow-up[a]		
		M	*SD*	*M*	*SEM*	R^2	*M*	*SEM*	R^2
					Mastery				
Errors	C	8.40	7.71	6.30	1.44	0.14	4.27[c b]	1.18	0.61
	I	7.56[c]	7.70	7.67[c]	1.50		6.81[c b]	1.18	
Time	C	109.80	23.93	88.57	4.69	0.03	79.81[c]	4.58	0.03
	I	97.33[c]	13.41	90.83[c]	5.02		81.49[c]	4.63	
					Instructional				
Errors	C	11.20	7.18	11.72	1.80	0.19	9.66[c]	1.33	0.03
	I	17.70	13.42	7.38	1.80		8.11	1.26	
Time	C	137.60	29.24	110.93	4.93	0.00	103.76[c]	7.28	0.00
	I	137.70	33.76	110.67	4.93		103.77	6.97	
					Frustrational				
Errors	C	18.89[c]	15.39	12.09[c]	2.20	0.06	10.72[d]	1.51	0.01
	I	24.80	11.83	9.70	2.07		9.60	1.32	
Time	C	156.67[c]	33.56	137.53[c]	6.59	0.20	118.93[d]	6.11	0.01
	I	177.60	45.75	126.80	6.17		117.11	5.40	

Note. C = control group ($N = 10$); I = intervention group ($N = 10$).
[a] Means adjusted for pre-intervention values. [b] The differences between the control and intervention group is found to be statistically significant at $p = .05$. [c] $N = 9$. [d] $N = 8$.

Table 5.5: *(Adjusted) means, standard deviations or errors, and R^2 for the Story Reading Test*

tests. The reading accuracy results revealed that, at posttest, the intervention group outperformed the control group with respect to the reading of 1-syllable and 2-syllable real words, and with respect to the reading of stories at the instructional level. If differences in initial (pseudo)word reading level were taken into account, it seemed that children that obtained an average or above average standard score on both the Dutch One-Minute-Test (standardized real word reading test) and the Klepel (standardized pseudoword reading test) (i.e., 6 out of 20 children, henceforth called the 'average' group), additionally outperformed their control group with respect to the reading of 3+4-syllable real words. At follow-up, the intervention group outperformed the control group with respect to the reading of 1-syllable and 2-syllable real words, and 1-syllable, 2-syllable and 3+4-syllable pseudowords. Moreover, the standard score obtained on the One-Minute-Test was higher for the intervention group than for the control group. The average group additionally outperformed their control group with respect to the reading of 3+4-syllable real words, and with respect to the reading of stories at the frustrational level. There was one counterintuitive result obtained in this study: stories at the mastery level were less accurately read by the intervention group than by the control group at both posttest and follow-up.

The reading speed results revealed that, at posttest, the intervention group was faster than the control group with respect to the reading of 1-syllable and 2-syllable real words. The average group additionally outperformed their control group with respect to the reading of stories at the frustrational level. At follow-up, the intervention group did better than the control group with respect to the reading of 1- and 2-syllable real words. However, 1-syllable, 2-syllable and 3+4-syllable pseudowords were read more slowly by the intervention group than by the control group at follow-up in order to be read more accurately. This could easily be explained by the observation that at pretest, when a pseudoword was too difficult for a child to read, a fast and wrong guess was made instead of a real try to decode the pseudoword. As a result, any attempt to really try to decode the pseudoword was slower than at pretest. However, there was no effect of intervention anymore with respect to the speed with which stories of different difficulty levels (mastery, instructional, frustrational) were read.

In sum, despite a very limited amount of training (4,5 hours on average) for the intervention group, the data demonstrated a beneficial influence of extra-curriculum practice with the reading tutor on the reading performance

of elementary school children, mainly with respect to real word reading and pseudoword reading. Immediately after the intervention period, reading accuracy and speed gains could be seen with respect to real word reading; four weeks after intervention, gains could be seen for both real word and pseudoword reading. With respect to story reading, a more continuous or larger amount of training is probably needed to preserve the obtained accuracy and speed gains at posttest, or to extend the effect of intervention from the average to the whole group.

Importantly, although this study's aim was to investigate the effect of intervention in a group of elementary school children with reading difficulties, results also indicate that a group of average performing children profited from individualized intervention by the reading tutor too. In comparison to the group of initially clinically performing children, they showed additional gains in reading accuracy and/or speed with respect to 3+4-syllable real words and stories at the frustrational level. However, since all 6 average performing children attended grade 2, one could wonder whether these additional gains are simply a characteristic of the total group of 2nd grade children ($N = 8$) which are the youngest group of children in this study. To answer this question, the 'group' x 'performance' interaction in the fixed part of the model was replaced by the 'group' x 'grade' interaction. As such it was found that this hypothesis can be true for the additional gain in reading accuracy with respect to the reading of 3+4-syllable real words at posttest, but not for the other (additional) gains found in the average performing or total group. This finding is in line with the results of Tressoldi and colleagues who stated that age does not make a difference with respect to fluency remediation in dyslexic children (Tressoldi, Lorusso, Brenbati, & Donini, 2007).

Moreover, an other important finding in this study is that, irrespectively of the children's initial word and pseudoword reading skills, or chronological age, not only gains in reading accuracy are observed, but also in reading speed. As such, the intervention provided by the reading tutor prototype (i.e., instructional-level non-repetitive text reading with phoneme-by-phoneme feedback) seems to be effective for children learning to read in an orthographic transparent language such as Dutch, in which speed rather than accuracy often forms the largest stumbling block for struggling readers.

However, some methodological limitations of this study must be taken into account when interpreting the data. The main limitation is the fact that the study's sample size is small and inevitably affects the power of the

study. An additional disadvantage of the study is that, given the small sample size, the age range is very large (2nd to 5th grade), making it difficult to disentangle the effects of (reading) age and reading development on intervention efficacy. Moreover, the heterogeneity of the group is even larger because of the fact that pretest word reading and pseudoword reading skills do not seem to be below average for all children (as such resulting in an average performing subgroup). Taken all this together, conclusions regarding the effect of reading age, reading development or reading ability on intervention efficacy must be handled with care.

However, despite the possibly doubtful power of this study, positive effects on reading accuracy ánd speed are obtained, and the speech synthesis implemented in the tutor is positively evaluated. Therefore, the way is paved for the implementation of the SPACE speech recognition module into the tutor, and the development of the fully automated reading tutor. An accurate speech recognizer makes it possible to track where the child is reading in the text and to accurately detect oral reading errors so that immediate and personally adapted feedback can be given on any new text the child needs to practice. As such, the motive to develop an automated reading tutor is not to let it replace any form of human support, but rather to let it be of use in situations where no human help and tutoring is available. Additional research will then be needed in order to evaluate the fully automated version of the reading tutor with both implemented automatic speech recognition and speech synthesis.

Notes

[1]Cleuren, L., Latacz, L., Kong, Y.-O., Vandermosten, M., Verhelst, W., Van hamme, H., & Ghesquière, P. (2008). Evaluation of a reading tutor with synthesized speech feedback for Dutch speaking elementary school children with reading difficulties. Manuscript submitted for publication.

[2]In contrast to the reading tutor discussed in the Introduction of this dissertation, no real speech recognition was implemented yet in the preliminary version of the reading tutor that was used in this study.

Chapter 6

General Conclusions

Chapter two of this dissertation discussed the development of the CHildren's Oral REading Corpus (CHOREC), and the evaluation of the quality of the annotations forming part of that corpus. Chapters three and four reported on a quantitative respectively qualitative analysis of oral reading errors extracted from the CHOREC corpus. Next, Chapter five reported on an intervention study that was carried out to evaluate a preliminary version of a computerized reading tutor using synthesized computer speech. To conclude, the current chapter summarizes the main findings of this dissertation and additionally highlights some important methodological findings and practical implications. It also discusses some limitations of the presented research, and provides some ideas for future research.

6.1 Main Findings

As discussed in the general introduction of this dissertation, the successful implementation of automatic speech recognition in the reading assessment process enables us to address all disadvantages associated with traditional paper-and-pencil methods for word decoding assessment: the oral reading assessment process can become automated and objective, and accurate (timing) measures of the child's reading performance can be obtained. At the same time it enables us to let a child really read aloud so that a mouse, keyboard, or touch screen response is no longer necessary when one wants to develop a computerized reading assessment tool. Second, within the context of reading intervention, accurate automatic speech recognition enables us to

convert a computer into a real reading tutor which can present reading material to a child and is able to really listen to a child's reading performance: it can track a child's progress and therefore follow as the child reads along, and can detect reading errors so that immediate feedback on these errors can be given.

Before a speech recognizer is able to accurately recognize children's reading (children's hesitations and oral reading errors included), it needs to have a lexicon, a language model, and an acoustic model at its disposal (Huang, Acero, & Hon, 2001). The lexicon is a dictionary that contains words and their possible phonetic transcriptions. The language model allows for the computation of probabilities of particular word sequences. In a reading assessment application, it is created from the knowledge of the sentence to be read and from predictions of likely oral reading errors. The acoustic model is a model that can be created by taking audio recordings of children's reading and their annotations, and by training the recognizer on these data to obtain statistical representations of the sounds that make up the different words (i.e., phones). To obtain a good acoustic model that captures as much variation as possible in sounding out a particular phoneme, a large annotated children's oral reading corpus is needed.

In other words, to enable a speech recognizer to accurately track a child's reading, and to accurately detect oral reading errors, it must be trained on a corpus that contains recordings of children's oral reading and accurate annotations of these recordings (with the latter usually being done manually and therefore being very labour-intensive and expensive). To meet this demand, the CHildren's Oral REading Corpus (CHOREC) was developed which contains 130 hours of recorded (by means of the reading tutor platform) and manually annotated read speech of 400 Dutch elementary school children with and without reading difficulties (Cleuren et al., 2006; Cleuren, Duchateau, et al., 2008). It adds to other existing (mostly English) children's speech corpora such as, for instance, the KIDS Corpus (Eskenazi, 1996), the OGI Kids' Speech Corpus (Shobaki, Hosom, & Cole, 2000), the PF_STAR British English Children's Speech Corpora (Batliner et al., 2005; Russell, D'Arcy, & Wong, 2004), and the TBALL Corpus (Kazemzadeh et al., 2005). The reading material that was used for the making of the CHOREC corpus consisted of the Real Word Reading Test (RWRT), the PseudoWord Reading Test (PWRT), and the Story Reading Test (SRT), all newly developed to be used in a computerized speech recognition based reading assessment tool.

The final version of the CHOREC corpus includes both information on utterances directly resulting from the child's efforts to read what was presented on the computer screen, as well as information on background noise and reading task-related and task-unrelated unforeseen utterances made by both the reading child and the examiner who was sitting next to the child during reading. More specifically, CHOREC provides information on the content of the reading task used for the making of the recordings, and on the timing of stimulus (i.e., individual (pseudo)words or paragraphs) presentation on the computer screen. It also provides a manual orthographic and broad-phonetic transcription of what the child actually read (hesitations and oral reading errors included), as well as an indication of the types of oral reading errors that were made and a description of the reading strategies that were used during reading.

In spite of the construction of a detailed annotation protocol, the thorough training and supervision of the different annotators, and the formal check and manual correction of the annotations afterwards, an analysis of annotator agreement and consistency in segmentation, transcription, and labelling was recommended (Cleuren, Duchateau, et al., 2008). Percentage agreement scores and kappa values both showed that agreement between annotations was substantial to almost perfect: percentage agreement scores varied between 86.4% and 98.6%, whereas kappa values varied between 0.72 and 0.97 (cf., Landis & Koch, 1977). These results indicated that the quality of the annotations was high and that CHOREC could be reliably used for the development of speech recognition systems, and in particular for the development of automatic speech recognition based reading assessment and intervention tools.

At the same time, the CHOREC corpus offered the unique possibility to accurately characterize the phenotypic reading performance pattern of Flemish Dutch speaking elementary school children with and without reading difficulties with respect to oral reading errors; information which is of interest for both speech technologists and reading researchers. The former can make use of such information to develop an accurate language model for the speech recognizer to optimize its performance. Without any language model, the speech recognizer's task would be to judge on its own whether or not an utterance is close enough to the target word to be read. This is a difficult task given the fact that it is difficult to quantify what an allowable deviation is, and given the fact that a huge amount of annotated children's reading would

be needed in order to obtain a fairly accurate decision. In contrast, the use of an error model (i.e., language model) enables the prediction of possible errors so that the task of the speech recognizer is narrowed down to find out which of the possible (correct or wrong) pronunciations was realized, which is an easier problem than the previously mentioned open-ended classification task. Moreover, not all error models will be equally likely for a particular word but will be assigned a prior probability. When a speech recognizer receives an acoustic signal to be recognized, it will make use of the acoustic model to search for a statistic model that best explains the acoustic data. As such, it can easily take prior probabilities into account by multiplying the posterior acoustic model probabilities with their prior language model probabilities.

Reading researchers can use the oral reading error data to find out more about oral reading processes in the orthographically fairly transparent Dutch language (cf., Seymour et al., 2003), in which reading can almost entirely be mastered on the basis of consistent grapheme-to-phoneme correspondence rules. This consistency stands in harsh contrast to the extreme inconsistencies found in the English orthography. As most reading research (research on oral reading errors included) has been conducted in English, a concern has emerged among reading researchers that findings may not be applicable to other languages with more consistent mappings between orthography and phonology (Protopapas & Skaloumbakas, 2007) . A language's orthographic transparency or lack of transparency can be associated with very substantial differences in reading development and early reading performance (e.g., Landerl et al., 1997; Seymour et al., 2003). As a result, a renewed focus on oral reading errors with respect to orthographically more transparent languages has emerged (e.g., Danielsson, 2001, 2002, 2003; Ellis et al., 2004; Goikoetxea, 2006; Gupta, 2004; Gupta & Jamal, 2006, 2007; Hoxhallari et al., 2004; Nergård-Nilssen, 2006; Seymour et al., 2003; Spencer & Hanley, 2003).

To contribute to the literature on the reading process in orthographically transparent languages, and, at the same time, to obtain information on children's oral reading behaviour that could be of interest to be implemented into the language model of the speech recognizer, both a quantitative and qualitative analysis of oral reading errors annotated in the CHOREC corpus was carried out (Cleuren, Kuppens, Duchateau, Van hamme, & Ghesquière, 2009a; Cleuren, Kuppens, Duchateau, Van hamme, & Ghesquière, 2009b). The main goal of the quantitative reading error analysis was to investigate

the associations between important child variables (i.e., (reading) age, word decoding level, gender) and reading task variables (i.e., relative text difficulty, lexical status, presence of context, word length), and the occurrence of reading errors during regular elementary school children's real word, pseudoword and text reading. The main goals of the qualitative reading error analysis were to find further evidence for the orthographic depth hypothesis with respect to the nature of reading in Dutch, and to further increase the understanding of the Dutch reading process in general by investigating the proportions of non-response errors, real word errors, and non-word errors made by Dutch speaking elementary school children of different age, gender, and word decoding level, during the reading of real words and stories.

The results of the quantitative analyses with respect to the task predictors replicated the results that were found in the orthographically opaque English language. More specifically, the results confirmed that, also in Dutch, pseudowords were harder to read (i.e., with lower accuracy) than unrelated real words, and the latter were harder to read than real words embedded in a story context. The data also confirmed that longer unrelated real words or pseudowords were read with lower accuracy than words or pseudowords with fewer syllables, and that stories at a child's individualized frustrational level induced relatively more reading errors than stories at a child's individualized mastery or instructional level. With respect to the child predictors, the results also mainly replicated the results that were found in the orthographically opaque English language. More specifically, the results confirmed that boys read less accurately than girls, that reading accuracy increased with increasing grade, that good word decoders obtained higher accuracy scores than average word decoders, that the latter were more accurate than poor word decoders, and that readers with a higher text reading level read with higher accuracy than readers with a lower text reading level. However, differences between groups did not always reach statistical significance. It was therefore suggested to use a more fine-grained accuracy measure, namely one that takes into account the total number of errors within one particular word or pseudoword (cf., infra).

The results of the qualitative analyses substantially strengthened the orthographic depth hypothesis with respect to the nature of reading in Dutch. Children seemed to be mainly relying on sublexical processing strategies to read unrelated real words, as they made much more non-word errors than real word or non-response errors when reading unrelated real words of differ-

ent word length. To read real words embedded in a story, children seemed to rely more strongly on lexical strategies; because of the presence of context, they made much more real word errors than non-word errors. As expressed by their small proportion of non-response errors, children also had greater confidence in trying to read difficult words when these were embedded in a story context. At the same time, the results clearly showed that, although the presence or absence of context could be associated with a particular predominant reading strategy, this never was an all-or-nothing phenomenon. Children combined both lexical and sublexical strategies when reading unrelated real words or real words in a story as indicated by their oral reading errors. Additionally, the results of the qualitative analyses provided new insights on the association between strategies used during reading and word length, relative text difficulty, grade, word decoding level, and gender. More specifically, the results showed that when real words became longer, children increasingly relied on sublexical strategies to read these words. At the same time, when real words became longer or the relative difficulty of stories increased, children became less confident in trying to read difficult words. The results also indicated that there was a developmental shift in the use of sublexical and lexical strategies to read words: the older children became, the more they relied on lexical strategies to read both unrelated real words and real words in a story. Also, the lower a child's decoding skills, the more he/she relied on lexical strategies to decode an unknown word during story reading. Finally, the results showed that girls were less likely than boys to use a sublexical strategy to read an unfamiliar word during unrelated real word reading.

As also discussed in the general introduction of this dissertation, the successful implementation of naturally sounding, highly intelligible, and phonetically correct computer speech into a computerized reading tutor enables this tutor to give oral instructions to the child, to encourage the child, to read to or read along with the child, and to provide different kinds of individualized speech feedback and decoding assistance when an error was made or help was asked for. It has also been demonstrated by various researchers that computerized reading intervention with speech feedback fosters word recognition skills and text reading fluency of novice readers and readers with reading difficulties (e.g., Davidson, Elcock, & Noyes, 1996; Elbro et al., 1996; Jiménez et al., 2007; Lovett, Barron, Forbes, Cuksts, & Steinbach, 1994; Re-

itsma, 1988; Spaai et al., 1991; van Daal & Reitsma, 2000; van Daal & van der Leij, 1992; Wise & Olson, 1998; Wise et al., 2000). Importantly, the use of synthesized speech (as opposed to digitized speech) ensures a program to be 'open' so that the reading tutor's users are not restricted to the reading material chosen and implemented by the authors of the reading tutor.

To evaluate a preliminary version of the computerized reading tutor with synthesized speech feedback (developed within the SPACE project), an intervention study was carried out (Cleuren, Latacz, et al., 2008). More specifically, the study was designed to investigate whether substantial improvements in reading accuracy and reading speed could be obtained in both average performing elementary school children and elementary school children with reading difficulties who worked with the reading tutor, and to investigate whether and how children's initial reading performance level had an influence on the reading tutor's intervention efficacy. At the same time, interest went out to the quality of the synthesized speech, and to children's reactions on the reading tutor's voice. To answer these questions, a randomized pretest-posttest matched-group design was used in which each child of a matched pair was randomly assigned to an intervention or control group. The intervention group received individualized intervention sessions that focussed on non-repetitive instructional-level text reading with phoneme-by-phoneme feedback when an error was made or help was asked for.

Based on the different opinions of the examiners involved in the study, and based on the lack of negative reaction with respect to this matter from the children that were trained with the reading tutor, it could be concluded that the speech synthesis sounded naturally, was clearly audible and well understandable. Children got quickly used to the reading tutor's voice and were very understanding with respect to the reading tutor's occasional mistakes. With respect to the reading tutor's effectiveness, despite a very limited amount of training for the intervention group, the data demonstrated a beneficial influence of extra-curriculum practice with the reading tutor on the reading performance of both average performing elementary school children and elementary school children with reading difficulties, mainly with respect to unrelated real word reading and pseudoword reading. Immediately after the intervention period, reading accuracy and speed gains could be seen with respect to unrelated real word reading; four weeks after intervention, gains could be seen for both unrelated real word and pseudoword reading.

6.2 Methodological Findings

Apart from the main practical realizations that were strived after, and apart from the theoretical findings that were explicitly searched for, the present dissertation also resulted in two interesting methodological findings.

First, as pointed out by Cucchiarini (1996) and repeated by Oller & Ramsdell (2006), accurately annotated speech recordings are not only of interest for the development and testing of speech technology applications; they are also employed in various fields of linguistics, including phonetics, phonology, dialectology, sociolinguistics, second-language teaching, speech pathology, and in studies of language development and language disorders. Moreover, as has been demonstrated in the present dissertation, a corpus of accurately annotated children's reading can be of value for reading researchers too. To assess the quality of such corpus annotations, different methods have been proposed in the literature, such as pairwise percentage agreement and the unweighted kappa statistic. Although the use of percentage agreement scores has been widespread, one of the major drawbacks of this agreement index is that it can be inflated by agreement by chance (Cucchiarini, 1996). In general, percentage agreement decreases as the number of categories increases. In contrast, the unweighted kappa statistic is a measure of agreement that does correct for agreement by chance (Cohen, 1960).

As both percentage agreement scores and unweighted kappa values were calculated to assess annotation agreement with respect to the CHOREC annotations, it could be confirmed that the sole interpretation of percentage agreement scores would have resulted in wrong conclusions with respect to the influence of degree of annotation detail on annotation agreement (Cleuren, Duchateau, et al., 2008). More specifically, the results showed that percentage agreement diminished when the annotation became more detailed: scores were always higher for the detection of reading errors than for orthographic transcriptions, and percentage agreement scores for the latter were always higher than for phonetic transcriptions. This was exactly the opposite as seen for the kappa values, and could be explained by the fact that, when calculating percentage agreement scores, there was no correction for agreement by chance.

Second, in the paper-and-pencil word or pseudoword reading assessment instruments typically being used in Flanders, a child is given a fixed amount of time and is asked to read a list of words or pseudowords as quickly and

as clearly as possible (e.g., Billiaert & Grysolle, 1998; Brus & Voeten, 1995; van den Bos et al., 1994; Verhoeven, 1995). Afterwards, the total amount of correctly read (pseudo)words is calculated. In case of a story reading test, the child is asked to read a story as quickly and as clearly as possible; afterwards, the time needed to complete the story and the total amount of errors made is calculated (e.g., Visser et al., 1998). In sum, to assess a child's reading accuracy, a score is calculated on the basis of a simple wrong versus correct dichotomy.

However, in an orthographically fairly transparent language such as Dutch, a more fine-grained accuracy measure can be argued for (Cleuren et al., 2009a). Because of the consistent mappings between graphemes and phonemes, children often rely on sublexical processing strategies to read difficult unrelated real words or pseudowords. Reading errors are the result of a left-to-right parse of a grapheme sequence which has failed because of an incomplete knowledge of grapheme-to-phoneme correspondences, or because of difficulties with pronunciation assemblage. As a result, more than one oral reading error can be made within a particular word or pseudoword, since the decoding process can fail in each of the separate grapheme-to-phoneme conversions that must be made to read that (pseudo)word. An accuracy measure that takes into account every single reading error made within a (pseudo)word (and not simply counts the number of erroneously read (pseudo)words) would much better capture a child's reading performance. Moreover, with respect to the quantitative analysis of oral reading errors carried out in the present dissertation, the use of a more fine-grained accuracy measure more likely would have led to significant differences between subgroups of children on the basis of (reading) age, reading level, and gender. However, future research is needed to verify this hypothesis.

6.3 Practical Relevance

The research presented in this dissertation also provided food for thought about two topics that are of practical relevance to reading assessment.

To start, the annotation of a word assessed as being read correctly versus a word assessed as being read incorrectly seems very straightforward. Consequently, an analysis of annotator agreement with respect to this classification problem in CHOREC (Cleuren, Duchateau, et al., 2008) resulted in an average kappa value of 0.80 which indicated an almost perfect agreement

between annotators (cf., Landis & Koch, 1977). However, at the same time, this kappa value also clearly indicated that, in some cases, reading accuracy judgments made by human judges do not agree. In other words, even experienced human judges using exactly the same assessment protocol do not show a completely perfect agreement with respect to their correct-incorrect judgements at the level of the whole (pseudo)word when asked to independently assess a particular child's reading performance on a particular paper-and-pencil oral reading task. Based on an investigation of the most common confusions between phonemes, it seemed that disagreements in CHOREC were probably largely attributable to the fact that the different annotators came from different Flemish regions. Because of this, their dialects differed and so did their judgements of vowel length and type. These results clearly emphasized the fact that examiner bias in the assessment of word decoding skills is a phenomenon to be aware of. On top of differences in familiarity with the child, or differences in the choice to do the assessment 'on line' or afterwards, examiners' different regional backgrounds seem to act as a source of serious examiner bias too. As such, these results underline the fact that an objective assessment of reading is in great demand.

Second, as much as being an important methodological finding for reading researchers, the use of a more fine-grained accuracy measure for readers of an orthographically fairly transparent language could be of interest for teachers and clinicians too. As these readers often rely on sublexical processing strategies to read difficult unrelated real words or pseudowords, valuable information on a child's reading accuracy would be obtained by counting the total number of oral reading errors within a particular word or pseudoword, since the decoding process could have failed in each of the separate grapheme-to-phoneme conversions that must be made to read that (pseudo)word. In contrast, a mere count of the number of erroneously read (pseudo)words would not provide information on how 'badly' the erroneously read (pseudo)words had been read. However, as conceded earlier, future research is needed to verify whether such a more fine-grained accuracy measure would indeed better differentiate between readers of different reading ability. Moreover, it must be acknowledged that such a measure would make the reading assessment process even more time-consuming, so that the development of computerized tools that can automate this process would become even more desirable.

6.4 Limitations

The oral reading error research presented in this dissertation has some limitations. A first limitation pertains to the fact that, in the first place, CHOREC was developed to be used in the development and testing of speech technology applications. Because of pragmatic considerations, the participating schools were selected to be within easy reach of the researchers involved in the development of the corpus. As a result, the procedure used for the sampling of the children did probably not result in a representative subgroup of the whole population of elementary school children. Therefore, the results of the quantitative and qualitative oral reading error analyses should be interpreted with some caution, and cannot simply be generalized to all Flemish elementary school children. Moreover, this sampling procedure resulted in an unbalanced group design which strengthened the importance of interpreting results with caution even more.

A second limitation pertains to the fact that there was a substantial amount of missing data. Data could be missing because of four reasons: (1) because of a bad recording, a particular test could not be annotated afterwards, (2) because of time restrictions, a particular test was not assessed, (3) because of a child's limited reading ability, a particular test was not assessed to prevent he/she from getting frustrated, and (4) because of the fact that a child's mastery and/or instructional story reading level was below AVI 1, or because his/her instructional and/or frustrational level was above AVI 9, a particular test could not be assessed. With respect to the qualitative analyses of oral reading errors, because of this substantial amount of missing data, 'task x child' or 'child x child' interaction effects were not discussed, as it was difficult to disentangle the effect of missing data within a particular subgroup from the effect of the characteristic on which the subgroup classification was based. In other words, on top of the fact that there was an unbalanced group design, it was never clear whether an interaction effect was really due to a particular combination of child and/or task characteristics or a result of the fact that the distribution of other child and/or task characteristics was not balanced due to missing data.

A third limitation pertains to the way a child's mastery, instructional, and frustrational level of text reading was operationalised. More specifically, a child's instructional respectively frustrational text reading level were defined as being one respectively two AVI levels higher than his/her mastery level of

text reading as assessed by the AVI test (Visser et al., 1998). However, it can be doubted whether the relation between the relative difficulty of a text for a particular child and the AVI level of a text (as determined by several word and text features), is that straightforward. Moreover, children could not obtain a mastery level of text reading above AVI 9, since this was the highest text reading level that was assessed. With respect to the quantitative analyses of oral reading errors, the lack of significant accuracy differences between stories at the mastery and instructional level could possibly be a side-effect of these missing data. With respect to the qualitative analysis of oral reading errors, it probably prevented us from finding any association between relative text difficulty and non-word or real word error proportions at the word level.

The intervention study to evaluate a computerized reading tutor presented in the present dissertation also had a major limitation which pertains to the fact that the study's sample size was small and inevitably affected the power of the study. Yet, this is the natural consequence of the fact that an intervention study asks for a highly intensive and time-consuming design. Moreover, given the small sample size, the age range is very large, making it difficult to disentangle the effects of (reading) age and reading development on intervention efficacy. The heterogeneity of the group is even larger because of the fact that pretest word reading and pseudoword reading skills do not seem to be below average for all children. Taken all this together, conclusions regarding the effect of reading age, reading development or reading ability on intervention efficacy must be handled with care.

6.5 Future Work

With this dissertation, we wanted to contribute to the realization of using automatic speech recognition and synthesis technology in the context of reading assessment and instruction. Although much work has already been done to achieve these ambitious goals, there is further progress to be made to come to a fully automated, speech technology based assessment and intervention of reading.

First, automatic speech recognition has been successfully used within the context of reading assessment and instruction to track a child's reading so that the child's progress through a reading task could be monitored, and feedback for both the child and the speech synthesizer could be generated on the basis of information provided by the tracker. Moreover, research has

demonstrated that a speech recognizer is able to fairly accurately classify second grade children into five performance groups (ranging from the best performing 25% to the worst performing 10% of children), and this with substantial agreement with a human classification (Duchateau et al., 2007). Children could automatically be classified either correctly or in an adjoining performance group on the basis of their unrelated real word or pseudoword reading performance, so that a follow-up by a human teacher or clinician was only needed for the 25% automatically selected worst performing children. Stimulated by these promising results, future work is needed to implement this technology into a web application so that schools al over Flanders can make use of this time-saving reading assessment method.

Second, although successfully being used to classify children into performance groups, the detection of reading errors at the individual word level by means of automatic speech recognition is not yet accurate enough to obtain an accurate score for a particular child on a particular task (Duchateau et al., 2007). To achieve this, the automatic reading error detection system should be further improved, for example by means of a further optimalization of the speech recognizer's language model. As has been mentioned in Chapter two, CHOREC also provides data on oral reading errors at the grapheme level at which four main categories of reading errors can be distinguished: sequential errors, substitution errors, deletion errors, and insertion errors (cf., Appendix C). An analysis of such errors at the subword level would provide information that can be used to further refine the speech recognizer's language model.

Third, apart from trying to further improve the automatic reading error detection system by means of a further optimalization of the speech recognizer's language model, the possibilities of automatic speech recognition to capture accurate timing measures must be explored too. Through the use of automatic speech recognition in the assessment of reading, new timing measures - that could not (easily) be collected through traditional paper-and-pencil or even computerized tests - can become available. Beck, Jia, & Mostow (2004), for instance, successfully used 'inter-word latency', which is the time that elapses between the reading of successive words in a text assessed by an automatic speech recognizer, as an automated measure of a student's reading ability. Likewise, the use of word latency and word production times (or a particular combination of these measures) as automated estimates of a child's reading performance need to be explored. Moreover, it must be explored whether these measures can also be used for a qualita-

tive evaluation of a particular child's strategies used during reading (cf., van der Schoot et al., 2004; van der Schoot, Licht, Horsley, & Sergeant, 2002). It could be hypothesized that slow externalized unit-by-unit decoding can be characterized by short latency times and long production times, whereas slow internalized unit-by-unit-decoding can be characterized by long latency times and short production times. Direct word recognition can likely be characterized by short latency times ánd short production times.

Fourth, in a sense, a speech recognizer that is trained on a non-dialect-specific corpus will suffer from some type of examiner bias too. This is due to the fact that both children participating in the making of such a corpus and collaborators in charge of such a corpus' annotations usually come from different regions. As a result, children have different dialect influences which are not captured by dialect-specific corpora and, consequently, are not captured by dialect-specific acoustic models. Instead, on the basis of non-dialect-specific data, only a general acoustic model can be built which takes into account as much variation as possible but is not able to take full allowance for children's dialect influences. Moreover, when annotators come from different regions too, and when their regional backgrounds are not matched to those of the participating children, annotators are also biased when making broad-phonetic transcriptions of what the child has actually read. Future research should therefore also focus on the development of dialect-specific children's speech corpora in which the annotations have been made by annotators having the same dialect influences.

Fifth, a whole new test battery to assess children's unrelated real word reading (RWRT), pseudoword reading (PWRT), and story reading performance (SRT) has been developed to be used in the making of CHOREC and to be used in a speech recognition based computerized assessment tool. However, before the widespread use of the RWRT, PWRT and SRT in a speech recognition based assessment tool can be recommended, it is necessary to obtain normative data for these instruments and to register whether these instruments are reliable and valid. Future research should address these issues.

Sixth, automatic speech synthesis within the context of reading instruction has been successfully used to provide phoneme-by-phoneme feedback whenever an error was made or help was asked for. However, two considerations for future improvements, which were not discussed in Chapter five, should be taken into account (cf., Duchateau et al., 2009). The first con-

sideration has to do with progressive and regressive assimilation processes in Dutch. These processes imply that a phoneme takes over characteristics of a preceding (progressive assimilation) or following (regressive assimilation) phoneme in a word or sentence. To obtain correctly pronounced whole-word and syllable-by-syllable feedback, it is important that these processes are accounted for by the speech synthesizer. However, when synthesizing phoneme-by-phoneme feedback, these processes should be neglected and phonemes should be produced as if they were standing alone. For example: in the Dutch word 'zakdoek' [*handkerchief*], the phoneme /k/ before the /d/ becomes voiced (/zAgduk/). However, when synthesising phoneme-by-phoneme feedback for this word; that phoneme needs to remain voiceless (/z/-/A/-/k/-/d/-/u/-/k/). The second consideration concerns the more specific phenomenon where a voiced phoneme in a word becomes voiceless at the end of the word when followed by a pause. When synthesising phoneme-by-phoneme feedback for such a word, the phoneme at the end of the word needs to remain voiceless. For example: in the Dutch word 'hond' [*dog*], the voiced phoneme /d/ becomes voiceless because it is placed at the end of the word. Thus, the word 'hond' is pronounced as /hOnt/ and not as /hOnd/; phoneme-by-phoneme feedback should not change this (/h/-/O/-/n/-/t/). Future research on automatic speech synthesis to be used in a reading tutor, should take these considerations into account.

Finally, in the first reading tutor prototype being evaluated in Chapter five, no real automatic speech recognition was used. Instead, by means of a so-called 'Wizard-of-Oz' method, a human researcher replaced the speech recognition module which made children think that the computer was really able to listen to them. Future research needs to be carried out to evaluate a second prototype of the reading tutor in which both automatic speech recognition and speech synthesis will be implemented. Moreover, automatic speech synthesis will also be used to read to or read along with the child, as well as to provide three different kinds of feedback: whole-word-feedback, syllable-by-syllable feedback, and phoneme-by-phoneme feedback.

Appendix A

Test Material

Inspired by the procedure used in the Three-Minutes-Test (Verhoeven, 1995), the Real Word Reading Test (RWRT) and the PseudoWord Reading Test (PWRT) both consisted of three lists of (pseudo)words of different length (i.e., different number of syllables). Inspired by the procedure used in the AVI test (Visser et al., 1998; Visser, 1997), the Story Reading Test (SRT) consisted of 9 stories of increasing technical difficulty level. In what follows, the content of each test as well as the selection of RWRT words and PWRT pseudowords, and the construction of the different SRT stories, will be described (in Dutch).

A.1 Constructie van de Real Word Reading Test (RWRT)

Ons doel bij het construeren van een nieuwe woordenleestest (RWRT) was het construeren van drie verschillende woordenlijsten:

- een lijst met éénlettergrepige (1LG) woorden,
- een lijst met tweelettergrepige (2LG) woorden,
- een lijst met drie- en vierlettergrepige (3+4LG) woorden.

A.1.1 Eénlettergrepige Woorden

Voor het samenstellen van de lijst met éénlettergrepige woorden, zijn we vertrokken van de éénlettergrepige woorden die gebruikt werden in de licen-

tiaatsthesissen van Verachtert (2001) en Schippers & Corens (2000). Dit resulteerde in 38 éénlettergrepige woorden. Voor elk van deze 38 woorden zijn we vervolgens - op basis van de 'Streeflijst woordenschat voor zesjarigen' (Schaerlaekens, Kohnstamm, & Lejaegere, 1999) - nagegaan door hoeveel procent van de Vlaamse én Nederlandse zesjarigen het woord gekend is. Enkel indien het woord door minstens 95% gekend is, werd het weerhouden; een maatregel die het aantal éénlettergrepige woorden terugbracht tot 19.

Omdat we echter in totaal over een lijst van 40 woorden wilden beschikken, was het nodig zelf nog 21 woorden bij te zoeken. De criteria die bij het samenstellen van deze woorden werden gebruikt, waren de volgende:

- het woord moet door minstens 95% van de Vlaamse én Nederlandse zesjarigen gekend zijn;
- de woorden worden zo gekozen dat zoveel mogelijk alle in het Nederlands voorkomende grafemen en grafeemclusters (en de daarmee corresponderende fonemen) in de woordenlijst opgenomen worden (hierdoor werden er ook twee woorden van de oorspronkelijke lijst van 19 woorden vervangen door een ander woord; we zochten nu dus 23 nieuwe woorden);
- de woorden worden zo gekozen dat de verdeling over de mogelijke woordstructuren (km, mk, mkm,...) zo analoog mogelijk is aan de verdeling in leeskaart 1 en 2 van de Drie-Minuten-Toets (Verhoeven, 1995).

Op deze manier bekwamen we een (wat woordstructuur en grafeem-/foneemverdeling betreft) goed uitgebalanceerde lijst met 40 gekende éénlettergrepige woorden. Als gevolg echter van ons strenge criterium dat zegt dat elk woord begrepen moet zijn door 95% van de Vlaamse én Nederlandse zesjarigen, werden volgende grafemen (of combinaties van) en de daarmee gepaard gaande fonemen niet in de lijst opgenomen:

c /k/; c /s/; q /k/; t /s/; x /ks/; y /E+/;
ch /S/; sj /S/; au /A+/; aai /aj/; oei /uj/

Om toch de belangrijkste van deze (combinaties van) grafemen en de daarmee gepaard gaande fonemen te kunnen opnemen, hebben we een aantal van deze 40 woorden (1 van de oorspronkelijke lijst van 19 en 6 van de zelf bijgezochte woorden) vervangen door woorden waarvoor:

- de 95%-grens verlaagd wordt naar 90%;
- toegestaan wordt dat de persoonsvorm van een in de streeflijst opgenomen werkwoord (waar wel nog het 95%-criterium voor geldt), gebruikt wordt.

Grafemen en de daarmee gepaard gaande fonemen die uiteindelijk *niet* in deze woordenlijst werden opgenomen, zijn:

c /s/; t/s/; q /k/; x /ks/; y /E+/

A.1.2 Tweelettergrepige Woorden

Ook voor het samenstellen van een lijst met tweelettergrepige woorden, zijn we vertrokken van de tweelettergrepige woorden die gebruikt werden door Verachtert (2001) en Schippers & Corens (2000). Dit gaf ons een lijst met 40 woorden. Na een analyse van deze woorden aan de hand van de 'Streeflijst woordenschat voor zesjarigen (95%)' (Schaerlaekens et al., 1999), weerhielden we 25 woorden als basis voor onze nieuwe woordenlijst. Een bijkomende opmerking die hierbij moet worden gemaakt, is dat het meervoud van een woord waarvoor de 95%-grens geldt, toegelaten werd.

Om tot 40 woorden te komen, was er dus een eigen aanvulling nodig van 15 woorden waarvoor we de volgende criteria hanteerden:

- het woord moet door minstens 95% van de Vlaamse én Nederlandse zesjarigen gekend zijn;
- de woorden worden zo gekozen dat zoveel mogelijk alle in het Nederlands voorkomende grafemen en grafeemclusters (en de daarmee corresponderende fonemen) in de woordenlijst opgenomen worden.

Als gevolg hiervan kwamen de volgende (combinaties van) grafemen en de daarmee gepaard gaande fonemen niet voor in de woordenlijst:

q /k/; x /ks/; y /E+/; ch /S/; sj /S/; uu /y/

Om toch de belangrijkste hiervan in de woordenlijst te kunnen opnemen, hebben we 2 van de oorspronkelijke 25 woorden vervangen door woorden waarvoor:

- de 95%-grens verlaagd wordt naar 88%.

(Combinaties van) grafemen en de daarmee gepaard gaande fonemen die uiteindelijk *niet* in deze woordenlijst werden opgenomen, zijn:

sj /S/; q /k/; x /ks/; y /E+/

Tot slot werd opgemerkt dat drie tweelettergrepige woorden uit deze lijst het meervoud zijn van drie woorden uit de éénlettergrepige woordenlijst. Het was dus wenselijk deze woorden te vervangen. Als criteria bij deze vervangingen golden:

- het woord moet door minstens 88% van de Vlaamse én Nederlandse zesjarigen gekend zijn;
- bij het vervangen van het woord mag de beoogde grafeem-/foneemspreiding niet in het gedrang komen.

A.1.3 Drie- en Vierlettergrepige Woorden

Als basis van onze lijst met drie- en vierlettergrepige woorden gebruikten we andermaal woorden uit de licentiaatsthesissen van Verachtert (2001) en Schippers & Corens (2000). Het gaf ons een basislijst van 41 woorden. Ook hier werd binnen deze woorden een selectie gemaakt op basis van de 'Streeflijst woordenschat voor zesjarigen' (Schaerlaekens et al., 1999). In plaats van een 95%-grens echter, gebruikten we voor de selectie ditmaal een 93%-grens waardoor er nog een lijst van 17 woorden overbleef. Verder dient hierbij ook opgemerkt te worden dat het meervoud van een woord waarvoor de 93%-grens geldt, toegelaten werd.

Voor de aanvulling tot een lijst van 40 woorden, hielden we rekening met de volgende voorwaarden:

- het woord moet door minstens 93% van de Vlaamse én Nederlandse zesjarigen gekend zijn;
- de woorden worden zo gekozen dat zoveel mogelijk alle in het Nederlands voorkomende grafemen en grafeemclusters (en de daarmee corresponderende fonemen) in de woordenlijst opgenomen worden.

De (combinaties van) grafemen en de daarmee gepaard gaande fonemen die dan in de bekomen lijst van 40 woorden niet werden opgenomen, zijn:

c /s/; q /k/; x /ks/; sj /S/; uu /y/; ei /E+/; ooi /oj/; oei /uj/; ieuw /iw/

Om de belangrijkste hiervan toch in de woordenlijst te kunnen opnemen, werden er een aantal belangrijke uitzonderingen op onze basisvoorwaarde (93%-grens) toegestaan waardoor 6 woorden (4 van de oorspronkelijke lijst en 2 van de zelf bijgezochte aanvulling) vervangen werden:

- de 93%-grens wordt verlaagd naar 87%;
- een woord mag een samengesteld woord zijn van twee gekende woorden (waarbij voor beide woorden de 93%-grens geldt);
- een woord mag het meervoud zijn van een gekend woord (waarvoor de 87%-grens geldt).

(Combinaties van) grafemen en de daarmee gepaard gaande fonemen die uiteindelijk *niet* in deze woordenlijst werden opgenomen, zijn:

sj /S/; q /k/; x /ks/

Tot slot werd opgemerkt dat vier drie- of vierlettergrepige woorden uit deze lijst afgeleid zijn (een verbuiging of een samenstelling) van vier woorden uit de tweelettergrepige woordenlijst. Het was dus wenselijk deze woorden te vervangen. Als criteria bij deze vervangingen golden:

- het woord moet door minstens 87% van de Vlaamse én Nederlandse zesjarigen gekend zijn;
- bij het vervangen van het woord mag de beoogde grafeem-/foneem-spreiding niet in het gedrang komen;
- het woord wordt vervangen door een woord met hetzelfde aantal lettergrepen.

Enkel het vierlettergrepige woord 'autorijden' werd behouden ondanks de aanwezigheid van het woord 'auto' in de lijst van tweelettergrepige woorden. Dit bij gebrek aan een goed 'au'-alternatief voor zowel 'auto' als 'autorijden'.

A.1.4 De Uiteindelijke Woordenlijsten

In Tabel A.1 worden de uiteindelijke woordenlijsten met éénlettergrepige, tweelettergrepige en drie- en vierlettergrepige woorden getoond.

	1LG	**2LG**	**3+4LG**
1	klas	water	omdraaien
2	mooi	mevrouw	poppenwagen
3	koe	drogen	konijnenhok
4	arm	winkel	elastiekje
5	groei	auto	ruziemaken
6	strand	schouders	teddybeer
7	bed	verhaal	dierentuin
8	eerst	koning	paddenstoelen
9	voor	moeilijk	verstoppertje
10	draai	speelplaats	wasmachine
11	sjaal	drinken	fototoestel
12	herfst	hoofdpijn	toiletpapier
13	duur	regen	vrachtwagen
14	straat	vliegtuig	buurmannen
15	leeuw	stoppen	vogelkooi
16	clown	opnieuw	olifant
17	hoek	gooien	schommelen
18	krant	sneeuwen	iedereen
19	hout	moeder	schoenenwinkel
20	vriend	liedje	knutselen
21	gauw	potlood	ophangen
22	chips	fietsbel	verjaardag
23	groen	vinger	sprookjesboek
24	feest	dichtbij	tandenborstel
25	reis	meisje	lucifer
26	jas	chauffeur	slaapkamer
27	huis	muziek	achterdeur
28	paard	waarom	ziekenhuis
29	vijf	scheuren	nieuwsgierig
30	muts	lawaai	afblijven
31	nieuw	zwemmen	kabouter
32	kind	vuurwerk	washandje
33	bang	appel	sneeuwwitje
34	oog	cola	goeiendag
35	zacht	kussen	vakantie
36	schoen	eerste	limonade
37	plas	circus	autorijden
38	neus	kleuren	eindelijk
39	knoop	voetbal	familie
40	plank	vlinder	chocolade

Table A.1: De woordenlijsten

A.2 Constructie van de PseudoWord Reading Test (PWRT)

Net zoals bij de constructie van een nieuwe woordenleestest, was ons doel bij het construeren van een nieuwe pseudowoordenleestest (PWRT) om drie verschillende pseudowoordenlijsten samen te stellen:

- een lijst met éénlettergrepige (1LG) pseudowoorden,
- een lijst met tweelettergrepige (2LG) pseudowoorden,
- een lijst met drie- en vierlettergrepige (3+4LG) pseudowoorden.

A.2.1 Constructie van Drie Pseudowoordenlijsten

Voor de drie pseudowoordenlijsten werden de pseudowoorden die werden gebruikt door Verachtert (2001) en Schippers & Corens (2000), als uitgangspunt genomen. Dit resulteerde voor de éénlettergrepige, de tweelettergrepige en de drie- en vierlettergrepige pseudowoorden in lijsten van respectievelijk 19, 22 en 19 pseudowoorden.

De 19 éénlettergrepige pseudowoorden werden aangepast op basis van het principe waarbij we de grafeem-/foneem-spreiding bij de pseudowoorden ongeveer gelijk wilden houden aan de grafeem-/foneem-spreiding bij de overeenkomstige woordenlijst. (Combinaties van) grafemen en de daarmee gepaard gaande fonemen die niet voorkomen bij de woordenlijsten komen dus ook niet voor bij de pseudowoordenlijsten. 7 van de oorspronkelijke éénlettergrepige pseudowoorden werden daarom aangepast.

Ook bij de tweelettergrepige en drie- en vierlettergrepige pseudowoorden werd een aanpassing uitgevoerd. De reden hiervoor was dat we ervoor gekozen hebben om zoveel mogelijk betekenisvolle woorddelen binnen pseudowoorden te vermijden. Er werden daarom respectievelijk 6 en 11 pseudowoorden aangepast.

Net zoals bij de woordenlijsten, wilden we ook voor de pseudowoorden beschikken over lijsten van 40 pseudowoorden. Er diende dus drie maal een aanvulling te gebeuren. Bij de aanvulling werd rekening gehouden met de volgende principes:

- de spreiding over de verschillende grafemen/fonemen wordt zoveel mogelijk gelijk gehouden aan de grafeem-/foneem-spreiding bij de bestaande

woorden;

- bij de éénlettergrepige pseudowoorden wordt de verdeling over de mogelijke woordstructuren (km, mk, mkm,...) zo analoog mogelijk gemaakt aan de verdeling bij de bestaande éénlettergrepige woorden;
- bij de tweelettergrepige en drie- en vierlettergrepige pseudowoorden wordt er verder op gewaakt dat zoveel mogelijk betekenisvolle woorddelen vermeden worden;
- bij de drie- en vierlettergrepige pseudowoorden wordt er voor gezorgd dat het aantal drielettergrepige en het aantal vierlettergrepige pseudowoorden gelijk wordt gehouden aan de verdeling bij de bestaande drie- en vierlettergrepige woorden.

De (combinaties van) grafemen en de daarmee gepaard gaande fonemen die dan in de bekomen lijsten van 40 pseudowoorden *niet* werden opgenomen, zijn:

- voor de éénlettergrepige pseudowoorden:

  c /k/; c /s/; q /k/; t /s/; x /ks/; y /E+/

- voor de tweelettergrepige pseudowoorden:

  c /s/; q /k/; x /ks/; y /E+/; ch /S/

- voor de drie- en vierlettergrepige pseudowoorden:

  c /k/; c /s/; q /k/; x /ks/; sj /S/

A.2.2 De Uiteindelijke Pseudowoordenlijsten

Aangezien het van belang was dat er niet getwijfeld werd aan de uitspreekbaarheid van een bepaald pseudowoord, werd een groep van 66 universiteitsstudenten gevraagd elk van de pseudowoorden op hun uitspreekbaarheid te beoordelen. Hiervoor beschikten ze per woord over een schaal van 1 tot 6: van 'heel moeilijk uitspreekbaar' tot 'heel goed uitspreekbaar'. Voor elk pseudowoord werd gevraagd om slechts 1 score te omcirkelen.

Pseudowoorden die in aanmerking kwamen voor een aanpassing waren:

- pseudowoorden die door minstens 1 beoordelaar een score '1' krijgen toegewezen;
- pseudowoorden waarvoor minder dan 50% van de beoordelaars minimum een score '4' gegeven heeft.

Bij de aanpassingen werd er voor gezorgd dat de beoogde grafeemspreiding behouden blijft. In Tabel A.2 worden de uiteindelijke pseudowoordenlijsten met éénlettergrepige, tweelettergrepige en drie- en vierlettergrepige pseudowoorden getoond.

A.3 Constructie van de Story Reading Test (SRT)

Bij de constructie van een nieuwe verhalenleestest (SRT) wilden we 9 teksten maken die overeenkomen met de AVI-niveaus 1 tot en met 9. Teksten van een verschillend AVI-niveau onderscheiden zich van elkaar op het vlak van woord-, zins- en tekstkenmerken en de leesindex A.

$$A = 195 - \left(\tfrac{2}{3}\ 100\ \tfrac{aantal\ lettergrepen}{aantal\ woorden}\right) - \left(2\ \tfrac{aantal\ woorden}{aantal\ zinnen}\right)$$

Verder hielden we bij de constructie van de nieuwe teksten ook rekening met de lengte van de AVI-verhalen van de bestaande AVI-toets (Visser et al., 1998): het aantal woorden en het aantal paragrafen werd zoveel mogelijk benaderd. Tot slot werd - zoals in de oorspronkelijke AVI-toets (Visser et al., 1998) - bij de eerste vier AVI-teksten (AVI 1-4) voor een begeleidende tekening gezorgd om het anticiperend lezen te bevorderen.

A.3.1 AVI 1

Tekstkenmerken:

- korte zinnen, één zin per regel;
- samengestelde zinnen verdeeld over twee regels komen voor (voor leesindex geteld als twee zinnen);
- hoofdletters kunnen voorkomen;
- leesindex A $>$ 100.

	1LG	**2LG**	**3+4LG**
1	fook	millen	ometuif
2	sooi	vocil	toejietsen
3	dau	geutel	oonwijlen
4	ant	sjobel	jattesiet
5	beeg	ierpieuw	nurudien
6	sprunt	walaan	stoenydaas
7	hool	erke	deuveltek
8	larst	haweel	juitonie
9	vout	saarweng	gevijdel
10	zwoei	gevicht	sidowaan
11	fam	eemde	spekkeraai
12	rachts	bepoud	wachteniek
13	vaap	orstalk	verpierik
14	sprieuw	veten	nappegreeuw
15	keng	gefouw	mantaroen
16	swoers	vurpaand	schielendaspen
17	doer	nizung	crobeklunker
18	plirt	fiewon	kabbestepen
19	jien	kneurem	verwarig
20	blard	vawaai	ooiebiekje
21	guul	strellen	fandelig
22	hoekt	zwieten	jalekrewen
23	neeuw	foetbans	smoralij
24	noork	oonste	zeekvlachine
25	vid	muider	kanaroe
26	zans	grijnken	toineetlijgen
27	leum	schielstaug	meitsegrok
28	haans	prilsood	kantelogsten
29	spaai	vloender	ondermind
30	sjalt	milste	choporatie
31	heik	veurder	zennebral
32	sank	kloeien	ijraspangen
33	roen	ulen	blottenduuf
34	frijk	orponk	girdofhaalder
35	eem	schodig	tobbermoeit
36	schard	ijpo	poentalschouden
37	grek	menuur	havedil
38	dron	spreikje	verbrakkertje
39	snaaf	hiffreeuw	gerauwejaak
40	stuid	wooien	hapeneren

Table A.2: De pseudowoordenlijsten

Woordtypen:

- éénlettergrepige woorden: mk, km, mkm;
- éénlettergrepige woorden met één medeklinkercombinatie kunnen voorkomen;
- de volgende woorden kunnen niet:
  - eindigend op -b;
  - eindigend op -nk of -ng;
  - eindigend op -dt;
  - eindigend op -ch of -cht;
  - eindigend op -eeuw of -ieuw;
  - eindigend op -auw of -ouw;
- de volgende woorden kunnen wel:
  - beginnende met een sch-;
  - met -aai of -ooi;
  - eindigend op -rd, -nd of -ld.

Layout:
Arial 23, interlinie 1, geen hoofdletters.

A.3.2 AVI 2

Tekstkenmerken:

- hoofdletters kunnen voorkomen;
- korte zinnen, één zin per regel;
- samengestelde zinnen worden nog per zin over twee regels verdeeld (voor de leesindex geteld als twee zinnen);
- leesindex A $>$ 100.

Woordtypen:

feest voor jan

nog een dag of vier.
dan is jan zes jaar.
ik geef je iets, zegt lies.
weet je al wat je wilt?

dat weet jan wel.
een bal of een beer.
een pen of een boek.
of een vis in een kom.

wat een lijst, roept lies.
ik kies wel iets.
en steek het in een doos, zegt ze.

nu is het feest.
jan is zes jaar.
hij heeft een kroon.
en er is taart en snoep.

lies geeft de doos aan jan.
geen bal of beer, geen pen of boek.
maar een kom met een vis.
zes jaar zijn, is heel fijn.

Table A.3: AVI1-tekst, A=117

- éénlettergrepige woorden:
  - nu ook eindigend op -b, -ch, -cht, dt, ng of nk;
  - twee medeklinkers vooraan en/of achteraan het woord;
  - drie medeklinkers vooraan of achteraan het woord;
  - de auw-, ouw-, eeuw- en ieuw-klank;
- tweelettergrepige of samengestelde woorden zonder spellingsmoeilijkheden;
- verkleinwoorden.

waar is de schoen?

het brood is op.
de kaas is ook bijna op.
mama zegt: ik ga naar de winkel.
ik haal een brood bij de bakker.
en daarna koop ik kaas.
wie gaat er mee?

ik ga mee, roept tom.
en ik haal poes, zegt an.
want poes gaat graag naar buiten.
mama stopt haar in een mand.

an staat al klaar en wacht.
ze heeft haar jas aan.
maar waar is tom naar toe?
tom zoekt zijn schoen.
waar is die schoen toch?

ik tel tot tien, roept mama.
tom zoekt en zoekt maar door.
waar blijf je? vraagt mama.
de schoen ligt onder zijn bed.
snel loopt hij de gang in.
oef, net op tijd.

Table A.4: AVI2-tekst, A=112

Layout:
Arial 23, interlinie 1, geen hoofdletters.

A.3.3 AVI 3

Tekstkenmerken:

- korte zinnen kunnen betekenisvol worden afgebroken en doorlopen op de volgende regel (de nieuwe zin begint wel nog op een nieuwe regel);
- leesindex A $>$ 100;

- voor de oorspronkelijke AVI3-teksten (3A en 3B) is de gemiddelde zinslengte (gZL): 6,6 - 6,8, maar we streven een gemiddelde zinslengte na van 4-6 woorden;
- gemiddelde woordlengte (gWL) < 1,2 lettergrepen.

Woordtypen:

- alle typen éénlettergrepige woorden;
- alle typen tweelettergrepige woorden;
- drie- en meerlettergrepige woorden zonder spellingsmoeilijkheden.

Layout:
Arial 23, interlinie 1, hoofdletters.

A.3.4 AVI 4

Tekstkenmerken:

- zinnen kunnen, betekenisvol afgebroken, doorlopen op de volgende regel;
- een gesproken zin begint nog steeds op een nieuwe regel;
- leesindex A > 100 (maar naar analogie met de oorspronkelijke AVI4-teksten (100/99) kan 'A = 100' ook) ;
- gemiddelde zinslengte (gZL): 5-7 woorden;
- gemiddelde woordlengte (gWL) < 1,2 lettergrepen (maar naar analogie met de oorspronkelijke AVI4-teksten kan = 1,2 lettergrepen ook).

Woordtypen:

- alle typen drie- en meerlettergrepige woorden;
- leenwoorden zonder lastige, afwijkende teken-klankcombinaties.

Layout:
Arial 19, interlinie 1, hoofdletters.

Goed nieuws

Els en Bert spelen tikkertje.
Ik krijg je wel, roept Bert.
Maar Els loopt de gang in.
Pak me dan, als je kan!
Bert rent snel achter haar aan.

Stop, hijgt Els, er is post.
Er ligt een brief in de bus.
Van wie is die? vraagt ze zich af.
Ook Bert denkt heel diep na.

Samen maken ze de brief open.
Er zit een grote, roze kaart in.
De kaart brengt goed nieuws.
Tante Hilde kreeg een kindje, een meisje!
Papa leest de kaart ook.

Wil je het kindje zien? vraag papa.
Ja, roepen Bert en Els.
Ze dansen rond de tafel.
Ze doen heel gek en wild.
En dan knutselen we een pakje, zegt papa.
Dat vinden ze een leuk plan.

Table A.5: AVI3-tekst, A=103, gWL=1.195, gZL=6.15

Lekke band

Het is zondag en Joris verveelt zich.
Dan hoort hij de bel.
Joris loopt naar de deur en doet open.

Er staat een meisje op de stoep.
Kan iemand een handje helpen? vraagt ze.
Mijn auto heeft een lekke band.

Ook moeder en vader komen nu kijken.
Vader wil de band wel vervangen.
Kijk maar goed, Joris, zegt hij stoer.
Eerst neemt hij de wieldop weg.
Dan haalt hij een krik uit de koffer.
Zo kan hij één kant van de auto optillen.
De lekke band schroeft hij los.
Dan kan de nieuwe band erop.
Ten slotte gaat de dop weer op het wiel.

De handen van vader zijn erg zwart.
En op zijn wang zit een veeg.
Maar dat is niet erg.
Joris kijkt hem vol trots aan.
En van moeder krijgt hij een zoen.
Blij rijdt het meisje weer naar huis.
Zo was het toch geen saaie dag!

Table A.6: AVI4-tekst, A=100, gWL=1.23, gZL=6,82

A.3.5 AVI 5

Tekstkenmerken:

- leesindex A: 99-94;
- gemiddelde zinslengte (gZL): 7 woorden (6-8 woorden);
- gemiddelde woordlengte (gWL): 1,27 lettergrepen (1,23-1,30).

Layout:
Arial 19, interlinie 1, hoofdletters.

A.3.6 AVI 6

Tekstkenmerken:

- leesindex A: 93-89;
- gemiddelde zinslengte (gZL): 8 woorden (7-9 woorden);
- gemiddelde woordlengte (gWL): 1,32 lettergrepen (1,29-1,35).

Layout:
Arial 18, interlinie 1, hoofdletters.

A.3.7 AVI 7

Tekstkenmerken:

- leesindex A: 88-84;
- gemiddelde zinslengte (gZL): 9 woorden (8-10);
- gemiddelde woordlengte (gWL): 1,36 lettergrepen (1,34-1,39).

Layout:
Arial 15, interlinie 1, hoofdletters, doorlopende tekst.

Verliefd

Tim is onze buurman.
Hij houdt van grapjes en is altijd vrolijk.
Wanneer we voetballen, doet hij graag mee.

Maar nu is hij verliefd.
Nu heeft hij geen zin meer in spelen.
Hij denkt alleen maar aan haar
en loopt met zijn hoofd in de wolken.
Ik denk dat dat normaal is,
als je pas verliefd bent.
En dat het een heel fijn gevoel is.

We stellen hem wel honderd vragen.
Ben je al met haar naar de film geweest?
Is ze knap en speelt ze graag voetbal?
Wanneer zien we haar een keer?

Haar naam is Liesbet, vertelt hij.
Ze is heel sportief, maar ze kan niet zwemmen.
Ze speelt wel goed basketbal.
En net als ik, houdt ze van muziek.

Tim vertelt dat ze niet kan lopen, ze zit in een rolstoel.
Haar twee benen zijn verlamd.
De bus nemen of ergens aanbellen,
is voor haar niet gemakkelijk.

Dan verklapt hij dat ze snel zullen trouwen.
En dat Liesbet onze buurvrouw wordt.
Dan zullen we haar vaak genoeg kunnen zien!

Table A.7: AVI5-tekst, A=94, gWL=1.29, gZL=7.82

Het winnende gedicht

Er is een cameraploeg op school vandaag.
Op de speelplaats staat een grote vrachtwagen.
Dat komt omdat het de week van de poëzie is.
Elk kind uit onze klas schreef daarom een gedicht.
En er is één gedicht dat wint.

In onze klas zijn twee mannen druk bezig.
Er staat een grote lamp en de vloer ligt vol met kabels.
Er komt nog iemand binnen met een luidspreker.
Hij zet hem op het bureau van de juf.
Dan gaat de lamp aan.
De reporter vertelt over de week van de poëzie.
Hij zegt dat één gedicht uitkiezen moeilijk was.

Toch is er slechts één winnaar.
Het gedicht dat wint, is dat van Anneleen.
De hele klas klapt voor haar.
Ze wordt een beetje rood.
Voor de camera leest ze haar gedicht aan ons voor.
Wanneer ze klaar is, klappen we opnieuw.
Iedereen staat om haar heen.

Die avond kijken we naar het nieuws.
We luisteren hoe Anneleen het gedicht voorleest.
Ze spreekt heel mooi en haar hoofd lijkt niet zo rood.
Wat zijn we trots op haar!

Table A.8: AVI6-tekst, A=90, gWL=1.32, gZL=7.70

Het leven bij de indianen

Vroeger leefden er heel veel soorten indianen. Die woonden in het bos, de bergen of in de woestijn. Soms waren het nomaden. Dat zijn mensen die geen vaste plaats hadden om te wonen. Ze moesten hun kamp steeds verplaatsen. Dat deden ze om de bizons te volgen. Die hadden ze nodig om te kunnen leven.

Voor ons is lezen en schrijven heel belangrijk. Maar voor de indianen was dat niet zo. Voor hen moest een jongen een goede jager en krijger worden. En een meisje moest goed kunnen koken. En dat leerden ze niet op school.

De vader, de grootvaders en de ooms vertelden de kleine jongen alles over de jacht en de oorlog. Ze toonden hem hoe je het spoor van een dier volgt en hoe je een boog opspant.

Meisjes deden niet mee aan de jacht of de oorlog. Ze leerden koken en het huishouden doen. Dat zou later immers hun taak zijn. Daarom hielpen ze hun moeder al heel jong met fruit plukken en hout zoeken. Ze zorgden ook vaak voor hun kleine broertjes en zusjes. Zo maakten ze zich klaar voor hun rol als moeder.

Had jij graag als indiaan geboren willen zijn? Of zou je je klas en de school dan missen?

Table A.9: AVI7-tekst, A=84, gWL=1.39, gZL=9.41

De maan

De maan is net zo oud als de aarde. Dat betekent dat ze vijf miljard jaar oud is. Net als de aarde is zij ontstaan uit gas en stof. Daaruit zijn ook de zon en alle planeten ontstaan.

De maan is het hemellichaam dat zich het dichtst bij de aarde bevindt. Sommige wetenschappers denken dat ze van de aarde is losgerukt. Eigenlijk weten we het niet zeker. Het is in elk geval wel zo dat de maan in de buurt van de aarde is ontstaan. Dat blijkt bijvoorbeeld uit bepaalde stoffen die ook op de aarde voorkomen.

Wanneer je 's avonds naar de maan kijkt, lijkt het net of je een gezicht kunt zien. Dat komt door de donkere vlekken op de maan. Die noemen we kraters. Dat zijn plekken waar lang geleden door de ruimte zwevende brokken zijn ingeslagen. Met een verrekijker kan je ze nauwkeurig bekijken.

Op de maan is er geen lucht. De ruimtevaarders die er rondlopen, moeten dus ruimtepakken aanhebben. Daarin zit de lucht die ze nodig hebben om te ademen. Ook moet het pak hen beschermen tegen de zon. Die kan in de ruimte fel schijnen omdat er geen lucht en wolken zijn. Het kan tot meer dan honderd graden warm worden op de maan!

Table A.10: AVI8-tekst, A=80, gWL=1.43, gZL=10.50

A.3.8 AVI 8

Tekstkenmerken:

- leesindex A: 83-79;
- gemiddelde zinslengte (gZL): 10 woorden (9-11);
- gemiddelde woordlengte (gWL): 1,41 lettergrepen (1,38-1,44).

Layout:
Arial 14, interlinie 1, hoofdletters, doorlopende tekst, aanhalingstekens.

De postduif

Mijn opa heeft wel twintig duiven. Hij geeft ze allemaal een naam. Ze wonen in een hok in de tuin. En ze krijgen elke dag granen, zaden en peulvruchten te eten. Die hebben ze nodig om voldoende energie te hebben en om de temperatuur van hun lichaam op peil te kunnen houden.

Opa vertelde me dat er veel soorten duiven bestaan. Eén van hen is de postduif. Dat is een duif die een heel grote afstand in één keer kan afleggen en gemakkelijk de weg naar huis terugvindt. En daar heeft die duif geen kompas of landkaart voor nodig. Hij weet de weg door naar de zon te kijken!

Daarom werden deze duiven in de tijd van de oude Romeinen en Grieken gebruikt om berichten te versturen. Toen had men immers nog geen auto of vliegtuig om post mee te vervoeren. De duiven droegen dan kleine berichtjes met zich mee in een kokertje aan hun poot.

Nu zorgt de postbode dat je brief in de juiste bus terechtkomt. Toch zijn veel mensen in Nederland en België lid van een postduivenvereniging. We noemen ze duivenmelkers. Maar ze gebruiken hun duiven alleen nog voor de sport. De duiven worden goed getraind om te kunnen meedoen aan wedstrijden. Als je duif als eerste thuiskomt, dan heb je gewonnen. En dan win je geld of een andere leuke prijs!

Table A.11: AVI9-tekst, A=74, gWL=1.48, gZL=11.25

A.3.9 AVI 9

Tekstkenmerken:

- leesindex A: 78-74 (hoewel de oorspronkelijke AVI9A-tekst een leesindex heeft van 72);
- gemiddelde zinslengte (gZL): 11 woorden (10-12);
- gemiddelde woordlengte (gWL): 1,46 lettergrepen (1,43-1,48).

Layout:
Arial 13, interlinie 1, hoofdletters, doorlopende tekst, aanhalingstekens.

Appendix B

Reading Strategy Labels

B.1 1 Attempt

- A word was read correctly during a first attempt: g or 'nothing' (segment is left empty)
- A word was read correctly but vowels were prolonged: vkg
- A word was read correctly but vowels were prolonged to such an extent that it was no longer regarded as correct: vkf
- A word was read wrongly during a first attempt: f
- A word was skipped: O
- Spelling behaviour that is not clearly audible occurs: *s
- An examiner read the word out loud: Vv
- The word was partly read out loud by an examiner: Vg

Remark: Vv and Vg are added to codes belonging to multiple attempts at reading a word. If no attempt was made, then Vv and Vg are used by themselves. If these codes occur, reading mistakes are no longer indicated. Vv is not written down if an examiner utterance serves to make the speech recogniser continue.

B.2 More Than 1 Attempt for 1 Word

- Last attempt:
  - If read correctly, the last attempt receives the coding g or vkg (cf., previous paragraph)
  - If read wrongly, the last attempt receives the coding vkf, f, O, *s, Vv or Vg (cf., previous paragraph)
    Exception: If a word was spelled out but NO attempt at synthesis occurred, the last attempt receives the coding lsg, lsf, ssg, ssf, *s, alg or alf (cf., following paragraph). Spelling out will be coded when a child spells out audibly, even he/she is whispering softly.
- Previous attempts:
  - A word was read correctly during the first attempt: g
  - A word was read correctly but vowels were prolonged: vkg
  - A word was read correctly but vowels were prolonged to such an extent that it was no longer regarded as correct: vkf
  - A word was read wrongly during a first attempt: f
  - A word was skipped: O
  - Spelling behaviour occurs and it is not clearly audible: *s
  - A word switched places with another word; the first word receives a (W). This may be in combination with coding for mistakes, e.g., ag(W).
  - An examiner read the word out loud: Vv
  - The word was partly read out loud by an examiner: Vg
  - The first letter or part of the word was read correctly: ag
  - The first letter or part of the word was misread: af
  - The last letter or part of the word was read correctly: alg
  - The last letter or part of the word was misread: alf
  - The word was spelled out correctly (letters): lsg
  - The word was spelled out wrongly (letters): lsf
  - The word was spelled out correctly (syllables): ssg
  - The word was spelled out wrongly (syllables): ssf

Appendix C

Reading Error Labels

C.1 Wrong Reading Strategy

C.1.1 Decoding

a (wrong or correct) syllable-by-syllable decoding that is not followed by an attempt at synthesized reading (e.g., voet...bal [*foot...ball*])

b (wrong or correct) letter-by-letter decoding that is not followed by an attempt at synthesized reading (e.g., b...a...l [*b...a...ll*])

c (wrong or correct) partial (any (combination of) letters, syllables) decoding that is not followed by an attempt at synthesized reading (e.g., v...oe...t...bal [*f...oo...t...ball*])

C.1.2 Direct Word Recognition

e an existent word is replaced (in 1 attempt) by another existent word that is orthographically alike (e.g., staart/straat [*cloud/could*])

f an existent word is replaced (in 1 attempt) by another existent word that is semantically alike (e.g., poes/kat [*cat/kitten*]) and/or semantically fits into the piece of text that is read

g an existent word is replaced (in 1 attempt) by a pseudoword (e.g., straat/skraat [*street/screet*])

h a pseudoword is replaced (in 1 attempt) by an existent word that is orthographically alike (remark: the whole word needs to be a real word) (e.g., foetbans/voetbal [*footbans/football*])

j unable to attribute a code for type of reading error

C.2 Sequence Errors

1. change in order of adjacent letters (e.g., droon/doorn [*board/broad*])
2. change in order of letters that are separated from each other by at least one other letter (e.g., stol/slot [*glare/grale*])
3. other sequence errors

C.3 Substitution Errors

C.3.1 Vowels

4. substitution of vowels (e.g., bal/bel [*ball/bell*])
5. substitution of checked by free vowels or vice versa (e.g., man/maan [*strict/streaked*]) (remark: substitutions of /E/, /@/ and /e/ also belong to this category)
6. incorrect substitution of steady state vowels borrowed from other languages by other (combinations of) vowels (e.g., au in chauffeur / oi in toiletpapier)
7. pronunciation of steady state vowels (orthographically represented as a combination of two different vowels) as if they were actually two vowels (e.g., au in auto; or vice versa, e.g., ignoring the diaeresis in poezie))

C.3.2 Consonants

9. mirroring of consonants: b/d/p/q/g, t/f, m/w

10. substitution of a voiced consonant by its voiceless counterpart or vice versa (e.g., /x//g, s/z , f/v, t/d, p/b, g/k) (remark: e.g., hond pronounced as /hOnd@/ can only be an error of this category if no extra syllable was added by the reader)

11. substitution of letters that are visually alike (e.g., c/G, m/n, k/h, t/l, l/k, r/z, v/w, y/j, h/b, f/s, y/ij, l/f)

12. incorrect assimilation of energy: substitution of /N/ by /nx/, /nG/, /Nx/ of /NG/, substitution of /Nk/ by /nk/ or /N/, substitution of /S/ by /sj/

13. wrong reading of loanwords:

    - substitution of /S/ by /x/, /k/ or /G/ (e.g., chocolade, machine)
    - substitution of /si/ by /ti/ (e.g., vakantie)
    - substitution of /s/ by /k/ (e.g., circus)
    - substitution of /x/ by /tS/ or /k/ + /h/ (e.g., chaos)

14. substitution of an original letter by an arbitrary other letter (e.g., knop/kjop [*board/boald*])

C.3.3 Vowel-Consonant

16. substitution of a consonant by a vowel or vice versa (often the letters are visually alike) (e.g., u-n / i-j / i-l)

C.3.4 Syllables

17. substitution of a syllable by another syllable or letter

C.4 Deletion Errors

C.4.1 Vowels

18. deletion of an arbitrary vowel

C.4.2 Consonants

20. deletion of a plural n or s, or deletion of a conjugational t or n

21. deletion of an r or l in a consonant cluster

22. deletion of a n or g in ng OR deletion of an n or k in nk

23. deletion of an arbitrary consonant (e.g., reis/rei, netjes/netje [*board/boar*])

24. deletion of ch, sc or s in sch OR deletion of h or c in ch

C.4.3 Syllables

25. deletion of the suffix -en or er (e.g., kinderen/kinder, schoenenwinkel/schoenwinkel [*traveler/travel*])

26. deletion of an arbitrary syllable (e.g., Gerrit/Gert [*lovingly/lovly*])

C.4.4 Other

28. the childs last reading attempt is the beginning of a word but then it gives up, does not get the time to correct itself, or is interrupted by the ASR and the reading ends in inaudible whispering (*s)

29. the word was not repeated entirely when correcting a previous reading attempt (e.g., voetbel...bal [*footbell...ball*])

C.5 Insertion Errors

C.5.1 Vowels

30. insertion of an arbitrary vowel

C.5.2 Consonants

32. erroneous insertion of a plural n or s OR a conjugational t or n

33. incorrect reading of sch when only one or two of the three graphemes are actually printed OR incorrect reading of ch when only c is printed

34. insertion of a n before a g or k OR insertion of a g or k after n so that the phonemes /N/ of /Nk/ are read in stead of a simple g, k or n

35. insertion of an arbitrary consonant (e.g., poes/poest, gisteren/gristeren [*start/starts*])

C.5.3 Syllables

37. wrong insertion of the suffix -er or -e[n/] (e.g., balk/balker [*bark/barker*])

38. insertion of an arbitrary syllable

Appendix D

Phonetic Symbols

Class	Example	Symbol	Class	Example	Symbol
Plosives	**p**ut	p	Sonorants	la**ng**	N
	bad	b		**m**at	m
	tak	t		**n**at	n
	dak	d		**l**at	l
	kat	k		**r**at	r
	goal	g		**w**at	w
				jas	j
Fricatives	**f**iets	f			
	vat	v	Free vowels	l**ie**p	i
	sap	s		b**uu**r	y
	zat	z		l**ee**g	e
	sjaal	S		d**eu**k	2
	rava**g**e	Z		l**aa**t	a
	li**ch**t	x		b**oo**m	o
	re**g**en	G		b**oe**k	u
	ge**h**eel	h			
			Sjwa	g**e**lijk	@
Checked vowels	l**i**p	I			
	l**e**g	E	Diphtongues	w**ij**s	E+
	l**a**t	A		h**ui**s	Y+
	b**o**m	O		k**ou**d	A+
	p**u**t	Y			

References

Allington, R. L., & McGill-Franzen, A. (1980, April). Word identification errors in isolation and in context: Apples vs. oranges. *The Reading Teacher*, 795-800.

Archer, N., & Bryant, P. (2001). Investigating the role of context in learning to read: A direct test of Goodman's model. *British Journal of Psychology*, *92*, 579-591.

Aro, M., & Wimmer, H. (2003). Learning to read: English in comparison to six more regular orthographies. *Applied Psycholinguistics*, *24*, 621-635.

Baddeley, A. D., Logie, R. H., & Ellis, N. C. (1988). Characteristics of developmental dyslexia. *Cognition*, *29*, 197-228.

Batliner, A., Blomberg, M., D'Arcy, S., Elenius, D., Giuliani, D., Gerosa, M., et al. (2005). The PF_STAR Children's Speech Corpus. In *Proceedings of Interspeech* (pp. 2761–2764). Lisbon, Portugal.

Bayerl, P. S., & Paul, K. I. (2007). Identifying sources of disagreement: Generalizability theory in manual annotation studies. *Computational Linguistics*, *33*(1), 3-8.

Beck, J. E., Jia, P., & Mostow, J. (2004). Automatically assessing oral reading fluency in a computer tutor that listens. *Technology, Instruction, Cognition and Learning*, *1*, 61-81.

Biemiller, A. (1970). The development of the use of graphic and contextual information as children learn to read. *Reading Research Quarterly*, *6*(1), 75-96.

Biemiller, A. (1979). Changes in the use of graphic and contextual information as functions of passage difficulty and reading achievement level. *Journal of Reading Behavior*, *11*(4), 307-318.

Billiaert, E., & Grysolle, R. (1998). *Leerlingvolgsysteem lezen: Analyse en handelen. Basisboek [Pupil's monitoring system for reading: Analysis and treatment. Basic manual].* Leuven/Apeldoorn, Belgium/The Netherlands: CSBO & Garant.

Blaxall, J., & Willows, D. M. (1984). Reading ability and text difficulty as influences on second graders' oral reading errors. *Journal of Educational Psychology*, *76*(1), 330-341.

Bodmann, S. M., & Robinson, D. H. (2004). Speed and performance differences among computer-based and paper-pencil tests. *Journal of Educational Computing Research*, *31*(1), 51-60.

Brouwer, R. H. M. (1963). Onderzoek naar de leesmoeilijkheid van Nederlandse proza [Research into the readability of Dutch prose]. *Paedagogische Studiën*, *10*, 454-464.

Brus, B. T., & Voeten, M. J. M. (1995). *Eén-Minuut-Test. Vorm A en B. Verantwoording en handleiding [One-Minute-Test. Form A and B. Manual].* Nijmegen, The Netherlands: Berkhout Testmateriaal B.V.

Chiappe, P., & Siegel, L. S. (1999). Phonological awareness and reading acquisition in English- and Punjabi-speaking Canadian children. *Journal of Educational Psychology*, *91*(1), 20-28.

Cicchetti, D. V. (1972). A new measure of agreement between rank ordered variables. In *Proceedings of the 80th Annual Convention of the American Psychological Association* (pp. 17–18). Honolulu, Hawaii.

Clariana, R., & Wallace, P. (2002). Paper-based versus computer-based assessment: Key factors associated with the test mode effect. *British Journal of Educational Technology*, *33*(5), 593-602.

Cleuren, L., Duchateau, J., Ghesquière, P., & Van hamme, H. (2008). Children's oral reading corpus (CHOREC): Description and assessment of annotator agreement. In *Proceedings of the 6th International Conference on*

Language Resources and Evaluation. Marrakech, Morocco. (8 pages, ISBN 2-9517408-4-0, EAN 9782951740846)

Cleuren, L., Duchateau, J., Sips, A., Ghesquière, P., & Van hamme, H. (2006). Developing an automatic assessment tool for children's oral reading. In *Proceedings of the International Conference on Spoken Language Processing* (pp. 817–820). Pittsburgh, USA.

Cleuren, L., & Ghesquière, P. (2008). AVI verandert: Wat met Vlaanderen? [AVI changes: What about Flanders?]. *Logopedie: informatiemedium van de Vlaamse Vereniging voor Logopedisten*, *21*(5), 46-48.

Cleuren, L., Kuppens, S., Duchateau, J., Van hamme, H., & Ghesquière, P. (2009a). A quantitative oral reading error analysis in elementary school children: Evidence from Dutch. Manuscript in preparation for submission.

Cleuren, L., Kuppens, S., Duchateau, J., Van hamme, H., & Ghesquière, P. (2009b). Oral reading errors of Dutch speaking elementary school children: A qualitative analysis at the word level. Manuscript submitted for publication.

Cleuren, L., Latacz, L., Kong, Y.-O., Vandermosten, M., Verhelst, W., Van hamme, H., et al. (2008). Evaluation of a reading tutor with synthesized speech feedback for Dutch elementary school children with reading difficulties. Manuscript submitted for publication.

Cohen, A. S. (1974). Oral reading errors of first grade children taught by a code emphasis approach. *Reading Research Quarterly*, *10*(4), 616-650.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Coltheart, M., Curtis, B., Atkins, P., & Haller, M. (1993). Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological Review*, *100*(4), 489-608.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, *108*(1), 204-256.

Crowe, L. K. (2003). Comparison of two reading feedback strategies in improving the oral and written language performance of children with language-learning disabilities. *American Journal of Speech-Language Pathology*, *12*, 16-27.

Cucchiarini, C. (1996). Assessing transcription agreement: Methodological aspects. *Clinical Linguistics and Phonetics*, *10*(2), 131-155.

Dahl-bäck, N., Jönsson, A., & Ahrenberg, L. (1993). Wizard of Oz studies: Why and how. In *Proceedings of the 1st International Conference on Intelligent User Interfaces* (pp. 193–200). Orlando, Florida, USA.

D'Angelo, K., & Mahlios, M. (1983, April). Insertion and omission miscues of good and poor readers. *The Reading Teacher*, 778-782.

Danielsson, K. (2001). Beginning readers' sensitivity to different linguistic levels: An error and correction analysis at the lexical, syntactic, and semantic levels. *Reading and Writing*, *14*, 395-421.

Danielsson, K. (2002). The impact of high versus low linguistic levels in Swedish beginning readers' oral reading of running texts. *Scandinavian Journal of Educational Research*, *46*(4), 427-457.

Danielsson, K. (2003). The relationship between grapheme-phoneme correspondences and reading errors in Swedish beginners' oral reading. *Scandinavian Journal of Educational Research*, *47*(5), 511-528.

Davidson, J., Elcock, J., & Noyes, P. (1996). A preliminary study of the effect of computer-assisted practice on reading attainment. *Journal of Research in Reading*, *19*(2), 102-110.

de Jong, P. F., & van der Leij, A. (2003). Developmental changes in the manifestation of a phonological deficit in dyslexic children learning to read a regular orthography. *Journal of Educational Psychology*, *95*(1), 22-40.

Dennis, I., & Evans, J. S. B. T. (1996). The speed-error trade-off problem in psychometric testing. *British Journal of Psychology*, *87*, 105-129.

de Wijs, A. (2008). Gaat AVI veranderen? De situatie in Nederland [Will AVI be changed? The situation in the Netherlands]. *Logopedie: informatiemedium van de Vlaamse Vereniging voor Logopedisten*, *21*(5), 41-45.

Dilley, L., Breen, M., Bolivar, M., Kraemer, J., & Gibson, E. (2006). A comparison of inter-transcriber reliability for two systems of prosodic annotation: RaP (Rhythm and Pitch) and ToBI (Tones and Break Indices). In *Proceedings of the International Conference on Spoken Language Processing* (pp. 317–320). Pittsburgh, USA.

Duchateau, J., Cleuren, L., Van hamme, H., & Ghesquière, P. (2007). Automatic assessment of children's reading level. In *Proceedings of the European Conference on Speech Communication and Technology* (pp. 1210–1213). Antwerp, Belgium.

Duchateau, J., Kong, Y.-O., Cleuren, L., Latacz, L., Roelens, J., Samir, A., et al. (2009). Developing a reading tutor: Design and evaluation of dedicated speech recognition and synthesis modules. *Speech Communication*, *51*(10), 985-994.

Ehri, L. C. (1991). Development of the ability to read words. In R. Barr, M. Kamil, P. Mosenthal, & P. Pearson (Eds.), *Handbook of reading research Volume II* (pp. 383–417). New York, NY, USA: Longman.

Elbro, C., Rasmussen, I., & Spelling, B. (1996). Teaching reading to disabled readers with language disorders: A controlled evaluation of synthetic speech feedback. *Scandinavian Journal of Psychology*, *37*, 140-155.

Ellis, N. C., & Hooper, A. M. (2001). Why learning to read is easier in Welsh than in English: Orthographic transparency effects evinced with frequency-matched tests. *Applied Psycholinguistics*, *22*(4), 571-599.

Ellis, N. C., Natsume, M., Stavropoulou, K., Hoxhallari, L., van Daal, V. H. P., Polyzoe, N., et al. (2004). The effects of orthographic depth on learning to read alphabetic, syllabic, and logographic scripts. *Reading Research Quarterly*, *39*(4), 438-468.

Eskenazi, M. S. (1996). KIDS: A database of children's speech. *Journal of the Acoustical Society of America*, *100*(4), 2759-2759.

Francis, H. (1984). Children's knowledge of orthography in learning to read. *British Journal of Educational Psychology*, *54*, 8-23.

Georgiou, G. K., Das, J. P., & Hayward, D. (2008). Revisiting the "simple view of reading" in a group of children with poor reading comprehension. *Journal of Learning Disabilities*, *42*(1), 76-84.

Goikoetxea, E. (2006). Reading errors in first- and second-grade readers of a shallow orthography: Evidence from Spanish. *British Journal of Educational Psychology*, *76*, 333-350.

Goodman, K. S. (1965). A linguistic study of cues and miscues in reading. *Elementary English Journal*, *42*, 39-44.

Goodman, K. S. (1969). Analysis of oral reading miscues: Applied psycholinguistics. *Reading Research Quarterly*, *5*(1), 9-30.

Goswami, U., Gombert, J. E., & de Barrera, L. (1998). Children's orthographic representations and linguistic transparency: Nonsense word reading in English, French, and Spanish. *Applied Psycholinguistics*, *19*(1), 19-52.

Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education*, *7*(1), 6-10.

Gupta, A. (2004). Reading difficulties of Hindi-speaking children with developmental dyslexia. *Reading and Writing: An Interdisciplinary Journal*, *17*, 79-99.

Gupta, A., & Jamal, G. (2006). An analysis of reading errors of dyslexic readers in Hindi and English. *Asia Pacific Disability Rehabilitation Journal*, *17*(1), 73-86.

Gupta, A., & Jamal, G. (2007). Reading strategies of bilingual normally progressing and dyslexic readers in Hindi and English. *Applied Psycholinguistics*, *28*, 47-68.

Hadwin, A. F., Winne, P. H., & Nesbit, J. C. (2005). Roles for software technologies in advancing research and theory in educational psychology. *British Journal of Educational Psychology*, *75*, 1-24.

Harding, L. M., Beech, J. R., & Sneddon, W. (1985). The changing pattern of reading errors and reading style from 5 to 11 years of age. *British Journal of Educational Psychology*, *55*, 45-52.

Hardré, P. L., Crowson, H. M., Xie, K., & Ly, C. (2007). Testing differential effects of computer-based, web-based and paper-based administration of questionnaire research instruments. *British Journal of Educational Technology*, *38*(1), 5-22.

Herrmann, J. A., Matyas, T., & Pratt, C. (2006). Meta-analysis of the nonword reading deficit in specific reading disorder. *Dyslexia*, *12*, 195-221.

Higgins, E. L., & Raskind, M. H. (2004). Speech recognition-based and automaticity programs to help students with severe reading and spelling problems. *Annals of Dyslexia*, *54*(2), 365-392.

Homan, S. P., Klesius, J. P., & Hite, C. (1993). Effects of repeated readings and nonrepetitive strategies on students' fluency and comprehension. *Journal of Educational Research*, *87*(2), 94-99.

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing: An Interdisciplinary Journal*, *2*(2), 127-160.

Horne, J. (2007). Gender differences in computerized and conventional educational tests. *Journal of Computer Assisted Learning*, *23*(1), 47-55.

Hoxhallari, L., van Daal, V. H. P., & Ellis, N. C. (2004). Learning to read words in Albanian: A skill easily acquired. *Scientific Studies of Reading*, *8*(2), 153-166.

Huang, X. D., Acero, A., & Hon, H.-W. (2001). *Spoken language processing: A guide to theory, algorithm, and system development*. NJ, USA: Prentice Hall.

Humphreys, G. W., & Evett, L. J. (1985). Are there independent lexical and nonlexical routes in word processing? An evaluation of the dual-route theory of reading. *Behavioral and Brain Sciences*, *8*(4), 689-739.

Irausquin, R. S., Drent, J., & Verhoeven, L. (2005). Benefits of computer-presented speed training for poor readers. *Annals of Dyslexia*, *55*(2), 246-265.

Jenkins, J. R., Fuchs, L. S., van den Broek, P., Espin, C., & Deno, S. L. (2003). Accuracy and fluency in list and context reading of skilled and RD groups. *Learning Disabilities Research and Practice*, *18*(4), 237-245.

Jiménez, J. E., Hernández-Valle, I., Ramírez, G., del Rosario Ortiz, M., Rodrigo, M., Estévez, A., et al. (2007). Computer speech-based remediation for reading disabilities: The size of spelling-to-sound unit in a transparent orthography. *The Spanish Journal of Psychology*, *10*(1), 52-67.

Katz, L., & Frost, R. (1992). The reading process is different for different orthographies: The orthographic depth hypothesis. In R. Frost & L. Katz (Eds.), *Orthography, phonology, morphology, and meaning* (pp. 45–66). Amsterdam, The Netherlands: Elsevier.

Kazemzadeh, A., You, H., Iseli, M., Jones, B., Cui, X., Heritage, M., et al. (2005). TBALL data collection: The making of a young children's speech corpus. In *Proceedings of Interspeech* (pp. 1581–1584). Lisbon, Portugal.

Kim, D.-H., & Huynh, H. (2008). Computer-based and paper-and-pencil administration mode effects on a statewide end-of-course English test. *Educational and Psychological Measurement*, *68*(4), 554-570.

Kuhn, M. R., & Stahl, S. A. (2003). Fluency: A review of developmental and remedial practices. *Journal of Educational Psychology*, *95*(1), 3-21.

LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology*, *6*(1), 293-323.

Laing, S. P. (2002). Miscue analysis in school-age children. *American Journal of Speech-Language Pathology*, *11*, 407-416.

Landerl, K. (2000). Influences of orthographic consistency and reading instruction on the development of nonword reading skills. *European Journal of Psychology of Education*, *3*, 239-257.

Landerl, K., & Wimmer, H. (2008). Development of word reading fluency and spelling in a consistent orthography: An 8-year follow-up. *Journal of Educational Psychology*, *100*(1), 150-161.

Landerl, K., Wimmer, H., & Frith, U. (1997). The impact of orthographic consistency on dyslexia: a German-English comparison. *Cognition*, *63*, 315-334.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174.

Latacz, L., Kong, Y.-O., Mattheyses, W., & Verhelst, W. (2006). Novel text-to-speech reading modes for educational applications. In *Proceedings of the ProRISC/IEEE Benelux Workshop on Circuits, Systems and Signal Processing.* Veldhoven, The Netherlands.

Latacz, L., Kong, Y.-O., & Verhelst, W. (2007). Unit selection synthesis using long non-uniform units and phonemic identity matching. In *Proceedings of the 6th ISCA Workshop on Speech Synthesis (SSW-6)* (pp. 270–275). Bonn, Germany.

Leeson, H. V. (2006). The mode effect: A literature review of human and technological issues in computerized testing. *International Journal of Testing*, *6*(1), 1-24.

Leslie, L. (1980). The use of graphic and contextual information by average and below-average readers. *Journal of Reading Behavior*, *12*(2), 139-149.

Leu, D. J. (1982). Oral reading error analysis: A critical review of research and application. *Reading Research Quarterly*, *3*, 420-437.

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schabenberger, O. (2006). *SAS for mixed models* (second ed.). Cary: SAS Institute.

Lovett, M. W., Barron, R. W., Forbes, J. E., Cuksts, B., & Steinbach, K. A. (1994). Computer speech-based training of literacy skills in neurologically impaired children: A controlled evaluation. *Brain and Language*, *47*, 117-154.

Lynch, A., Fawcett, A. J., & Nicolson, R. I. (2000). Computer-assisted reading intervention in a secondary school: An evaluation study. *British Journal of Educational Technology*, *31*(4), 333-348.

MacArthur, C. A., Ferretti, R. P., Okolo, C. M., & Cavalier, A. R. (2001). Technology applications for students with literacy problems: A critical review. *The Elementary School Journal*, *101*(3), 273-301.

Macaruso, P., Hook, P. E., & McCabe, R. (2006). The efficacy of computer-based supplementary phonics programs for advancing reading skills in at-risk elementary students. *Journal of Research in Reading*, *29*(2), 162-172.

McCoy, K. M., & Pany, D. (1986). Summary and analysis of oral reading corrective feedback research. *Reading Teacher*, *39*(6), 548-554.

McDonald, A. S. (2002). The impact of individual differences on the equivalence of computer-based and paper-and-pencil educational assessments. *Computers and Education*, *39*, 299-312.

Miller, J. W., & Isakson, R. L. (1978). Contextual sensitivity in beginning readers. *The Elementary School Journal*, *78*(5), 325-331.

Moscicki, E. K., & Tallal, P. (1981). A phonological exploration of oral reading errors. *Applied Psycholinguistics*, *2*, 353-367.

Mostow, J., & Aist, G. (2001). Evaluating tutors that listen: An overview of project LISTEN. In K. Forbus & P. Feltovich (Eds.), *Smart machines in education: The coming revolution in educational technology* (pp. 169–234). Cambridge, MA, USA: MIT Press.

Mostow, J., Beck, J., Winter, S. V., Wang, S., & Tobin, B. (2002). Predicting oral reading miscues. In *Proceedings of the 7th International Conference on Spoken Language Processing* (pp. 1221–1224). Denver, USA.

Nergård-Nilssen, T. (2006). Developmental dyslexia in Norwegian: Evidence from single-case studies. *Dyslexia*, *12*, 30-50.

NICHHD. (2000). *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction.* Washington, DC, USA: U.S. Government Printing Office. (NIH Publication No. 00-4769)

Oller, D. K., & Ramsdell, H. L. (2006). A weighted reliability measure for phonetic transcription. *Journal of Speech, Language, and Hearing Research*, *49*, 1391-1411.

Pany, D., & McCoy, K. M. (1988). Effects of corrective feedback on word accuracy and reading comprehension of readers with learning disabilities. *Journal of Learning Disabilities*, *21*(9), 546-550.

Patel, T. K., Snowling, M. J., & de Jong, P. F. (2004). A cross-linguistic comparison of children learning to read in English and Dutch. *Journal of Educational Psychology*, *96*(4), 785-797.

Perkins, V. L. (1988). Feedback effects on oral reading errors of children with learning disabilities. *Journal of Learning Disabilities*, *21*(4), 244-248.

Pitt, M. A., Johnson, K., Hume, E., Kiesling, S., & Raymond, W. (2005). The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication*, *45*, 89–95.

Protopapas, A., & Skaloumbakas, C. (2007). Traditional and computer-based screening and diagnosis of reading disabilities in Greek. *Journal of Learning Disabilities*, *40*(1), 15-36.

Protopapas, A., Skaloumbakas, C., & Bali, P. (2008). Validation of unsupervised computer-based screening for reading disability in Greek elementary grades 3 and 4. *Learning Disabilities: A Contemporary Journal*, *6*(1), 45-69.

Rashotte, C. A., & Torgesen, J. K. (1985). Repeated reading and reading fluency in learning disabled children. *Reading Research Quarterly*, *20*(2), 180-188.

Raven, J. C., Court, J. H., & Raven, J. (1992). *Standard progressive matrices.* Oxford, UK: Oxford Psychologists Press.

Reitsma, P. (1988). Reading practice for beginners: Effects of guided reading, reading-while-listening, and independent reading with computer-based speech feedback. *Reading Research Quarterly*, *23*(2), 219-235.

Reynolds, M., & Besner, D. (2008). Contextual effects on reading aloud: Evidence for pathway control. *Journal of Experimental Psychology*, *34*(1), 50-64.

Roberts, J. K., & Monaco, J. P. (2006, April). Effect size measures for the two-level linear multilevel model. *Paper presented at the annual meeting of the American Educational Research Association*. San Francisco, USA.

Russell, M. J., D'Arcy, S. M., & Wong, L. P. (2004). Recognition of read and spontaneous children's speech using two new corpora. In *Proceedings of the International Conference on Spoken Language Processing* (pp. 1473–1476). Jeju Island, Korea.

Samuels, S. J. (2002). Reading fluency: Its development and assessment. In A. Farstrup & S. Samuels (Eds.), *What research has to say about reading instruction* (3rd ed., pp. 166–184). Newark, DE, USA: International Reading Association.

Savage, R., Stuart, M., & Hill, V. (2001). The role of scaffolding errors in reading development: Evidence from a longitudinal and a correlational study. *British Journal of Educational Psychology*, *71*, 1-13.

Schaerlaekens, A., Kohnstamm, D., & Lejaegere, M. (1999). *Streeflijst woordenschat voor zesjarigen. Derde herziene versie gebaseerd op nieuw onderzoek in Nederland en België. [Target word list for six-year-olds. Third revision based on new research in the Netherlands and Belgium].* Lisse: Swets & Zeitlinger.

Schippers, K., & Corens, J. (2000). *Collectieve schriftelijke toetsen voor het technisch lezen. Constructie en validering van de lexicale decisieproef en de test woordkettingen. [Collective written tests to assess technical reading. Construction and validation of the lexical decision test and the word chains test].* Masters thesis, Katholieke Universiteit Leuven.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*(4), 523-568.

Seymour, P. H. K., Aro, M., & Erskine, J. M. (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology*, *94*(2), 143-174.

Shobaki, K., Hosom, J.-P., & Cole, R. A. (2000, October). The OGI Kids' speech corpus and recognizers. In *Proceedings of the Sixth International Conference on Spoken Language Processing* (pp. 258–261). Beijing, China.

Siegel, L. S. (2003). Basic cognitive processes and reading disabilities. In H. L. Swanson, K. R. Harris, & S. Graham (Eds.), *Handbook of learning disabilities* (pp. 158–181). New York, NY, USA: The Guilford Press.

Sim, G., & Horton, M. (2005). Performance and attitude of children in computer based versus paper based testing. In *Proceedings of the World Conference on Educational Multimedia, Hypermedia and Telecommunications* (pp. 3610–3614). Montreal, Canada.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics*, *24*(4), 323-355.

Spaai, G. W. G., Ellermann, H. H., & Reitsma, P. (1991). Effects of segmented and whole-word sound feedback on learning to read single words. *Journal of Educational Research*, *84*(4), 204-213.

Spencer, L. H., & Hanley, J. R. (2003). Effects of orthographic transparency on reading and phoneme awareness in children learning to read in Wales. *British Journal of Psychology*, *94*, 1-28.

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, *21*(4), 360-407.

Strickland, D. S. (2002). The importance of effective early intervention. In A. E. Farstrup & S. J. Samuels (Eds.), *What research has to say about reading instruction* (3rd ed., pp. 69–86). Newark, DE, USA: International Reading Association.

Stuart, M., & Coltheart, M. (1988). Does reading develop in a sequence of stages? *Cognition*, *30*, 139-181.

Szezsulski, P. A., & Manis, F. R. (1987). A comparison of word recognition processes in dyslexic and normal readers at two reading-age levels. *Journal of Experimental Child Psychology*, *44*, 364-376.

Thaler, V., Ebner, E. M., Wimmer, H., & Landerl, K. (2004). Training reading fluency in dysfluent readers with high reading accuracy: Word specific effects but low transfer to untrained words. *Annals of Dyslexia*, *54*(1), 89-113.

Thomson, M. (1978). A psycholinguistic analysis of reading errors made by dyslexics and normal readers. *Journal of Research in Reading*, *1*(1), 7-20.

Thorstad, G. (1991). The effect of orthography on the acquisition of literacy skills. *British Journal of Psychology*, *82*(4), 527-537.

Tressoldi, P. E., Lorusso, M. L., Brenbati, F., & Donini, R. (2007). Fluency remediation in dyslexic children: Does age make a difference? *Dyslexia*, *14*, 142-152.

van Daal, V. H. P., & Reitsma, P. (2000). Computer-assisted learning to read and spell: Results from two pilot studies. *Journal of Research in Reading*, *23*(2), 181-193.

van Daal, V. H. P., & van der Leij, A. (1992). Computer-based reading and spelling practice for children with learning disabilities. *Journal of Learning Disabilities*, *25*(3), 186-195.

van den Bos, K. P., Spelberg, H. C. L., Scheepstra, A. J. M., & de Vries, J. R. (1994). *De Klepel: Pseudowoordentest [The Klepel: Pseudoword test].* Nijmegen, The Netherlands: Berkhout.

van der Schoot, M., Licht, R., Horsley, T. M., Aarts, L. T., van Koert, B., & Sergeant, J. A. (2004). Inhibitory control during sentence reading in dyslexic children. *Child Neuropsychology*, *10*(3), 173-188.

van der Schoot, M., Licht, R., Horsley, T. M., & Sergeant, J. A. (2002). Fronto-central dysfunctions in reading disability depend on subtype: Guessers but not spellers. *Developmental Neuropsychology*, *22*(3), 533-564.

Verachtert, P. (2001). *Computergestuurde leesdiagnostiek. De mogelijke rol van de computer in de diagnostiek van dyslexie. [Computer based reading assessment. The possible role of the computer in the assessment of dyslexia].* Masters thesis, Katholieke Universiteit Leuven.

Verhoeven, L. (1995). *Drie-Minuten-Toets. Handleiding [Three-Minutes-Test. Manual].* Arnhem, The Netherlands: CITO.

Visser, J. (1997). *Lezen op AVI-niveau. De schaling, normering en validering van de nieuwe AVI-toetskaarten. [Reading at AVI level. The scaling, standardization, and validation of the new AVI test cards].* 's Hertogenbosch, The Netherlands: KPC Groep.

Visser, J., van Laarhoven, A., & ter Beek, A. (1998). *AVI Toetspakket. Handleiding [AVI testing package. Manual].* 's Hertogenbosch, The Netherlands: KPC groep.

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2008). Comparibility of computer-based and paper-and-pencil testing in K-12 reading assessments. *Educational and Psychological Measurement*, *68*(1), 5-24.

Wanzek, J., & Vaughn, S. (2007). Research-based implications from extensive early reading interventions. *School Psychology Review*, *36*(4), 541-561.

Weber, R.-M. (1968). The study of oral reading errors: A survey of the literature. *Reading Research Quarterly*, *4*(1), 96-119.

Weber, R.-M. (1970). A linguistic analysis of first-grade reading errors. *Reading Research Quarterly*, *5*, 427-451.

Werker, J. F., Bryson, S. E., & Wassenberg, K. (1989). Toward understanding the problem in severely disabled readers. Part 2: Consonant errors. *Applied Psycholinguistics*, *10*, 13-30.

Wilkinson, K. M., & Hennig, S. (2007). The state of research and practice in augmentative and alternative communication for children with developmental/intellectual disabilities. *Mental Retardation and Developmental Disabilities Research Reviews*, *13*, 58-69.

Wimmer, H. (1993). Characteristics of developmental dyslexia in a regular writing system. *Applied Psycholinguistics*, *14*, 1-33.

Wimmer, H. (1996). The nonword reading deficit in developmental dyslexia: Evidence from children learning to read German. *Journal of Experimental Child Psychology*, *61*, 80-90.

Wimmer, H., & Goswami, U. (1994). The influence of orthographic consistency on reading development: Word recognition in English and German children. *Cognition*, *51*, 91-103.

Wimmer, H., & Hummer, P. (1990). How German-speaking first graders read and spell: Doubts on the importance of the logographic stage. *Applied Psycholinguistics*, *11*, 349-368.

Wise, B. W., Cole, R., van Vuuren, S., Schwartz, S., Snyder, L., Ngampatipatpong, N., et al. (2005). Learning to read with a virtual tutor: Foundations to literacy. In C. Kinzer & L. Verhoeven (Eds.), *Interactive literacy education: Facilitating literacy environments through technology* (pp. 31–75). New York, NY, USA: Lawrence Erlbaum.

Wise, B. W., & Olson, R. K. (1998). Studies of computer-aided remediation for reading disabilities. In C. Hulme & R. M. Joshi (Eds.), *Reading and*

spelling: Development and disorders (pp. 473–487). Mahwah, NJ, USA: Lawrence Erlbaum.

Wise, B. W., Ring, J., & Olson, R. K. (2000). Individual differences in gains from computer-assisted remedial reading. *Journal of Experimental Child Psychology*, *77*, 197-235.

Yoon, T.-J., Chavarria, S., Cole, J., & Hasegawa-Johnson, M. (2004). Inter-transcriber reliability of prosodic labeling on telephone conversation using ToBI. In *Proceedings of the International Conference on Spoken Language Processing* (pp. 2729–2732). Jeju Island, Korea.

Ziegler, J. C., & Goswami, U. (2006). Becoming literate in different languages: Similar problems, different solutions. *Developmental Science*, *9*(5), 429-453.