Clozing in on readability:

How linguistic features affect and predict text comprehension and on-line processing

Published by
LOT
Trans 10
3512 JK Utrecht
The Netherlands

phone: +31 30 253 6111

e-mail: lot@uu.nl
http://www.lotschool.nl

Cover illustration: cat-crouch-iStock_000007472694Small

ISBN: 978-94-6093-277-9
NUR 616

Clozing in on readability:

How linguistic features affect and predict text comprehension and on-line processing

Leesbaarheid ontrafeld:

Hoe linguïstische kenmerken tekstbegrip en tekstverwerking beïnvloeden en voorspellen

(met samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor
aan de Universiteit Utrecht
op gezag van de rector magnificus, prof.dr. G.J. van der Zwaan,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen
op vrijdag 6 april 2018
des middags te 2.30 uur

door

Suzanne Kleijn
geboren op 16 oktober 1988 te Nieuwegein

Promotor: Prof.dr. T.J.M. Sanders
Copromotor: Dr. H.L.W. Pander Maat

The research presented here was part of the LIN-project ('Readability index for Dutch') funded by NWO ('The Netherlands Organisation for Scientific Research'), Cito ('The Dutch Institute for Educational Measurement') and Nederlandse Taalunie ('The Dutch Language Union') [NWO grant number 321 89 002].

Table of contents

Acknowledgements

60 texts, in 4 versions, 11519 answered cloze tests, 397658 answered cloze gaps, 549737 fixations, 3107 students, and over 400 text features… For a dissertation on readability, the story of how it all came together is much more about numbers than about words: the incredible amount of data but also the incredible amount of hours that were necessary to collect, clean-up and analyze these data, let alone to write it all down. It also took some incredible people to make this happen.

First of all my amazing supervisors: Henk Pander Maat and Ted Sanders. Thank you for choosing me to do this wonderful project with you.

Henk, I know you don't remember but we met during a course you gave in my bachelors. Perhaps it was destiny because that course was called: *Tekstontwerp en begrijpelijkheid* 'Text design and comprehensibility'. And here we are 10 years later, this time working together on text comprehension.

Ted, I don't know whether you remember, but I believe the first time we spoke was when I had signed up for the Research Master Linguistics. I remember you asking me what my long term goals were and whether I was planning to do a PhD after the master. I said that that was not my intention, but since I also never intended to do a research master I also said "but never say never". I looked it up and that meeting took place in April 2010. I think neither of us imagined that this is where we ended up 8 years later. I'm glad I kept an open mind.

Ted and Henk, thank you for your support and for always making time for me (during working hours, weekends and probably many an evening), and even when personal and work circumstances made that almost impossible. I really appreciate it. Thank you for showing me the big picture when I started to drown in the details. I never once thought of quitting even when the work kept piling on with no end in sight, and that was largely thanks to you two! I greatly enjoyed working with both of you and look forward to working with you in future projects. Although perhaps, we should stay away from any projects involving connectives for a while, just to keep the peace :).

I also want to thank our project partners Cito and the Nederlandse Taalunie for co-funding this project, as well as the LIN-project team (Antal van den Bosch, Anton Béguin and Johan van Hoorde), our computational and programming experts (Rogier Kraf, Ko van der Sloot, Maarten van Gompel, Martijn van der Klis and Florian Kunneman), all the student assistants and people at Cito who helped collect and score our data, and especially Servaas Frissen, who probably never wants to hear the word "clozetoets" again, thank you for all your help and perseverance in collecting the cloze data. It took a while but we got there in the end! Thank you for

sticking with us. And of course a special thanks to all the schools and students who participated without whom we would be nowhere.

I want to thank all my (ex-)colleagues from UiL-OTS and afd. Taalbeheersing who have made (and continue to make) my time at the Trans very enjoyable and a great experience. Among whom my fellow PhDs and Post-docs: Hans Rutger Bosker, Anne-France Pinget, Monica Koster, Renske Bouwer, Louise Nell, Merel Scholman, Naomi Kamoen, Gerdineke van Silfhout, Heidi Klockmann, Björn ‘t Hart, Anne van Leeuwen, Brigitta Keij, Jolien Scholten, Carolien van den Hazelkamp, Alexia Guerra Rivera, Dominique Blok, Marjolein van Egmond, Marijn Struiksma, Hannah de Mulder, Hiske Feenstra, Ileana Grama, Anja Goldschmidt, Maartje Schulpen, Andrea Santana C., Yipu Wei, Mirjam Hachem, Myrthe Bergstra, Lotte Hendriks, Anna Sara Romøren, Eva Poortman, Zenghui Liu, Hanna de Vries, Stavroula Alexandropoulou, Suzanne Bogaerds-Hazenberg, Yuan Xie and Nina Sangers.

Iris Mulders, thank you for all your help in and out of the eye-tracking lab. I still remember the red, foldaway, cart filled to the brink with the ‘not-so-portable’ EyeLink system, surrounded by a number of bags and boxes which all needed to come with us in order to run the eye-tracking experiments... It was insane, but also thanks to Chris van Run and Alex Manus, we made it work!

Huub van den Bergh, if you hadn’t shown me how to work with MLwin, I would probably still be waiting for my statistical models to converge. Thanks for all the tips and tricks!

A very special thanks to my office mates for most of my PhD-years: Marloes Herijgers and Jet Hoek, thank you for putting up with me and for being there through ups and downs; sharing frustrations as well as a lot of laughs.

And to all my friends outside of the university: Non, Nienk, Dree, Marlies, Anna, Patrick, Dennis, Twan, and of course my fellow Losers: Mief, Paultje and my lovely paranymphs Alvie and Janiet (Vette sjizzle!). Thank you for your support and for keeping me distracted with great evenings, dinners, drinks, trips, talks and jokes. Even if we don’t see each other that often, it always feels like no time has passed at all when I do see you. I’m very grateful for you all! And I promise: I’ll make it up to you! No more granny-actions and hermit-behaviors: it is time to celebrate!

I also want to thank Eva Kieboom and my (Ex-)Tonecollectors (René, Bert, Jos, Maut, Sarah, Egbert). Thanks for being my musical outlet!

Last but definitely not least, I want to thank my family: ma, pa, Emil and Anneke. Thank you for giving me a stable foundation to build my future on and for your endless encouragement and support. I know I can always count on you. And Emil, I know this dissertation is not going to convince you that Alpha-research can also be ‘science’, but I guess we have to agree to disagree on that ;).

And Dennis, Hofstadter’s law has been proven right once again:

> *It always takes longer than you expect, even when you take into account Hofstadter's Law. — Douglas Hofstadter*

Suzanne
February 2018

1 Introduction

The first readability formulae were developed almost 100 years ago (DuBay, 2006), but since then the need for objective measures of readability seems only to have increased. These days not just educators and publishers are concerned with matching readers and texts. Governments, organizations and companies are interested as well, as new regulations compel them to communicate clearly and openly with citizens and customers. The need for automated tools that help them assess the difficulty of text is great, in the Netherlands as well as in other countries. Unfortunately, many of the existing tools for Dutch are neither build nor empirically validated by research (Jansen, 2005; Jansen & Boersma, 2013; Kraf, Lentz & Pander Maat, 2011). Their validity is questionable and readability assessments offered by these tools should be regarded with caution. Fortunately, recent developments in computational linguistics and ongoing experimental work on text comprehension and discourse processing create possibilities to change this situation. It is now possible to automatically analyze texts and to compute complex linguistic features with the press of a button (cf. Coh-Metrix by Graesser, McNamara, Louwerse & Cai, 2004; T-Scan by Pander Maat et al., 2014). Moreover, because experimental studies have provided us with new insights on how text characteristics affect readers, we can develop new and more valid indices to approximate text difficulty. With these new insights and technologies we can build better tools that can assess whether a text is suitable for a certain reader or even build diagnostic tools that can point out potential problems within a text.

These new developments, together with the lack of a Dutch readability tool, formed the inspiration for the LIN-project: 'LeesbaarheidsIndex voor het Nederlands' [Dutch readability index].[1] The main goal of this project is to build a validated automated readability assessment tool for Dutch which is insightful for researchers as well as for the general public. The LIN-tool is aimed at reading levels of adolescents but has the potential to extend to adult readers of the Dutch-speaking population.

Obviously, several steps must be taken in order to create such an instrument. First, we need a tool that can automatically extract and compute linguistic features from Dutch text. Once the features are extracted, they need to be

[1] The LIN-project was funded by NWO ('The Netherlands Organisation for Scientific Research'), Cito ('The Dutch Institute for Educational Measurement') and Nederlandse Taalunie ('The Dutch Language Union') [NWO grant number 321-89-002].

validated and calibrated. We need to know how these features affect readers within the target population and at which levels these features become problematic. For this step we need to collect behavioral data[2] on how our target population understands texts. We need actual data to show how our readers comprehend texts and how this process is influenced by linguistic features. Next, we can select the factors that best predict the data and as the final step we can build the LIN-tool. Scholars, programmers and experts from Utrecht University, Radboud University, Cito ('The Dutch Institute for Educational Measurement') and Nederlandse Taalunie ('The Dutch Language Union') worked together to accomplish this goal.

As part of the LIN-project, the dissertation presented here is mainly focused on investigating the effects of linguistic features on text readability and on collecting the empirical data on which the LIN-tool will be based. In the remainder of this introduction, we will discuss problems and limitations associated with prior readability research and how we addressed these issues in the design of our study. The introduction will end with a chapter overview and a reading guide for the remaining chapters of this dissertation.

1. Problems associated with (traditional) readability research

Readability research became increasingly popular in the mid-20th century with researchers like Dale and Chall (1948), Flesch (1948), Klare (1963) and Bormuth (1969). However, in the seventies researchers started questioning its scientific foundations, and critique on readability research has been fierce ever since (Anderson & Davison, 1988; Bailin & Grafstein, 2001; 2016; Bruce, Rubin & Starr, 1981; Davison & Kantor, 1982; Duffy & Kabance, 1982; Jansen & Lentz, 2008; Kintsch & Vipond, 1979; Klare, 1976a; Noordman & Vonk, 1994; Redish & Selzer, 1985; among others). Below we will discuss some of these critiques. For expository reasons, we divide them in three groups: criticisms focusing on the lack of causal relevance of predictors, criticisms focusing on the disregard of reader characteristics and criticisms focusing on statistical issues. However, before we turn to these issues we must first note a crucial difference between two areas of application: *readability prediction* and *readability improvement* (cf. Klare, 1984).[3] This distinction lies at the core of many critiques.

[2] Throughout this dissertation we will refer to any objective measurement of comprehension or processing ease as behavioral data. This does not include self-reports (i.e., judgments) of the reader.

[3] Klare refers to this distinction as 'prediction' versus 'production'.

1.1 Prediction versus improvement

Readability prediction concerns assessing whether a text is appropriate for a certain target population. It is therefore very popular with educators and publishers, who use it to select and categorize materials. Readability improvement, on the other hand, concerns identifying how a text can be improved to match the target population. Writers can check whether their text is at the level they presume and if not adapt it until it is. These two areas of application pose different requirements. For prediction it is not crucial to be able to explain why a text suits the reader or not, but for improvement it is. Furthermore, text improvement depends on stylistic and structural text features that can be modified while retaining the text's content. Readability prediction, on the other hand, may use all kinds of features, including for instance the text topic or the text's author.

The difference between prediction and improvement can also be illustrated by distinguishing two different levels of complexity: *conceptual difficulty* and *stylistic difficulty*.[4] Conceptual difficulty is the difficulty level of the *message* which the text is trying to convey (i.e., the content). For example, a text on 'genetic engineering' will probably contain concepts that are harder to understand than the concepts in a text on 'picking daisies'. *Stylistic difficulty* concerns the *manner* in which the message is conveyed. Even an easy concept can be written down in a way it becomes incomprehensible (Bormuth, 1966). While readability prediction is concerned with both components, readability improvement is only concerned with stylistic difficulty. This is because the conceptual difficulty of a text is a given: it cannot be 'improved'. A writer has certain ideas about what he wants the reader to know. Concepts cannot simply be 'left out' because they might be hard to understand.[5] In this sense, improving text readability is limited to decreasing the stylistic difficulty of the text.

Although the predictors used in readability formulae may look like stylistic features, they have been designed for readability prediction and not for improvement. So when writers use readability formulae for improvement purposes (i.e., writing to the formula or fooling the formula; Bruce & Rubin, 1988; Klare, 1976a; Noordman & Vonk, 1994), it is uncertain whether this will actually help make the texts more readable. Especially, because predictors do not require causal relevance in order to be effective.

---

[4] See also the dualistic approach to style (Leech & Short, 2007).

[5] Of course a text can be simplified conceptually to accommodate different types of readers, for example when a text for 5th-graders is adapted for 4th-graders, but then the writer has a different idea about what the 4th-graders should know compared to what the 5th-graders should know.

1.2 Lack of causal relevance

A particularly frequent criticism of traditional research is that the used predictors are not causally relevant to comprehension problems (Anderson & Davison, 1988). At best, they correlate with the true, underlying causes (Kintsch & Vipond, 1979). The predictors used in readability formulae lack intelligence: they do not reflect syntactic, semantic and especially not structural features that actually cause text difficulty. This is problematic for a number of reasons. For prediction purposes, these 'shallow predictors' will yield wrong predictions in certain circumstances. A popular predictor like 'sentence length' is in fact an unintelligent measure of sentence complexity. Due to its correlation with sentence structure, sentence length is a relatively strong predictor of readability (Bailin & Grafstein, 2016; Davison et al., 1980; Gough, 1966). However, when length is kept constant as in Sentence (1) and (2) below, it becomes clear that there is no causal relation between sentence length and readability. Both sentences contain the same 18 words, but most readers will have more trouble understanding (1) than (2). Nevertheless, based on sentence length, no difference would be predicted between (1) and (2).

(1) The inspector decided, because Bob had failed to send in the appropriate form, to reject the claim.

(2) The inspector decided to reject the claim, because Bob had failed to send in the appropriate form.

A second problem with such shallow predictors is that they suggest that texts can be improved by crude, superficial repair work like chopping sentences in half, replacing long words by shorter (semi-)synonyms or by deleting words. The lack of a causal relationship between linguistic features and readability is why writing to the formula does not always improve readability. Or as Klare (1984) noted: "*merely shortening words and sentences to improve readability is like holding a lighted match under a thermometer when you want to make your house warmer*" (pp.717-718). Even more problematic is the fact that some 'repairs' have negative consequences. For example, splitting up sentences can damage semantic and/or structural ties (Davison & Kantor, 1982; Liu, Kemper & Bovaird, 2009). When the two clauses in (3) are split up into three separate sentences (4), the reader is left to infer how these sentences relate to each other. But traditional readability formulae do not go beyond the word and sentence level, so they are insensitive to intersentential dependencies (Bailin & Grafstein, 2001; Davison & Kantor, 1982; Sanders & Noordman, 1988). These types of 'repairs' can thus result in a text which is in fact harder to understand than the original (Land, Sanders & Van den Bergh, 2008).

(3) Because he had to work at night to support his family, Paco often fell asleep in class.

(4) Paco had to make money for his family. Paco worked at night. He often went to sleep in class.

(Ross, Long & Yano, 1991, p.2)

Especially discourse features seem to suffer from shallow adaptations, because there are no predictors that cover discourse markers, coherence relations and global text structure. This in spite of the fact that many comprehension studies have provided empirical support for the importance of coherence and cohesion in text comprehension (Britton & Gülgöz, 1991; Davison & Kantor, 1982; Degand & Sanders, 2002; Kintsch & Vipond, 1979; McNamara, Kintsch, Butler Songer & Kintsch, 1996; Liu et al., 2009; Sanders, 2001).

1.3 Disregard for the role of the reader

Traditionally readability has been viewed from the perspective of the text. However, a text's difficulty is not just determined by characteristics of the text; text and reader characteristics interact (Kintsch & Vipond, 1979; McNamara et al., 1996; Meyer, 2003; Noordman & Vonk, 1994; Oakland & Lane, 2004). Different readers have different needs and they bring different sets of skills, experience and knowledge to the table. As a result, they respond differently to texts and text features. For example, while connectives help readers with low levels of prior knowledge, they do not help readers with high levels of prior knowledge (Kamalski, Sanders & Lentz, 2008; McNamara et al., 1996; McNamara, 2001).

Only a few readability studies have tried to take reader characteristics into account (e.g., Mikk & Elts, 1999; Zakaluk & Samuels, 1988). Usually, readability is a game of averages and individual differences are discarded (see Section 1.4). In addition, readers are dynamic. They can adopt different strategies of reading, depending on the text genre and their reading goals. Unfortunately, it is often unclear for what type of text a readability formula gives valid predictions. This means that differences between text genres and reader goals are also ignored.

Another area in which readers are often neglected is in the kind of data chosen to empirically validate readability predictors. In order to determine the relative weight of predictors, readability studies need comprehension or processing data from the target population. These data can be collected by presenting a large number of texts to a large number of people and measuring how well they understand these texts (cf. Bormuth, 1969; Crossley, Dufty, McCarthy & McNamara, 2007; Dascalu et al., 2014; Nelson, Perfetti, Liben & Liben, 2012;

Staphorsius, 1994). But collecting such data takes time and effort, and as a result only few studies calibrate predictors using actual behavioral data. As an alternative, some studies have used crowdsourcing technology to collect human judgments of readers (Crossley, Skalicky, Dascalu, McNamara & Kyle, 2017; De Clercq & Hoste, 2016; De Clercq et al., 2014; Pitler & Nenkova, 2008). Readers are presented with two different texts and have to judge which text they find easier to understand. Although this technique enables researchers to quickly gather information from a large number and diverse set of participants, self-reported comprehension is not the same as actual comprehension (Cromley & Azevedo, 2006; Glenberg, Wilkinson, Epstein, 1982; Kintsch & Vipond, 1979). Still, these studies do use insights from the target population. Many studies, however, use proxies like expert judgments or graded corpora. Graded corpora contain texts that have been indexed by experts on for instance U.S. grade level or L2-level (Collins-Thompson, 2014; Crossley, Allen, & McNamara, 2012; Feng, Jansche, Huenerfauth, Elhadad, 2010). Expert judgements can be problematic because experts do not always agree when judging text on readability and there is no way of knowing whether the judgments of these experts line up with the actual readability level for the target reader. In fact, studies have shown that experts do a poor job of predicting actual reader problems (De Jong & Lentz, 1996; Lentz & De Jong, 1997; Feng, 2010). Furthermore, some graded corpora incorporate readability measures when indexing their texts. This creates a bias towards these measures and results in circular reasoning when these corpora are used in readability research (see Collins-Thompson, 2014).

1.4 Statistical issues

Even if researchers collect comprehension data from target readers, there are statistical issues due to the way these data are handled. One problem is that in most readability studies all measures are averaged over readers and over (parts of) the text. Many readability studies claim to explain between 60% and 80% of the variance in text comprehension scores, but these proportions are a vast overestimation of the strength of the predictors (Anderson & Davison, 1988). Readability studies aggregate comprehension scores over participants which eliminates the variance in individual scores and inflates the predictive power. In addition, it blinds researchers to interactions between text and reader characteristics, since they only see mean text scores. It is also common to average values for linguistic features within texts. A feature's mean value is used to represent the entire text. This causes a systematic underestimation of the variance within texts. Furthermore, for improvement purposes, it makes localizing the problematic regions in a text much harder.

A second statistical issue is that readability research often uses texts that have a wide variety of difficulty levels, subjects and/or genres. This increases the

explanatory power of the linguistic features in comparison to texts that are more similar. Rodriguez and Hansen (1975) used Bormuth's predictors (1966; 1969) and showed that when texts and readers are more restricted, the explained observed variance dropped by 50%.

1.5 Previous readability research for Dutch texts

Research and development of Dutch readability tools has been subject to most of the critiques mentioned above (De Clercq & Hoste, 2016; De Jong & Schellens, 1995; Hacquebord & Lenting-Haan, 2012; Jansen, 2005; Jansen & Boersma, 2013; Jansen & Lentz, 2008; Jansen & Woudstra, 1979; Kraf et al., 2011; Renkema, 1982; Sanders & Noordman, 1988; Staphorsius, 1994; Van Oosten, Tanghe & Hoste, 2010)[6]. For adult readers, the most popular formula is the Flesch-Douma ((5); Douma, 1960). Douma revised the Flesch formula to make it suitable for Dutch. Because Dutch sentences and words were approximately 10% longer than English equivalents, Douma reduced the strength of the word and sentence length components by 10%. This revision lacked any empirical support.

(5) Reading ease Flesch-Douma = 206.835 - 0.77 * word length - 0.93 * sentence length

A readability formula which was the product of extensive research is the CLIB-formula (6) of Staphorsius (1994). The CLIB is designed for primary school children and is based on almost 10 years of research. The formula was calibrated and validated by collecting cloze scores of children (ages 7 – 12) on a collection of 240 texts. The tests were distributed equally over children of different reading proficiency levels and each child took two different cloze tests.

(6) CLIB = 46 - 6.603 * word length + 0.474 * percentage highly frequent words - 0.365 * Type/Token ratio + 1.425 * inverse sentence length

The CLIB was the product of thorough research and a strong empirical design. Nevertheless, it has its limitations. While it does well explaining average text scores ($r^2 = 0.72$), Kraf and Pander Maat (2009) showed that when the individual variance of the children is taken into account, the explained variance is cut in half. In addition, it is only applicable to children between 7 and 12 years of age. CLIB was not validated for older readers.

[6] This does not mean that all Dutch studies employ a similar methodology. We can differentiate between studies that use a traditional approach using regression modelling (e.g., Staphorsius, 1994) and studies using advanced machine learning techniques (e.g., De Clercq & Hoste, 2016), see also Chapter 6.

2. Our approach to readability

2.1 Improved extraction of linguistic features

While in the 20th-century we were limited to simple shallow linguistic features like 'letters per word' and 'words per sentence', with the boom of computational linguistics we can now choose from a wide range of complex measures and the number of features is ever increasing. We are able to assign words to word-categories ('Part-of-Speech Tagging'), conduct syntactic and morphologic analyses, perform memory-based word prediction, and model language using sophisticated techniques. These technologies open doors to compute a wide range of measures, like the density of different types of conjunction, dependency lengths and the number of personal references within a text. In addition, we have access to a number of linguistic databases which offer information regarding words and word combinations. Frequency lists of words and lemmas, lists of concrete nouns and adjectives, classifications of connectives, and words that indicate situation model dimensions, are only some of the lists easily available to us now.

For English, strides were made with the creation of Coh-Metrix (Graesser et al., 2004; McNamara, Graesser, McCarthy & Cai, 2014) a computational tool primarily designed to measure indices of coherence but which also includes indices that measure lexical and syntactic complexity. Coh-Metrix and subsequent spin-offs have shown their success in identifying differences between text genres (Graesser & McNamara, 2011), L2-simplified and authentic texts (Crossley, Louwerse, McCarthy & McNamara, 2007); and also in predicting readability judgments (Crossley et al., 2017) and comprehension scores (Crossley, Dufty et al., 2007). For Dutch, developments have lagged behind a little and automatic analysis was fragmented across tools and databases. Recently, these tools were brought together in a single analysis tool named 'T-Scan' (Kraf & Pander Maat, 2009; Pander Maat et al., 2014). T-Scan is an analysis tool which automatically extracts more than 400 linguistic features from text. Enriched by insights from discourse studies, it currently provides features on *lexical complexity*, *sentence complexity*, *referential and relational coherence*, *concreteness*, *person-oriented writing* and *(word) probabilities*. Furthermore, T-Scan returns measures calculated over the whole text, but also returns measures on word, sentence and paragraph levels. As a result, T-Scan provides data that shows variance within texts and that enables us to locate (potential) problems within a text. What T-Scan does not return is a prediction of readability. Currently, T-Scan is purely an extraction tool: it only describes texts. Tools like T-Scan can be used for comparisons (between sentences, texts, genres, etc.) but not for judgments concerning readability.

2.2 Validation of linguistic features

Once the linguistic features are extracted, they have to be calibrated to determine which features at which levels are problematic for readers. For this purpose we collect behavioral data from our target population. The approach we put forward differs from previous research on four major points.

In contrast to traditional research, we do not limit ourselves to one aspect of readability. Not only do we assess *text comprehension* (i.e., the 'off-line reading product') but we also measure *processing ease* (i.e., the 'on-line reading process'). In our view, a text has a high level of readability if a reader both understands a text and is able to do so with a reasonable amount of effort. To collect the off-line comprehension data, we use a specially designed type of cloze test. To collect the on-line data, we use eye movement registration.

Secondly, while we do want to know *what* makes a text difficult to process, we specifically want to know what makes a text difficult for *which reader*. We want to know whether effects of linguistic features can be generalized across readers. We therefore include low-proficiency and high-proficiency readers from different ages in our study. In addition, standardized tests will be administered to approximate their reading ability which can be used as a reader characteristic within our analyses. By letting readers read multiple texts, we get a better understanding of differences between readers and how these differences interact with text difficulty.

Thirdly, even though readability prediction does not necessarily require a causal relation between a predictor variable and text comprehension (see Section 1.1) we are interested in causal effects of linguistic features. Preferably, our predictors are causally relevant, as is substantiated by experimental and theoretical evidence. For readability improvement purposes it is important to see how much effect linguistic features actually have when text content is kept equal. For instance, lexical complexity has often been found an important determinant of text difficulty. However, reducing the lexical complexity of a text tends to have smaller effects on comprehension than readability formulae predict, mainly because lexical complexity is highly dependent on text content. It is important to find out whether these effects can be generalized over large numbers of texts and readers.

The only way to verify that a linguistic feature has a causal effect on readability is to test it in a controlled experiment: "*...only the independent variable of interest is changed in order to assess its effect. This means that content (what is said) should be held constant while structure (how it is said) is varied.*" (Klare, 1976a, p.133). To this end our design includes three large scale experimental studies. Our texts are carefully manipulated on one of three text features: lexical complexity, syntactic complexity and coherence marking. This results in two versions of each text. The content as well as other stylistic features are not manipulated. As a result, text versions are at the same level of conceptual difficulty,

but they differ in stylistic difficulty. This design allows us to differentiate between causal effects of features and correlational relationships with readability. In addition, our experiments will show how large these causal effects really are and how much stylistic interventions can improve readability.

The final difference with traditional readability studies is that behavioral data will not be averaged over readers and texts. The extraction tool T-Scan enables us to analyze text on different levels and our two validation measures - cloze and eye movement - are also able to offer us localized data. These data will be analyzed using multilevel analysis techniques. These techniques give us the opportunity to account for reader and text variance, instead of ignoring it.

3. Methodological design of the study

Our approach led to an integrated methodological design which includes experimental studies as well as correlational studies.

Our materials included two types of authentic texts, both informative and both important for adolescent readers: 1) educational textbook texts written especially for secondary education, and 2) public information texts written for the general public but containing information that was relevant for adolescents. 146 educational texts ranging 300 to 400 words were collected from history, geography, Dutch language and economy textbooks. 120 public information texts were collected from websites and brochures of Dutch government institutions and government affiliated organizations. The text topics were related to health (e.g., diabetes, donor registration), the environment (e.g., pest control), public safety (e.g., criminal justice), and other socially important matters. Out of this collection of texts we quasi-randomly selected 30 educational textbook and 30 public information texts. We used a quasi-random selection procedure to ensure a diverse sample of texts. The values of linguistic features had to vary across texts to investigate the relationships between features and readability. To make sure there was some diversity within the sample we adapted the selection procedure of Liu, Kemper and Bovaird (2009, p. 656). Liu et al. divided their 200 texts along two dimensions by their quartile scores. This resulted in a matrix of 16 cells. One text was randomly selected from each cell. We adapted this procedure by adding another dimension and using tertile scores rather than quartile scores. This resulted in a 3^3-matrix with 27 cells.

Because our manipulations would automatically create variance in lexical complexity, syntactic complexity and relational coherence of our texts, we chose three other text features as dimensions: referential coherence, concreteness and personal style. Referential coherence has proven to be an important feature in

discourse processing (Kintsch & Vipond, 1979), but like concreteness, it is hard to manipulate referential coherence in natural texts without altering its content or altering other stylistic features at the same time. By using these features as dimensions, we made sure that our text selection includes texts that intrinsically vary on concreteness and referential coherence. The final dimension, personal style, was chosen because it was observed that in both the educational textbooks and in the public information texts a number of texts were written in a very personal style, while others were not and some were written in a mixed style. We thought personal style to be a relevant feature, as it has also been suggested as a predictor in prior readability research (e.g., Flesch, 1943).

All texts were analyzed with T-Scan and within each genre the texts were divided into tertiles. The 27 cells in our matrix were filled with two or more texts. One text was randomly selected out of every cell. Three more texts were randomly selected to bring the total of texts to 30 per genre. Next, the texts were divided over three manipulation conditions: lexical complexity, syntactic complexity and coherence marking. Texts were randomly assigned to a manipulation condition and not assigned based on their potential for that particular manipulation. Although no requirements were imposed with regard to the strength of the manipulations, we did impose a minimum number of manipulations that had to be possible in the text.[7] If a text did not reach this minimum number, another text was randomly selected from the same cell in the matrix.

Two versions of each text were created: a relatively easy version in which readability was supposedly increased and a relatively difficult version in which readability was supposedly reduced.[8] The three different manipulation conditions formed three separate experimental studies focusing on causal effects of linguistic features. The data of the experiments was combined and provided the data for the correlational study (see Table 1).

Table 1: Distribution of texts over genre and comprehension studies

	Lexical complexity (Ch 3)	**Syntactic complexity (Ch 4)**	**Coherence marking (Ch 5)**	**Text complexity (Ch 6)**
Educational text	10	10	10	30
Public information text	10	10	10	30
Total	20	20	20	60

[7] This number depended on the manipulation condition. For instance, in the lexical complexity condition we aimed to manipulate 20% of the content words. We allowed a 10% deviation, meaning that 18 to 22 percent of the content words had to be manipulated.

[8] The easy and difficult text versions of our 60 texts will be made available via DANS (Data Archiving and Networked Services; https://dans.knaw.nl/). A list of the original texts can be found in Appendix 1.

The sixty texts – in two versions – were turned into cloze test and presented to a total of 2926 eighth through tenth grade students enrolled in different levels of the Dutch educational system. After clean-up procedures, the final dataset included cloze scores of 2749 students (see Table 2). For 2403 of these students standardized reading ability scores were available. Each student filled out 4 cloze tests. All cloze data was collected by our partner Cito.

Table 2: Distribution of participants over Grade and Level of education[9] for comprehension studies

	Pre-voc. (low)	Pre-voc. (medium)	Pre-voc. (high)	General education	Pre-uni.	Total
Grade 8	77	277	491	144	178	1167
Grade 9	166	277	316	246	222	1227
Grade 10	-	-	-	209	146	355
Total	243	554	807	599	546	2749

A selection of the texts was also used to collect eye-tracking data. We selected four lexically and four syntactically manipulated public information texts, once again in two versions (see Table 3) and presented them to 181 ninth grade students (see Table 4). Each student read all eight texts, but only in one text version.

Table 3: Distribution of texts over eye-tracking studies

	Lexical complexity (Ch 3)	Syntactic complexity (Ch 4)	Coherence marking (Ch 5)	Text complexity (Ch 6)
Public information text	4	4	-	8

Table 4: Distribution of participants over Level of education for eye-tracking studies

	Pre-voc. (low)	Pre-voc. (medium)	Pre-voc. (high)	General education	Pre-uni.	Total
Grade 9	-	47	-	54	80	181

4. Chapter overview and reading guide

The studies presented in this dissertation are designed to function as part of the overall project as well as stand-alone investigations of readability. Chapters can be read individually and certain information is repeated across chapters to facilitate stand-alone reading. However, references to the overall project and other chapters are made to facilitate integrated reading and to show how each chapter fits in with the other chapters in this dissertation.

[9] These five levels are ordered from practice oriented education to theoretical oriented education. For more information on the Dutch educational system see EP-Nuffic (2015).

In Chapter 2 we present our main method of investigation: a new cloze procedure called the 'Hybrid Text Comprehension cloze' (HyTeC-cloze). This method was especially developed for the LIN-project in order to test comprehension differences between texts, texts versions and readers in a reliable way. We start out with a review of different comprehension methods, after which we focus our attention to cloze procedures and introduce the HyTeC-cloze. The chapter ends with an evaluation of the procedure using empirical data collected for the LIN-project.

In Chapters 3 and 4 we zoom in on two linguistic features that have a strong history in readability studies. They are the strongest predictors in readability formulae: lexical complexity and syntactic complexity. Although lexical and syntactic features are strong predictors of text difficulty, it is questionable whether reducing these types of complexity improves readability of the text. Chapters 3 and 4 both include a cloze and an eye-tracking experiment, focusing on comprehension and processing ease effects respectively. In these experiments we test whether modifying lexical and syntactic features of a text can improve its readability and whether these effects can be generalized across multiple texts and adolescents differing in reading proficiency.

In Chapter 5 we focus on coherence, a concept that has become an important concept in reading research in the last decades, but which has been found to be difficult to integrate in readability studies (e.g., Kintsch & Vipond, 1979; Noordman & Vonk, 1994). With the development of tools that include linguistic indices of coherence – like Coh-Metrix (Graesser et al., 2004) and T-Scan (Pander Maat et al., 2014) – inclusion of coherence indices in readability formulae has become a real possibility. One of these indices, connectives, have been found to increase processing ease. When appropriate connectives are added between text segments, readers process the second segment faster than when the connective is left out (Van Silfhout, Evers-Vermeul, Mak & Sanders, 2014). Because of the large number of on-line processing studies that have investigated connectives, we limit our investigation to the off-line effects of connectives. Results of off-line studies have been less robust and it may be that connectives only positively affect on-line processing of text.

Finally, in Chapter 6 we turn our attention to readability prediction. Leaving the experimental setups behind, we pull all our collected data for a correlational study. Linguistic features are validated based on the data of chapters 3, 4 and 5. The subset of features that yields the best prediction of text comprehension will be presented as the Utrecht Readability model ('U-Read').

2 The Hybrid Text Comprehension cloze: Validity and reliability

The question of how to assess text comprehension is one that does not have a simple answer. Whether the assessment is for research or educational purposes, each testing situation is slightly different and hence the 'best' assessment method has to meet different practical and theoretical requirements each time. In this chapter we present a new cloze procedure which we believe to be applicable to assess text comprehension when examining a large number of texts and/or readers. As a measure of text comprehension, the cloze test is not widely accepted and often critiqued (e.g., Klein-Braley & Raatz, 1984; Pearson & Hamm, 2005; Shanahan, Kamil &Webb Tobin, 1982). However, many of these critiques are only valid for standard cloze tests in which every X^{th} word has been deleted or extend to more accepted assessment methods as well. As we will show, the standard cloze test is just one type of cloze and critiques are not always valid for other cloze types. After a short review of widely accepted methods for text comprehension assessment, we will turn to cloze tests and the particular cloze test configuration that we propose: the Hybrid Text Comprehension cloze test (HyTeC-cloze). We will evaluate this procedure using data collected for the LIN-project.

1. Comprehension measures

1.1 Comprehension questions

Text comprehension is most often assessed with comprehension questions. These questions are designed to assess whether key concepts and relations within the text have been understood. There are however many types of questions and not all implementations are necessarily of high quality and validity (Graesser, Ozuru & Sullins, 2010; Kintsch & Vipond, 1979). For instance, questions vary in the level of abstractness of the required information (from concrete 'entities' to highly abstract 'themes'; Mosenthal, 1996), in the cognitive processes that are involved (e.g., recognition, application, evaluation; Bloom, 1956) or in the level of mental representation they assess (surface code, text-base or situation model; Kintsch, 1998; 2012). Construction of these questions is time consuming: texts have to be analyzed to determine the key concepts and questions have to be carefully formulated and tested to make sure that they are interpreted as intended. If the answer format is multiple-choice, response options have to be constructed as well. A well-constructed

test, however, can offer valuable insights into which concepts and relations within a text are understood and to what extent.

Comprehension questions are often used in standardized tests of reading comprehension. Overall, researchers and teachers seem to have much faith in these tests, even though they have often been found to measure different constructs (Cutting & Scarborough, 2006; Keenan, Betjemann & Olson, 2008), are often highly passage independent (Coleman, Lindstrom, Nelson, Lindstrom & Gregg, 2010; Keenan & Betjemann, 2006), do not agree when identifying readers with reading disabilities (Keenan & Meenan, 2014) and correlate .31 to .79 at best (Cutting & Scarborough, 2006; Keenan et al., 2008). Most of these tests use multiple-choice comprehension questions. Advantages of this answer format are the speed and objectivity of rating. The disadvantages are that a correct understanding of the text or even reading the text is not always necessary. Test takers are able to perform well above chance on multiple-choice tests without reading a single word of the text (Coleman et al., 2010; Keenan & Betjemann, 2006; Hanna & Oaster, 1980; Katz, Lautenschlager, Blackburn & Harris, 1990). The answer can be deduced by looking at the semantics of the question in combination with world knowledge. Of course, the difficulty depends on the distractors used in the test as well. Distractors should be on the right level, meaning they should not be too easy but also not too hard (e.g., distractor and key are so similar that discrimination hinges on an incredibly small nuance in meaning). In addition, test takers may have to deal with distractors that they would not have imagined themselves (Alderson, 2000; Ozuru, Best, Bell, Witherspoon & McNamara, 2007). Changing the answer option to 'open ended' does eliminate these problems, but rating the questions is more time consuming and without a strict scoring format, the reliability of the test can suffer due to subjectivity of the rater.

1.2 Recall and think-aloud procedures

To circumvent the use of questions which may introduce their own source of difficulty, test makers can use recall or think-aloud procedures. These methods do not suffer from interference from questions or response options and are therefore viewed as a 'purer' measure of comprehension (e.g., Alderson, 2000). In free-recall procedures, readers are asked to read a text and to verbalize afterwards what they remember of the text without looking back. The recall transcript shows what a reader has remembered, but also how that information is structured in memory.

In think-aloud procedures, readers are instructed to verbalize their thoughts while reading. Think-aloud transcripts are often regarded as a process measure to show how comprehension is reached and to reveal whether certain inferences were made or not. Large disadvantages of both methods are that scoring the transcripts is very labor intensive and scoring can be very subjective. Extensive scoring protocols

are vital and raters have to be trained to follow the protocols (Millis, Magliano & Todaro, 2006).

1.3 Ordering, sorting and mental model tasks

Over the last 25 years, researchers have started to develop comprehension measures that are focused on assessing the deepest level of comprehension: the situation model (e.g., Britton & Gülgöz, 1991; Kamalski, 2007; McNamara, Kintsch, Songer & Kintsch, 1996; Van Silfhout, 2014). This is thought to be the richest mental representation of a text because linguistic information is integrated with world knowledge. Ordering tasks require readers to put phrases or events in the correct order. Sorting tasks require readers to organize key concepts from the text into groups of similar concepts. For example, if a text discusses multiple causes and solutions to global warming, all causes should be grouped together and all solutions should be grouped together. Mental modeling tasks are similar, but a schema or diagram is provided to structure the grouping (Kamalski, 2007). Problematic for all these tasks is that multiple categorizations may be acceptable or categorizations may be only partially correct. Furthermore, these tasks cannot be used on just any text: they are not widely applicable.

1.4 Cloze tests

Cloze tests come in various forms. They have in common that "*bits of some discourse are omitted and the task set the examinee is to restore the missing pieces.*" (Oller & Jonz, 1994, p. 19). Subsequently, the answers of the examinee are scored. The score can be seen as a measure of the readability of the text, but also as a measure of the reading ability of the examinee (e.g., O'Toole & King, 2011; Taylor, 1953). As such, cloze tests have been popular in readability studies as well as in language testing endeavors. Regardless of its popularity, the cloze test is not widely accepted among scholars as a valid measure of text comprehension. The validity of this method has long been and still is under discussion (e.g., Brown, 2013; Chen, 2004; Gellert & Elbro, 2013; Greene, 2001; Kobayashi, 2002ab; 2004; Oller & Jonz (Eds.), 1994; O'Toole & King, 2010; 2011; Trace, Brown, Janssen & Kozhevnikova, 2017). Critics believe that cloze is not sensitive to intersentential constraints and that it predominantly measures lower-order skills (i.e., grammatical and linguistic knowledge; Alderson, 1979a; Kintsch & Yarbrough, 1982; Klein-Braley & Raatz, 1984; Pearson & Hamm, 2005; Shanahan et al., 1982). They claim cloze is not a valid measure of text comprehension because it does not measure discourse level representation.

Advocates of the cloze challenge this view and claim that a large percentage of cloze gaps does require information processing across sentences (Brown, 1983; Chihara, Oller, Weaver, & Chávez-Oller, 1977; Cziko, 1983; Henk,

1982; Jonz, 1994; among others). Brown (1983) showed that 56 to 70 percent of cloze items in standard cloze tests are cohesive items (following the classification of Halliday & Hasan, 1976). In addition, Jonz (1994) found that on average 32% of the gaps requires information beyond sentence borders. Finally, analyses of item difficulty show that passage-level variables influence the difficulty of individual cloze gaps (Chávez-Oller, Chihara, Weaver & Oller, 1985; Kobayashi, 2002a; Trace et al.; 2017). All these findings support the stance that standard cloze tests measure beyond sentence boundaries.

An important reason why the discussion surrounding the validity of the cloze is still going on is that there is no such thing as 'the cloze test'. There are in fact many different types of cloze tests. The way the test is configured will strongly influence the validity of the test as a measure of text comprehension (Alderson, 2000; Watanabe & Koyama, 2008). In order to determine the true validity of a particular cloze test, we need a systematic classification of cloze. This classification scheme will be presented in the next section.

2. Types of cloze tests

Classification schemes of cloze tests are rarely complete (see Watanabe & Koyama, 2008). The most systematic classification available to us is by Staphorsius (1994), who uses nine characteristics to describe his cloze test. We supplemented his list with the characteristics 'Deletion ratio' and 'Scoring' (Table 1). Most of the characteristics interact. For instance, deletion ratio influences the number of gaps and the answer format influences how deletions are marked. We will discuss these characteristics together.

Table 1: Cloze test characteristics (adapted from Staphorsius, 1994, p.145)

Characteristics	
1.	Deletion strategy
2.	Deletion ratio
3.	Deletion distance
4.	Number of gaps
5.	Number of starting points
6.	Excluded text segments
7.	Excluded words
8.	Pre-cloze or post-cloze testing
9.	Open or closed answer format
10.	Marking of deletion
11.	Scoring

2.1 Deletions

The first and foremost characteristic is the *deletion strategy*. Words can be deleted following a random, a mechanical or a rational strategy.

2.1.1 Random deletion strategy

With *random deletion* every word in a text receives a number and with the help of random lists or computer programs, a set of numbers is selected and the corresponding words are turned into gaps. During the whole procedure, every word has the same chance of being selected. The biggest advantage of this deletion strategy is that it is free from text and experimenter biases. Its greatest drawback is that in completely random procedures an experimenter has no control at all over the distribution of the gaps over the text. As a result, gaps can succeed each other immediately (resulting in less immediate context to fill in those gaps) and gaps can be unevenly distributed throughout the text. This strategy is therefore not used very often.

2.1.2 Mechanical deletion strategy

The *mechanical deletion strategy* is very popular because of its relative ease of implementation and its 'objective' nature (i.e., low experimenter bias). With mechanical deletion, every X^{th} word is deleted. This is usually the fifth word, but there are studies that go as far as every 18^{th} word (see Watanabe & Koyama, 2008). Since the deletion distance is fixed, gaps cannot immediately succeed each other and they are dispersed at regular intervals throughout the whole text. As with random deletion, it does not matter what type of word is deleted: in principle all words can be sampled. Of course exceptions can be made. Often the title and first sentence are left intact to provide the reader with at least some contextual support. Also, very unpredictable words like numbers and proper names are usually left alone.

Apart from these exceptions, experimenter bias is reduced to one element: the starting point. With a deletion ratio of five, the experimenter can delete the first, sixth, eleventh word etcetera, but can also start with the second, third or even fifth word. So with a deletion ratio of five, it is possible to create five cloze versions of the same text. In several studies it was found that these cloze versions do not differ in the proportion of lexical, syntactic and cohesive items that they sample (Bachman, 1982; Brown, 1983; Jonz, 1994, O'Toole & King, 2010).[1] However, this does not necessarily mean that the cloze versions are of similar difficulty. Shifting the starting point changes the whole test, because a different sample of words is drawn from the text (Brown, 1993). One lexical item is not necessarily equal to

[1] Note that this finding is to be expected. When the test is long enough the sample will approach the actual distribution of word types in the text.

another lexical item and the same holds for syntactic and cohesive items. Therefore, changing the starting point of a cloze test may influence the score of an examinee, simply because the individual items differ in their level of difficulty. In a study with 556 participants, O'Toole and King (2010) found that cloze scores and cloze score distributions of the same text differ significantly when different starting points are used. Thus, predictions of readability or reading ability may differ depending on the specific cloze version that is selected. Similar results were previously found by Brown (2002), Alderson (1979a), and Porter (1978). Cloze versions of the same text can thus lead to significantly different results. Therefore, a single version of a mechanically clozed text might not represent the actual difficulty level of a text or reading ability of a reader. To minimalize this 'error', it is advisable to use more than one cloze version per text. Ideally one should use as many versions as possible. That is: for a deletion ratio of five one should use five versions, each with a different starting point (Bormuth, 1969; O'Toole & King, 2010). In that case every word in the text is sampled. Depending on the deletion ratio this may not be practical, in which case just a couple of versions are selected (e.g., Porter, 1978; Staphorsius, 1994).

2.1.3 Rational deletion strategy

The third deletion strategy is *rational deletion*, in which the experimenter selects which words will be deleted. Selection can be limited to a specific grammatical class (e.g., articles, prepositions or verbs) or based on a specific hypothesis (e.g., which type of information is needed to fill in the gap; Bachman, 1985; Gellert & Elbro, 2013; Levenston, Nir and Blum-Kulka; 1984). Therefore, the rational deletion strategy is often used when experimenters want to measure specific skills or have specific ideas about what types of items measure 'true' text comprehension. By tailoring the cloze test, a lot of noise can be eliminated from the data. For example, one criticism of mechanical and random cloze tests is that they contain a lot of function words and that these words represent grammatical knowledge of the examinee rather than text comprehension. Such items can be reconstructed with grammatical knowledge alone; the discourse context and the comprehensibility of the text do not factor into it. Therefore, some researchers only use content words as gaps in their cloze test. As such, they largely eliminate the noise created by grammatically solvable items and are left with a 'purer' measure of text comprehension. Gellert and Elbro (2013) and Levenston et al. (1984) selected items that show coherence (e.g., pronominal references, connectives). Bachman (1985), on the other hand, selected words based on which information was necessary to reconstruct the word. He maximized the number of items that needed information across clauses or sentences. The drawback of rational procedures is that the classification of each potential item can be subjective, as is the selection of one

potential item over another. Jonz's analysis of Bachman's gaps was significantly different from Bachman's original analysis (Jonz, 1994). If the rationale of a rational cloze test is not specified to the letter or not sufficiently grounded in theory, the sample becomes very subjective.

2.2 Lay-out and answer format

When the gaps are selected, there are different formats in which the test can be presented. Firstly, the examinee may be asked to read the original non-clozed text first. Only afterwards – usually after a delay – he is presented with a cloze version of that same text. This is called a *post-cloze test,* since the cloze follows the reading of the intact text. However, most tests are *pre-cloze tests:* the first time the examinee sees the text, the gaps are already in place.

Secondly, the answer format of cloze tests can be different. A test can be closed or open. In open tests, examinees have to produce the words that were deleted themselves. In closed tests, examinees are given possible answers from which they have to choose the correct option. This can be done in a standard multiple-choice format (i.e., for each gap, a limited number of options is given) or by putting all answers below the text in a random order (i.e., answers to all gaps serve as distractors). In addition, examinees can be supported by how the deletion is marked. Usually gaps are indicated by blanks with a fixed length (1a), but it is possible to let the length of the blank be determined by the general length (1b) or specific (1c) length of the intended word.

(1) a. The _____ was shining. The children were _____ in the garden.
b. The _____ was shining. The children were _____ in the garden.
c. The _ _ _ was shining. The children were _ _ _ _ _ _ _ in the garden.

2.3 Scoring format

Related to the answer format is the scoring format. Contrary to closed test like multiple-choice tests, open tests can be scored in different ways. The most efficient way of scoring is *exact scoring*. Only originally deleted words – usually including spelling errors – are acceptable in this scoring format. *Semantic scoring* (or *acceptable word scoring*) allows originally deleted words but also semantically correct alternatives. The acceptability of alternative answers is usually scored according to the *global* appropriateness criterion, which means that the answer has to fulfill "*all the contextual requirements of the entire discourse context in which it appears*" (Oller & Jonz (Eds.), 1994, p.416). In contrast, when the *local*

appropriateness criterion is adhered to the answer only has to fulfill the contextual requirements of the immediate sentence.

Exact and semantic scoring are usually dichotomously coded (0 = false, 1 = correct), but a weighed coding can also be used. Miller and Coleman (1967) awarded 3 points for an exact replication, 2 points for a semantic alternative and 1 point when the answer was of the same class as the deleted word. Another scoring form that uses weighed coding is *clozentropy scoring*. Clozentropy "*logarithmically weights the acceptable answers according to their frequency in a native speaker pretest*" (Brown, 1980, p.311). Clozentropy scoring is only used in second language testing.

2.4 Advantages of cloze tests

Cloze tests have specific advantages compared to the other assessment methods mentioned in Section 1. Firstly, they are relatively easy and fast to create and can be used on a wide range of texts. For studies with a large number of materials, this is a big advantage. Furthermore, cloze tests are very suitable to systematically investigate the difficulty of texts. All texts are mutilated in the same way and the difficulty of the items directly transfers from the text (Klare, 1976a). Thus, text difficulty can be assessed by comparing cloze scores of one person on different texts. This is hard to do with standard comprehension questions. They are not directly comparable to one another because every question is different. For a valid comparison, it is important that questions are of equal difficulty. When you ask difficult questions about an easy text and easy questions about a difficult text, the resulting scores do not reflect text difficulty. Another advantage is that while it is hard to ask even 10 intelligent questions about a 300 word text, with cloze we get many testing points which are also spread out over the entire text. That is, cloze tests cover more parts of the text than comprehension questions generally do. Finally, cloze is also more resilient to experimenter biases when it is used in experimental studies. While questions are often designed to be sensitive enough to pick up differences between text versions (i.e., if a question does not 'work', it is altered or removed), cloze tests designed to measure text comprehension are not. In that sense, cloze tests scores provide a more honest representation of the effect of the manipulation.

Of course, all advantages hinge on the right configuration of the cloze test, which is why we developed a new cloze procedure. We believe the HyTeC-cloze procedure is less susceptible to problems previously reported in the literature.

3. Introducing the Hybrid Text Comprehension cloze

The Hybrid Text Comprehension cloze (HyTeC-cloze) was developed as an alternative to the standard cloze and other standard comprehension assessment measures. We configured the HyTeC-cloze to be:

- a valid and reliable measure of text comprehension that is less sensitive to intrasentential (local) constraints than standard cloze tests;
- sensitive to comprehension differences between texts, text versions and test takers with different reading abilities;
- applicable to a wide range of texts;
- easy and fast to create;
- mirroring the difficulty of the text without confounding text difficulty with question difficulty;
- suitable for high and low proficiency test takers.

An overview of the HyTeC-cloze procedure is given in Table 2. The construction manual can be found in English in Appendix 2 and in Dutch in Appendix 3. The rationale behind the procedure is explained in the section below.

Table 2: Standard cloze procedure (Oller & Jonz (Eds.), 1994) versus HyTeC-cloze procedure

Characteristics		**Standard cloze**	**HyTeC-cloze**
1.	Deletion strategy	Mechanical	Mechanical-Rational
2.	Deletion ratio	20%	10%
3.	Deletion distance	Fixed: every 5th word	Varied: at least 1 word in between
4.	Number of gaps (per 300 words)	60	30
5.	Number of starting points	3 to 5	2
6.	Excluded text segments	- Title - First sentence	- Title - First sentence
7.	Excluded words	- Proper names - Numbers	- Locally predictable words - Guess words
8.	Pre-cloze or post-cloze testing	Pre-cloze	Pre-cloze
9.	Open or closed answer format	Open	Open
10.	Marking of deletion	Fixed length marking	Fixed length marking
11.	Scoring	Exact + spelling errors	Semantic + spelling errors

3.1 Deletion strategy

The HyTeC-cloze procedure is partially named after its deletion strategy. It employs a hybrid strategy: *mechanical-rational deletion*. Both mechanical and rational strategies have strong and weak points. Rational deletion allows us to limit the types of words that are sampled, resulting in a cleaner measurement of text comprehension. In addition, rational deletion has a big advantage when it comes to experimental studies. It allows for direct comparisons between experimental versions of the same text. Consider a text that is syntactically manipulated like in (2) and we mechanically cloze this sentence with starting at the 5th word with a fixed deletion distance of five. This means that in (2a) the words *growing, burned* and *wounds* disappear (see (3a)). On the other hand, if we do the same for (2b) the words *its, new,* and *part* are deleted (see (3b)). This would not be a fair comparison between two text versions, because a new level of variance is added on top of another. We would not be able to distinguish between the effects of our manipulation and possible effects of the difference in gaps. Cloze version and text version would be confounded.

(2) a. The tree will, by growing new bark over the burned part, heal its own wounds.
b. The tree will heal its own wounds by growing new bark over the burned part.

(3) a. The tree will, by [……] new bark over the [……] part, heal its own [……].
b. The tree will heal [……] own wounds by growing [……] bark over the burned [……].

(Adapted from Davison & Kantor, 1982, p.192)

In contrast, in rational procedures the same words are naturally deleted in both versions since they are chosen by the experimenter. For example:

(4) a. The [……] will, by growing new bark over the burned [……], heal its own [……].
b. The [……] will heal its own [……] by growing new bark over the burned [……].

If an effect is found in (4), it cannot depend on the difference in gaps, but only on the syntactic manipulation. Of course the order of the gaps may occasionally change, but not the gaps themselves or the available context. For experimental research it is vital that text version and cloze version are not confounded.

Then again, a completely rational procedure also has its drawbacks. The words that are sampled are selected by the experimenter who may (unintentionally) prefer certain words over others or subjectively views parts of the text as more important than others. The result might be a test that does not 'mirror' the text completely. That is: the selection may give "*a false impression of the nature of a text*" (Alderson, 1979a, p.226). On the other hand, objectively mirroring a complete text is one of the strengths of mechanical deletion. Mechanical deletion objectively samples all parts of the text and reflects its overall difficulty. By combining both strategies we capitalize on their strengths.

The HyTeC-cloze procedure starts off with a rational selection of possible cloze gap candidates (Step I; see Appendix 2). The experimenter decides which words are candidates for deletion. The mechanical strategy comes into play in the next phase of the cloze-production process (Step II) when it is decided which specific words out of all the options are deleted. For example, if we only allow content words we would first select all content words (Step I). These words are all candidates, but only a sample of them can actually be used in a test (see Section 3.4). Thus in Step II, it is mechanically decided which of these candidates are to be used in the actual test. This is done in the same way as we would in standard mechanical deletion procedures. Depending on the chosen deletion ratio we could for example take only every 4th candidate. We thus remove the arbitrary (and potentially biased) choice of selecting one candidate over another.

3.2 Deletion ratio and number of gaps

In mechanical cloze tests, it has become standard practice to use a deletion ratio of 1 in 5. Increasing the context surrounding a specific gap does not seem to influence the score (Alderson, 1979b, MacGinitie, 1961; Rankin & Thomas, 1980; Taylor, 1956).[2] A ratio of 1 in 5 is not realistic for rational cloze tests, because we have a lot less candidates for deletion. In addition, if the texts under investigation are manipulated, we are left with even less candidates since words that differ between versions can obviously not become gaps. Another consideration is that for certain participant groups – like young children – it is advisable to use a lower ratio (Robinson, 1981; Staphorsius, 1994). If the ratio is too high for the participant, it will result in a floor effect and no variance can be observed.

Following Greene (2001) and Bachman (1985), ratios between '1 in 9' and '1 in 11' are reasonable adaptations. To test whether this holds for Dutch and for our intended age group (i.e., adolescents), we used a ratio of 1 in 10 in one pretest and a

[2] This finding only seems to hold for analyses where just the items that are present in all deletion ratio versions are compared. Some scholars have found an effect of deletion ratio on relative total scores, but the direction of this effect seems unpredictable (see Alderson, 1979b).

ratio of 1 in 12 in another. Means and standard deviations of the total scores indicated that both were reasonable ratios for our Dutch adolescents. Since there was no difference between the ratios of 1 in 10 and 1 in 12, for the HyTeC-cloze a ratio of 1 in 10 was chosen. This deletion ratio results in more items and therefore in a higher number of observations per text. Given texts of 300 to 400 words, each text will contain between 30 and 40 gaps.

3.3 Included and excluded words

The gap selection for the HyTeC-cloze is based on the two heuristics:

Heuristics[3]

1. Gaps cannot be too locally predictable; words that can be reconstructed purely by use of rules of grammar or knowledge of usage conventions are not good candidates. They do not rely on discourse level comprehension (e.g., Oller & Jonz (Eds.), 1994).

2. Gaps cannot be too unpredictable; words that can only be reconstructed when the test taker has the necessary prior world knowledge are not good candidates. They can only be guessed (i.e., 'extra-textual knowledge'; Bachman, 1985; Levenston et al., 1984).

As is standard procedure in all cloze tests, a gap corresponds to one word only. This word may be a compound, as long as it is written as one word (e.g., *policeman* but not *ice-cream-flavored;* cf. Bormuth, 1966). Abbreviations are not allowed unless they are abbreviations of names (e.g., USA) in which case the general rule for names is followed (see Section 3.3.2).

3.3.1 Heuristic 1: predictable words

Following from the first heuristic, words that can be filled in by grammatical knowledge alone are not selected as candidate gaps. These are mainly function words like articles, prepositions and auxiliary verbs. However, not all function words are excluded. Function words that mark referential cohesion and discourse coherence are prime candidates for testing comprehension at the discourse level (see also Alderson, 1979a). Anaphoric pronouns are allowed since they show referential coherence and are often intersentential. In addition, they can often be replaced by

[3] Note that although there are similarities with Bachman's levels of constraint (within clause; across clause but within sentence; across sentence; extra-textual; Bachman, 1985), our classification does not center around clause or sentence boundaries.

their antecedent (e.g., '*Peter was very tired. He/Peter slept until twelve o'clock.*'). Relative and interrogative pronouns are not allowed since they only operate locally.

Furthermore, we allow most conjunctions and conjunctive adverbs. Most of the markers of coherence relations measure text comprehension on an intersentential level. Therefore, they are prime candidates when testing text comprehension. An exception is made for the coordinating conjunct '*and*'. This conjunct often does not function intersententially and is usually very predictable. That is why it is excluded.

Lastly, sometimes the predictability of words lies in their combination. When words are part of common expressions, phrasal verbs or antonym pairs, they become highly locally predictable. Even without context, most readers will know by convention that the answer in (5) must be *'time'* and that in (6) the answer is probably '*bad*'. These words do not make good candidates.

(5) Once upon a [……]

(6) Good and [……]

3.3.2 Heuristic 2: unpredictable words

Following from the second heuristic, words that are not cued by the context at all are also not good candidates. These words are 'guess words' and depend solely on extra-textual knowledge (see also Bachman, 1985; Levenston et al., 1984). We defined five types of guess words:

1. Technical terms
2. Proper names
3. Units of measurement (e.g., hour, centimeter, year)
4. Cardinal directions (e.g., north, west)
5. Numbers

As Oller and Jonz (1994) note about technical terms: "*Such items, if they were deleted, would normally generate little or no variance and could not therefore contribute significantly to the quality of the test.*" (p.4). The same can be said about the other types of guess words. Even when we accept every answer as long as it is in the same ballpark, the risk of zero or low variation is high. Our pretest confirmed this. When a date was chosen as gap, none of the participants was able to fill in an acceptable answer (e.g., another date). They did not even guess; they all left these gaps blank.

For technical terms and names, the guess factor is only present when they are used just once. For example, in the sentence "*This is called ADHD.*" the word *ADHD* can only be filled in if the reader has prior knowledge on the subject or the

text itself. However, if the term is repeated in the text, like in: "*This is called ADHD. ADHD can be controlled by diet and medicine.*", then the second *ADHD* can be inferred from the discourse. These words are usually co-referential. Therefore, only the first time a term or name is mentioned it does not qualify as a potential gap candidate. However, the second, third and X^{th} mention do qualify. This exception is not made for the other types of guess words, since they are, generally, not used co-referentially.

3.3.3 Word repetitions

Based on the heuristics, a lot of words are excluded. We are left with what we believe to be a 'purer' measure of text comprehension than in standard cloze tests. But there is a downside to excluding a large number of words. Especially with a small number of candidates, some candidates might be overrepresented in the sample. In one of our pretests, a lemma that occurred 7 times in the text ended up as a gap 5 times and by chance all instances ended up in the same cloze version. Therefore, in the final procedure a limit was set for lemma repetitions. The maximum number of repetitions of a lemma within a cloze test is equal to the ratio of repetitions present in the candidate sample. So, if one lemma makes up 10% of the candidates, that lemma can be chosen as a gap three times in a cloze test with 30 gaps (see Appendix 2, Step II). Adhering to a relative limit rather than an absolute limit prevents overrepresentation of a lemma in the sample while still allowing a text feature like lemma repetition to be mirrored in the cloze gap selection (see also Section 3.1).

3.3.4 Reliability of candidate selection

When the candidate selection procedure leaves room for multiple interpretations, experimenter bias creeps in. For instance, while using the same procedure Jonz (1994) came to a different sample of items than Bachman. Bachman's procedure proved to be unreliable in this respect. To avoid this, in the HyTeC-cloze procedure restrictions are mainly formulated as to leave no room for doubt. The procedure leaves less room for experimenter bias by specifying which types of words do not adhere to the heuristics. However, when it comes to deciding if a word combination is a common expression, the line gets blurry. Jonz's notes show that in half of the cases that he disagreed with Bachman, the reason was that he thought that the item was a collocation or a "*multi-part lexical item*" while apparently Bachman did not (Jonz, 1994, p.321). The same problem could threaten the reliability of the HyTeC-cloze procedure. The reliability of the candidate selection procedure was pretested to rule out any interpretation errors or other problems. A student assistant, who was unfamiliar with cloze procedures, followed the procedure and selected all possible candidates for deletion from three different texts. Agreement between his

selection and our own selection was 96% (Cohen's Kappa =.93).[4] The results of the pretest showed that the selection procedure was clear and reliable.

3.4 Cloze versions

In Section 2.1.2 we discussed how changing the starting point of a mechanical cloze test can influence the results. The sampled words are different and the precise sample that is used influences the outcome of the test. This is also a pitfall for a mechanical-rational cloze test. Our pretests showed that the selection of possible candidates for cloze gaps results in more candidates than necessary. We found that in all trial texts, it was possible to construct at least two different cloze versions. Sometimes it was even possible to construct up to five different cloze versions of the same text. If we only select one version out of all possible samples, by chance we might end up with a biased sample. Similar to standard mechanical cloze tests, we can limit that effect by creating multiple versions. The number of possible versions is determined by the length of the text, combined with the number of candidates (see Appendix 2, step II). For example, a text of 300 words requires 30 gaps (i.e., 10%). If there are 120 candidates we can create four unique cloze tests, each sampling different words. Candidates are distributed over versions by counting them off: candidate 1 ends up in version 1, candidate 2 in version 2, …, candidate 5 in version 1. Out of the possible versions, two will be randomly selected to participate in the study.

3.5 Answer format and scoring

The HyTeC-cloze tests are open pre-cloze tests with fixed length blanks. The answer is therefore not cued in any way. The advantages of a multiple-choice answer format – e.g., the speed and objectivity of rating, and the fact that test takers do not have to 'produce' but only have to 'recognize' the correct answer – do not outweigh the disadvantages in our opinion. Multiple-choice answers can be guessed and distractors or cues can confuse test takers rather than help them (e.g., Abraham & Chapelle, 1992; Alderson, 2000). We want test takers to base their answer on the text and the interaction with the text alone. We do not want them to be 'disturbed' or limited by cues and distractors. Their answers should stem from their own representation of the text and not be mediated by the test maker's representations.

Results on the cloze tests will be scored semantically following the global appropriateness criterion. In addition, misspelled words are accepted.[5] Most scholars agree that semantic scoring has higher face validity than exact scoring. When measuring text comprehension, it seems highly illogical to fault a reader for filling

[4] Calculated over 3 different texts, 1052 words total.

[5] This includes typing errors when the test is administered via computer.

in an acceptable answer (e.g., a synonym) rather than the exact word. The reader is clearly able to understand the text to a degree that allows him to fill in a 'good' answer: how could that be wrong? Nevertheless, many scholars keep using the exact method due to its ease. They are supported by findings that the correlation between exact and semantic scores is so high that it seems pointless to score semantically (mean r = .99; as reported by Staphorsius, 1994). Indeed, in our pretests we also found high correlations (range .759 to .873), but we are hesitant to ignore the distinctions between scoring methods. First of all, our correlations are not nearly as high as reported by Staphorsius.[6] In addition, it is unknown whether these correlations hold for all types (or combinations of) items, as well as for texts and readers of all levels. For instance, Brown found that semantic scoring (AC) led to higher scores for both high and low proficiency L2 students, but that on average "*the high proficiency students benefited more from the use of the AC scoring than did the low proficiency students.*" (Brown, 2002, p.96). O'Toole and King (2011) also warn against semantic scoring, since it may lead to an underestimation of text difficulty and an overestimation of reading competence. Although O'Toole and King make a valid point with regard to anchoring and floor/ceiling effects, in our view the total reverse can be said for exact scoring. That is: it may lead to an overestimation of text difficulty and an underestimation of reading competence. Given the studies of Brown (2002) and O'Toole and King (2011), it seems ill-advised to generalize over scoring methods. Although high correlations have been found, exact scoring is not equivalent to semantic scoring. Furthermore, the reliability estimates for semantic scoring are generally higher than those for exact scoring. In a meta-analysis of 223 reliability estimates (across 24 ESL/EFL-studies), Watanabe and Koyama (2008) found a mean reliability estimate of .74 for semantic scoring (k = 97) compared to .64 for exact scoring (k = 122). Moreover, reliability estimates for semantic scoring were more stable, ranging from .60 to .97 while exact scoring ranged from .14 to .99. In a small-scale L1 test with six cloze tests we found similar results: the mean reliability estimate for semantic scoring was .68 (range .60 to .77) compared to .57 for exact scoring (range .42 to .66). Our results indicate that also in native language testing, semantic scoring results in more reliable results. Thus, from a theoretical as well as from a statistical point of view semantic scoring is more suited for measuring text comprehension than exact scoring. That is why, although it is much more time consuming, answers will be scored semantically.

3.5.1 Scoring procedure

In studies with a large number of participants and/or texts it can be very efficient to administer the tests via computer. Answers can be collected automatically. If the

[6] This is to be expected since we do not allow predictable (closed class) grammatical items.

same answer is given ten times, it only has to be scored once. This saves time and it makes scoring more reliable in that when participants give the same answer they will also receive the same score. To limit judgment biases, each unique answer will be scored by two independent judges. When the judges disagree, a third judge makes the final call. The result of this procedure is a scored list with answers that have been given on each gap. The resulting list will be used to score all given answers automatically. This list is never finalized, since every new application of the cloze tests will inevitably result in new unique answers that have to be scored. However, after the first application the number of unique answers will decrease drastically for each new application.

4. Evaluation of the HyTeC-cloze procedure

In this final section we evaluate the HyTeC-cloze procedure on the basis of the results of a large-scale study. The procedure was used to collect text comprehension data for the readability index for Dutch (see Chapter 1). This study included 60 texts in two text and two cloze versions. The texts ranged from 300 to 420 words, so cloze tests contained 30 to 42 cloze gaps. The cloze tests were administered to 2926 Dutch secondary school students in grades 8 through 10. Most students filled in a total of 4 cloze tests (divided over two sessions). Students were enrolled in different levels of the Dutch education system, ranging from the lowest pre-vocational level ('vmbo-bb') to pre-university level ('vwo')[7]. This data was used to address: 1. what the cloze measures, 2. how scores correlate with other measures, 3. the sensitivity of the HyTeC-cloze test, 4. the internal reliability, and other validity threats such as data loss.

4.1 Semantic versus exact scoring method

The data was scored using the semantic scoring procedure outlined in Section 3.5. We also scored the data using the exact scoring method so we could compare their results. The exact and semantic scores correlated highly (r_s = .862; p < .001). However, they did not correlate as highly as previously reported (cf. Section 3.5). Furthermore, the correlation was not completely stable. The relation decreased in strength going from the lowest level of education to the highest (from .859 to .789) and varied between individual cloze tests (from .637 to .951; see Table 3). For completeness purposes we will report findings for both the semantic scoring method

[7] The Dutch system distinguishes multiple levels of education. Going from practice oriented education to theoretical oriented education, the levels included in the study are: vmbo-bb, vmbo-kb, vmbo-gt, havo and vwo.

and the exact scoring method wherever possible, but the semantic score outperformed the exact score in all tests we present here.

Table 3: Summary Spearman's rho correlations calculated over cloze test versions (k = 120)

Correlation	Mean r_s	SD r_s	Median r_s	Minimum r_s	Maximum r_s
Semantic scoring - Exact scoring	.848	.061	.855	.637	.951

4.2 Level of measurement

The HyTeC-cloze test is designed to measure text comprehension. Other cloze tests have been criticized because their gaps seem to rely more on local linguistic predictability (on the basis of grammatical knowledge and knowledge of collocations) than intersentential, context or passage dependent comprehension. We believe that the configuration of the cloze test will determine to what extent it measures comprehension beyond the sentence level and whether only grammatical or probability information is used to answer the gaps. The results below suggest that we were fairly successful in our undertaking.

4.2.1 Local predictability of cloze gaps

Since predictable words are not included as gaps in the HyTeC-cloze (see Section 3.3.1), the relation between local predictability and cloze scores should be weak. Two analyses were done to check this claim.

First, we examined whether the HyTeC-cloze procedure was successful in selecting cloze gaps that were not highly locally predictive. If the procedure was successful, the words that were used as cloze gaps should have a lower local probability compared to words that were not turned into cloze gaps. T-Scan – a tool for automatic Dutch text complexity analysis (Pander Maat et al., 2014) – was used to determine the forward log-probability (probability of $Word_N$ given $Word_{N-2}$ and $Word_{N-1}$) and backward log-probability (probability of $Word_N$ given $Word_{N+1}$ and $Word_{N+2}$) of all words.[8] The probabilities of words that were used as cloze gaps were significantly lower than the probabilities of words that were not used as cloze gaps (see Table 4; Forward probability: $U = 77859698.500$; $z = -67.300$; $p < .001$; $r = -.31$; Backward probability: $U = 75133694.500$; $z = -69.842$; $p < .001$; $r = -.32$). Reversing the $\log_{10}$-transformation shows us that the words used as cloze gaps are on average almost 23 times less probable (based on the 2 words preceding them) compared to non-clozed words. Based on the 2 words following them, they are 34

[8] T-Scan uses WOPR (http://ilk.uvt.nl/wopr) to model backward and forward trigram probabilities. The models were trained on the newspaper-section of the SoNaR-corpus (Oostdijk, Reynaert, Hoste & Van den Heuvel, 2013). Unseen items are estimated through Good-Turing smoothing (see Van den Bosch & Berck, 2009).

times less probable compared to the non-clozed words. The selection of non-locally predictive words was thus successful.

Table 4: Means, standard deviations and medians for Forward and Backward log-probability

Words	**Forward log-probability**			**Backward log-probability**		
	Mean	**SD**	**Median**	**Mean**	**SD**	**Median**
Overall (N=46274)	-2.468	1.652	-2.091	-2.461	1.790	-2.190
Not a cloze gap (N=38456)	-2.239	1.573	-1.826	-2.202	1.722	-1.864
Cloze gaps (N=7818)	-3.594	1.566	-3.528	-3.734	1.556	-3.804

Next, the probability measures were entered in a logistic regression to see how much variance they can explain as predictors of the cloze scores. Combining the measures gives us a window of 4 words surrounding the cloze gap. Together forward and backward log-probability only explained 2.4% of the variance observed in the semantic scores and 7.3% for the exact scores (see Table 5). The difference between semantic and exact scores was to be expected. The probability measures indicate the probability of the exact word that was deleted in the text, not the probability of all semantically correct words that could occur there. Therefore, we expected the probability measures to explain more variance for exact scores than for semantic scores. Yet, even for exact scores the explained variance is low and it is therefore very unlikely that the HyTeC-cloze test only measures local level predictability.

Table 5: Explained variance by log-probability measures at item level

Scoring method	**Model**	**Explained variance (r^2_N)**
Semantic	Forward and backward probability	.024
	Forward probability	.022
	Backward probability	.018
Exact	Forward and backward probability	.073
	Forward probability	.068
	Backward probability	.046

4.2.2 Correlations with other measures

Correlations of the cloze scores with other text comprehension measures or with standardized ability tests, can give us an idea of the convergent validity of our cloze test: do the tests measure the same construct? For the majority of the students, standardized reading ability and vocabulary scores were available. The summed semantic cloze scores[9] correlated on average .606 with the reading ability scores and

[9] The summed scores were calculated by adding up the scores of the cloze gaps for each participant for each cloze test. Because cloze tests had a different number of gaps, the summed scores were normalized to a 30-gap test.

.604 with the vocabulary scores (see Table 6).[10] The exact cloze score correlated slightly lower (reading ability: *mean* r_s= .564; vocabulary: *mean* r_s= .558). Given that well-established, standardized tests of reading ability have been found to correlate moderately with each other – between .31 and .79 (Cutting & Scarborough, 2006; Keenan et al., 2008) – a mean correlation of .60 with an even higher median suggests that the HyTeC-cloze test does not underperform compared to other established reading ability tests.

Table 6: Summary Spearman's rho correlations calculated over cloze test versions (k = 120)

Correlation	Mean r_s	SD r_s	Median r_s	Minimum r_s	Maximum r_s
Semantic score - Reading ability	.606	.137	.621	.161	.856
Semantic score - Vocabulary	.604	.125	.609	.220	.839
Exact score - Reading ability	.564	.154	.587	.047	.832
Exact score - Vocabulary	.558	.143	.569	.105	.870

In addition to the standardized test scores, we have some data that can indicate how our participants would have performed had we used the same texts but a different assessment method. A selection of 8 texts was used in an eye-tracking study. Within this study, text comprehension was assessed with 8 multiple-choice questions per text. 181 ninth-grade students answered these questions after reading the texts. The students were not able to look back in the text when answering the questions. Mean scores were calculated per text and education level and then compared to the corresponding mean cloze scores. The multiple-choice score correlated .525 (p =.008) with the semantic cloze scores and.389 (p = .060) with the exact cloze scores.

4.3 Sensitivity of the test

A good assessment method has to be sensitive to known differences in comprehension levels (i.e., known-group validity). The readers in the sample are enrolled in different levels of education and differ in age. Furthermore, half of the texts that were clozed were taken from educational textbooks written for different education levels and grades; the other half were public information texts. Thus, we have strong reasons to believe there should be a lot of variance in the sample: between students ánd between texts. The HyTeC-cloze must be sensitive enough to show these types of variance. It must be able to discriminate between students with different reading abilities and it must also be able to discriminate between different texts and in case of experimental studies, between text versions.

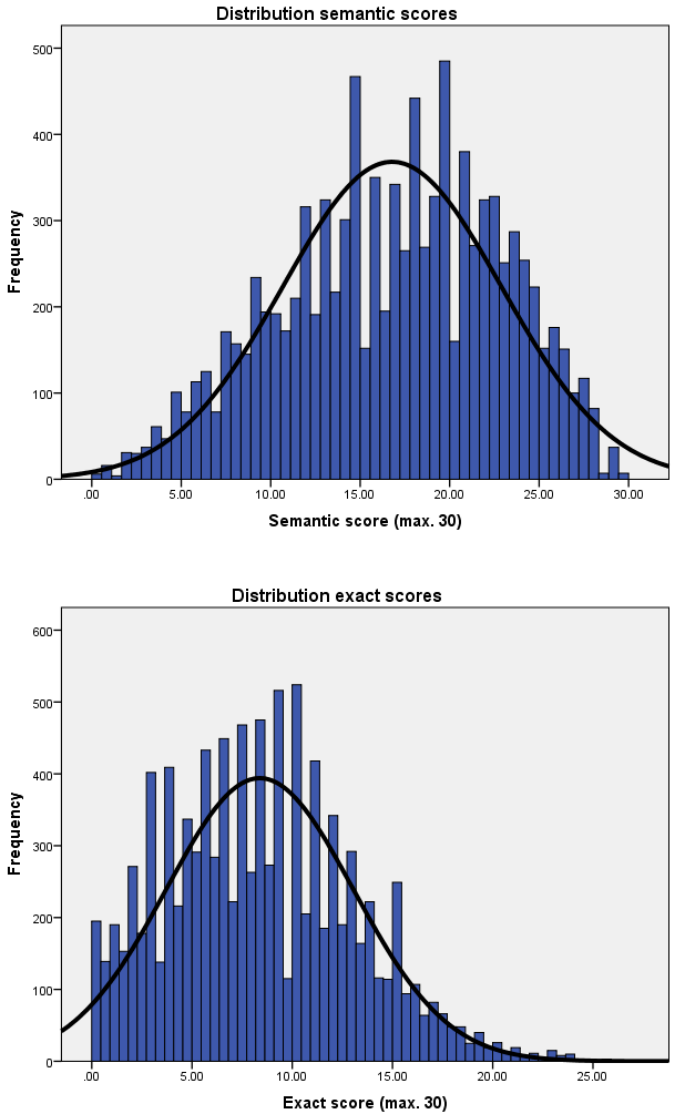
First, we investigated the overall amount of variance in the cloze scores by plotting the frequency distributions. If the data was normally distributed and there

[10] The standardized reading ability and vocabulary scores correlated .569 with each other, partially explaining the similar correlation with the cloze score.

was no evidence of either floor or ceiling effects, we could proceed to investigate the expected variance between groups.

As shown in Figure 1, the frequency distribution of the semantic scores was close to normal with a mean score of 16 out of 30 items correctly answered. The mean exact score was of course lower (semantic score = exact answers + semantically correct answers) and the distribution had a heavier left-tail. The figures show score distributions of all students. When we compared distributions of the different education levels in the sample, we found that exact scoring was particularly problematic for the lowest levels of the Dutch education system with many observations that were zero (or close to) zero. Again, semantic scores were more normally distributed and showed variation especially in the lower education levels.

Figure 1: Frequency distribution of semantic and exact summed scores



A lot of variance exists in the sample, but can we attribute this variance to known differences between students and/or texts? Linear mixed effects modeling was used to answer this question. The data was hierarchically structured: students are nested in schools and cloze tests are nested in texts and semi-crossed with students. This structure was tested in a stepwise procedure. *Level of education* and *Grade* were introduced as fixed factors. The final model is shown in Table 7.

All factors improved the fit of the model. As expected, scores varied between students, texts and a little bit between cloze versions.[11] The student variance was diminished by introducing the fixed factors *Level of education* and *Grade*. The differences between all education levels were significant and in the expected direction: when the education level increased, so did the cloze score. The same relation was observed for *Grade*: students in higher grades scored higher than those in the lower grades. These results show that the HyTeC-cloze was sensitive to known differences in the sample.

Table 7: Final model semantic score

Random effects	**Estimates**	**Standard deviation**		
School	1.562	1.250		
School: Student	5.838	2.416		
Text	10.586	3.254		
Text: Cloze version	1.483	1.218		
Residual	8.357	2.891		
Fixed effects	**Estimates**	**Standard error**	**T-value**	***p***
Intercept	9.836	0.544	18.097	<.001
Education level: pre-voc. low	0[a]			
Education level: pre-voc. medium	2.620	0.257	10.207	<.001
Education level: pre-voc. high	4.934	0.271	18.200	<.001
Education level: general	7.555	0.341	22.139	<.001
Education level: pre-uni.	9.866	0.323	30.547	<.001
Grade 8	0[a]			
Grade 9	1.159	0.149	7.771	<.001
Grade 10[b]	1.982	0.307	6.462	<.001

[a] Set as reference level. [b] Unbalanced, no 10th grade pre-vocational students were present in the sample.

In addition to different texts, the data also included different text versions. Texts were manipulated in one of three ways: 1. Words were substituted for less or more familiar alternatives, 2. Syntactic dependency lengths were increased or decreased, 3. Connectives were removed or added. Each manipulation was expected to influence text comprehension. In separate analyses, we tested whether the HyTeC-cloze was sensitive to these subtle manipulations of text difficulty. Since these analyses will be extensively discussed in Chapters 3, 4 and 5, we will not go into detail here. We will only point out that the HyTeC-cloze was sensitive to these differences. Still, not all effects were observed in the summed cloze score. For some manipulations it was necessary to zoom in on the gaps directly surrounding the manipulation to find a significant effect. So, although the HyTeC-cloze is sensitive enough to detect text version differences, it will depend on the specific manipulation

[11] This confirms the importance of using multiple cloze versions of the same text.

whether it is visible in the summed cloze score or whether the effect is localized to specific gaps.

4.4 Internal reliability

For each cloze test version (k = 120), internal reliability was measured using Cronbach's alpha. A summary of the results is given in Table 8. Semantic scoring and exact scoring are both relatively reliable, but semantic scoring is systematically more reliable. In addition, semantic scoring is more stable across cloze tests and never drops below .70. These scores are high, especially given the fact that many studies have reported dramatically low alphas for cloze tests (Brown, 2013; Watanabe & Koyama, 2008).

Table 8: Summary internal reliability calculated over cloze test versions

Scoring method	Mean α	SD α	Median α	Minimum α	Maximum α
Semantic	.828	.038	.831	.707	.899
Exact	.738	.075	.742	.519	.894

4.5 Response rates and data loss

The validity of any test is threatened when test takers do not answer seriously or when they do not answer at all. When either happens, the measurement does not necessarily represent the student's true ability. We explored our data to see whether these threats were present.

4.5.1 Blank cloze gaps

As with many constructed-response tests, response rates for cloze items tend to be lower than those for multiple-choice items. Leaving a gap blank is tempting because it takes more effort to fill in a gap than to circle an answer. In the present study the cloze tests were administered digitally on computers but students were not obligated to fill in every cloze gap. They could leave gaps blank just like in paper-based tests. Of course they were instructed not to leave gaps blank and to guess if necessary. Nevertheless, in 9.3% of the cases students left the cloze gap blank. It is highly likely that these gaps were left blank for various reasons (see Hashkes & Koffman 1982 as cited in Cohen, 1984). It could mean that: 1. the student did not know the answer, 2. the student was unmotivated, 3. the student ran out of time, or 4. the student simply did not see it (although this is rather unlikely because the gaps were given a distinct color and were underlined to make them stand out, see Appendix 4). We would like to differentiate between "not knowing" and the other options in our analyses, because "not knowing" can be interpreted as a wrong answer whereas other options are real missing cases. Hence, we conducted a qualitative exploration of the blanks. We investigated the number of blanks per student per cloze test, the location of the blanks (e.g., at the end of the test), and the face-validity of the

answers to surrounding gaps (serious answers or not). Most gaps seem to be left blank because the student did not know the answer to that particular gap. These blanks were dispersed throughout the test and surrounded by serious answers. 1/3rd of the gaps was probably left blank because the student was not motivated enough to continue or ran out of time. The surrounding gaps were left blank as well or throughout the test the student filled in mostly nonsense answers. Out of all the blanks, only these cases are considered to be real cases of data loss.

4.5.2 Non-blank cloze gaps

Blank gaps are of course easy to spot in the data. But students had an entire keyboard available to them and they did use it. Besides serious answers they filled in symbols (dots, hyphens, dashes, question marks), strings of letters (aaaaaaaa, adfgd, x) and words that are clearly not serious answers (unicorn, Santa, curse words).[12] As with blanks, we must discriminate between cases in which the answer could mean 'I don't know' and other options. Again, we investigated the surrounding gaps to guide our interpretation. Some cases were very clear. A couple of students used succeeding cloze gaps to write out exactly how they felt about taking the test: "I hate this test" and "This is stupid". Others used the gaps to indicate their more general lack in motivation. "I'm just filling in words trying to finish this test so I don't have to stay after class." These cases were of course removed.

4.5.3 Total data loss

Based on the considerations listed above, 9.66% of the filled out cloze tests was identified as a possible threat to the validity and these tests were removed from the data. The percentage of data loss was higher for pre-vocational education than for higher education levels, but decreased in higher grades. This could mean that the lowest education levels – at least in grade 8 – were frustrated with the test, but this finding may also be a reflection of a general lack of motivation to read (see Land, 2009). We compared the standardized readability scores of the students that were removed to those of the students that remained in the dataset. There was no significant difference between groups ($F(1,4756) = 0.028$; $p = .866$). The data loss was unbiased in this respect and did not result in a sample that underrepresented low-ability students.

[12] During the scoring procedure we asked our judges to mark these cases, so that we could check them out later.

5. Discussion

Cloze is a popular assessment method in readability studies and language proficiency testing but it has never been widely accepted as a valid measure of text comprehension. According to its critics, cloze tests are "*beset with problems*" (Klein-Braley & Raatz, 1984, p.134) and should be avoided at all costs. The critics' biggest concern seems to be that gaps can be answered correctly without understanding the text. They claim that localized, low-level processing is enough to fill in the gaps successfully and that it is not necessary to integrate sentences into a discourse level representation. Yet, even mechanical cloze tests contain a large number of cloze gaps that depend on the integration of information across sentences and rational cloze tests can be designed to explicitly target higher-order processes. In fact, most 'problems' with cloze do not hold for all types of cloze tests and can be successfully addressed in cloze design. In this paper we presented such an improved cloze procedure: the HyTeC-cloze. This hybrid cloze procedure combines the strengths of mechanical and rational cloze tests into a valid and reliable measure of text comprehension. The rational strategy is used to exclude words that do not rely on text level comprehension from becoming cloze gaps (e.g., articles, copula, multi-word expressions, and guess words). The remaining words in the text are candidates for deletion and a sample of them is randomly selected via mechanical selection. This procedure results in a cloze test that has a very low sensitivity to local predictability (only 2.4% explained variance) and is still fast and easy to produce since it does not require an extensive analysis of the texts. Furthermore, the HyTeC-procedure is widely applicable. It can be used to assess a wide range of texts without confounding text difficulty with question difficulty and is suitable for test takers of high and low ability. Most importantly, our results show that it matches and sometimes even outperforms standardized tests of reading ability when it comes to validity and reliability. These qualities, together with its sensitivity to discriminate between texts, text versions and readers, make the Hybrid Text Comprehension cloze an appealing method for experimental and correlational studies.

3 Generalizing lexical effects across texts and readers

The relationship between lexical complexity and text difficulty is a rather obvious one: knowing the words in a text is a logical prerequisite for understanding it (e.g., Perfetti & Stafura, 2014). In fact, lexical complexity is the strongest predictor of readability (Bailin & Grafstein, 2016). Texts containing long or unfamiliar words are harder to understand and to process than texts with short or familiar words (Crossley, Allen & McNamara, 2012; Crossley, Louwerse, McCarthy & McNamara, 2007; Dale & Chall, 1948; Flesch, 1948; Staphorsius, 1994). These correlations are strong and seem to suggest that the difficulty of a text can be effectively reduced by making different lexical choices. However, experimental studies have shown that is not so easy. Whether lexical manipulations affect comprehension seems to depend on the number of difficult words, their relevance and the level of difficulty of the substituted word given the word knowledge of the reader (Freebody & Anderson, 1983ab; Stahl, 1991; 2003b; Stahl, Jacobson, Davis & Davis, 1989). If a reader does not know the words or only has a limited understanding of the words, it will be harder to grasp what the writer is trying to convey. The higher the number of complex words in a text, the higher the demand that is placed on the reader. Yet, a full understanding of all the words in a text is not necessary. Readers are able to overcome unknown words and they even learn new words from text (Nagy, Anderson & Herman, 1986; Nagy, Herman & Anderson, 1985; Stahl, 1991; 2003b). Readers can infer or derive word meaning using other sources than lexical knowledge (e.g., morphology, grammar, context, world knowledge).

While limited word knowledge may not always lead to a decrement in comprehension, it will often impact processing. Less familiar words will be less extensively represented and connected in memory than familiar words. Activating the meaning of these words will take more effort and, if inferential and derivational processes are required, the demand placed on mental resources will be even larger. Lexical complexity effects on online-processing have been observed in lexical decision tasks, in rapid naming and in reading times (Chaffin, Morris & Seely, 2001; Rayner, 1998; 2009; Schilling, Rayner, Chumbley, 1998; White, Drieghe, Liversedge & Staub, 2016; Williams & Morris, 2004). Unfamiliar words take more time to process than familiar words, even when confounding factors like word length are kept constant.

However, the robustness and generalizability of these processing and comprehension results is yet unclear. Comprehension studies generally test only a small number of texts on a limited range of readers. Processing studies often focus

on processing words and sentences in isolation instead of embedded in continuous text. The generalizability of these results to more normal reading situations is quite low. Both for comprehension and processing, we will study lexical complexity effects in a broader range of contexts.

1. Word knowledge, lexical choice and simplification

1.1 Word knowledge

Word knowledge is a multidimensional concept (Cutler, 1983; Stahl, 1991). It encompasses the form of a word (incl. orthography, phonology) but also its meaning (i.e., the concept it refers to). Full knowledge of a word presupposes that form and meaning are linked in memory. It also includes "*an understanding of the core meaning of a word and how it changes in different contexts*" (Stahl, 2003a, p.19). Both form and meaning contribute to the difficulty level of a word. Readers may be familiar or unfamiliar with: the word, the concept it refers to or both. The *conceptual difficulty* denotes how complex the underlying concept is (Nagy & Hiebert, 2010). Genetic engineering, for example, is inherently more complex than grocery shopping. The *stylistic difficulty* denotes the difficulty level of the form – the specific word – that is used to convey that concept (e.g., 'begin' versus 'commence').[1] Stylistic difficulty can be reduced by substituting one word for another, more familiar word. In contrast, conceptual difficulty is constant given the content of the text: it cannot be 'improved'. Of course a text can be altered conceptually to accommodate different types of readers, for instance when a text for 5th-graders is adapted for 4th-graders. In that case, however, the writer has a different idea about what the 4th-graders should know compared to what the 5th-graders should know and in essence writes a different text. Concepts can be explained differently or more extensively, but this could also be seen as a stylistic adaptation since the concepts that are explained do not change.

Both conceptual difficulty and stylistic difficulty are relative concepts: they largely depend on the reader. Word knowledge reflects experience and world knowledge (Bailin & Grafstein, 2016; Stahl & Nagy, 2006). What is easy for one reader might be difficult for another and vice versa. Football fans will know the names of different plays, while those same names (and the concepts they refer to) will be unknown by others. Despite these individual fluctuations, lexically complex words are less likely to be known by readers than less complex words.

[1] See also the dualistic approach to style (Leech & Short, 2007).

1.2 Lexical choice and simplification

There are two ways to decrease the lexical complexity of a text. One is to simplify the content and the other is to simplify the language (Honeyfield, 1977). The term 'simplification' is used to refer to both actions, which in practice frequently coincide (Davies & Widdowson; 1974; Honeyfield, 1977). Simplification is common practice in education. Materials are adapted to fit the proficiency level of the students (for instance, for beginning second language learners). Information that does not serve the educational goal is deleted or altered. However, in other situations it may not be possible to change the content. A writer has certain ideas about what she wants the reader to know. Altering information or leaving it out all together is not an option.

In such contexts, the only way to improve readability is to reduce a text's stylistic difficulty. To change the stylistic difficulty without altering the meaning of a text is a difficult task, however. The choice for a specific word is functionally motivated and influenced by many factors, including content, writer, target audience, theme, goal, text structure, genre and medium. For example, educational books teach their readers new concepts, which includes teaching them new vocabulary as well (Johnson & Otto, 1982). These lexical items cannot be altered, no matter how difficult they may be. In addition, lexical choices go hand-in-hand with choices in syntactic structure, the use of fixed expressions and other types of collocations (Crossley, Louwerse et al., 2007; Davies & Widdowson, 1974; Davison & Kantor, 1982). Adapting the lexical complexity regularly affects other text features as well (Crossley, Louwerse et al., 2007; Davies & Widdowson; 1974; Davison & Kantor, 1982; Honeyfield, 1977). For example, Honeyfield (1977) observed that lexical items are often replaced with paraphrases. The paraphrase in (2) may reduce the lexical complexity of (1), but results in a more complex syntactic structure. It becomes a toss-up which feature impacts readability the most.

(1) A series of misfortunes

(2) Circumstances which were beyond his control

(Honeyfield, 1977, p.433)

Some studies have been more or less successful in avoiding such confounds. Freebody and Anderson (1983ab) substituted 1/3rd, 1/4th and 1/6th of their content words for an intuitively less complex alternative and investigated the effects on free recall, summary, sentence recognition and sentence verification. All measures showed a tendency for texts with complex words to receive lower comprehension scores than texts with less complex words. However, the complexity effect did not reach significance on every occasion. Johnson and Otto (1982) also were unable to find a significant effect. Still, they only changed 5% of the words and their texts

were overall very conceptually complex due to a high number of technical terms which were not altered. On the other hand, Stahl and colleagues (1989) used texts without technical terms and found a clear effect of lexical complexity. They substituted 1 in every 6 content words with words that were above the complexity level of their 6th-grade students (see Examples (3) and (4)). Lexical complexity influenced the number of recalled elements and the order in which the 6th-graders recalled the elements. They took this as an indication of the readers' trouble in building a coherent text representation. A less coherent representation may in turn have led to the lower number of recalled elements. However, to create this effect, they needed to substitute words with words from word lists created for 8th-grade students. In addition, the resulting passage in (4) seems somewhat unnatural or inconstant in style compared to (3). The inserted difficult vocabulary does not seem to match the simple vocabulary in other sentences. One would expect a regular writer to be more consistent and after reading '*adversaries*' one may expect the writer to use '*allies*' rather than '*friends*' in '*Now they needed friends*'.

(3) "Perhaps the Shami have killed him," he thought. That would be very bad. His village already had too many enemies. Now they needed friends. They had traded with the Shami for the whole year, and all had gone well.

(4) "Perhaps the Shami have killed him," he speculated. That would be very bad. His village already had too many adversaries. Now they needed friends. They had bartered with the Shami for the whole year, and all had been satisfactory.

(Stahl et al., 1989, p.32)

Conversely, lexical simplification can have unwanted consequences. Words with low-lexical complexity (Low-LC) often have more ambiguous meanings (Davies & Widdowson; 1974; Crossley et al., 2012), while high-lexical complexity words (High-LC) are more specific and may invoke connotations that are not included in their simplistic counterpart. Certain enrichments of the text representation may be lost when an easier alternative is used in a text. An example is given by Stahl (2003b) who shows that it depends on the context whether substitution of '*debris*' by '*trash*' leads to a loss of meaning. Debris is a specific type of trash that is caused by an accident. In (5) this is vital for understanding what is going on, in (6) it is not.

(5) He stepped over the debris/trash on his way to the car. He heard someone quietly moaning inside.

(6) He stepped over the debris/trash on this way to the car. He put his key into the ignition and took off.

(Stahl, 2003b, p.245)

Simplifying the vocabulary is thus a difficult task and may not always lead to a better understandable text. Or perhaps better said: it may not lead to a better text for every reader. For readers who understand the implication of '*debris*' in (5), replacing it by '*trash*' can have a negative effect on their comprehension. On the other hand, for readers who do not know what '*debris*' means, '*trash*' may lead to a relatively better understanding of the situation. We must be aware of the potentially adverse effects of lexical simplification; especially when texts are read by readers that differ in skills and experience.

1.3 Lexical complexity in on-line processing

Even when lexical complexity does not affect comprehension, effects may be observed in how readers process a text. Word features associated with lexical complexity have been found to influence readers' processing times. Word length (Rayner, 1998), morphology (Bertram, 2011; Pollatsek & Hyönä, 2006), word familiarity (Chaffin et al., 2001; Williams & Morris, 2004), word predictability (Kennedy, Pynte Murray & Paul, 2013) and word frequency (Kennedy, Pynte, Murray & Paul, 2013; White et al., 2016; Williams & Morris, 2004) drive eye movements (for a review see Rayner, 2009). While low level visual features like word length and word spacing primarily influence *where* to move the eyes while reading (e.g., Leyland, Kirkby, Juhasz, Pollatsek & Liversedge, 2013), other features like frequency and familiarity more often influence *when* the eyes move (Rayner, 2009). High-LC words are processed slower than Low-LC words (Chaffin et al., 2001; Rayner, 2009, Schilling et al., 1998; Warren, Reichle & Patson, 2011; Williams & Morris, 2014, White et al., 2016) and High-LC words receive more regressive fixations and rereads than Low-LC words (Warren et al., 2011; Williams & Morris, 2004). Lexical complexity effects have been observed in eye movements while reading, but also in lexical decision and word naming tasks (Schilling et al., 1998; Kuperman, Drieghe, Keuleers, & Brysbaert, 2013).

Unfortunately, many studies that have investigated the effects of lexical complexity on processing have focused on isolated word and sentence processing. Target words are embedded in sentences that are controlled for confounds like length and syntax. The context is often very limited. Usually, sentences are used that start out with a neutral context, followed by the target word and end with a slightly

supportive or at least plausible context (e.g., Williams & Morris, 2004). Yet, Chaffin et al. (2001) showed that readers actively use context to infer the meaning of an unfamiliar target word. Readers pay more attention to the context when the context is informative with regards to the meaning of the target word. In these situations, lexical complexity effects are not confined to the processing of just the target words. Contexts can guide expectations and they provide information that can help infer word meanings.

Radach, Huestegge and Reilly (2008), in particular, showed the importance of presenting stimuli in context. They compared reading times on declarative sentences that were either presented in isolation or in natural passages of six lines. The presentation format not only affected reading times – initial times slower in isolation, but more rereading in passages – but format also interacted with the effects of word frequency. The effect of word frequency was smaller in the passage condition.[2] It may be that the context provided additional support or that readers adopt a different strategy when they encounter infrequent words in passages compared to isolated sentences.

Another limitation of using isolated sentences is that they only include one target word per sentence. While this makes for a clean investigation of lexical complexity for the target word, it rules out any cumulative effects of previously encountered difficult words. As readability studies show, a difficult word rarely comes alone. While readers may be able to overcome a single unknown word, when the number increases so does the demand placed on mental resources and inferring their meaning quickly becomes harder; especially because the available context diminishes at the same time. A single difficult word may be easy to overcome, but if difficult words keep coming the impact on processing might be much larger. It is thus important to look at the cumulative effect of difficult words as well (Stahl, 2003b).

1.4 Present study

Until now, comprehension studies have compared the effects of lexical simplification using only small numbers of texts and limited groups of readers. Both the choice of text and the reader will influence the relative effects of lexical complexity. Texts with a high number of low lexically complex words will generally not benefit from lexical simplification, while texts with high lexically complex words will potentially become easier once the lexical complexity is reduced. However, what is considered to be 'low lexical complexity' will depend on the reader. It is therefore important to investigate the effects of lexical complexity using

[2] Radach et al. (2008) only present analyses of frequency effects for first pass reading times. However, passage reading had significant longer rereading times. It may be that reading times are equal in sum, but distributed differently over early and late processes.

a wider range of texts and readers. This holds for comprehension as well as for processing studies. For processing studies it is also important to study the effects of lexical complexity of longer texts and not just of words or sentences presented in isolation.

In the present study, we extend the existing body of work and examine the generalizability of lexical complexity effects on text comprehension (Experiment 1) and effects on on-line processing (Experiment 2). In Experiment 1 we systematically manipulate the complexity of twenty texts, and present these texts to hundreds of readers varying in reading proficiency. In Experiment 2 we select four texts and present them once again to a wide range of readers. Together the results will show to what extent lexical complexity affects readability and how lexical complexity effects hold up in 'the real world'. That is, for a variety of texts and readers.

2. Experiment 1: Effects on text comprehension

Experiment 1 is designed to test the hypothesis that lexical simplification will increase text comprehension. We predict that readers will have a relatively higher level of knowledge of a low lexical complex word compared to their level of knowledge of a high lexical complex word. However, the exact difference may differ between readers and the difference may be extremely small for some readers. With this in mind we choose to manipulate lexical complexity across a wide spectrum, rather than manipulate lexical complexity as a dichotomous construct (i.e., easy vs. difficult). We are thus examining the effects of relative lexical complexity.

2.1 Method

2.1.1 Participants

Thirty-two Dutch secondary schools participated with in total 786 students from grades 8, 9 and 10 (age: 13 - 16). Testing was introduced as part of the regular school curriculum. The students were enrolled in different levels of secondary education: in pre-university education ('vwo'), general secondary education ('havo') or pre-vocational education ('vmbo-gt', 'vmbo-kb' or 'vmbo-bb').[3] The distribution of participants over grades and education levels is given in Table 1. Grade 10 students of pre-vocational education were not included in the study because pre-vocational students are graduating in grade 10.

[3] These five levels are ordered from theoretical oriented education to practice oriented education. For more information on the Dutch educational system see EP-Nuffic (2015).

Table 1: Distribution of participants over grades and education levels

	Pre-voc. (low)	Pre-voc. (medium)	Pre-voc. (high)	General education	Pre-uni.	Total
Grade 8	39	70	123	29	84	345
Grade 9	89	103	55	27	65	339
Grade 10	-	-	-	57	45	102
Total	128	173	178	113	194	786

2.1.2 Materials

Twenty texts were selected from a collection of Dutch authentic texts (see Chapter 1). The texts were randomly selected and not selected based on their potential for manipulation success. Ten texts were taken from educational textbooks on history, geography, Dutch language and economics. These texts were written especially for students in secondary education. The other ten texts were public information texts which discussed matters related to health (e.g., the flue), legislation (e.g., maximum no. of working hours) and public safety (e.g., crime). These texts were written for the general public but were also relevant for Dutch adolescents. All texts were 300 to 400 words long and did not contain figures or tables.

Manipulation. The twenty texts were manipulated to create text versions with relatively low lexical complexity and text versions with relatively high lexical complexity. Previous studies have used one of three strategies to perform such adaptations. The first is to use word lists which contain words that should be known by the target population. For instance, Stahl et al. (1989) used word lists for 6th-graders and 8th-graders. They substituted original words in 6th-grade texts for words on the 8th-grade list. Another option is to use frequency lists (cf. Johnson & Otto, 1982). Words with high frequencies are replaced for words with lower frequencies. The third strategy is to base substitutions on intuition (cf. Freebody & Anderson, 1983ab). We choose to follow the intuitive approach for two reasons. Firstly, word and frequency lists have their limitations: they are not sensitive to lexical ambiguity (Stahl, 2003b), they do not always reflect the experience of a particular target audience (Keuleers, Brysbaert & New, 2010), they miss words that are known by all but not written about that often (e.g., 'tricycle' see Breland, 1996; Williams & Morris, 2004), they overestimate the difficulty of compound nouns (Anderson & Davison, 1988; Stahl, 2003b) and do not take into consideration the dispersion of words across genres and subject domains (Nagy & Hiebert, 2010). Secondly, we are interested in manipulating lexical complexity over a range of words and texts. We are interested in relative changes in complexity and do not consider lexical complexity to adhere to a strict limit in which words below a certain limit are 'easy' and words that are above are 'difficult'.

The lexical complexity of the twenty texts was manipulated following three guidelines:

I. Only the stylistic difficulty of the texts is manipulated. The content of the text cannot be removed, added upon or altered in any way. This includes terminology/content specific vocabulary that can be viewed as part of the content (e.g., '*Lebensraum*').

II. Texts must remain natural and must reflect a unified style. Archaic or stilted language is to be avoided (see also Klare 1976a).

III. Confounding factors must be kept in check (e.g., word length, syntactic structure, cohesion, lexical diversity). We avoid changing these factors, or if changes are necessary balance them out over text versions (i.e., if an extra word is added in case of one manipulation; we try to add a word in the other version in another manipulation). Lexical diversity and cohesion can be controlled by manipulating every instance of a word in both text versions. We change collocations as a whole or leave them alone.

The original texts were used as a starting point. Each text was manipulated in two directions resulting in a version with relatively low lexical complexity (Low-LC version) and a version with relatively high lexical complexity (High-LC version) compared to each other. The versions were created by replacing 20% of the content words.[4] Because the texts were manipulated in two directions, words from the original text could end up in the High-LC version (see (7)) or in the Low-LC version (see (8)). Occasionally original words were replaced in both versions (see (9)). Each manipulation was reviewed by at least one other researcher to control for experimenter bias.

(7) Original: *Rabiës is een infectieziekte die de hersenen aantast.*
"Rabies is an infectious disease that impairs the brain."
Low-LC: *Rabiës is een infectieziekte die de hersenen beschadigt.*
"Rabies is an infectious disease that damages the brain."
High-LC: *Rabiës is een infectieziekte die de hersenen aantast.*
"Rabies is an infectious disease that impairs the brain."

[4] We aimed for 20% of the content words, but allowed a 10% deviation. The number of manipulations had to lie between 18 and 22 percent. If a text did not reach the minimum number, another text was randomly selected from the collection of texts to take its place.

(8) Original: *Iemand heeft genoeg gespaard om er een tijdje tussenuit te kunnen.*
"Someone has saved enough to take a break for a while."
Low-LC: *Iemand heeft genoeg geld om er een tijdje tussenuit te kunnen.*
"Someone has enough money to take a break for a while."
High-LC: *Iemand heeft voldoende middelen om er een tijdje tussenuit te kunnen.*
"Someone has sufficient means to take a break for a while."

(9) Original: *Dit gebeurt door een rechercheteam.*
"This is done by a detective squad."
Low-LC: *Dit wordt gedaan door een rechercheteam.*
"This is done by a detective squad."
High-LC: *Dit wordt uitgevoerd door een rechercheteam.*
"This is conducted by a detective squad."

The examples above also illustrate that not all manipulations are equally strong. Some differences between the versions are rather small and some are larger.[5] Some may have an effect on all readers, while others will have a limited effect or no effect at all on comprehension. This difference in potential strength reflects the normal situation in which lexical simplification takes place in practice. A word that is used in the Low-LC version is not per definition 'easy'. It is only hypothesized to be easier than its counterpart in the High-LC version. That is why a word that is used as a 'High-LC' alternative in one text may be a 'Low-LC' alternative in another situation.

Quantitative checks. After the texts were manipulated, we used a quantitative approach and examined how the manipulation affected text features. All text versions were analyzed with T-Scan, a tool which automatically extracts over 400 text features from Dutch texts (Pander Maat et al., 2014). T-Scan was used to see which text features were affected and to check that no unwanted confounds were present. A summary of these checks is given in Appendix 5.

[5] Note that some nuance is lost in the translation.

The word frequency of each text version was calculated using the SUBTLEX-NL word frequency list (Keuleers et al., 2010).[6] As predicted, the manipulation affected the word frequency metrics. The overall word frequency of Low-LC texts was higher than the frequency of High-LC texts ($F(1,38) = 7.353$, $p = .010$); see Table 2). Manipulated words in the Low-LC text versions were on average 13.5 times more frequent than their equivalent in the High-LC text version. In addition, Low-LC texts contained a higher percentage of words that ranked among the top 1000, 5000 or 10.000 most frequent words in SUBTLEX-NL (Top1000: $F(1,38) = 6.974$, $p = .012$; Top5000: $F(1,38) = 7.004$, $p = .012$; Top10000: $F(1,38) = 5.568$, $p = .024$). Figure 1 shows the mean word frequency for all texts and text versions in the study. The materials cover a fairly wide range of frequencies ranging from 3.58 to 4.81.[7] Figure 1 also shows how the mean frequency of each Low-LC text version relates to the High-LC version and the original text. Note that the absolute difference between Low-LC and High-LC versions varies between texts. In addition, for some texts the frequency of the original text lies in the middle of the Low-LC and High-LC versions' frequencies while for other texts it lies closer to one of the versions. This illustrates the bidirectional nature of the manipulations and how manipulations depended on the specific text.

No differences were found between text versions in concreteness, type-token-ratio or syntactic structures, which shows that we were able to avoid confounds seeping into our manipulations. Manipulated words in the Low-LC version did have a slightly shorter word length (6.75 letters vs. 7.83 letters) compared to manipulated words in the High-LC version ($F(1,1729) = 49.827$, $p < .001$), but this difference was not significant at text level ($F(1,38) = 1.659$, $p = .206$).

Table 2: Mean word frequency of Low-LC and High-LC words per billion words log-transformed

Lexical complexity	**Mean**	**SD**
Low-LC	4.560	1.109
High-LC	3.428	0.972

[6] The SUBTLEX-NL frequency list is based on a collection of subtitles. The word frequency in subtitles is believed to be a closer approximation of word frequency in everyday language than corpora based on edited texts, especially for certain participant groups (Keuleers et al., 2010; Pander Maat & Dekker, 2016). Given the age of our participants, we prefer this corpus above other available Dutch corpora.

[7] Cf. Pander Maat and Dekker (2016) found a mean frequency of 4.37 (SD = 0.39) for their Dutch genre corpus which includes 10 different genres ranging from gossip-columns to popular science articles. The lowest text frequency in their corpus was 3.29 and the highest 5.38. Our manipulated versions are well within these limits.

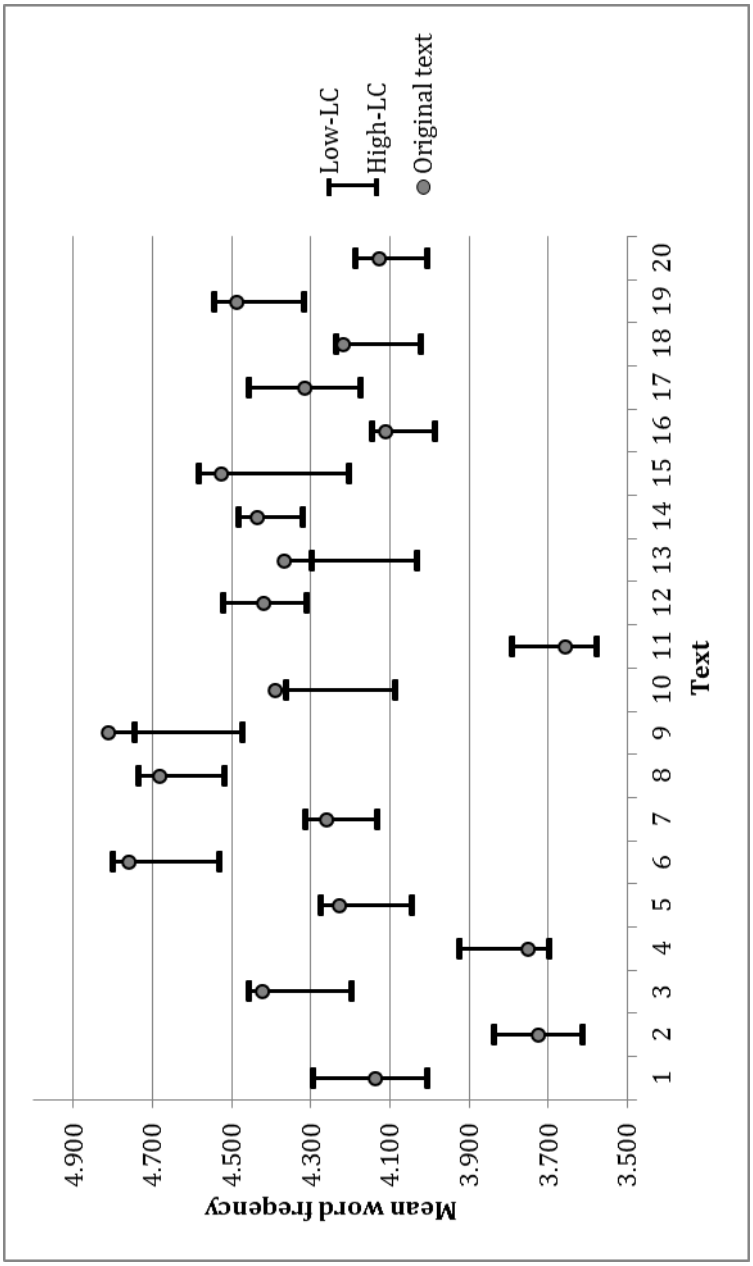


Figure 1: Mean word frequency per text and text version per billion words log-transformed [a]

[a] *Calculated over the whole text, including words that were not manipulated.*

2.1.3 Measures

Comprehension assessment. The texts were transformed into cloze tests following the Hybrid Text Comprehension cloze procedure (HyTeC-cloze; see Chapter 2). This hybrid cloze procedure combines the strengths of mechanical and rational cloze tests into a valid and reliable measure of text comprehension. The rational strategy is used to exclude words that do not rely on text level comprehension from becoming cloze gaps. This includes words that can be reconstructed using only grammatical knowledge or knowledge of usage conventions (e.g., articles and multi-word expressions). Also excluded are words that can only be guessed, such as names and numbers. All other words in the text are candidates for deletion, except for the words that were altered as part of the manipulation. The remaining candidates are divided over different cloze versions via mechanical selection. Two cloze versions were randomly selected to serve in the study. In total 10% of the words were deleted. Depending on the text length, the cloze tests contained 30 to 40 cloze gaps. The same words were deleted in the Low-LC and High-LC text version.

Reading ability. Standardized reading ability scores were made available for all students. Two different tests were used to measure Reading ability. Although scores of both tests were mapped to the same scale, analyses showed that scores for one of the tests were consistently higher. To control for this complication, the factor Reading test was included in the analyses.

2.1.4 Design

This study is part of a larger scale project in which the difficulty of 60 texts was assessed among 2926 participants. Participants were randomly assigned to this part of the study.

The experiment was set up following a matrix sampling design (e.g., Gonzalez & Rutkowski, 2010). Each participant was given four different cloze tests: one cloze test of a Low-LC educational text, one of a Low-LC public information text, one of a High-LC educational text and one of a High-LC public information text. To balance out possible order effects, each combination of cloze tests was presented in two orders.

2.1.5 Procedure

All testing took place at the participating schools. The tests were administered by the school teachers in classroom settings. Cloze tests were presented digitally on computers. Participants filled in the cloze gaps on the screen. To fill in all four cloze tests, participants took part in two sessions of 45 minutes. Schools scheduled all sessions themselves over the course of a couple weeks.

2.1.6 Scoring procedure and data-clean up

The answers to each cloze gap were dichotomously scored (1 = correct; 0 = incorrect) according to the acceptable word scoring procedure (see Chapter 2). Following the acceptable word scoring procedure, not just originally deleted words were scored as correct but semantically correct alternatives were given the same score (including spelling errors and typos). The acceptability of alternative answers was judged by the global appropriateness criterion which means that the answer had to fulfill "*all the contextual requirements of the entire discourse context in which it appears*" (Oller & Jonz, 1994, p.416). Each answer was scored by two independent judges from a pool of 16 judges. When the judges disagreed, a third judge made the final decision. All judges received a short training to familiarize them with the scoring procedure.

10% of the data was removed because students repeatedly gave non-serious answers or did not answer the cloze gaps at all. Cases where students occasionally failed to fill in a gap were regarded as incorrect answers rather than missing answers. A separate analysis of the data showed that the results did not change when these cases were excluded from the dataset. The final dataset contained 107375 cases within 3161 cloze tests.

2.1.7 Analyses

It is common practice in comprehension studies to use summed cloze scores in analyses and treat the scores as a continuous variable. However, at response level the data is binary: a correct or an incorrect answer. Aggregating binomial data is not recommended because binomial data do not necessarily follow a normal distribution. Aggregation reduces error variance and increases the chance of a type I error (Quené & Van den Bergh, 2008). The data was therefore analyzed at the response level: each answer to a cloze gap represented a case.

The data was analyzed using generalized linear mixed effect modeling (GLMM) with a logit link. Observations were nested within students and texts, with students nested in schools. After the random structure was introduced to the model, the fixed predictors were added to the model following a stepwise procedure.

Descriptions of all the predictors are given in Table 3.

Table 3: Descriptions of predictors used in Experiment 1

Predictor	Description	Levels
Lexical complexity	Text version: high or low lexical complexity	2 levels
Education level	Level of education in which the student is enrolled	5 levels
Grade	Grade in which the student is enrolled	3 levels
Reading ability	Reading ability score on Dutch reading ability test (centered and standardized)	Continuous
Reading test	Reading test used to test reading ability	2 levels
Genre	Educational text or public information text	2 levels

2.2 Results

Table 4 shows the mean percentage of gaps that was answered correctly per Lexical complexity text version, Education level and Grade. The final model is presented in Table 5.

The final model revealed main effects for Lexical complexity, Education level, Grade,[8] Reading ability and Reading test in the expected directions. Odds of a correct answer were higher in the low complexity text version, higher for students enrolled in higher levels of education or grades, and higher for students with high reading ability scores. However, Lexical complexity interacted with the education level of the student. All students benefitted from a text with lower lexical complexity, but the size of the effect differed (see Figure 2). Differences in the probability of a correct answer ranged from 1.81% to 5.45%, depending on the education level. The effect seems to follow an inverted U-shape, with the largest effect sizes for pre-vocational medium and pre-vocational high students, and smaller effects for education levels at the low end of the spectrum (pre-vocational low students) and high end of the spectrum (general and pre-university students)[9]. General students showed the smallest effect, but even for them the effect of Lexical complexity was significant (Z = 2.135, p = .033). Genre was also a significant predictor, but did not interact with Lexical complexity. Public information texts scored lower than texts taken from educational textbooks.

[8] Since the data was unbalanced – no 10th-grade pre-vocational students participated – a separate analysis was run without 10th-grade students. This analysis did not change the interpretation of the final model presented in Table 13.

[9] Although the effect size seems to increase from general education to pre-university education, this difference was not significant (Z = -1.489, p = .136).

Table 4: Average probability of a correctly answered cloze gap per Education level, Grade and Lexical complexity text version

		Lexical complexity	
Education level	**Grade**	**Low-LC**	**High-LC**
Pre-vocational (low)	Grade 8	33.39%	30.68%
	Grade 9	34.45%	31.69%
Pre-vocational (medium)	Grade 8	38.60%	34.46%
	Grade 9	44.74%	39.18%
Pre-vocational (high)	Grade 8	48.13%	41.11%
	Grade 9	54.83%	49.94%
General	Grade 8	52.90%	51.42%
	Grade 9	58.48%	55.94%
	Grade 10	66.14%	64.16%
Pre-university	Grade 8	65.39%	60.95%
	Grade 9	68.15%	64.98%
	Grade 10	72.62%	68.79%

Table 5: Final model cloze data

Random effects	**Estimates**	**SE**	***Z***	***p***	
School	0.000	0.000			
School: Student	0.253	0.015	16.867	<.001 [a]	
Text	0.122	0.018	6.777	<.001 [a]	
Fixed effects	**Estimates**	**SE**	***Z***	***p***	**Odds ratio**
Intercept	0.243	0.088	2.761	.006	1.275
Lexical complexity: Low-LC	0.146	0.025	5.840	<.001	1.157
Lexical complexity: High-LC	0 [b]				1.000
Education level: pre-voc. low	-0.672	0.103	-6.524	<.001	0.511
Education level: pre-voc. medium	-0.595	0.087	-6.840	<.001	0.552
Education level: pre-voc. high	-0.431	0.090	-4.789	<.001	0.650
Education level: general	-0.111	0.082	-1.354	.176	0.895
Education level: pre-uni.	0 [b]				1.000
Grade 8	0 [b]				1.000
Grade 9	0.096	0.054	1.778	.075	1.101
Grade 10 [c]	0.234	0.104	2.250	.024	1.264
Reading ability	0.330	0.031	10.645	<.001	1.391
Reading test: R	0 [b]				1.000
Reading test: V	-0.480	0.101	-4.752	<.001	0.619
Genre: Public information	0 [b]				1.000
Genre: Educational textbook	0.429	0.032	13.406	<.001	1.536
Low-LC * pre-voc. low	-0.051	0.041	-1.244	.214	0.950
Low-LC * pre-voc. medium	0.066	0.036	1.833	.067	1.068
Low-LC * pre-voc. high	0.079	0.040	1.975	.048	1.082
Low-LC * general	-0.067	0.045	-1.489	.136	0.935
Low-LC * pre-uni.	0 [b]				1.000

[a] One-sided. [b] Set as reference level. [c] Unbalanced, no 10th grade pre-vocational students were present in the sample.

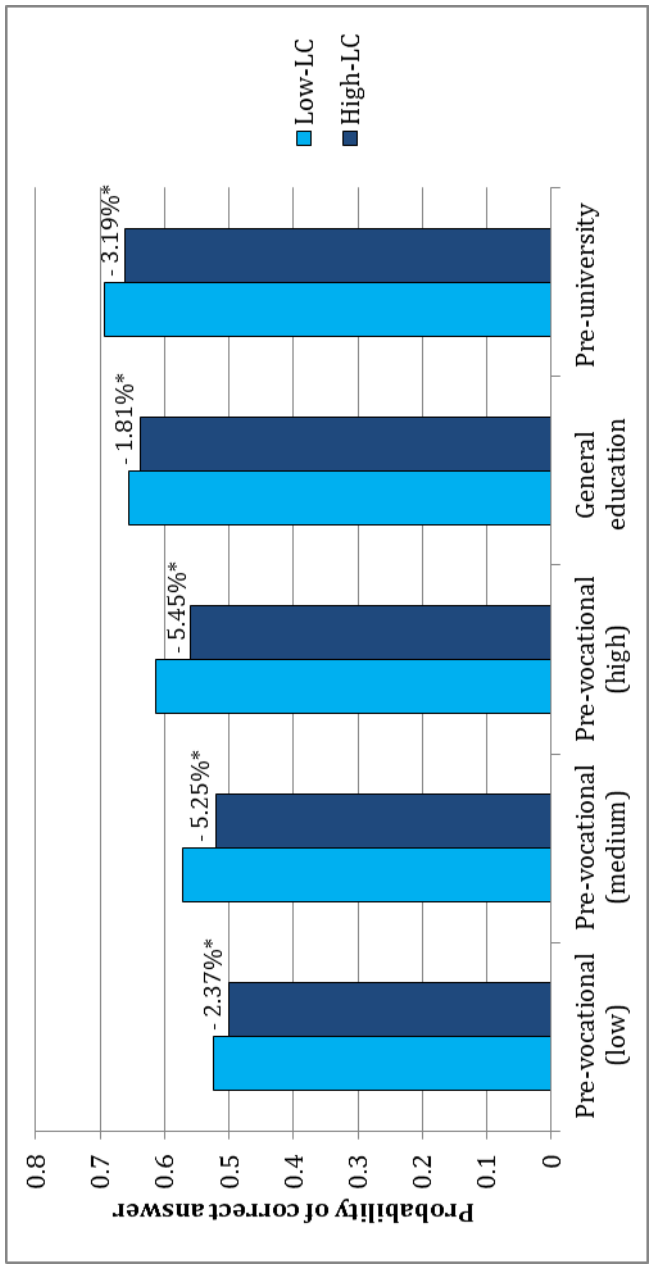


Figure 2: Final model estimated interaction effect Lexical complexity and Education level[10]

* *Significant* ($p < .05$).

[10] Estimates are set on the reference levels for Reading ability, Reading test and Genre (see Table 5). They do not reflect overall mean scores. For instance, the probability of a correct answer will be much lower for an average pre-vocational low student, since their Reading ability score is much lower than the centered mean.

2.3 Discussion

The results of Experiment 1 confirm that decreasing the lexical complexity of a text increases comprehension. Cloze scores were higher in the low lexical complexity condition than in the high lexical complexity condition. This finding replicates previous results of Freebody and Anderson (1983ab), Stahl et al. (1989) and Stahl (1991). In addition, the scale of this study enables us to generalize the effect over texts. Using multilevel modeling, we took into account the variance that exists between texts and between readers, and found an overall effect of lexical complexity. Although Experiment 1 also revealed an effect of genre – public information texts were more difficult than texts from educational textbooks – this effect did not interact with the effect of Lexical complexity. The lexical simplification was equally successful in both genres. It makes sense that the educational texts were easier, since they were written especially for the students while public information texts were written for the general public.

In contrast to previous studies, comprehension was measured using the HyTeC-cloze test. Cloze tests are not yet widely accepted as valid measures of text comprehension (e.g., Klein-Braley & Raatz, 1984; Pearson & Hamm, 2005; Shanahan, Kamil & Webb Tobin, 1982). However, the HyTeC-cloze test was especially designed to measure text comprehension and has proven to be a more reliable and valid alternative to traditional cloze tests and even to some standardized tests of reading (see Chapter 2). In addition, a decisive advantage of cloze testing is that it removes question difficulty as a confounding factor which hinders generalization over texts (Klare, 1976a).

Stahl and colleagues (1989) also used cloze testing in Study 2. While they also found that decreasing lexical complexity increased cloze scores, their effect was only significant for deleted function words. Stahl and colleagues used standard cloze tests in which content words and function words have an equal chance to be deleted. The HyTeC-cloze tests, on the other hand, samples mainly content words and only some types of function words (like subject pronouns). The inability of Stahl and colleagues to find an effect for content words could also be related to their scoring procedure. Stahl and colleagues used an exact scoring procedure in which only exact replacements of the originally deleted words count. This scoring procedure can result in floor effects which can hide the effect of the manipulation, especially for content words. Stahl et al. report means of 14.6 out of 49 content words and 28.5 out of 63 function words correctly answered, which may indicate a less than optimum range to find an effect on content words.

Another finding of Experiment 1 is that the effect size of lexical simplification differs between readers of different education levels. Low-level and high-level readers scored only 1% or 2% higher, while medium level readers scored more than 5% higher on Low-LC text versions. This inverted U-shape indicates that

there is an optimum benefit from lexical simplification. This finding is not surprising considering the way we manipulated lexical complexity. Readers will only benefit from lexical simplification if they have less knowledge of a High-LC word compared to the alternative Low-LC word. If knowledge of the Low-LC word is at the same level as the knowledge for the High-LC word, there is no benefit of simplification. Because we manipulated complexity across a wide spectrum, some manipulations will benefit a student's understanding while others will not. The number of beneficial manipulations is likely to be smaller at the outer ends of the readers' skill distribution. For low-level students a relatively large number of Low-LC words were still too difficult and for high-level students a large number of High-LC words were already easy enough. We expect that if we would include even lower-level students in Experiment 1, the results of these students would show even smaller benefits or if their level is low enough would show no benefit at all. Conversely, once a reader has surpassed a certain level even the High-LC texts are not that hard and so the benefit will also decrease. While this interaction shows that the size of the lexical complexity effect may vary between readers, we need to emphasize that lexical simplification did benefit students at all levels. Hence, it is a robust effect even though its size may be modest.

3. Experiment 2: Effects on text processing

Experiment 1 showed that lexical complexity influences comprehension. Decreasing the lexical complexity of the texts resulted in higher comprehension scores but not to the same extent for all readers. In Experiment 2 we further our investigation and focus on how lexical complexity influences on-line processing. While off-line comprehension measures show how well a reader understood the text, they do not reveal how much effort it took to get to this level of comprehension. If it took readers more effort to process the high complexity text in order to build a coherent representation of the text, this should be reflected in on-line measures. We predict that High-LC words will require more processing time compared to words in the Low-LC condition.

As discussed in Section 1.3, processing experiments on lexical complexity have primarily tested the lexical complexity effect in isolated sentences. In this experiment we examine the effect using a selection of texts from Experiment 1. By presenting lexical manipulations within the same text, we gain insight in how readers handle lexical complexity in close to normal reading situations while still keeping stylistic and conceptual difficulty separate.

3.1 Method

3.1.1 Participants

In total 181 Dutch ninth grade students participated in the study (119 female; 62 male). 80 participants were enrolled in pre-university education, 54 in general education and 47 in pre-vocational medium education. Eye movement data of 20 participants was removed from the analyses because either the calibration procedure failed or because the registration proved to be unstable. These participants were only included in the comprehension data analysis. All remaining participants had normal or corrected to normal vision.

3.1.2 Materials

Four texts were selected from the public information texts used in the cloze study (see Section 2.1.2). Stimuli were presented with black letters on a white computer screen. Because the texts were too long to present on one screen, they were split up into four or five screens. To avoid unnatural text breaks, breaks either coincided with paragraph endings or other natural break points. The segmentation was kept constant over text versions.

Comprehension questions. Each text was followed by eight multiple-choice comprehension questions targeted at information presented in the text. Questions were designed to measure the understanding of the main points of the texts, rather than the meaning of the manipulated words in particular. They did not include manipulated words, except in two cases. In these instances the words in the question were matched to the words in the text version. Questions were presented one by one on the computer screen.

Reading ability. Standardized reading ability scores were available for all but one student. A one-way Anova revealed that pre-university students had higher reading ability scores than pre-vocational and general students, but that there was no difference between pre-vocational and general students. However, standard deviations were large and some pre-vocational students performed better than pre-university students.

3.1.3 Design

The texts were divided over two lists, as a Latin-Square. As a result, participants read every text but only in one condition. Half of the participants read the texts in the reversed order.

3.1.4 Apparatus

The eye movements of the participants were recorded with a desktop eye-tracker: the SR Research EyeLink 1000. The eye-tracker recorded the position of the right pupil via a Logitech QuickCam Pro 5000 webcam at a rate of 500Hz. A remote setup with target sticker was used, allowing participants to move their head slightly. Accuracy of this eye-tracker is 0.5 degrees. Stimuli were presented on a 17 inch computer screen (1280x1024px).

3.1.5 Procedure

Participants took part in two sessions of approximately half an hour each spread over two days. In one session they read the four lexically manipulated texts discussed in the present study, in the other session they read four syntactically manipulated texts used for another study (see Chapter 4). The order was balanced between participants. Recording took place in a private room at the participating schools. Each session started with an oral instruction during which the equipment and procedure were explained. The participants were instructed to read each text at their own pace, but to make sure that they could answer comprehension questions at the end of the text.

The instruction was followed by a 13-point calibration and validation procedure. Participants fixated on a sequence of dots which appeared on various locations on the computer screen. After a successful calibration and validation sequence the testing started with a practice text and three practice questions to familiarize the participant with the procedure.

Each text fragment started with a single dot on the screen which indicated the location of the first word of the fragment. When the participant fixated on the dot, the dot vanished and the fragment appeared. To progress to the next text fragment participants pressed the 'next' button on the button-box. To answer questions participants pressed the button on the button box which corresponded to their answer. Participants could not look back in the text while answering the questions and could not go back and revise their answers. There was no time limit.

3.1.6 Data preparation and clean-up

Eye movement data. Fixations were checked and assigned to their corresponding word using Fixation 0.1.0.15 (Cozijn, 1994). Track losses were removed from the data, as well as regions that contained blinks and regions that did not receive a fixation at all. For each manipulated word, six duration measures were calculated: *First fixation duration* (FF), *First pass gaze duration* (FP), *First pass total gaze duration* (TG), *Second pass gaze duration* (SP), *Regressing path duration (RP)* and *Total fixation duration* (TF). Descriptions of all measures can be found in Table 6.

In addition, two fixation patterns were calculated: *Saccade-in* (S-in; see Figure 3a) and *Saccade-out* (S-out; Figure 3b).[11]

Table 6: Descriptions of eye-tracking measures

Measures	**Description**
First fixation duration (FF)	Duration of the first fixation within a region in first pass.
First pass gaze duration (FP)	Summed duration of all fixations and intermittent saccades within a region in first pass before the eyes leave the region (either regressively or progressively).
First pass total gaze duration (TG)	Summed duration of all fixations and intermittent saccades within a region in first pass before the eyes leave the region progressively.
Second pass gaze duration (SP)	Summed duration of all fixations and intermittent saccades within a region in second pass.
Regression path duration (RP)	Summed duration of all fixations and intermittent saccades within a region in first pass plus regressions to previous regions before the eyes leave the region progressively.
Total fixation duration (TF)	Summed duration of all fixations on the region (including second, third… n^{th} pass).
Saccade-in (S-in)	Direction from which the eyes traveled *prior* to fixating on the region (i.e., location of fixation $_{N-1}$). 1 = Regression; Fixation on the region is preceded by a fixation from a successive region. 2 = Refixation; Fixation on the region is preceded by another fixation on that same region 3 = Progression; Fixation on the region is preceded by a fixation on a previous region.
Saccade-out (S-out)	Direction in which the eyes traveled *after* fixating on the region (i.e., location of fixation$_{N+1}$). 1 = Regression; Fixation on the region is followed by a regressive fixation on a previous region. 2 = Refixation; Fixation on the region is followed by another fixation on that same region 3 = Progression; Fixation on the region is followed by a progressive fixation on a successive region.

[11] Note that although in case of a refixation S-in and S-out measure the pattern between the same two fixations, they are compared to another set of regressions and progressions.

Figure 3: Illustration fixation patterns
Figure 3a: Illustration saccade-in to manipulated word

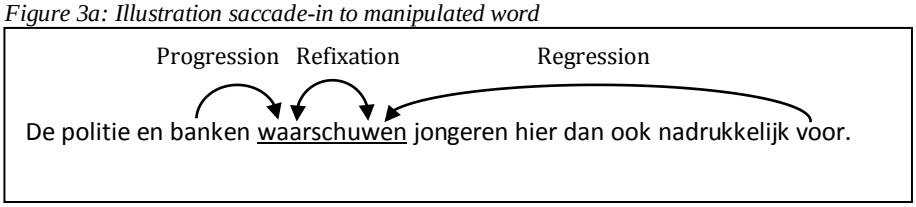
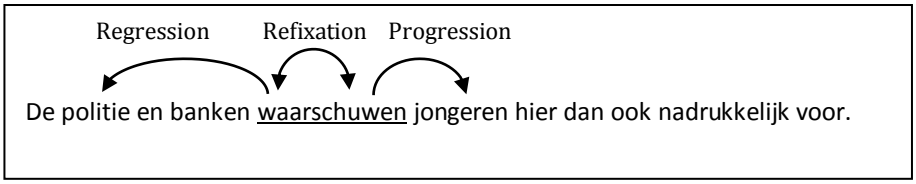


Figure 3b: Illustration saccade-out from manipulated word



Comprehension data. Responses to the comprehension questions were scored as correct or incorrect. Each question had only one correct answer. 1.2% of the data had to be removed because participants indicated to have accidently pressed the button before reading the whole question.

3.1.7 Analyses

The eye movement data was analyzed at word level. Each manipulated word pair was regarded as a separate item and measures were calculated for each word. Word pairs were nested within sentences and texts. This dependency was modeled in the random effect structure of the models. Observations were also crossed between students and texts.

The following fixed predictors were added to the model: Lexical complexity, Education level, Reading ability and Word length. Descriptions of these predictors are given in Table 7. Word length was introduced to the models to confirm that the effect of lexical complexity remained even when corrected for the length of the word.[12] The final models are presented in the results section.

[12] Manipulated words were slightly longer in the High-LC version than in the Low-LC version (see Section 2.1.2). By adding word length to the models, this slight imbalance was controlled.

Table 7: Descriptions of predictors used in Experiment 2

Predictor	Description	Levels
Lexical complexity	Text version: high or low lexical complexity	2 levels
Education level	Level of education in which the student is enrolled	3 levels
Reading ability	Reading ability score on Dutch reading ability test (centered and standardized)	Continuous
Word length	Length of the word in letters (centered)	Continuous

3.2 Results

3.2.1 Duration measures

A log-transformation was carried out on all duration measures to normalize the distributions (Baayen, 2008). The final models of the duration measures are shown in Table 9. All duration measures showed the same pattern with main effects for Lexical complexity, Education level and Reading ability in the expected directions. Durations were longer in the High-LC text version than in the Low-LC version (See Figure 4; $\chi^2_{FF}(1) = 62.312$, $p < .001$; $\chi^2_{FP}(1) = 150.048$, $p < .001$; $\chi^2_{SP}(1) = 56.996$, $p < .001$; $\chi^2_{TG}(1) = 258.558$, $p < .001$; $\chi^2_{RP}(1) = 178.169$, $p < .001$; $\chi^2_{TF}(1) = 332.104$, $p < .001$). Main effects for Education Level were also found (See Figure 5; $\chi^2_{FF}(2) = 19.064$, $p < .001$; $\chi^2_{FP}(2) = 27.817$, $p < .001$; $\chi^2_{SP}(2) = 25.206$, $p < .001$; $\chi^2_{TG}(2) = 25.228$, $p < .001$; $\chi^2_{RP}(2) = 12.973$, $p = .002$; $\chi^2_{TF}(2) = 9.917$, $p = .007$). Pre-vocational students read slower than general and pre-university students, but the contrast for general and pre-university students did not reach significance in any of the duration measures. In addition, there was a significant interaction between Lexical complexity and Education level in the late measures *Second pass gaze duration* ($\chi^2_{SP}(2) = 8.053$, $p = .018$) and *First pass total gaze duration* ($\chi^2_{TG}(2) = 6.205$, $p = .045$), and a trend towards an interaction in *Total fixation duration* ($\chi^2_{TG}(2) = 4.532$, $p = .104$). The effect of Lexical complexity was larger for Pre-vocational students than for Pre-university and General students (see Figure 6). Main effects were also found for Reading ability and Word length. Students with higher Reading ability scores read faster compared to students with lower scores ($\chi^2_{FF}(1) = 5.621$, $p = .018$; $\chi^2_{FP}(1) = 9.301$, $p = .002$; $\chi^2_{SP}(1) = 3.940$, $p = .047$; $\chi^2_{TG}(1) = 10.106$, $p = .001$; $\chi^2_{RP}(1) = 9.433$, $p = .002$; $\chi^2_{TF}(1) = 9.906$, $p = .002$. Longer words had longer reading times ($\chi^2_{FF}(1) = 8.292$, $p = .004$; $\chi^2_{FP}(1) = 602.985$, $p < .001$; $\chi^2_{SP}(1) = 120.357$, $p < .001$; $\chi^2_{TG}(1) = 786.522$, $p < .001$; $\chi^2_{RP}(1) = 486.760$, $p < .001$; $\chi^2_{TF}(1) = 670.650$, $p < .001$).

Table 8: Means and standard deviations duration measures at word level in milliseconds

Measure	Lexical complexity	Pre-vocational	General	Pre-university
First fixation duration	Low-LC	290 (164)	256 (133)	256 (131)
	High-LC	294 (174)	264 (140)	260 (140)
First pass gaze duration	Low-LC	372 (278)	320 (225)	312 (217)
	High-LC	388 (305)	335 (235)	325 (236)
Second pass gaze duration	Low-LC	380 (316)	330 (222)	324 (213)
	High-LC	533 (401)	408 (268)	393 (344)
First pass total gaze duration	Low-LC	434 (341)	384 (297)	380 (302)
	High-LC	456 (366)	403 (308)	401 (329)
First pass regression path duration	Low-LC	504 (502)	444 (504)	438 (585)
	High-LC	525 (514)	466 (527)	463 (617)
Total fixation duration	Low-LC	412 (314)	352 (258)	340 (249)
	High-LC	434 (344)	369 (267)	358 (274)

Table 9: Final models for eye-tracking duration measures at word level in log10(ms)
Table 9a: Final model First fixation duration

First fixation duration				
Random effects	*Estimates*	*St.Dev*		
Student	0.00213	0.04609		
Text	0.00003	0.00538		
Text: Sentence	0.00030	0.01736		
Text: Sentence: Word pair	0.00201	0.04486		
Residual	0.03272	0.18089		
Fixed effects	*Estimates*	*SE*	*t*	*p*
Intercept	2.399	0.009	279.593	<.001
Lexical complexity: Low-LC	-0.022	0.003	-7.894	<.001
Lexical complexity: High-LC	0[a]			
Education level: Pre-voc. (medium)	0.037	0.010	3.735	<.001
Education level: General	-0.003	0.010	-0.275	.784
Education level: Pre-uni.	0[a]			
Reading ability	-0.010	0.004	-2.371	.019
Word length	0.002	0.001	2.880	.004

[a] *Set as reference level.*

Table 9b: Final model First pass gaze duration

First pass gaze duration				
Random effects	*Estimates*	*St.Dev*		
Student	0.00377	0.06136		
Text	0.00000	0.00000		
Text: Sentence	0.00074	0.02728		
Text: Sentence: Word pair	0.00425	0.06516		
Residual	0.03746	0.19354		
Fixed effects	*Estimates*	*SE*	*t*	*p*
Intercept	2.427	0.011	223.528	<.001
Lexical complexity: Low-LC	-0.037	0.003	-12.249	<.001
Lexical complexity: High-LC	0[a]			
Education level: Pre-voc. medium	0.065	0.013	4.974	<.001
Education level: General	0.009	0.013	0.749	.455
Education level: Pre-uni.	0[a]			
Reading ability	-0.017	0.005	-3.050	.003
Word length	0.020	0.001	24.556	<.001

[a] *Set as reference level.*

Table 9c: Final model Second pass gaze duration

Second pass gaze duration				
Random effects	*Estimates*	*St.Dev*		
Student	0.00361	0.06010		
Text	0.00057	0.02388		
Text: Sentence	0.00101	0.03171		
Text: Sentence: Word pair	0.00371	0.06089		
Residual	0.05073	0.22524		
Fixed effects	*Estimates*	*SE*	*t*	*p*
Intercept	2.394	0.022	109.714	<.001
Lexical complexity: Low-LC	-0.067	0.017	-3.928	<.001
Lexical complexity: High-LC	0[a]			
Education level: Pre-voc. medium	0.118	0.022	5.323	<.001
Education level: General	0.011	0.021	0.523	.601
Education level: Pre-uni.	0[a]			
Reading ability	-0.015	0.008	-1.985	.049
Word length	0.026	0.002	10.971	<.001
Low-LC * Pre-voc. medium	-0.062	0.026	-2.384	.017
Low-LC * General	0.010	0.026	0.368	.713
Low-LC * Pre-uni.	0[a]			

[a] *Set as reference level.*

Table 9d: Final model First pass total gaze duration

First pass total gaze duration				
Random effects	*Estimates*	*St.Dev*		
Student	0.00505	0.07108		
Text	0.00007	0.00814		
Text: Sentence	0.00049	0.02210		
Text: Sentence: Word pair	0.00460	0.06780		
Residual	0.03661	0.19135		
Fixed effects	*Estimates*	*SE*	*t*	*p*
Intercept	2.460	0.012	196.901	<.001
Lexical complexity: Low-LC	-0.044	0.005	-9.729	<.001
Lexical complexity: High-LC	0[a]			
Education level: Pre-voc. medium	0.079	0.015	5.166	<.001
Education level: General	0.012	0.015	0.823	.412
Education level: Pre-uni.	0[a]			
Reading ability	-0.020	0.006	-3.179	.002
Word length	0.023	0.001	28.045	<.001
Low-LC * Pre-voc. medium	-0.016	0.007	-2.215	.027
Low-LC * General	0.001	0.007	0.121	.904
Low-LC * Pre-uni.	0[a]			

[a] *Set as reference level.*

Table 9e: Final model Regression path duration

Regression path duration				
Random effects	*Estimates*	*St.Dev*		
Student	0.00581	0.07625		
Text	0.00011	0.01060		
Text: Sentence	0.00000	0.00000		
Text: Sentence: Word pair	0.00592	0.07693		
Residual	0.05505	0.23462		
Fixed effects	*Estimates*	*SE*	*t*	*p*
Intercept	2.518	0.014	184.449	<.001
Lexical complexity: Low-LC	-0.049	0.004	-13.348	<.001
Lexical complexity: High-LC	0[a]			
Education level: Pre-voc. medium	0.056	0.016	3.416	<.001
Education level: General	0.009	0.016	0.570	.569
Education level: Pre-uni.	0[a]			
Reading ability	-0.021	0.007	-3.071	.003
Word length	0.022	0.001	22.063	<.001

[a] *Set as reference level.*

Table 9f: Final model Total fixation duration

Total fixation duration				
Random effects	*Estimates*	*St.Dev*		
Student	0.00648	0.08052		
Text	0.00013	0.01157		
Text: Sentence	0.00128	0.03576		
Text: Sentence: Word pair	0.00556	0.07459		
Residual	0.04402	0.20981		
Fixed effects	*Estimates*	*SE*	*t*	*p*
Intercept	2.517	0.015	170.728	<.001
Lexical complexity: Low-LC	-0.053	0.005	-11.011	<.001
Lexical complexity: High-LC	0[a]			
Education level: Pre-voc. medium	0.054	0.017	3.124	.002
Education level: General	-0.000	0.017	-0.026	.979
Education level: Pre-uni.	0[a]			
Reading ability	-0.022	0.007	-3.147	.002
Word length	0.023	0.001	25.897	<.001
Low-LC * Pre-voc. medium	-0.015	0.008	-2.036	.042
Low-LC * General	-0.002	0.007	-0.281	.779
Low-LC * Pre-uni.	0[a]			

[a] *Set as reference level.*

Figure 4: Estimated main effect of Lexical complexity for First fixation duration, First pass gaze duration and Regression path duration

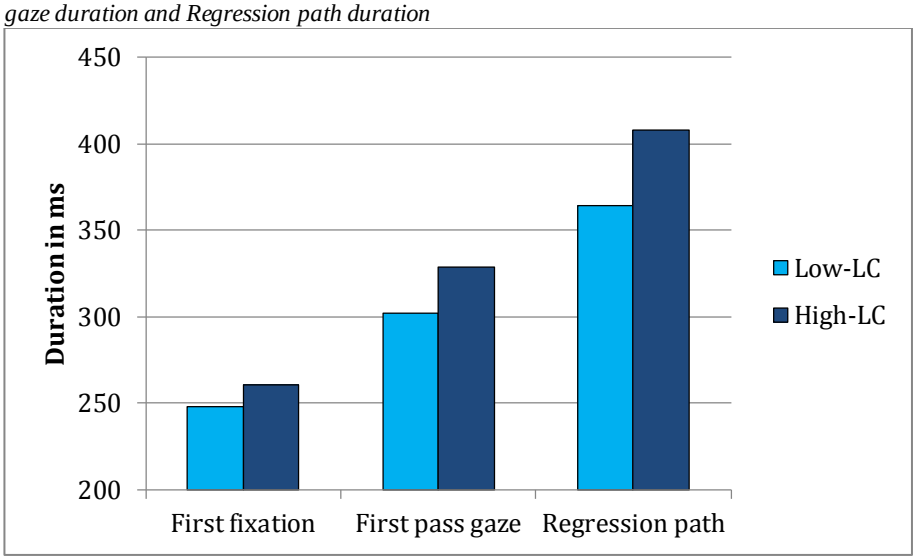
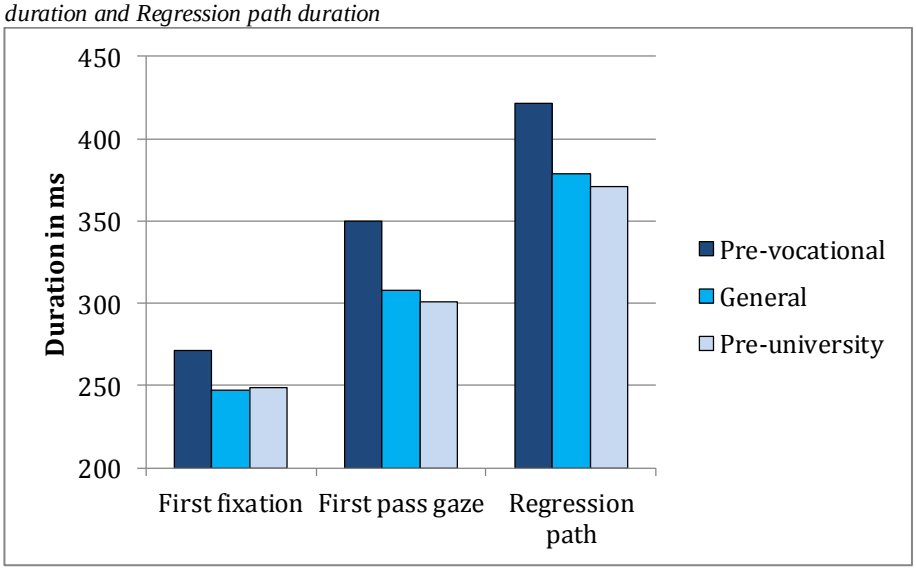


Figure 5: Estimated main effect of Education level for First fixation duration, First pass gaze duration and Regression path duration



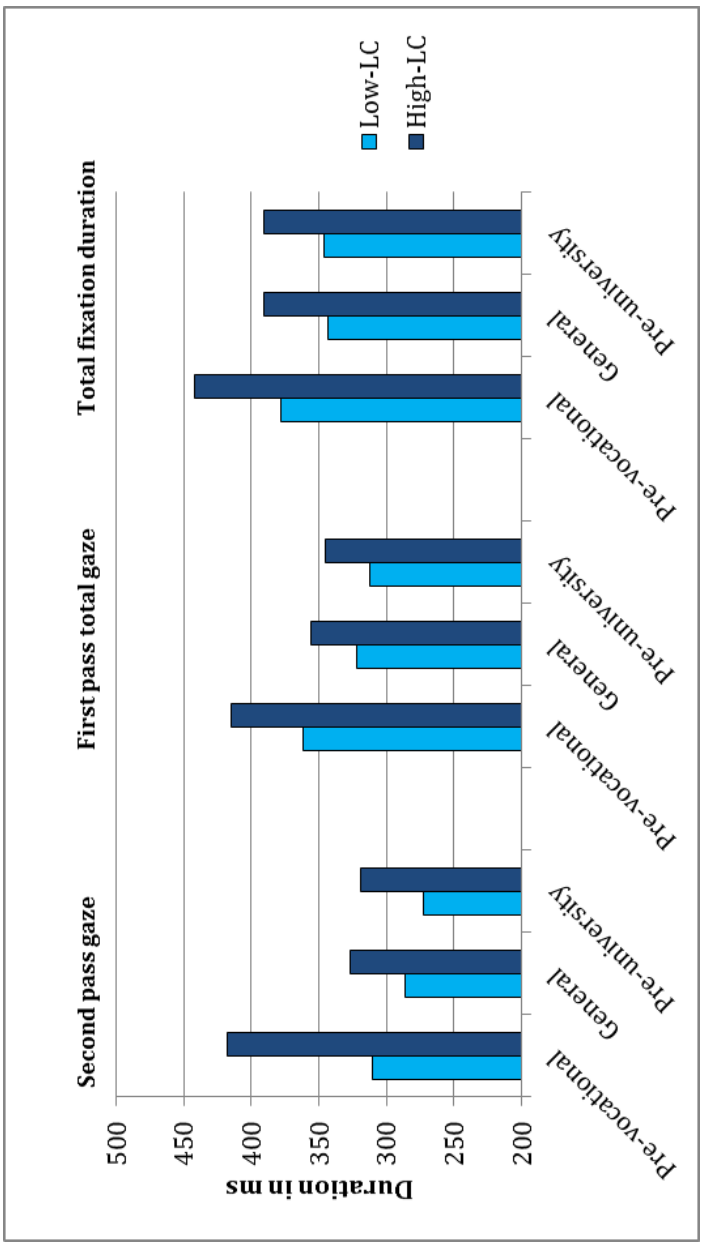


Figure 6: Estimated interaction effect of Lexical complexity and Education level for Second pass gaze duration, First pass total gaze duration and Total fixation duration

3.2.2 Fixation patterns

Percentages for the fixation pattern measures are given in Table 10 and the final models are given in Table 11. The random structure had to be flattened in order for the models to convergence. Text and Sentence levels were therefore dropped.

All predictors proved to be significant in predicting the Saccades-in and Saccades-out of manipulated words (Lexical complexity: $F_{S\text{-}in}(2,32.929) = 9.449$, $p < .001$; $F_{S\text{-}out}(2,32.439) = 7.817$, $p < .001$; Education level: $F_{S\text{-}in}(4,32.929) = 2.450$, $p = .044$; $F_{S\text{-}out}(4, 32.439) = 2.982$, $p = .018$; Reading ability: $F_{S\text{-}in}(2,32.929) = 3.761$, $p = .023$; $F_{S\text{-}out}(2,32.439) = 3.938$, $p = .020$; Word length: $F_{S\text{-}in}(2,32.929) = 1927.009$, $p < .001$; $F_{S\text{-}out}(2,32.439) = 443.731$, $p < .001$). The final models of the Saccade-in and Saccade-out show that Lexical complexity did not affect the number of regressive fixations toward or from a manipulated word.[13] However, lexical complexity did affect the number of refixations compared to the number of progressive fixations. High-LC words were more often immediately refixated, compared to Low-LC words. Lexical complexity did not interact with Education level. Pre-vocational students made fewer regressions and more refixations than pre-university students. Students with high Reading ability scores were less likely to refixate a word compared to students with low scores. Word length was also significant ($F(2,32.439) = 443.731$, $p < .001$). Long words were more likely to be refixated and had a higher chance of being followed by a regression than shorter words.

Table 10: Mean percentages for fixation patterns at manipulated word

Measure	Lexical complexity	Pattern	Pre-vocational	General	Pre-university
S-IN	Low-LC	Regression	10.80%	11.10%	12.59%
		Refixation	21.58%	19.27%	17.57%
		Progression	67.62%	69.63%	69.83%
	High-LC	Regression	10.91%	11.27%	11.77%
		Refixation	28.47%	25.19%	23.28%
		Progression	60.62%	63.53%	64.94%
S-OUT	Low-LC	Regression	12.74%	14.02%	14.86%
		Refixation	21.78%	19.65%	17.89%
		Progression	65.49%	66.33%	67.24%
	High-LC	Regression	11.13%	13.43%	14.38%
		Refixation	28.88%	25.66%	23.54%
		Progression	60.00%	60.91%	62.08%

[13] Note that fixation patterns only describe successive fixation pairs of 2 fixations; so, fixations after fixating on a manipulated word or fixations prior to fixating on a manipulated word. If a manipulated word is part of a regression path but not the target of the first regressive fixation, it is not reflected in this measure.

Table 11:Final models fixation patterns
Table 11a: Final model Saccade-in

Saccade-in				
1. PROGRESSION[a] VS. REGRESSION				
Random effects	*Estimates*	*SE*	*z*	*p*
Student	0.139	0.023	6.129	<.001[b]
Word pair	0.732	0.098	7.434	<.001[b]
Fixed effects	*Estimates*	*SE*	*t*	*p*
Intercept	-2.002	0.091	-21.884	<.001
Lexical complexity: Low-LC	-0.045	0.037	-1.215	.224
Lexical complexity: High-LC	0[c]			
Education level: Pre-vocational	-0.073	0.091	-0.802	.423
Education level: General	-0.095	0.087	-1.083	.279
Education level: Pre-university	0[c]			
Reading ability	-0.020	0.038	-0.535	.592
Word length	0.035	0.006	5.695	<.001
2. PROGRESSION[a] VS. REFIXATION				
Random effects	*Estimates*	*SE*	*z*	*p*
Student	0.151	0.023	6.663	<.001[b]
Word pair	1.421	0.172	8.244	<.001[b]
Fixed effects	*Estimates*	*SE*	*t*	*p*
Intercept	-2.726	0.114	-23.869	<.001
Lexical complexity: Low-LC	-0.144	0.033	-4.319	<.001
Lexical complexity: High-LC	0[c]			
Education level: Pre-vocational (medium)	0.259	0.090	2.869	.004
Education level: General	0.095	0.087	1.088	.277
Education level: Pre-university	0[c]			
Reading ability	-0.102	0.038	-2.710	.007
Word length	0.363	0.006	61.908	<.001

[a] Progression is taken as the reference level. Positive estimates denote a decrease in progressions and an increase of the contrasted pattern. [b] One-sided. [c] Set as reference level.

Table 11b: Final model Saccade-out

Saccade-out				
1. PROGRESSION[a] VS. REGRESSION				
Random effects	*Estimates*	*SE*	*Z*	*p*
Student	0.167	0.026	6.410	<.001[b]
Word pair	0.806	0.105	7.666	<.001[b]
Fixed effects	*Estimates*	*SE*	*t*	*p*
Intercept	-1.747	0.100	-17.529	<.001
Lexical complexity: Low-LC	0.001	0.036	0.024	.981
Lexical complexity: High-LC	0[c]			
Education level: Pre-vocational (medium)	-0.229	0.096	-2.376	.018
Education level: General	-0.048	0.092	-0.525	.600
Education level: Pre-university	0[c]			
Reading ability	-0.031	0.040	-0.786	.432
Word length	0.021	0.008	2.477	.013
2. PROGRESSION[a] VS. REFIXATION				
Random effects	*Estimates*	*SE*	*Z*	*p*
Student	0.118	0.018	6.605	<.001[b]
Word pair	0.203	0.031	6.559	<.001[b]
Fixed effects	*Estimates*	*SE*	*t*	*p*
Intercept	-2.189	0.072	-30.367	<.001
Lexical complexity: Low-LC	-0.122	0.031	-3.871	<.001
Lexical complexity: High-LC	0[c]			
Education level: Pre-vocational (medium)	0.185	0.080	2.314	.021
Education level: General	0.086	0.078	1.109	.267
Education level: Pre-university	0[c]			
Reading ability	-0.092	0.034	-2.728	.006
Word length	0.231	0.008	29.719	<.001

[a] Progression is taken as the reference level. Positive estimates denote a decrease in progressions and an increase of the contrasted pattern. [b] One-sided. [c] Set as reference level.

3.2.3 Comprehension questions

Analysis of the answers to the multiple-choice questions revealed main effects for Lexical complexity and Education Level in the expected directions (Table 13). Performance on the comprehension questions was better in the Low-LC version. Students enrolled in general and pre-university education performed better than pre-vocational students. Reading ability was not significant when Education level was in the model. No interactions were found between Lexical complexity, Education Level or Reading ability.

Table 12: Percentage correctly answered multiple-choice questions per text version and Education level

Lexical complexity	Pre-vocational	General	Pre-university
Low-LC	57.24%	68.72%	72.03%
High-LC	53.10%	65.96%	69.65%

Table 13: Final model multiple-choice questions

Random effects	Estimates	SE	*Z*	*p*	
Student	0.103	0.026	3.962	<.001 [a]	
Text	0.008	0.008	1.000	.159	
Fixed effects	**Estimates**	**SE**	***Z***	***p***	**Odds ratio**
Intercept	0.823	0.078	10.551	<.001	2.277
Lexical complexity: Low-LC	0.138	0.057	2.421	.015	1.148
Lexical complexity: High-LC	0 [b]				1.000
Education level: pre-voc. medium	-0.686	0.091	-7.538	<.001	0.504
Education level: general	-0.168	0.089	-1.888	.059	0.845
Education level: pre-uni.	0 [b]				1.000

[a] *One-sided.* [b] *Set as reference level.*

3.3 Discussion

Experiment 2 showed that processing lexically complex words in natural contexts requires more processing time compared to words that are less lexically complex. This effect was observed in early reading times (*First fixation duration*, *First pass gaze duration*) and in late measures (*First pass total gaze duration*, *Second pass gaze*, *Regression path duration*, *Total fixation duration*). Analyses of the fixation patterns showed that manipulated words in the High-LC texts were more often immediately refixated than their counterparts in the Low-LC texts. A fixation on a High-LC word was more often followed by another fixation on that same word. Lexical complexity did not influence regressions. The longer reading times in the late measures were not caused by High-LC words evoking more regressions and High-LC words were not the immediate target of more regressions. This last finding is not in line with previous results of Williams and Morris (2004) and Chaffin et al. (2001). Both studies found more regressions into lexically complex words. We believe the differences in the experimental setups caused readers to adapt different reading strategies.

First of all, Williams and Morris (2004) presented lexically complex words in isolated sentences and Chaffin et al. presented their words in a limited two-sentence-context. In both studies it is clear that the sentence the reader is seeing is all the information he is going to get. Furthermore, readers are instructed that they have to read for understanding since a comprehension question can follow the sentence. It is highly likely that this question will involve the complex word. In our experiment, however, the words are embedded in a vast context. The upcoming

sentences may fill in some of the blanks concerning the complex word and may show that the complex word is not very relevant in light of the whole text (i.e., not worth spending extra time on). It was also less likely that our questions would target a specific complex word since there were many of them and the size of the text would make it very inefficient to reinspect every single complex word. In fact, 'skipping' difficult words – as in, not getting hung up on difficult words unless the context indicates they are vital for good comprehension – can be a very good reading strategy according to many scholars (Pressley & Allington, 2015). In addition, our manipulation of lexical complexity was not extreme. The High-LC words were not made-up words and not very infrequent. Most words should be known by our participants, only to a lesser extent than Low-LC words. Thus, High-LC words may take more time to retrieve from memory and to integrate them into a mental representation of the text, but the retrieved meaning may be good enough to progress through the text.

The results of Experiment 2 show an interaction between Lexical complexity and Education level, but only in the late duration measures. In the early measures, in the fixation patterns and in the comprehension scores, only main effects for Lexical complexity and Education level were found. It seems that initially students of all levels processed High-LC words the same. At first high-level students benefit as much from lexical simplification as low-level students do. However, when High-LC words are reread, high-level students require less time than low-level students.

Overall, general students and pre-university students read faster and performed better on the comprehension questions than pre-vocational students. Although general education students tended to read slower and answer less questions correctly compared to pre-university students, the difference did not reach significance. This may be due to a biased selection since the general education students and pre-university students who participated were all part of combination classes. As a result, the general education students may perform better than general education students who are not taught alongside pre-university students. A comparison between the 9th-grade general education students who participated in Experiment 1 and the general education students who participated in Experiment 2 confirmed this suspicion. General education students from Experiment 2 had significantly higher standardized reading ability scores than general education students in Experiment 1 ($t(136) = -4.024$, $p < .001$).

4. General Discussion

In two experiments we investigated the effect of lexical complexity on text comprehension and on-line processing. In contrast to earlier studies, we used a large number of texts and presented these texts to readers varying in reading skill. Twenty texts were randomly selected and manipulated to create a low lexical complexity version and high lexical complexity version of each text. Twenty percent of the content words were substituted for words that were intuitively less familiar or more familiar than the equivalent words used in the other text version. The alterations were conservative and were controlled for confounding factors like content, word length and syntactic structure. In Experiment 1, HyTeC-cloze tests were used to measure how lexical complexity affected text comprehension. In Experiment 2, eye-tracking was used to measure processing times for manipulated words embedded in natural text.

The results show that lexical simplification has a positive effect on comprehension and on on-line processing times. Reducing the lexical complexity improved the probability of a correct answer on the cloze items up to 5 percent. Although the size of the effect depended on the educational level of the students, students of all levels benefited from lexical simplification. This finding is further supported by the multiple-choice question results of Experiment 2. It replicated the effect of lexical complexity on comprehension found in Experiment 1, both in size and in direction. Only the interaction with education level did not reach significance, which could be related to either the smaller number of participants or the lower number of items used in Experiment 2.

Lexical simplification also reduces the time that is necessary to process words. Early and late duration measures showed shorter processing times for Low-LC words than for High-LC words. This effect was larger in the late measures for low-level students than for high-level students, but initially all students benefitted equally from simplification.

Looking at the size of the lexical complexity effect, the effect on comprehension may seem small. The odds of a correct answer were only 1.16 times higher in the Low-LC text versions than in the High-LC text versions (1.15 for multiple-choice questions in Experiment 2). For a cloze test of 30 gaps, this means that only one or two more gaps are answered correctly in the Low-LC version. Although the lexical complexity effect was highly significant, its practical significance and therefore the real gain for readability is small. Yet, given the fact that not all manipulations were presumed to be equally strong, 1.16 odds can be regarded as quite a lot. Our texts varied in their level of lexical complexity and the extent to which they allowed lexical simplification. In addition, this effect is not confounded with changes in content, syntactic structure or other factors. Therefore,

our findings represent the effects of lexical simplification without various forms of inflation. This is the overall effect that can be achieved by changing just the lexical stylistic difficulty of a random text.

We are unaware of any lexical complexity study that used 20 texts that were randomly selected and not picked out because of their potential for lexical simplification. Due to the scale of this study, we are confident that lexical simplification affects readability and that previous findings are not restricted to the particular stimuli used in these studies. We now know that lexical simplification has small but robust positive effects on both aspects of readability: comprehension and processing ease.

4 How Syntactic Dependency Length affects readability

In the previous chapter we focused on lexical complexity and how it affects comprehension as well as on-line processing. In this chapter we focus on how syntactic complexity – in particular syntactic dependency length (SDL) – affects readability. Traditionally, readability research has focused on sentence length as an index of syntactic complexity (Bailin & Grafstein, 2016; Dale & Chall, 1948; Flesch, 1948; Klare, 1963; 1974; Staphorsius, 1994). Although sentence length is strongly correlated with complexity, length and complexity are not equivalent (Bailin & Grafstein, 2016; Davison et al., 1980; Gough, 1966). As a result, splitting up long sentences has little effect on the readability of a text (Coleman, 1962; Duffy & Kabance, 1982; Johnson & Otto, 1982) and it can even result in reduced readability (Davison & Kantor, 1982).

As is apparent in Example (1) below, long sentences are not necessarily difficult. This sentence, taken from the popular children's book Stuart Little, contains 107 words but is much easier to understand than the shorter sentence in (2). Although (1) contains more information than (2), it has a transparent syntactical structure making it easier for a reader to see how elements are connected. Example (2) below contains only 9 words but is extremely difficult due to a double center-embedding.

(1) In the loveliest town of all, where the houses were white and high and the elm trees were green and higher than the houses, where the front yards were wide and pleasant and the back yards were bushy and worth finding out about, where the streets sloped down to the stream and the stream flowed quietly under the bridge, where the lawns ended in orchards and the orchards ended in fields and the fields ended in pastures and the pastures climbed the hill and disappeared over the top toward the wonderful wide sky, in this loveliest of all towns Stuart stopped to get a drink of sarsaparilla.

(E.B. White, Stuart Little)

(2) The man the woman the boy saw knew died.

(Kemper & Kemtes, 2002, p.80)

The syntactic structure of a sentence is supposed to help readers understand who did what to whom. It lets readers know how elements are connected and guides them in

building a coherent mental representation of the text. Complex structures like in (2) introduce an extra level of complexity and lower the readability of the text. Especially, since the same information can be related using an easier syntactic structure, as in (3).[1]

(3) The boy saw the woman who knew the man who died.

(Kemper & Kemtes, 2002, p.80)

The ease with which a text can be read is one aspect that determines a text's readability. The other aspect is how well it is understood (see Chapter 1). Yet, many studies that have investigated the effects of syntactic complexity on readability have focused solely on processing ease (Bailin & Grafstein, 2016). By comparing processing times of supposedly easy and complex structures, they have tried to determine the cognitive mechanisms underlying syntactic complexity. These studies have helped develop processing models that predict exactly where processing delays will occur within a sentence. For the most part, they do not predict whether comprehension of the sentence will be affected.

For readability purposes it is important to investigate both processing ease and comprehension to see how syntactic complexity affects readers. Readers may need more time to process complex syntactic structures, but this does not have to result in diminished comprehension. In the present study we examine the effect of complexity on on-line processing as well as on off-line comprehension. Furthermore, we present sentences in natural contexts in order to test the actual effects of syntactic complexity on these aspects of readability. This is necessary if we want to generalize our findings to normal reading situations. Studies that have embedded syntactically complex sentences in supportive contexts have already shown that these contexts reduce processing delays or that they make delays disappear completely (Crain & Steedman, 1985; Davison & Lutz, 1985; Grodner, Gibson & Watson, 2005; Kaiser & Trueswell, 2004; Levy, Fedorenko, Breen & Gibson, 2012). If this also holds for extensive contexts, reducing the syntactic complexity of a text may only have a limited effect on readability. In order to test this, we investigate the syntactic complexity of texts that readers can encounter in everyday life.

[1] Although these sentences describe the same situation, we note that focus has shifted from the man to the boy. Within a discourse context, such a change may be unwanted.

1. Theoretical approaches to syntactic complexity and syntactic dependency length

Through the years there have been many theories explaining why certain syntactic structures are more difficult to understand than others (e.g., *Active filler hypothesis*, Clifton & Frazier, 1989; *Dependency Locality Theory*, Gibson, 1998; 2000; *Surprisal*, Hale, 2001). These theories are traditionally divided into memory-based approaches and expectation-based approaches, but more recent work suggests that a complete theory of syntactic complexity will require a combination of both approaches (Demberg & Keller, 2008; Futrell & Levy, 2017; Levy, Fedorenko & Gibson, 2013; Levy & Keller, 2013; Nicenboim, Vasishth, Gattei, Sigman & Kliegl, 2015; Vasishth & Drenhaus, 2011).

1.1 Memory-based versus expectation-based approaches

According to memory-based approaches (e.g., *Dependency Locality Theory*, Gibson, 1998; 2000; *Activation and cue-based retrieval theory*, Vasishth & Lewis, 2006; *Similarity-based interference theory*, Gordon, Hendrick, Johnson & Lee, 2006), syntactic complexity results from limitations of working memory. Successful syntactic parsing requires integration of elements over the course of the sentence. Complex syntactic structures require more computational resources for storage, retrieval and integration of elements compared to simple structures. A key concept underlying these accounts is 'locality'. Processing costs are low when elements that have to be integrated are adjacent (e.g., syntactic heads and their dependents like verb and subject). When elements are separated by other elements (i.e., 'non-local dependencies'), the processing costs go up. The unresolved dependency – and any additional unresolved dependencies it may cross – must be kept active in working memory for a longer period of time and is subject to decay and interference. This explains why center-embeddings like (2) are more difficult than right-branching structures like (3).

(2) [The man [the woman [the boy saw] knew] died.]

(3) [The boy saw the woman] who knew the man] who died.]

Memory-based approaches are supported by studies that have found increased reading times for non-local dependencies, particularly at the point where the processing cost is at its highest (Bartek, Lewis, Vasishth & Smith, 2011; Gibson, 1998; 2000; Grodner & Gibson, 2005). For instance, Grodner and Gibson (2005) manipulated the distance between the subject of a relative clause and the embedded verb (4) - (6). Reading times for the verb '*supervised*' increased as the distance

between the subject and the verb increased. This increase in processing times is referred to as the 'locality effect'.

(4) The administrator who the nurse supervised scolded the medic…

(5) The administrator who the nurse from the clinic supervised scolded the medic…

(6) The administrator who the nurse who was from the clinic supervised scolded the medic…

However, increasing the distance between a head and its dependent does not always result in a delay at the point of the dependency resolution. In certain circumstances, increasing the distance between a dependent and the head facilitates processing of the syntactic head (Konieczny, 2000; Levy & Keller, 2013; Vasishth & Lewis, 2006). This 'anti-locality effect' has been predominantly found in SOV languages like German and Hindi where head-final structures are common. In head-final structures the head is easier to process because the intervening material guides predictions about what the head will be.

The facilitating effect of intervening materials is not directly in line with memory-based accounts (but see Vasishth & Lewis, 2006), but it is in line with expectation-based accounts (Hale, 2001; Levy, 2008; MacDonald & Christiansen, 2002). According to these accounts, computational resources are allocated according to expectations. Readers use linguistic knowledge and world knowledge to predict what element is coming next. As a reader progresses through a text or sentence, new information comes in, guiding and refining expectations. When expectations are met, integration of a new word into the existing structure runs smoothly. Conversely, if expectations are not met, processing difficulties will occur.

Levy and Keller (2013) found evidence indicating that expectation and memory factors interact. They manipulated the location of a dative NPs and PP adjuncts (7). Adding the dative NP to the relative clause facilitated processing of the target region (a vs. c), but when the PP adjunct was also moved to the relative clause, the effect changed direction and a locality effect was found (c vs. d). Thus, their anti-locality effect turned into a locality effect when the distance between the head and its dependent was extremely large.

(7) a. Nachdem der Lehrer [*zur zusätzlichen Ahndung des mehrfachen Fehlverhaltens*]^{adj} [*dem ungezogenen Sohn des fleißigen Hausmeisters*]^{dat} den Strafunterricht verhängte, hat der Mitschüler, der den Fußball versteckt hat, die Sache bereinigt.

b. … hat der Mitschüler, der [*zur zusätzlichen Ahndung des mehrfachen Fehlverhaltens*]^{adj} den Fußball versteckt hat, …
c. … hat der Mitschüler, der [*dem ungezogenen Sohn des fleißigen Hausmeisters*]^{dat} den Fußball versteckt hat, …
d. … hat der Mitschüler, der [*zur zusätzlichen Ahndung des mehrfachen Fehlverhaltens*]^{adj} [*dem ungezogenen Sohn des fleißigen Hausmeisters*]^{dat} den Fußball versteckt hat, …

(Levy & Keller, 2013, p.212)

Besides intervening material, the preceding context of a sentence can also guide expectations when processing complex syntactic structures. Processing delays have been found to decrease or disappear when complex sentences are embedded in supportive contexts (Davison & Lutz, 1985; Grodner et al., 2005; Kaiser & Trueswell, 2004; Levy et al., 2012). For instance, Grodner and colleagues (2005) presented restricted relative clauses (8) and non-restrictive relative clauses (9) with or without a supportive context. The target region ('*dog bit*') was read slower in restrictive clauses than in non-restrictive clauses when it was presented without context, but was read faster when it was presented with context. In (8), a restrictive relative clause is more likely because by itself, '*the postman*' is ambiguous. In (9), on the other hand, '*the postman*' is not ambiguous and a relative clause of any kind is less likely.[2]

(8) [A vicious dog bit a postman on the leg and another postman on the arm.]^{context}
The postman that the dog bit on the leg needed seventeen stitches and had a permanent scar from the injury.

(9) [A vicious guard dog bit a postman and a garbage man.]^{context}
The postman, who the dog bit on the leg, needed seventeen stitches and had a permanent scar from the injury.

(Grodner et al., 2005, p.281)

[2] However, as Grodner et al. (2005) note '*the dog bit*' is repetitive in (9) which may have caused a penalty and thus a delay. A passive construction ('*who was bitten*') may have been more appropriate to test the effect of context.

1.2 Processing non-local dependencies

Memory-based and expectation-based approaches do not always make similar predictions on where within sentences processing delays occur. Memory-based approaches claim non-local dependencies increase processing demands and predict delays – and no speed-up effects – in non-local dependencies. Expectation-based approaches can predict both delays and speed-up effects, which may occur at different locations in the sentence. While integration of a final head may be facilitated, the intervening material may be less expected. As a result, processing delays can occur at these locations rather than at the point of the unresolved dependency. Furthermore, corpus-based evidence suggests that local-dependencies are more common than non-local dependencies. Even in languages with free word order, there is a preference to minimize dependency lengths (Futrell, Mahowald & Gibson, 2015; Liu, Xu & Liang, 2017; Temperley, 2007). This suggests that local-dependencies are more likely than non-local dependencies and thus expectations may be higher for local-dependencies in general (Liu et al., 2017).

Both memory-based and expectation-based accounts thus seem to predict that sentences with non-local dependencies are overall more difficult to understand than sentences with local dependencies. Long dependencies require more computational resources and are less expected than short dependencies. Unfortunately, most parsing studies do not report processing times for regions other than their target verb and do not report processing times for the entire sentence. As a result, it is unknown whether anti-locality or locality effects win out in the end. For readability research, this is an important question. If these effects cancel each other out, it cannot be said that the readability of the text is negatively affected by non-local dependencies. In that case the readability will not increase by reducing the number of non-local dependencies.

A second open question is when does a non-local dependency become problematic? Memory-based approaches especially predict that as the distance between a head and a dependent increases, processing costs go up. So, if there is only one element between a head and a dependent, this is easier to overcome than when there are two.

The easiest way to count the length of a dependency is to count the number of words intervening between a syntactic head and its dependent: Syntactic Dependency Length (SDL).[3] When the SDL is zero, the head and dependent are adjacent (i.e., local dependency). If the SDL is more than zero, there are words intervening between the head and its dependent (i.e., non-local dependency). Levy

[3] This is of course a simplified view. For instance, the Dependency Locality Theory counts new discourse elements and not individual words (although Gibson notes that processing other kinds of elements may also come at some cost, 2000, p.107).

and Keller (2013), for example, increased dependency lengths by 6 or 12 words, while Grodner and Gibson (2005) increased the distance with only 3 or 5 words (see Section 1.1).

It is conceivable that small increases in SDL are less harmful than larger ones. But the size of the SDL increase (further referred to as 'delta') may not be the sole factor. An increase of 5 words may have a different effect in a sentence that is already rather difficult to progress, while an increase of 5 words in a simple sentence may not affect a reader that much. Some evidence for this can be found in the study of Bartek et al. (2011). Bartek et al. (2011) replicated the study of Grodner and Gibson (2005) in an eye-tracking experiment. They compared an adjacent dependency with dependencies separated by a prepositional phrase of 3 words or a relative clause of 5 words. These dependencies were added to a matrix clause (see (10)) or a center-embedded relative clause (see (11)). In the matrix clause condition Bartek et al. (2011) found no difference in processing delay between adding 3 or 5 words, while in the embedded clause condition 5 words resulted in a bigger delay than 3 words. Although the delta was the same in both conditions, the results were not. In the embedded clause condition the reader must also integrate the object with the verb ('*the administrator*' with '*supervised*'). So even when the SDL between '*the nurse*' and '*supervised*' is zero, the distance between '*the administrator*' and '*supervised*' is not. The base level to which the SDL increase is added is already higher in the embedded clause condition compared to the matrix clause condition. In sum, increasing the maximum SDL of a sentence from 5 to 10 may not have the same effect as increasing the maximum SDL from 15 to 20.

(10) The nurse [∅/from the clinic/who was from the clinic] supervised the administrator…

(11) The administrator who the nurse [∅/from the clinic/who was from the clinic] supervised…

1.3 Comprehending non-local dependencies

Theoretically, syntactically complex sentences should affect comprehension if the reader is unable to parse the sentence correctly or if a processing overload occurs. If processing non-local dependencies takes more computational resources, processing will break down when these resources are unavailable (Gibson, 1998). In addition, if more resources are allocated to syntactic parsing, comprehension may be affected because less resources are available for deeper levels of processing (e.g., inference generation) and comprehension is limited to surface-code or text-base level representation (Kintsch, 1992). When a complex sentence is part of a text, decreased

comprehension can even extend beyond the sentence since what is read becomes part of the reader's mental representation (Bailin & Grafstein, 2016). If the representation is incomplete or incorrect, integration of new sentences is also compromised (Poulsen & Gravgaard, 2016). On the other hand, the surrounding context can also offer support and help to correct incorrect interpretations of the sentence (Bailin & Grafstein, 2016).

Evidence for decreased comprehension in non-local dependencies is mainly restricted to experiments comparing subject- and object-extracted relative clauses (SRC vs. ORC). Readers make more mistakes in thematic role identification and score lower on verification tasks for ORCs than for SRCs (e.g., Holmes & O'Regan, 1981; King & Just, 1991, Murphy, 2013; but cf. Grodner & Gibson, 2005; Fedorenko, Piantadosi & Gibson; 2012). Other studies did not find evidence that non-local dependencies reduced comprehension. Levy et al. (2012) compared in situ RCs with RCs that were extraposed across prepositional phrases and verbs. They did not find an effect on verification statements. Renkema (1989) presented discontinuous sentences in natural contexts of 6 sentences. He also did not find an effect on verification statements.

One explanation for these null-results is the fact that both studies used skilled, experienced readers. Multiple studies have shown that individual differences – in particular in working memory capacity – affect how readers process and comprehend complex sentences (King & Just, 1991; Norman, Kemper & Kynette, 1992; Stine-Morrow, Ryan & Leonard, 2000). Readers with lower working memory capacity are more likely to encounter processing overloads which will be detrimental to comprehension.

1.4 Individual differences in syntactic parsing

Even in seemingly homogeneous participant groups, individual differences in skills, cognitive abilities, world knowledge or experience influence how readers process and understand text (Farmer, Misyak & Christiansen, 2012; Just & Carpenter, 1992; Kuperman & Van Dyke, 2011; Perfetti, 1997; Traxler et al., 2012; Van Dyke, Johns & Kukona, 2014). Both memory-based accounts and expectation-based accounts implicitly assume that individual differences will affect the way sentences are processed (see Nicenboim et al., 2015). For memory-based approaches, an obvious factor when parsing complex sentences is working memory capacity (WMC). Readers with low WMC require more time to process complex sentences than high-WMC readers and will more often fail to comprehend the sentence correctly (King & Just, 1991; Fedorenko, Gibson & Rohde, 2006; Nicenboim et al., 2015). Complex sentences can also force low-WMC readers into shallow processing, resulting in shorter processing times but lower comprehension (Nicenboim et al.,

2015; Nicenboim, Logačev, Gattei & Vasishth, 2016; Ferreira, Bailey & Ferraro, 2002).

Expectation-based accounts also assume individual differences influence syntactic parsing, but these differences are related to differences in experience. Experienced readers will make more accurate and more refined predictions compared to less experienced readers. In addition, these readers have probably faced similar syntactic structures before and are therefore more likely to succeed. For instance, Street and Dąbrowska (2010) found that increasing the experience of people who had problems with understanding passives improved their comprehension of this structure dramatically. The effect was still present 12 weeks after training.

However, it is unclear whether WMC, experience or any individual characteristic directly influences the mechanisms underlying syntactic parsing and non-local dependency resolution. The individual abilities associated with reading success show high inter-correlations, which makes it difficult to determine the unique contributions of different factors (Kuperman & Van Dyke, 2011; Traxler et al., 2012; Van Dyke et al., 2014). Regardless of the exact source of individual differences, it is important to include reader characteristics in experiments and to monitor for possible interactions between text and reader characteristics.

2. Experiment 1: Effects of SDL on text processing

In Experiment 1 we use eye-tracking to investigate the effects of non-local syntactic dependencies on on-line text processing. We hypothesize that sentences with longer syntactic dependency lengths will require more time to process than sentences with shorter SDLs. In addition, the size of this slowdown effect is likely to differ depending on both the skills of the reader and the size of the dependency length increase.

We focus on the overall effects for the reader. It is beyond the scope of this experiment to differentiate between memory-based and expectation-based accounts of syntactic parsing. Instead, we focus on the net result of locality and anti-locality effects associated with increasing SDL in natural contexts. If these effects cancel each other out, the difference in processing costs will be zero and readability is not affected. Conversely, if readers need more time to process Long-SDL sentences, this means readability is negatively affected by long SDLs because it takes more effort for readers to process the text. We use real texts to approximate the effect of SDL in natural reading situations.

2.1 Method

2.1.1 Participants

In total 181 Dutch ninth-grade students participated in the study (119 female; 62 male). 80 participants were enrolled in pre-university education ('vwo'), 54 in general education ('havo') and 47 in pre-vocational medium education ('vmbo-kb').[4] Eye movement data of 23 participants was removed from the analyses because either the calibration procedure failed or because the registration proved to be unstable. These participants were only included in the comprehension data analysis. All remaining participants had normal or corrected to normal vision.

2.1.2 Materials

Four texts were quasi-randomly selected from a Dutch corpus.[5] The texts were informative texts written for the general public by the Dutch government and government affiliated organizations. The text topics were: diabetes, smoke alarms, citizen's arrests and drinking water. The texts were 300 to 420 words long and did not contain figures or tables. The texts were randomly assigned to this study and not selected based on their potential for manipulation.

Manipulation. In order to change the syntactic dependencies of the texts without changing the texts' meaning, we varied the word order of sentences within the texts. Although Dutch does not have totally free word order, it does allow limited scrambling of constituents (see Unsworth, 2005) and rearrangement of subordinate and main clauses resulting in left-branching structures. By rearranging the words in a sentence, we can decrease or increase the distance between a head and the dependent without altering the content of the sentence.[6] For example, in (12) the prepositional phrase '*aan een opsporingsambtenaar*' can be placed before or after the verbal phrase that contains the main verb. If it is placed before the verbal phrase (12b) the distance between the modal verb '*dient*' and the main verb '*overgedragen*' increases, as does the distance between the verb modifier '*onverwijld*' and '*overgedragen*'.

[4] The Dutch education levels are ordered from theoretical oriented education to practice oriented education. For more information on the Dutch educational system see EP-Nuffic (2015).

[5] The texts were selected from a collection of 120 public information texts (see Chapter 1).

[6] Rearranging constituents and clauses can affect interpretation and focus. We made sure that shifts did not alter the meaning of the sentence or text. Each manipulation was reviewed by at least one other researcher.

(12) a. De aangehoudene dient onverwijld overgedragen te worden aan een opsporingsambtenaar (politie).
The detainee needs immediately handed over to be to a criminal investigator (police).
'The detainee needs to be handed over to a criminal investigator (police) immediately.'

b. De aangehoudene dient onverwijld aan een opsporingsambtenaar (politie) overgedragen te worden.
The detainee needs immediately to a criminal investigator (police) handed over to be.
'The detainee needs to be handed over to a criminal investigator (police) immediately.'

(13) a. Onderzoekers weten nog niet precies hoe het komt dat iemand diabetes type 1 krijgt.
Researchers know yet not exactly how it happens that someone diabetes type 1 gets.
'Researchers do not exactly know yet why someone gets diabetes type 1.

b. Hoe het komt dat iemand diabetes type 1 krijgt, weten onderzoekers nog niet precies.
How it happens that someone diabetes type 1 gets, know researchers yet not know.
'Researchers do not exactly know yet why someone gets diabetes type 1.

Each text was manipulated to create a text version with relatively short syntactic dependencies (Short-SDL version) and a version with relatively long syntactic dependencies (Long-SDL version). In contrast to previous studies we did not split sentences or add more constituents to sentences to manipulate SDL. Such procedures create confounds of SDL with factors like sentence length, number of propositions, coherence marking, and changes in meaning. By changing just the word order, such confounds do not exist in our materials. The word order was changed in 1/3rd of the sentences (39 out of 119 sentences). Which dependencies were manipulated depended on the possibilities offered by the particular sentence (e.g., subject - verb, verb - complement).

Afterwards, the manipulations were quantitatively checked by comparing the maximum SDL of the Short-SDL and Long-SDL sentences.[7] The maximum SDL was higher for Long-SDL sentences than for Short-SDL sentences ($t(38) = 10.186$, $p < .001$; see also Table 1).[8]

Since the manipulated sentences varied in length and structure, both the SDL in the Short-SDL sentences (further referred to as 'base level') and the difference between lengths in the Short- and Long-SDL versions (further referred to as 'delta') varied across sentences. Figure 1 presents the base level and delta of each manipulated sentence. As this figure illustrates, SDL is manipulated across a wide spectrum.

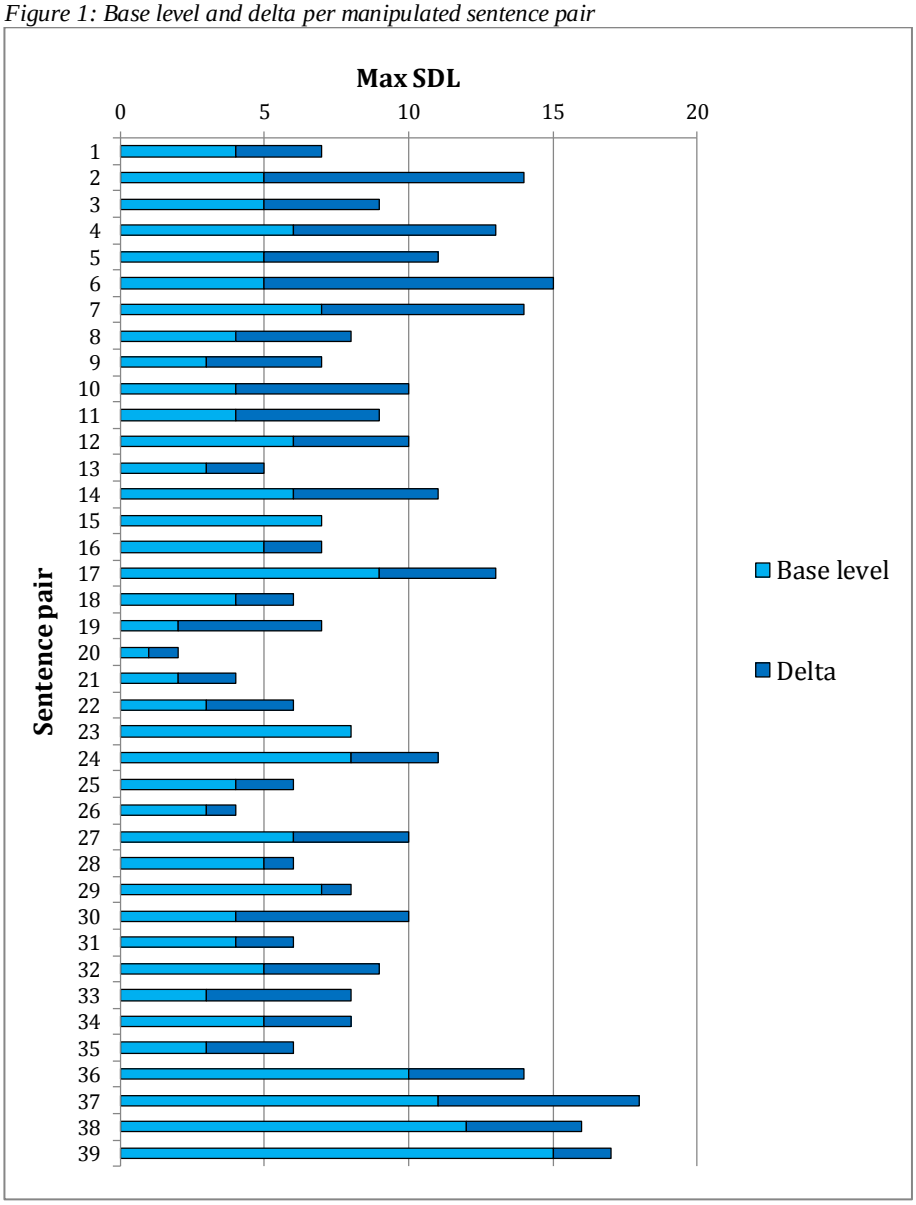
Table 1: Means and standard deviations for maximum syntactic dependency length per SDL version (manipulated sentences only)

	Max SDL	
SDL-version	**Mean**	**SD**
Short-SDL	5.46	2.89
Long-SDL	9.23	3.78

[7] The maximum syntactic dependency lengths were calculated manually to avoid miscalculations due to parsing errors. However, manually calculated SDLs correlated .899 with values calculated by the Dutch automatic text analysis tool T-Scan (Pander Maat et al., 2014).

[8] Appendix 6 presents a graphical overview of the mean maximum SDL of all text versions of the four texts, including the mean maximum SDL of the original text.

Figure 1: Base level and delta per manipulated sentence pair



Comprehension questions. Each text was followed by eight multiple-choice comprehension questions. Questions were designed to measure the understanding of the main points of the texts, rather than the meaning of the manipulated sentences in particular.

Reading ability. Standardized reading ability scores were available for all but one student. A one-way Anova revealed that pre-university students had higher reading ability scores than pre-vocational and general students, but that there was no difference between pre-vocational and general students. However, standard deviations were large and some pre-vocational students performed better than pre-university students.

2.1.3 Design

The texts were divided over two lists, as a Latin-Square. As a result, participants read every text but only in one condition. Half of the participants read the texts in the reversed order.

2.1.4 Apparatus

The eye movements of the participants were recorded with a desktop eye tracker: the SR Research EyeLink 1000. The eye tracker recorded the position of the right pupil via a Logitech QuickCam Pro 5000 webcam at a rate of 500Hz. A remote setup with target sticker was used, allowing participants to move their head slightly. Accuracy of this eye tracker is 0.5 degrees. Stimuli were presented on a 17 inch computer screen (1280x1024px).

Stimuli were presented with black letters on a white computer screen. Because the texts were too long to present on one screen, they were split up into three or four screens. To avoid unnatural text breaks, breaks either coincided with paragraph endings or other natural break points. Presentation of the manipulated sentences was kept as identical as possible between text versions. The shifts in word order naturally resulted in specific words ending up at slightly different positions or lines on the screen from one text version to the next, but the number of line breaks within the sentence was kept equal and each sentence began and ended at the same location on screen.

2.1.5 Procedure

Participants took part in two sessions of approximately half an hour each spread over two days. In one session they read the four syntactically manipulated texts discussed in the present study, in the other session they read four lexically manipulated texts used for another study (see Chapter 3). The order was balanced between participants. Recording took place in a private room at the participating schools. Each session started with an oral instruction during which the equipment and procedure were explained. The participants were instructed to read each text at their own pace, but to make sure that they could answer comprehension questions at the end of the text.

The instruction was followed by a 13-point calibration and validation procedure. Participants fixated on a sequence of dots which appeared on various locations on the computer screen. After a successful calibration and validation sequence the testing started with a practice text and three practice questions to familiarize the participant with the procedure.

Each text fragment started with a single dot on the screen which indicated the location of the first word of the fragment. When the participant fixated on the dot, the dot vanished and the fragment appeared. To progress to the next text fragment participants pressed the 'next' button on the button-box. To answer questions participants pressed the button on the button box which corresponded to their answer. Participants could not look back in the text while answering the questions and could not go back and revise their answers. There was no time limit.

2.1.6 Data preparation and clean-up

Eye movement data. Fixations were checked and assigned using Fixation 0.1.0.15 (Cozijn, 1994). Track losses were removed from the data, as well as regions that contained blinks. Six measures were calculated for the manipulated sentences: *First pass gaze duration* (FP), *First pass total gaze duration* (TG), *Second pass gaze duration* (SP), *Regressing path duration (RP), Total fixation duration* (TF), *Regression probability* (REG). Descriptions of these measures are given in Table 2.

Table 2: Descriptions of eye-tracking measures

Measures	**Description**
First pass gaze duration (FP)	Summed duration of all fixations and intermittent saccades within a sentence in first pass before the eyes leave the sentence (either regressively or progressively).
First pass total gaze duration (TG)	Summed duration of all fixations and intermittent saccades within a sentence in first pass before the eyes leave the sentence progressively.
Second pass gaze duration (SP)	Summed duration of all fixations and intermittent saccades within a sentence in second pass.
Regression path duration (RP)	Summed duration of all fixations and intermittent saccades within a sentence in first pass plus regressions to previous sentences before the eyes leave the sentence progressively.
Total fixation duration (TF)	Summed duration of all fixations on the sentence (including second, third… n^{th} pass).
Regression probability (REG)	Presence of a regressive fixation from the sentence to a previous sentence in first pass (1= regressive fixation; 0 = no regressive fixation).

Comprehension data. Responses to the comprehension questions were scored as correct or incorrect. Each question had only one correct answer. 1.2% of the data had to be removed because participants indicated to have accidently pressed the button before reading the whole question.

2.1.7 Analyses

Duration measures were analyzed using linear mixed effect modeling with sentence pairs nested in texts and crossed with students. Regression probability and comprehension were analyzed using generalized mixed effect modeling with a logit link since these measures represented binomial data (see Quené & Van den Bergh, 2008).

Goal of the analysis was to investigate the effect of increased syntactic dependency lengths on sentence processing times. Reader characteristics (Education level and Reading ability) were included to test for possible interactions between the manipulation and the skillset of the reader. In addition, since we work with natural texts our SDL manipulations differ in strength (see Section 2.1.2). We included two metrics (Delta and Base level) to test whether the size of the effect depended on the strength of the manipulation.[9] Small deltas may not produce as big an effect as large deltas, especially for proficient readers. On the other hand, if the base level is already high, even a small delta might tip the scales and produce large effects.

Descriptions of all included factors are given in Table 3.

Table 3: Descriptions of factors used in Experiment 1

Factor	Description	Levels
SDL-version	Text version: Long-SDL or Short-SDL	2 levels
Education level	Level of education in which the student is enrolled	3 levels
Reading ability	Reading ability score on Dutch reading ability test (centered and standardized)	Continuous
Delta	Delta in maximum SDL in short and long version (small ≤ 4 vs. large ≥ 5)	2 levels
Base level	Maximum SDL for Short-SDL sentence (low ≤ 4 vs. high ≥ 5)	2 levels

[9] Values of deltas and base levels were not equally distributed (i.e., few values in the right tail and some large outliers). To reduce the influence of outliers and the skewed distribution, values were collapsed into high and low values. The cutoff point of 5 was determined by looking at the median maximum SDL of 120 public information texts and 146 educational textbook texts which were collected for the LIN-project (see Chapter 1). The median maximum SDL of these texts was 5. Main effects for Delta and Base level were only included in combination with interaction terms, since main effects would be characteristic of the particular sentence and not of our manipulation. Main effects were dropped when the interactions were not significant.

2.2 Results

2.2.1 Eye movement measures

Table 4 and Table 5 show the means and regression probability for the eye movement measures. Final models are presented in Table 6.

The analysis revealed that SDL-version was a significant factor for *First pass total gaze* ($X^2(1) = 8.872$, $p = .003$), *Regression path* ($X^2(1) = 6.099$, $p = .014$) and *Total fixation duration* ($X^2(1) = 7.975$, $p = .005$). Durations were longer in the Long-SDL version compared to the Short-SDL version. No effect of SDL-version was found in *First pass gaze duration*, *Second pass gaze duration* or *Regression probability* ($p > .130$). SDL-version did not interact with Education level, Reading ability, Delta or Base level.

Education level was a significant factor for all duration measures, except for *Second pass gaze* (FP: $X^2(2) = 14.527$, $p < .001$; TG: $X^2(2) = 14.451$, $p < .001$; RP: $X^2(2) = 13.020$, $p = .001$; TF: $X^2(2) = 12.77$, $p = .002$). Pre-university students read faster than pre-vocational students and general students. There was no difference between pre-vocational students and general students. Reading ability did not improve the model once Education level was included.

Table 4: Means and standard deviations duration measures for manipulated sentences in milliseconds

Measure	SDL-version	Pre-vocational	General	Pre-university
First pass gaze duration	Short-SDL	3614 (2189)	3504 (2307)	3098 (1857)
	Long-SDL	4051 (2170)	3447 (2092)	3135 (1846)
Second pass gaze duration	Short-SDL	3941 (2142)	2960 (1705)	2850 (1616)
	Long-SDL	3884 (2059)	3380 (1750)	3254 (1742)
First pass total gaze duration	Short-SDL	3972 (2049)	3767 (2190)	3384 (1769)
	Long-SDL	4359 (2052)	3813 (1924)	3420 (1763)
First pass regression path duration	Short-SDL	4030 (2063)	3821 (2233)	3483 (1866)
	Long-SDL	4426 (2094)	3917 (2040)	3509 (1850)
Total fixation duration	Short-SDL	3746 (1932)	3570 (2127)	3224 (1751)
	Long-SDL	4143 (1989)	3602 (1815)	3250 (1682)

Table 5: Regression probability for manipulated sentences in percentages

Measure	SDL-version	Pre-vocational	General	Pre-university
Regression probability	Short-SDL	9.06%	8.73%	10.60%
	Long-SDL	7.60%	9.89%	9.16%

Table 6: Final models for eye-tracking duration measures at sentence level in log10(ms)
Table 6a: Final model First pass total gaze duration

First pass total gaze duration				
Random effects	*Estimates*	*St.Dev*		
Student	0.009	0.095		
Text	0.021	0.144		
Text: Sentence	0.012	0.110		
Residual	0.013	0.113		
Fixed effects	*Estimates*	*SE*	*t*	*p*
Intercept	3.585	0.062	57.841	<.001
SDL-version: Short-SDL	0[a]			
SDL-version: Long-SDL	0.014	0.005	2.979	.003
Education level: Pre-voc. medium	0[a]			
Education level: General	-0.032	0.023	-1.411	.160
Education level: Pre-uni.	-0.075	0.020	-3.686	<.001

[a] *Set as reference level.*

Table 6b: Final model Regression path duration

Regression path duration				
Random effects	*Estimates*	*St.Dev*		
Student	0.009	0.094		
Text	0.012	0.109		
Text: Sentence	0.020	0.143		
Residual	0.015	0.121		
Fixed effects	*Estimates*	*SE*	*t*	*p*
Intercept	3.590	0.062	58.302	<.001
SDL-version: Short-SDL	0[a]			
SDL-version: Long-SDL	0.012	0.005	2.470	.014
Education level: Pre-voc. medium	0[a]			
Education level: General	-0.029	0.023	-1.265	.208
Education level: Pre-uni.	-0.071	0.020	-3.466	<.001

[a] *Set as reference level.*

Table 6c: Final model Total fixation duration

Total fixation duration				
Random effects	*Estimates*	*St.Dev*		
Student	0.010	0.098		
Text	0.012	0.108		
Text: Sentence	0.020	0.142		
Residual	0.014	0.118		
Fixed effects	*Estimates*	*SE*	*t*	*p*
Intercept	3.560	0.061	58.156	<.001
SDL: Short-SDL	0[a]			
SDL: Long-SDL	0.014	0.005	2.824	.005
Education level: Pre-voc. medium	0[a]			
Education level: General	-0.033	0.023	-1.418	.158
Education level: Pre-uni.	-0.073	0.021	-3.474	<.001

[a] *Set as reference level.*

2.2.2 Comprehension questions

Analysis of the answers to the multiple-choice questions revealed only a significant main effect for Education level (Table 8). Students enrolled in general and pre-university performed better than pre-vocational students. Neither SDL-version nor Reading ability proved to be significant factors and no interactions were found between these factors and Education level.

Table 7: Percentage correctly answered multiple-choice questions

SDL-version	**Pre-vocational**	**General**	**Pre-university**
Short-SDL	58.25%	75.90%	79.21%
Long-SDL	60.40%	76.50%	77.68%

Table 8: Model multiple-choice questions

Random effects	**Estimates**	**SE**	***Z***	***p***	
Student	0.119	0.030	3.967	<.001[a]	
Text	0.490	0.349	1.404	.080[a]	
Fixed effects	**Estimates**	**SE**	***Z***	***p***	**Odds ratio**
Intercept	0.361	0.359	1.006	.314	1.435
SDL: Short-SDL	0[b]				
SDL: Long-SDL	0.012	0.060	0.200	.841	1.012
Education level: pre-voc. medium	0[b]				
Education level: general	0.796	0.104	7.654	<.001	2.217
Education level: pre-uni.	0.925	0.096	9.635	<.001	2.522

[a] *One-sided.* [b] *Set as reference level.*

2.3 Discussion

In Experiment 1 we increased the syntactic dependency lengths (SDL) of sentences embedded in authentic texts. SDL was manipulated across different syntactic structures and lengths. Our results show that increasing the SDL of a sentence results in longer sentence reading times. This effect was observed in *First pass total gaze duration*, *Regression path duration* and *Total fixation duration*, but not in *First pass gaze duration* and *Second pass duration*. Syntactic dependency length also did not affect the probability of a regression. Post-hoc exploration of these measures revealed that approximately 8% of the observations in *First pass gaze* only contained 1 or 2 fixations and that a regression was more likely when the fixation count was low (β = -0.488, SE = 0.027, $\chi^2(1) = 328.051$, $p < .001$, Exp(β) = 0.614). Thus, it seems that in 8% of the cases students did not read the entire sentence in one pass. They may have regressed to a previous sentence before reaching the end of the manipulated sentence or their first fixation on the sentence may have been the result of an overshot saccade. Since our manipulations are located midway through or at the end of the sentence, these students would not have reached the manipulation in FP. A post-hoc analysis was conducted using only observations with more than 2 fixations in FP. In this analysis SDL-version was a significant factor. In line with results on other duration measures, FP was longer for Long-SDL sentences (β = 0.013, SE = 0.006, t = 2.297, p = .022). It seems that this effect was obscured in the initial analysis due to the noise within the measure.

Against expectations, the increase in reading time was not moderated by either reader characteristics or the strength of the manipulation (i.e., Delta and Base level). Low-level readers were not more affected by our manipulation than high-level readers. The fact that the effect was independent from the strength of the manipulation is more in line with expectation-based accounts than with memory-based accounts. However, most of our manipulations were rather modest; especially compared to Levy and Keller (2013) who used sentences of 29 words with subordinate and relative clauses, and increased SDL with 6 or 12 words at a time. We increased SDLs with a maximum of 10 words (mean = 4) and did so in fairly easy, short sentences with few subordinate clauses and no relative clauses. The difference between smaller and larger manipulations may not have been large enough to show increased effects. Presumably, even less-able readers could overcome the additional SDL rather quickly. However, even the smaller manipulations caused readers to slow down. This result contradicts the statement that memory demands must be fairly high to result in observable locality effects (Levy & Keller, 2013; Shain, Van Schijndel, Futrell, Gibson & Schuler, 2016).

Our results show that it takes more time to process a sentence when a syntactic head and its dependent are further apart. We do not know where exactly this delay occurs. Our design does not allow us to conduct reliable localized

analyses on word level syntactic processing. Increasing the SDL of the sentences has led most syntactic heads to end up at the end of the sentence. Since sentence final words elicit more regressions than non-final words in general (Rayner, Kambe & Duffy, 2000; Warren, White & Reichle, 2009), this confound would interfere with our analysis (see also Nicenboim et al., 2015). Such a confound was unavoidable to preserve the ecological validity of the materials. As a result, we are unable to determine the exact time course of the delay found on sentence reading times and cannot distinguish between accounts of localized syntactic processing (i.e., memory-based vs. expectation-based accounts). We can say that overall, increasing the SDL of a sentence places a higher demand on processing than decreasing the SDL, even when differences in SDL are small.

The delay in reading time was not accompanied by a decrement in comprehension. Increasing the SDL of a text did not affect the students' performance on multiple-choice questions. It seems students were able to overcome the additional processing costs associated with increased SDLs. These results are in line with results from Renkema (1989) and Levy et al. (2012), who also did not find comprehension to be affected.

3. Experiment 2: Effects of SDL on text comprehension

Experiment 1 confirmed that syntactic dependency length influences the way readers process sentences. It did not show an effect on comprehension. However, the multiple-choice questions used in Experiment 1 were designed to measure overall text comprehension and focused on the most important information in the text. Questions did not specifically target the manipulated sentences. So, Experiment 1 showed that comprehension of the main points in the text was not affected. Comprehension of the manipulated sentences may very well have been lower, but the task was not sensitive enough to detect this. In order to further examine SDL effects on comprehension, we need to use a more localized measure of comprehension.

In Experiment 2 we use a different comprehension measure: the Hybrid Text Comprehension cloze test (see Chapter 2). The HyTeC-cloze offers an equally distributed measurement of comprehension over the entire text, sampling both manipulated and non-manipulated sentences. By using this measure we can investigate whether SDL affects just the manipulated sentences or whether it extends beyond sentence borders.

We also increase the statistical power of the experiment. Since Experiment 1 did not reveal any interactions between SDL and the reader characteristics, we

increase the variance in reader characteristics by including students from grades 8 through 10 and enrolled in 5 rather than 3 different levels of education. This expands the range of differences between students and should make any interaction between reader characteristics and SDL easier to detect.

3.1 Method

3.1.1 Participants

Twenty-nine Dutch secondary schools participated with in total 824 students from grades 8, 9 and 10 (age: 13 - 16). Testing was introduced as part of the regular school curriculum. The students were enrolled in different levels of secondary education: in pre-university education ('vwo'), general secondary education ('havo') or pre-vocational education ('vmbo-gt', 'vmbo-kb' or 'vmbo-bb'). Grade 10 students of pre-vocational education were not included in the study because pre-vocational students are graduating in grade 10. In addition, we must note that pre-vocational low students were slightly underrepresented in the sample (see Table 9).

Table 9: Distribution of participants over Grade and Level of education

	Pre-voc. (low)	**Pre-voc. (medium)**	**Pre-voc. (high)**	**General education**	**Pre-uni.**	**Total**
Grade 8	18	110	170	61	40	401
Grade 9	36	95	52	67	79	332
Grade 10	-	-	-	61	35	100
Total	54	205	222	189	154	824

3.1.2 Materials

The four public information texts used in Experiment 1 were supplemented with 6 more public information texts and 10 texts taken from educational textbooks (subjects: history, geography, Dutch language and economics).[10] We included two genres to increase the diversity of the materials, both in terms of the range of syntactic structures and in terms of combinations with other stylistic factors and content.[11] The educational texts were written especially for students in secondary education. The texts were randomly assigned to this study and not selected based on their potential for manipulation.

The additional texts were manipulated in the same way as the four texts in Experiment 1. The twenty texts contained a total of 180 manipulated sentence pairs. Long-SDL sentences had higher maximum SDLs than Short-SDL sentences (t(179)

[10] The educational textbook texts were quasi-randomly selected from a collection of 146 educational textbook texts (see Chapter 1).

[11] See Pander Maat and Dekker (2016) and Pander Maat (2017) for analyses of stylistic factors across genres in Dutch.

= 17.424, p < .001; see Table 10). In addition, the SDL was higher for sentences from public information texts than for sentences from educational textbooks (F(1,359)) = 8.504, p = .004). However, there was no interaction with SDL-version (F < 1). The difference in SDL between the Short-SDL and Long-SDL sentences was the same for both genres.[12]

For eight sentence pairs, the manipulation did not alter the maximum SDL of the sentence, but only affected shorter SDLs within the sentence. For this reason, these sentence pairs were not included in the localized analysis.

Table 10: Means and standard deviations for maximum syntactic dependency length per SDL-version and genre (manipulated sentences only)

	Total		**Public information texts**		**Educational textbook texts**	
SDL-version	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Short-SDL	5.53	3.96	6.18	3.66	4.96	4.15
Long-SDL	9.71	4.85	10.49	4.29	9.01	5.23

3.1.3 Measures

Comprehension assessment. The texts were transformed into cloze tests following the Hybrid Text Comprehension cloze procedure (HyTeC-cloze; see Chapter 2). This hybrid cloze procedure combines the strengths of mechanical and rational cloze tests into a valid and reliable measure of text comprehension. The rational strategy is used to exclude words that do not rely on text level comprehension from becoming cloze gaps. This includes words that can be reconstructed using only grammatical knowledge or knowledge of usage conventions (e.g., articles and multi-word expressions). Also excluded are words that can only be guessed, such as names and numbers. All other words in the text are candidates for deletion. The candidates are divided over different cloze versions via mechanical selection. Two cloze versions were randomly selected to serve in the study. In total 10% of the words were deleted. Depending on the text length, the cloze tests contained 30 to 42 cloze gaps. The same words were deleted in the Short-SDL and Long-SDL text version.

Reading ability. Standardized reading ability scores were made available for all students. Two different tests were used to measure Reading ability. Although scores of both tests were mapped to the same scale, analyses showed that scores for one of the tests were consistently higher. To control for this complication, the factor Reading test was included in the analyses.

[12] Appendix 6 presents a graphical overview of the mean maximum SDL of all text versions of the twenty texts, including the mean maximum SDL of the original text.

3.1.4 Design

This study is part of a larger scale project in which the difficulty of 60 texts was assessed among 2926 participants. Participants were randomly assigned to this part of the study.

The experiment was set up following a matrix sampling design (e.g., Gonzalez & Rutkowski, 2010). Each participant was given four different cloze tests: one cloze test of a Short-SDL educational text, one of a Short-SDL public information text, one of a Long-SDL educational text and one of a Long-SDL public information text. To balance out possible order effects, each combination of cloze tests was presented in two orders.

3.1.5 Procedure

All testing took place at the participating schools. The tests were administered by the school teachers in classroom settings. Cloze tests were presented digitally on computers. Participants filled in the cloze gaps on the screen. To fill in all four cloze tests, participants took part in two sessions of 45 minutes. Schools scheduled all sessions themselves over the course of a couple weeks.

3.1.6 Scoring procedure and data-clean up

The answers to each cloze gap were dichotomously scored (1 = correct; 0 = incorrect) according to the acceptable word scoring procedure (see Chapter 2). Following the acceptable word scoring procedure, not just originally deleted words were scored as correct but semantically correct alternatives were given the same score (including spelling errors and typos). The acceptability of alternative answers was judged by the global appropriateness criterion which means that the answer had to fulfill "*all the contextual requirements of the entire discourse context in which it appears*" (Oller & Jonz, 1994, p.416). Each answer was scored by two independent judges from a pool of 16 judges. When the judges disagreed, a third judge made the final decision. All judges received a short training to familiarize them with the scoring procedure.

10% of the data was removed because students repeatedly gave non-serious answers or did not answer the cloze gaps at all. Cases where students occasionally failed to fill in a gap were regarded as incorrect answers rather than missing answers. A separate analysis of the data showed that the results did not change when these cases were excluded from the dataset. The final dataset contained 108132 cases within 3114 cloze tests. 40241 of these cases were embedded in manipulated sentences.

3.1.7 Analyses

Cloze scores were not aggregated before analysis. The data was analyzed at the response level using generalized linear mixed effect modeling (GLMM) with a logit link. Each case represents an answer to an individual cloze gap. Observations are nested within students and texts, with students nested in schools.

Two separate analyses were performed: a text level analysis and a localized analysis. Goal of the first analysis was to see if increasing the syntactic dependency lengths within a text influences overall cloze performance. All cloze gaps were included in this analysis (i.e., cloze gaps in manipulated sentences as well as cloze gaps in sentences that are not manipulated). Reader characteristics were included to test for possible interactions between the manipulation and the skillset of the reader. Finally, the text feature Genre was included in the analysis to test the generalizability of effects across genres.

It was considered likely, however, that any effect of our manipulation would be limited to the manipulated sentences and would not affect performance on all cloze gaps. Therefore, we ran the same analysis using just the scores from manipulated sentences. An additional goal of this analysis was to test the relative strength of the manipulations. The manipulated sentences differ in their base SDL level and their SDL delta (see Section 2.1.2). Hypothetically, increasing the SDL from 5 to 10 may not have the same effect as increasing the SDL from 1 to 6. In addition, small deltas may not affect comprehension, while larger deltas will. These hypotheses were explored by including 2-way-interactions and 3-way-interactions of SDL-version with Base level and Delta in the analysis.

Descriptions of all factors are given in Table 11.

Table 11: Descriptions of factors used in Experiment 2

Factors	Description	Levels
SDL-version	Text version: short or long sentence dependency lengths	2 levels
Education level	Level of education in which the student is enrolled	5 levels
Grade	Grade in which the student is enrolled	3 levels
Reading ability	Reading ability score on Dutch reading ability test (centered and standardized)	Continuous
Reading test	Reading test used to test reading ability	2 levels
Genre	Educational text or public information text	2 levels
Delta	Delta in maximum SDL in short and long version (small ≤ 4 vs. large ≥ 5)	2 levels
Base level	Maximum SDL for Short-SDL sentence (low ≤ 4 vs. high ≥ 5)	2 levels

3.2 Results

3.2.1 Text level analysis

Table 12 shows the mean percentage of gaps that was answered correctly per SDL-version, Education level and Grade. The final model is presented in Table 13

and shows significant main effects for the factors SDL-version, Education Level, Grade, Reading test, Reading ability and Genre, and two significant interactions between SDL-version and Genre, and between SDL-version and Reading ability. SDL-version did not interact with Education level or Grade.

As expected, students enrolled in higher levels of education, higher grades and with higher reading ability scores performed better than students in lower levels, grades or with lower reading ability scores. Students also performed better in the Short-SDL condition than in the Long-SDL condition, but only in public information texts. The effect of SDL-version was not significant for texts taken from education textbooks. Public information texts were overall more difficult than the education texts, as indicated by the main effect of genre.

SDL-version also interacted with Reading ability. Readers with high reading ability scores were more affected by the SDL-manipulation than readers with low reading ability scores.

Table 12: Mean probability of a correctly answered cloze gap per Education level, Grade and SDL-version[13]

		SDL-version	
Education level	**Grade**	**Short-SDL**	**Long-SDL**
Pre-vocational (low)	Grade 8	41.38%	39.20%
	Grade 9	34.69%	32.72%
Pre-vocational (medium)	Grade 8	39.70%	40.49%
	Grade 9	43.52%	41.45%
Pre-vocational (high)	Grade 8	47.56%	46.45%
	Grade 9	57.08%	54.41%
General	Grade 8	54.61%	54.16%
	Grade 9	63.41%	61.05%
	Grade 10	67.45%	66.60%
Pre-university	Grade 8	69.03%	69.73%
	Grade 9	72.30%	71.90%
	Grade 10	71.19%	70.44%

[13] The 8th-grade pre-vocational (low) students performed exceptionally well on the cloze tests and had relatively high reading ability scores. However, there were only 18 8th-grade pre-vocational (low) students in the sample.

Table 13: Final model text level analysis cloze data

Random effects	**Estimates**	**SE**	***Z***	***p***	
School	0.030	0.014	2.143	.016[a]	
School: Student	0.232	0.014	16.571	<.001[a]	
Text	0.023	0.005	4.600	<.001[a]	
Fixed effects	**Estimates**	**SE**	***Z***	***p***	**Odds ratio**
Intercept	-0.345	0.130	-2.654	.008	0.708
SDL: Short-SDL	0[b]				
SDL: Long-SDL	-0.070	0.020	-3.500	<.001	0.932
Education level: pre-voc. low	0[b]				
Education level: pre-voc. medium	0.190	0.088	2.159	.031	1.209
Education level: pre-voc. high	0.464	0.095	4.884	<.001	1.590
Education level: general	0.647	0.118	5.483	<.001	1.910
Education level: pre-uni.	0.921	0.120	7.675	<.001	2.512
Grade 8	0[b]				
Grade 9	0.118	0.047	2.511	.012	1.125
Grade 10[c]	0.176	0.093	1.892	.058	1.192
Reading ability	0.312	0.031	10.065	<.001	1.366
Reading test: R	0[b]				
Reading test: V	-0.443	0.105	-4.219	<.001	0.642
Genre: Public information	0[b]				
Genre: Educational textbook	0.353	0.028	12.607	<.001	1.423
Long-SDL * Educational textbook	0.062	0.029	2.138	.033	1.064
Long-SDL * Reading ability	-0.037	0.014	-2.643	.008	0.964

[a] One-sided. [b] Set as reference level. [c] Unbalanced, no 10th grade pre-vocational students were present in the sample.

3.2.2 Localized analysis

The final model of the localized analysis is presented in Table 15. The analysis revealed significant main effects for SDL-version, Education level, Grade, Reading ability, Reading test, and Genre that were in line with the text level analysis. The additional factors Delta and Base level were also significant, indicating that sentence pairs with large deltas or high base levels were overall harder to comprehend than sentence pairs with small deltas or low base levels. However, the interaction between these factors was also significant, showing that the effects of Delta and Base level were not completely additive and the combined effect was smaller.

More importantly, there was a significant interaction between SDL-version and Delta. The probability of a correct answer was higher in the Short-SDL sentences than in the Long-SDL sentences and this effect was even larger for sentences with large deltas. However, SDL-version also interacted with Genre. SDL-version had less effect in Educational textbook text, compared to public information texts. The combination of these interactions shows us the following pattern: when the Delta is large, students perform worse in the Long-SDL version in both genres although the effect is stronger for public information texts. If the Delta

is small, no effect is observed for Educational textbook text while for Public information texts the difference is still significant and in the expected direction (see Figure 2).

Table 14: Mean probability of a correctly answered cloze gap per Education level, Grade and SDL-version (manipulated sentences only)

		SDL-version	
Education level	**Grade**	**Short-SDL**	**Long-SDL**
Pre-vocational (low)	Grade 8	37.83%	31.88%
	Grade 9	30.68%	27.67%
Pre-vocational (medium)	Grade 8	35.54%	36.61%
	Grade 9	41.38%	36.96%
Pre-vocational (high)	Grade 8	45.40%	41.83%
	Grade 9	54.69%	51.26%
General	Grade 8	50.57%	50.84%
	Grade 9	60.89%	57.61%
	Grade 10	64.49%	63.93%
Pre-university	Grade 8	67.07%	66.16%
	Grade 9	70.59%	68.90%
	Grade 10	70.01%	66.96%

Table 15: Final model localized analysis cloze data

Random effects	**Estimates**	**SE**	***Z***	***p***	
School	0.024	0.011	2.182	.015[a]	
School: Student	0.194	0.015	12.933	<.001[a]	
Sentence pair	0.054	0.008	6.750	<.001[a]	
Fixed effects	**Estimates**	**SE**	***Z***	***p***	**Odds ratio**
Intercept	-0.218	0.132	-1.652	.099	0.804
SDL: Short-SDL	0[b]				
SDL: Long-SDL	-0.109	0.037	-2.946	.003	0.897
Education level: pre-voc. low	0[b]				
Education level: pre-voc. medium	0.249	0.093	2.677	.007	1.283
Education level: pre-voc. high	0.561	0.099	5.667	<.001	1.752
Education level: general	0.730	0.118	6.186	<.001	2.075
Education level: pre-uni.	1.007	0.123	8.187	<.001	2.737
Grade 8	0[b]				
Grade 9	0.154	0.049	3.143	.002	1.166
Grade 10[c]	0.190	0.093	2.043	.041	1.209
Reading ability	0.302	0.032	9.438	<.001	1.353
Reading test: R	0[b]				
Reading test: V	-0.446	0.098	-4.551	<.001	0.640
Genre: Public information	0[b]				
Genre: Educational textbook	0.258	0.035	7.371	<.001	1.294
Delta: small	0[b]				
Delta: large	-0.372	0.041	-9.073	<.001	0.689
Base level: low	0[b]				
Base level: high	-0.233	0.030	-7.767	<.001	0.792
Long-SDL * Educational textbook	0.146	0.046	3.174	.002	1.157
Long-SDL * Large delta	-0.113	0.043	-2.628	.009	0.893
Large Delta * High base level	0.125	0.045	2.778	.005	1.133

[a] One-sided. [b] Set as reference level. [c] Unbalanced, no 10th grade pre-vocational students were present in the sample.

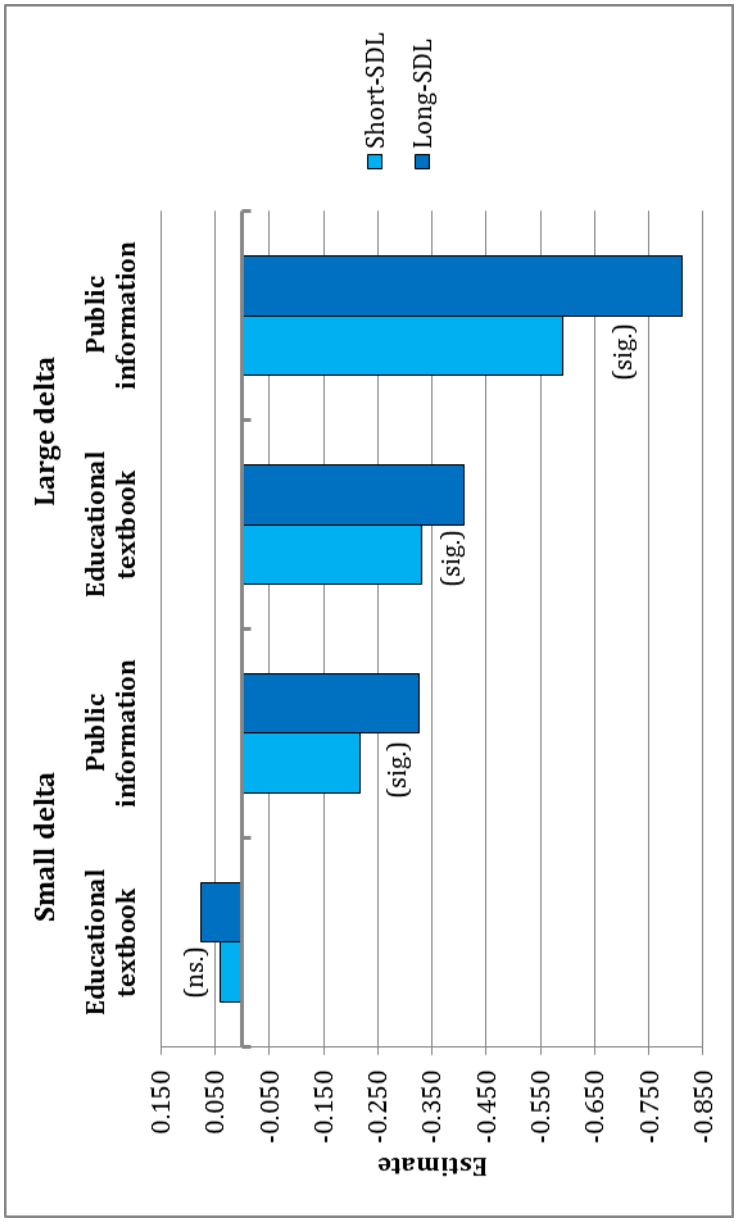


Figure 2: Pattern resulting from double interaction between SDL-version and Genre, and SDL-version and Delta

3.3 Discussion

In contrast to Experiment 1, the results of Experiment 2 show that syntactic dependency length can affect comprehension. Cloze scores were lower when syntactic dependency lengths increased. However, this effect is moderated by several factors. For public information texts, cloze scores were lower when syntactic dependency lengths were increased regardless of the size of the SDL increase. For educational textbook texts, the effect was only present when the SDL increase was large. The localized analysis showed that the interaction of SDL-version and Genre could not be explained completely by differences in the Delta and Base levels of the manipulations. The main effect of Genre shows that educational texts are generally less difficult than public information texts. It seems that other factors within the educational texts make it easier to understand the text and reduce the effect of increased syntactic dependency lengths. Apparently, readers were able to compensate for a slightly higher SDL, but less so when processing more demanding text. We are unaware of other studies that have found similar results. Some converging evidence for these assumptions can be found in two eye-tracking studies of Bartek et al. (2011). The first experiment we already discussed in Section 1.2. Examples of the materials are repeated below as (14) and (15). In this experiment Bartek et al. found locality effects at the verb, but reading times for the distance by 3 and 5 words only differed in the embedded condition (15) and not in the matrix clause condition (14). In the second experiment Bartek and colleagues altered the lexical complexity of the materials (shorter, more frequent words) and reduced possible interference by using animate referents as subjects and inanimate referents as objects (see (16) and (17)). The syntactic structures were the same as in their earlier experiment. Again, they found locality effects at the verb. Only now there was no difference between adding 3 or 5 words in either condition. So, processing delays were diminished when sentences were less complex (i.e., with regard to other stylistic features like lexical complexity). It is likely that comprehension follows a similar trend, with syntactic complexity easier to overcome when the text is less complex on other dimensions.

(14) The nurse [∅/from the clinic/who was from the clinic] supervised the administrator…

(15) The administrator who the nurse [∅/from the clinic/who was from the clinic] supervised…

(16) The child [∅/from the school/who was from the school] played the sports that were…

(17) The sports that the child [∅/from the school/who was from the school] played were…

The effect of our SDL-manipulation was also moderated by the reading ability of the student. Against expectation, good readers were more affected by the manipulation than poor readers but only in the text level analysis. The interaction between SDL-version and Reading ability did not reach significance – nor showed a trend towards significance – in the localized analysis ($p = .223$). This indicates that the interaction was carried by answers to cloze gaps that were not embedded in manipulated sentences. Post-hoc analysis confirmed that for students with high reading ability scores the effect of SDL-version spilled over to other sentences. It is likely that good readers tried to integrate information across sentence boundaries to increase their understanding of the text, while poor readers did not. When SDL was increased and comprehension of the manipulated sentence was diminished, comprehension of other sentences also suffered. When SDL was reduced, good students were able to benefit more by using their higher level of comprehension of manipulated sentences to increase comprehension in other sentences. Alternatively, reducing the SDL may have freed up computational resources that skilled readers were able to use to better understand the text.

4. General Discussion

In two experiments we investigated the effect of syntactic dependency length (SDL) on how readers process and comprehend texts. A total of twenty texts were randomly selected and manipulated to create a text version with shorter syntactic dependencies and a text version with longer syntactic dependencies. We compared minimal sentence pairs that only differed in dependency length. We respected the ecological validity of the materials and presented the manipulated sentences in their natural context. In contrast to earlier studies, we included a wide range of syntactic

structures that readers encounter in everyday life and presented these texts to readers varying in reading proficiency.

Experiment 1 focused on the effects of SDL during on-line processing. Our results show that even a subtle increase in SDL causes readers to slow down when reading normal sentences in context. Readers need more time to process a Long-SDL sentence compared to a Short-SDL sentence. While Experiment 1 showed no effect of SDL on overall comprehension, Experiment 2 showed that higher SDLs can affect comprehension. Increased SDLs resulted in lower scores on the HyTeC-cloze tests. However, the effect was moderated by the size of the increase (i.e., SDL Delta), the complexity of the text and the reading ability of the reader. Text and reader characteristics influence whether the additional complexity of non-local dependencies results in diminished comprehension or not. When our texts were easy enough, readers could overcome small increases in SDL. On the other hand, in more challenging texts small increases were already problematic. This effect was not influenced by reader characteristics. Only comprehension of non-manipulated sentences was influenced by an interaction between reader and text version. For skilled readers the effect of SDL spilled over into non-manipulated sentences and they profited from decreased SDL in manipulated and non-manipulated sentences. Less-skilled readers did not show a similar benefit in non-manipulated sentences.

This study is one of the few that actually showed an effect of syntactic simplification on comprehension. Even with far more liberal syntactic adaptations, most experiments have turned out null-effects (e.g., Bailin & Grafstein, 2016; Duffy & Kabance, 1982; Johnson & Otto, 1982; Ulijn & Strother, 1990). This study is the first that we know of to document comprehension effects of SDL increases across a wide range of syntactic structures. Thus, decreasing syntactic dependency lengths tends to have a positive effect on processing ease ánd on comprehension. Although not all texts and all readers are equally affected, it seems good practice to reduce syntactic dependency lengths where possible. That is: in situations where long dependencies are not functional. Long dependencies can shift focus and highlight the importance of certain information (Renkema, 1989; cf. (2) and (3) above). Reducing the SDL in such cases is probably not beneficial for the readability of the text. As any adaptation intended to increase readability, it should be handled with care.

5 Comparing effects of connectives across coherence relations, texts and readers

In the previous two chapters we discussed how lexical and syntactic factors can affect readability. These factors have a strong tradition in readability studies. The final factor we will investigate in an experimental setup has not: coherence. While coherence has become a key concept in discourse studies (Sanders & Canestrelli, 2012), its importance has only slowly seeped through in readability studies (Bailin & Grafstein, 2016; Davison & Kantor, 1982; Graesser, McNamara, Louwerse & Cai, 2004; Sanders & Noordman, 1988). Kintsch and Vipond already argued for the inclusion of factors of coherence within readability formulae back in 1979, but progress on that front has been slow. Kintsch and Vipond (1979) believed that "*the absence of a theory of text structure and text processing*" made it impossible at that time to capture the true readability of a text (p.334). To date, theories on text structure and comprehension have come a long way. Yet, implementing these theories into the field of readability has proven to be a daunting task. One reason for this is that it is difficult to reliably measure coherence. Coherence is in essence a phenomenon that is located in the mind of the reader (Graesser et al., 2004; Sanders & Pander Maat, 2006; Sanders, Spooren & Noordman, 1992; Sanford, 2006; Zwaan & Rapp, 2006), which makes it by definition difficult to capture in an objective measure. As a result, coherence can only be approximated in readability studies by using linguistic cues that signal coherence (cf. Crossley, Skalicky, Dascalu, McNamara & Kyle, 2017; Feng, Elhadad & Huenerfauth, 2009; Feng Jansche, Huenerfauth & Elhadad, 2010; Pitler & Nenkova, 2008). The linguistic expression of coherence is often referred to as 'cohesion' although both terms are sometimes used interchangeably (Grasser et al., 2004; Sanders & Pander Maat, 2006).

The inclusion of coherence factors in readability research is especially important for another reason as well: when writers and editors indiscriminately apply readability formulae, cohesion tends to suffer (Davison & Kantor, 1982; Graesser et al., 2004; Honeyfield, 1977; Land, Sanders & Van den Bergh, 2008). When writers try to 'fool the formula' (Davison & Kantor, 1982; Noordman & Vonk, 1994), they often split up sentences to increase the readability rating of their text. Such adaptations do not necessarily benefit comprehension (Davison & Kantor, 1982; Duffy & Kabance, 1982; Johnson & Otto, 1982; Klare, 1984). For instance, in (1) the multi-clause sentence is split up into several short, independent sentences. As a result, the word '*because*' disappears. This is unfortunate since '*because*' is a cohesive device. It signals a coherence relation between the two clauses. The three

sentences in (1b) seem to describe three separate facts about Paco rather than a coherent story about why Paco is tired in class (cf. (1a)).

(1) a. Because he had to work at night to support his family, Paco often fell asleep in class.
b. Paco had to make money for his family. Paco worked at night. He often went to sleep in class.

(Ross, Long & Yano, 1991, p.2)

Another example of the effects of splitting sentences is given in (2). In (2b) the adverbial clause is split off. As a result, it is less clear in (2b) that growing new bark is the way the tree heals compared to (2a). Such revisions can be harmful to the readability of the text, but can be a side effect of the indiscriminative application of traditional readability formulae.

(2) a. If given a chance before another fire comes, the tree will heal its own wounds by growing new bark over the burned part.
b. If given a chance before another fire comes, the tree will heal its own wounds. It will grow new bark over the burned part.

(Davison & Kantor, 1982, p.192)

Halliday and Hasan (1976) have presented an extensive description of cohesive devices that can indicate coherence. Devices like '*because*' are connectives. Connectives can make relations between text segments explicit, both at a local level and at a global level (Pander Maat & Sanders, 2006). They signal that there is a relation between two segments and they signal the type of relation (e.g., causal, temporal, contrast). When a relation is not marked by a connective or another cohesive device, the reader is left to infer the relation in order to establish coherence. As such, connectives can help readers create a coherent mental representation of the text. Connectives have primarily been found to speed up integration of the upcoming segment (Cozijn, Noordman & Vonk, 2011; Kleijn, Mak & Sanders, 2011; Maury & Teisserenc, 2005; Van Silfhout, Evers-Vermeul, Mak & Sanders, 2014; Van Silfhout, Evers-Vermeul & Sanders, 2015; among others). Connectives thus increase reading ease, which is one aspect of readability (see Chapter 1). It is less clear whether connectives also increase the second aspect of readability: comprehension. Studies on the influence of coherence markers on text comprehension do not show a consistent pattern. Some studies show that linguistic marking of coherence relations improves the mental text representation, as is apparent in better answers on specific questions (Loman & Mayer, 1983) and better performance on probe recognition tasks (Millis & Just, 1994), while others find no effect on free recall (Meyer, 1975). Some authors have even claimed that in certain situations connectives can have a

negative influence on text comprehension (McNamara, Kintsch, Butler Songer, & Kintsch, 1996; Millis, Graesser & Haberlandt, 1993). In the present study we will further examine what effects connectives have on text comprehension and whether these effects can be generalized over texts and readers.

1. Coherence marking and comprehension

In order to understand a text, readers must create a coherent representation. Readers must understand how clauses and paragraphs relate to each other. Because connectives make these relations explicit, they should facilitate this process. By signaling the relation, connectives increase the chance that the reader notices the relation. Furthermore, it increases the chance that the reader interprets the relation correctly.

1.1 Effects of connectives and coherence marking on comprehension

There is much evidence for the facilitating effects of connectives during on-line processing. The dominant view has become that connectives give processing instructions: they instruct the reader how to process the upcoming segment and how to relate it to a previous one (e.g., Sanders & Spooren, 2007). As such, connectives make it easier to integrate information connected by a coherence relation (Cain & Nash, 2011; Canestrelli, Mak & Sanders, 2013; Cozijn, Noordman & Vonk, 2011; Kamalski, Lentz, Sanders & Zwaan, 2008; Kleijn et al., 2011; Koornneef & Sanders, 2013; Maury & Teisserenc, 2005; Sanders & Noordman, 2000; Van Silfhout et al., 2014; 2015). This facilitation is immediate: primarily the words just after the connective are read faster in the explicit condition compared to the same words in the implicit condition.

For comprehension, studies show less consistent results. Connectives can facilitate comprehension (Degand & Sanders, 2002; Degand, Lefèvre & Bestgen, 1999; Loman & Mayer, 1983; Lorch & Lorch, 1986; Millis & Just, 1994; Sanders, Land, Mulder, 2007), but can also reduce comprehension when an inappropriate connective is used (Cain & Nash, 2011; Millis et al., 1993; Murray, 1997). Most studies, however, show no effect of connectives or cohesive devices (Murray, 1995; Sanders & Noordman, 2000) or report mixed results with facilitation occurring only in certain cases and not in others (McNamara et al., 1996; McNamara, 2001; O'Reilly and McNamara, 2007; Ozuru, Dempsey & McNamara, 2009; Van Silfhout, Evers-Vermeul & Sanders, 2014; Kamalski, Sanders & Lentz, 2008; Land et al., 2008; Spyridakis & Standal, 1987). Even when cohesion is increased by adding connectives in combination with other cohesive devices – e.g., making referential

chains more explicit, eliminating ellipses, adding elaborative content and/or adding advanced organizers – increasing cohesion does not always result in increased comprehension. For instance, Freebody and Anderson (1983a) increased cohesion by strengthening cohesive ties (e.g., argument overlap) and by adding connectives and cue phrases to their text fragments. However, they did not find an effect of cohesion on recall, summarization or verification.

More recently, Van Silfhout and colleagues (2014; 2015) manipulated the presence of connectives in history texts. They found that the connectives facilitated comprehension of pre-vocational and pre-university 8th-grade students, but only when comprehension was measured with inference questions that directly tapped the understanding of the coherence relations. They also found facilitation of connectives on situation-model questions (i.e., sorting task and timeline task) in one experiment, but were unable to replicate these results in a second experiment. No effect was found on their verification statements. Similarly, Land, Sanders and Van den Bergh (2008) found facilitating effects of connectives for multiple-choice questions that targeted both connections between text segments and facts, but performance on situation-model questions increased only for the lowest-level readers. However, in contrast to Van Silfhout, in Land et al. the presence of connectives was confounded with a manipulation of layout. Texts with connectives were presented in a normal, continuous layout format; text without were presented with each sentence starting on a new line. Van Silfhout showed that readers were differently affected by the combination of these two features (Van Silfhout, Evers-Vermeul & Sanders, 2014).

There are multiple reasons for the discrepancies in the results on comprehension (see also Spyridakis, 1989ab; Degand & Sanders, 2002; Sanders et al., 2007). First of all, texts vary both in their content and stylistic difficulty across studies. The text under investigation will determine which coherence relations are present to begin with and how conceptually difficult these relations are. The particular text will therefore strongly influence the results, especially when studies only investigate the effects of coherence marking in one text. It is important that studies use internal replications to get more reliable results. By presenting readers with multiple texts, the results will depend less on one specific text and this will increase the generalizability of the findings. For instance, Linderholm et al. (2000) increased the cohesion in a difficult and in an easy text. They found that effects of coherence marking only increased comprehension for their difficult text and not for their easy text. Similarly, Spyridakis and Standal (1987) only found effects of connectives for their medium and high level texts and not for their low level text. Effects of coherence marking thus strongly depend on the text under investigation.

Secondly, individual differences between readers moderate the results leading to differences in the strength and the direction of comprehension effects (see Kamalski, Sanders & Lentz, 2008; McNamara et al., 1996; McNamara, 2001; Land et al., 2008; O'Reilly & McNamara, 2007; Ozuru et al., 2009; Van Silfhout et al.,

2014; 2015). Readers with more prior knowledge seem to benefit less from a highly cohesive text than from a low cohesive text, presumably because a low cohesive text forces them to actively engage and to process the text on a deeper level. Conversely, readers with low levels of prior knowledge benefit the most from a highly cohesive text (McNamara et al., 1996; McNamara, 2001; Kamalski, Sanders & Lentz, 2008). Reading skills have also been found to interact with coherence marking effects. Ozuru et al. (2009) and O'Reilly and McNamara (2007) found that readers with high reading proficiency benefit more from coherence marking than readers with low reading proficiency. Reading skills are thus required to take full advantage of coherence marking. Indeed, teaching students to recognize coherence markers can result in better text comprehension (Vennink, 2014). On the other hand, Land et al. (2008) found that low-level students benefit more from coherence markers than higher level students. However, Land et al. used texts that were written especially for the low-level students. With respect to content and style, these texts were probably too easy for higher-level students for them to experience the full effect of coherence marking. Thus, the effect of coherence marking is affected by the interaction between reader and text.

A third reason for the diverging results is that comprehension has been measured using a large variety of tasks (e.g., free, cued or prompted recall, verification statements, bridging inference questions, situation-model questions, cloze, and summarization). Not all measures are equally sensitive to connective manipulations. Methods like free recall may be too global to measure the local effects of coherence markers. The most consistent results are found "*when comprehension questions are specifically directed towards explicitly marked relations*" (Sanders et al., 2007, p.230). However, even similar methodologies are difficult to compare because they are constructed or scored differently, and they usually depend on the text under investigation. Situation-model questions are particularly difficult to compare across studies and texts. These questions are completely custom-made for a particular text and the text will determine whether timeline, sorting or schematic situation-model questions are most appropriate.

1.2 Effects of coherence types

Up to now, we have regarded coherence as if all relations are equal and affect readers in the same way. There are in fact many different types of coherence relations such as additive (3), temporal (4), contrast (5) and causal relations (6). Although there are far more detailed taxonomies of coherence relations available in literature, in this chapter we will focus on these four main categories and differences between these categories.

(3) John went hiking and Bill went to the cinema.

(4) John cooked supper. Afterwards, Bill did the dishes.

(5) John likes to go fishing, but Bill hates the outdoors.

(6) Bill cleaned up the kitchen because John left a mess.

Although the effect of linguistic marking on different coherence relations has received little attention in comprehension studies, it is likely that the linguistic marking of some relations will have more impact than the marking of others. Many comprehension studies have investigated different connectives and other coherence markers simultaneously (Beck, McKeown, Sinatra, & Loxterman, 1991; Freebody & Anderson, 1983a; McNamara et al., 1996; Meyer, 1975; Vidal-Abarca & Sanjose, 1998). As a result, we do not know whether different connectives have similar effects on comprehension. There are, however, strong reasons to expect that the effect of a connective will differ depending on the type of relation it marks. For one, coherence relations differ in their relative cognitive complexity (see the cognitive approach to coherence relations, Sanders, Spooren & Noordman, 1992; 1993; Sanders & Spooren, 2007; 2009). Simple relations are acquired first and followed by increasingly complex relations. For instance, children learn positive relations before contrastive relations, and additive relations before causals (Evers-Vermeul & Sanders, 2009; 2011; Spooren & Sanders, 2008; Van Veen, 2011). In addition, complex relations require more time to process: subjective relations take more time to process than objective ones (Canestrelli et al., 2013; Traxler, Bybee & Pickering, 1997).

According to this cognitive approach, additive relations are considered to be the simplest relations (Sanders, 2005). In a way they also represent the weakest connection, since – from a truth-conditional point of view – the additive relation is already implied by the juxtaposition of two text segments and they offer the least information on how the text might continue. Temporal relations introduce a temporal ordering and are therefore slightly more difficult. Contrast and causal relations are considered more complex and stronger connected relations. Meyer and Freedle (1984) found that texts with additive structures are recalled worse than texts with causal or contrastive structures. Similar results have been found by Mulder (2008) on verification statements and by Sanders and Noordman (2000) on recall and verification statements. The question is whether coherence marking interacts with this difference in complexity. Weak, low-complexity relations may benefit less from marking than strong, high-complexity relations. Connectives carry with them all the information concerning the relation. Additive connectives like '*and*' and '*furthermore*' are less restrictive. Koornneef and Sanders (2013) found in their

completion task that an additive connective elicited a much wider range of coherence relations than a causal or contrast connective. Therefore, additive relations might not supply very clear processing instructions, compared to other connectives. However, Sanders and Noordman (2000) found no interaction between coherence marking and coherence type (list vs. problem-solution) on reading time, recall or verification statements. Murray (1995) found mixed results. In his first experiment he found an interaction between coherence type and coherence marking. Introducing an appropriate connective increased cued recall for causal relations, but decreased recall for additive relations. No effect was found for contrast relations. However, this pattern was not replicated in Murray's second experiment.

The cognitive complexity of coherence relations may not be the only factor influencing comprehension effects. In a later paper, Murray proposes that coherence relations are marked according to the 'Continuity principle': "*This principle states that readers have a bias toward interpreting sentences in a narrative as following one another in a continuous manner. As readers progress through a narrative, they assume that the events will follow in a linear fashion. And when this occurs, reading is relatively easy. Continuity can be conveyed easily via additive or causal relations. When a reader encodes a text event that is discontinuous in the absence of a marker or indication of the discontinuity, reading is more difficult.*" (1997, p. 228).[1] According to the Continuity principle, connectives will benefit comprehension and processing the most when they mark a discontinuous relation. This also includes connectives that mark a non-linear order of events (e.g., '*because*'). Even additive relations are not necessarily completely continuous. The second segment can elaborate on the first – in which case it continues the topical focus of the prior segment – but the segments may also be parallel as in a list. In a list, the second segment introduces a new point that refers back to a higher level topic or referent (Knott, Oberlander, O'Donnell & Mellish, 2001; Pander Maat, 2001; 2002).[2] In this sense it does not flow from the first segment in the same way as elaborations. Hence, list relations may not be as continuous as elaboration relations.

According to the continuity principle, connectives can be left out in continuous relations without causing too much problems because these relations are more expected. Indeed, Segal, Duchan and Scott (1991) found that the majority of successive sentences in narrative texts are continuous. Readers expect successive

[1] Zwaan and colleagues have presented similar views on processing continuous event structures versus discontinuous event structures (Zwaan, Langston & Graesser, 1995; Zwaan & Radvansky, 1998)

[2] Additive relations may also share 'joint relevance' where together the two segments answer one topic question ("*What did you do this summer?*" "*I painted my house and visited family.*"; Pander Maat, 2001).

text segments to be continuous and if segments are discontinuous, they will benefit from signals that guide their expectations to overcome the disruption. Additional evidence comes from corpus research. Asr and Demberg (2012) studied the frequency distribution of relations in the Penn Discourse Tree Bank corpus (Prasad et al., 2008). The frequency with which a relation was explicitly marked differed between coherence types. Drawing on the *Continuity hypothesis* and the *Causality-by-Default hypothesis* (Sanders, 2005), they hypothesized that causal relations and relations that did not disrupt the continuity of the text (e.g., not contrastive and not in backward order) would be explicitly marked less often than other relations.[3] Their findings showed support for both hypotheses. Thus it seems that certain relations are more in line with readers' expectations than others; in particular temporal succession, forward causality and elaboration. If so, marking these relations should have less effect than marking an unexpected relation.

In sum, there is evidence from acquisition, processing and corpus studies, that types of coherence relations are different and that different predictions may be formulated for the effects of linguistic marking of these relations on comprehension. It is likely that some relations will benefit more from explicit marking than other relations. Or viewed from another perspective: some relations may depend more strongly on explicit marking than others in order to be correctly understood.

1.3 The optionality of connectives

The presence of a connective is not always optional. Adding or removing connectives can be restricted for different reasons. Firstly, relations can become uninterpretable when a connective is removed. In (7) it is possible to remove the contrast connective '*but*' and still get a contrastive interpretation because of the syntactic parallelism and the semantic contrast (tall-short). In (8) on the other hand, it is not possible to get a contrastive reading without the connective. Contrast relations that are not evident by the use of negation or lexical contrast may become very hard or impossible to grasp when the connective is removed. In such cases, the connective is not optional.

(7) a. John is tall <u>but</u> Bill is short.
b. John is tall. Bill is short.

(8) a. John is tall but he's no good at basketball.
b. #John is tall. He's no good at basketball.

(Lakoff, 1971, p.131)

[3] Cf. Taboada (2009) on what constitutes an implicit relation.

Secondly, connectives must mark the relation intended by the writer. Adding random connectives will not help the reader to create a coherent mental representation. The use of an inappropriate connective is not beneficial for comprehension because the relation cannot be interpreted (Cain & Nash, 2011; Millis et al., 1993; Murray, 1997). No connective seems to be better than the wrong connective. Furthermore, adding a connective has the propensity to change the meaning of a relation (e.g., Kleijn et al., 2011).

Finally, although an implicit relation may be uninterpretable without a suitable connective, this does not mean that connectives should be added everywhere. Connectives also draw attention to the relation by signaling important relations. If all relations are marked, it may seem that every relation is equally important. On the other hand, when no relation is marked texts can become overly fragmented and the normal flow of the text is disrupted (see (1b) above). Corpus studies indicate that between 27% and 44% of the relations is signaled by a connective or cue phrase[4] (Mann & Taboada, 2017; Taboada, 2006). Of course, the 'implicit' relations can be marked with other devices (e.g., syntax, semantics and pragmatic mechanisms; Halliday & Hasan, 1976; Taboada, 2009). In fact, it may be that no relation is truly implicit. Every relation has certain cues to guide readers; otherwise they would not be able to establish the relation (Taboada, 2009). However, these cues may be very subtle and not as clear cut as connectives, which are the prototypical linguistic markers of coherence relations.

1.4 Present study

Although the research on connectives and other cohesive devices is extensive, most studies have investigated the effects on comprehension using only one or two texts at a time. In addition, manipulations were often very liberal in the sense that they enhanced cohesion in any way possible, including adding information not included in the low-cohesive version (e.g., Beck et al., 1991; Freebody & Anderson, 1983a; Loman & Mayer, 1983; McNamara et al., 1996; Meyer, 1975; Ozuru et al., 2009; Vidal-Abarca & Sanjose, 1998). Based on these studies, it is hard to generalize the effects of connectives on the readability of texts. In the present study we therefore follow more recent work of Van Silfhout and colleagues (2014; 2015) in which cohesion manipulations were limited to adding connectives or cue phrases to authentic texts. While Van Silfhout et al. focused on so-called positive relations – that is additive, temporal and causal relations – we also include contrastive relations. We hypothesize contrast connectives to be indispensable for creating coherence (Haberlandt, 1982). From a cognitive complexity as well as a continuity perspective, contrast relations should benefit strongly from coherence marking. Another

[4] E.g., '*on the other hand*', '*as a result*', '*That is why*'.

difference with Van Silfhout's studies is that we vastly increase the number of texts and readers to investigate the generalizability of the results.

We examine whether twenty authentic Dutch texts benefit from adding additive, temporal, causal and contrast connectives. These texts are randomly selected from a larger corpus and differ both in the number of coherence relations and in the types of coherence relations that they contain. These texts are presented to readers varying in reading proficiency to see whether effects of connectives can be generalized over texts and readers. In contrast to previous studies, text comprehension is not measured with text-based or bridging-interference questions, but with Hybrid Text Comprehension cloze tests (see Chapter 2). The HyTeC-cloze procedure makes it possible to directly compare effects across texts because performance is dependent on text difficulty without interference from question difficulty (cf. Klare, 1976a). Furthermore, the HyTeC-cloze makes it possible to simultaneously investigate effects on comprehension on a global text level and on a local level (e.g., targeting specific manipulations).

2. Method

2.1 Participants

Thirty-five Dutch secondary schools participated with in total 794 students from grades 8, 9 and 10 (age: 13 - 16). Testing was introduced as part of the regular school curriculum. The students were enrolled in different levels of secondary education: in pre-university education ('vwo'), general secondary education ('havo') or pre-vocational education ('vmbo-gt', 'vmbo-kb' or 'vmbo-bb').[5] The distribution of participants over Grades and Level of education is given in Table 1. Grade 10 students of pre-vocational education were not included in the study because pre-vocational students are graduating in grade 10.

Table 1: Distribution of participants over Grade and Level of education

	Pre-voc. (low)	**Pre-voc. (medium)**	**Pre-voc. (high)**	**General education**	**Pre-uni.**	**Total**
Grade 8	24	72	175	54	45	370
Grade 9	34	50	96	63	71	314
Grade 10	-	-	-	75	35	110
Total	58	122	271	192	151	794

[5] These five levels are ordered from theoretical oriented education to practice oriented education. For more information on the Dutch educational system see EP-Nuffic (2015).

2.2 Materials

Twenty texts were selected from a collection of Dutch authentic texts (see Chapter 1). This collection contained 146 educational textbook texts and 120 public information texts. The texts were randomly selected and not selected based on their potential for manipulation success. Ten texts were taken from educational textbooks on history, geography, Dutch language and economics. These texts were written especially for students in secondary education. The other ten texts were public information texts which discussed matters related to health (e.g., donor registration), the environment (e.g., pest control) and regulations (e.g., obtaining a scooter license). These texts were written for the general public but were also relevant for Dutch adolescents. All texts were 300 to 410 words long and did not contain figures or tables.

Low-coherence marking (Low-CM) and high-coherence marking (High-CM) text versions were created by manipulating approximately 1/3rd of the coherence relations (N=193).[6] These relations were explicitly marked by a connective in the High-CM version and were left implicit in the Low-CM version. Relations were causal (9), contrastive (10), additive (11) or temporal (12). An example of a complete text is given in Appendix 8. Connectives were only left out if they were optional. That is, if leaving the connective out did not alter the interpretation of the relation or the flow of the text (see Section 1.3). The text versions were kept as close to the original as possible and we only adapted coherence relations that were already present in the original texts. As a result, it depended on the text at hand which types of coherence relations were manipulated (i.e., causal, contrast, additive or temporal relations). Table 2 presents the overall frequency of the manipulations per coherence relation. Half of the manipulated relations were causal, including objective as well as subjective causal relations in forward and in backward order. There were only a few temporal relations. Contrast and additive relations were well represented in the sample. Additive relations included list and elaboration relations.

[6] If a text did not reach this minimum number of manipulations, another text was randomly selected from the collection of texts to take its place. If more manipulations were possible, preference was given to the more complex relations. A comparison of the number of connectives in the original text versus the number of connectives in the Low-CM and High-CM text versions can be found in Appendix 7.

Causal relation

(9) Als de productiestructuur inderdaad verbetert, kan het land op de wereldmarkt beter concurreren met andere landen. Daardoor kan de export van het land stijgen en de import van het land afnemen.

'If the production structure improves, the country can compete better with other countries on the global market. As a result, the county's export can increase and the country's import can decrease.

Contrast relation

(10) Minderjarigen kunnen vanaf hun twaalfde hun wens in het Donorregister laten opnemen. Ouders of voogden hoeven hiervoor géén toestemming te verlenen. Maar als minderjarigen instemmen met donatie en voor hun zestiende overlijden, kunnen ouders of voogden alsnog weigeren.

'Minors can record their preference in the Donor Registers from age twelve. Parents or guardians do not have to consent for this. But if minors agree to become a donor and die before they are sixteen, parents or guardians can still refuse.'

Additive relation

(11) Als de brandweer wordt gealarmeerd door een brandmelder heeft dit veel consequenties. Ten eerste rukt de brandweer met spoed uit naar het meldadres. Dat brengt verkeersrisico's met zich mee. Ten tweede, als de brandweer uitrukt voor een nodeloze alarmering, is ze op dat moment niet beschikbaar voor andere wel noodzakelijke, hulpverlening.

'There are a lot of consequences when the fire department is notified by a fire alarm. First of all, the fire department rushes to the location of the fire alarm. This can result in dangerous traffic situations. Secondly, if the fire department responds to a false alarm, she will not be available to provide aid in a genuine rescue situation.'

Temporal relation

(12) In de late middeleeuwen zien we steeds meer steden in Europa. Handelaren vormden handelsgemeenschappen op plaatsen waar relatief rijke afnemers zaten, zoals een adellijk hof, een militaire vesting of een klooster. Deze handelsgemeenschappen trokken vervolgens ambachtslieden aan.

'In the late Middle ages, we find more and more cities in Europe. Tradesmen formed trading communities at locations with rich customers, like a noble house, a military stronghold or a monastery. Subsequently, these communities attracted craftsmen.'

Table 2: Frequencies of manipulated coherence relations

Coherence relation	Frequency
Causal	100
Contrast	42
Additive	42
Temporal	9
Total	193

2.3 Measures

2.3.1 Comprehension assessment

The texts were transformed into cloze tests following the Hybrid Text Comprehension cloze procedure (HyTeC-cloze; see Chapter 2). This hybrid cloze procedure combines the strengths of mechanical and rational cloze tests into a valid and reliable measure of text comprehension. The rational strategy is used to exclude words that do not rely on text level comprehension from becoming cloze gaps. This includes words that can be reconstructed using only grammatical knowledge or knowledge of usage conventions (e.g., articles and multi-word expressions). Also excluded are words that can only be guessed, such as names and numbers. All other words in the text are candidates for deletion, except for the connectives that were added as part of the manipulation. The remaining candidates are divided over different cloze versions via mechanical selection. Two cloze versions were randomly selected to serve in the study. In total 10% of the words were deleted. Depending on the text length, the cloze tests contained 30 to 41 cloze gaps. The same words were deleted in the Low-CM and High-CM text version.

2.3.2 Reading ability

Standardized reading ability scores were made available for all students. Two different tests were used to measure Reading ability. Although scores of both tests were mapped to the same scale, analyses showed that scores for one of the tests were consistently higher. To control for this complication, the factor Reading test was included in the analyses.

2.4 Design

This study is part of a larger scale project in which the difficulty of 60 texts was assessed among 2926 participants. Participants were randomly assigned to this part of the study.

The experiment was set up following a matrix sampling design (e.g., Gonzalez & Rutkowski, 2010). Each participant was given four different cloze tests: one cloze test of a Low-CM educational text, one of a Low-CM public information text, one of a High-CM educational text and one of a High-CM public information

text. To balance out possible order effects, each combination of cloze tests was presented in two orders.

2.5 Procedure

All testing took place at the participating schools. The tests were administered by the school teachers in classroom settings. Cloze tests were presented digitally on computers. Participants filled in the cloze gaps on the screen. To fill in all four cloze tests, participants took part in two sessions of 45 minutes. Schools scheduled all sessions themselves over the course of a couple weeks.

2.6 Scoring procedure and data-clean up

The answers to each cloze gap were dichotomously scored (1 = correct; 0 = incorrect) according to the acceptable word scoring procedure (see Chapter 2). Following the acceptable word scoring procedure, not just originally deleted words were scored as correct but semantically correct alternatives were given the same score (including spelling errors and typos). The acceptability of alternative answers was judged by the global appropriateness criterion which means that the answer had to fulfill "*all the contextual requirements of the entire discourse context in which it appears*" (Oller & Jonz, 1994, p.416). Each answer was scored by two independent judges from a pool of 16 judges. When the judges disagreed, a third judge made the final decision. All judges received a short training to familiarize them with the scoring procedure.

10% of the data was removed because students repeatedly gave non-serious answers or did not answer the cloze gaps at all. Cases where students occasionally failed to fill in a gap were regarded as incorrect answers rather than missing answers. The final dataset contained 99735 cases within 2861 cloze tests.

2.7 Analyses

Cloze scores were not aggregated before analysis. The data was analyzed at the response level using generalized linear mixed effect modeling (GLMM) with a logit link. Each case represents an answer to an individual cloze gap. Observations are nested within students and texts, with students nested in schools.

Two separate analyses were performed: a text level analysis and a relation level analysis. Goal of the first analysis was to see if increasing the number of connectives within a text influences overall cloze performance. All cloze gaps were included in this analysis (i.e., cloze gaps in manipulated sentences as well as cloze gaps in sentences that were not manipulated). Reader characteristics were included to test for possible interactions between the manipulation and the skillset of the reader. Finally, the text feature Genre (educational textbook or public information) was included in the analysis to test the generalizability of effects over genres.

Given that most effects in prior research have been found on questions that targeted the relation itself (Section 1.1), it is likely that effects of the manipulation are localized to the segments surrounding the connective and will not affect performance on all cloze gaps. Therefore, we ran another analysis using only cloze scores from sentences in which a connective was added or removed ('host sentences'). Although the span of the connectives might have exceeded the host sentence, we expect any effect of coherence marking to be the strongest here.

A second goal of this relation level analysis was to test whether the size of the effect is moderated by the type of coherence relation. Adding a contrast connective may not have the same effect as adding an additive connective. This hypothesis was explored by including a 2-way-interaction of Coherence Marking with Connective type in the analysis. We expect the strongest facilitation for contrast relations and slightly less facilitation for causal relations since causal relations are generally more expected than contrast relations (see Section 1.2). We are unsure whether additive relations will facilitate comprehension. On the one hand we hypothesize that marking these relations will have less facilitating effect than causal relations, because they are a continuation of the text (and thus expected). In addition, additive connectives are not very informative regarding the content of the upcoming segment. On the other hand, because such continuations are more often unmarked, marking the relation may have an adverse effect. Highlighting a relation that should not necessarily receive such focus could disrupt comprehension rather than increasing it.

Finally, an exploratory analysis was performed using further subcategorization of the coherence types. As discussed in Section 1.2, coherence types may not be as homogeneous as they may appear and further differentiation may be preferable. For instance, our causal relations include objective as well as subjective relations in continuous (i.e., forward) and in discontinuous (i.e., backward) order. In addition, our additive relations include list as well as elaboration relations. We explored whether subcategorization of our coherence types revealed other patterns regarding linguistic marking.

Descriptions of all factors are given in Table 3.

Table 3: Descriptions of factors

Factor	**Description**	**Levels**
Coherence marking	Text version: high or low coherence marking	2 levels
Education level	Level of education in which the student is enrolled	5 levels
Grade	Grade in which the student is enrolled	3 levels
Reading ability	Reading ability score on Dutch reading ability test (centered and standardized)	Continuous
Reading test	Reading test used to test reading ability	2 levels
Genre	Educational text or public information text	2 levels
Coherence type	Type of manipulated coherence relation	5 levels

3. Results

3.1 Text level analysis

Table 4 shows the mean percentage of gaps that was answered correctly per text version, Education level and Grade. The analysis revealed main effects for Education level, Grade, Reading ability, Reading test and Genre, but not for Coherence marking. Odds of a correct answer increased with Education level, Grade and Reading ability score. Performance on educational texts was better than on public information texts. The model including Coherence marking is presented in Table 5.

Table 4: Mean probability of a correctly answered cloze gap per text version, Education level and Grade in percentages

		Coherence marking	
Education level	**Grade**	**High-CM**	**Low-CM**
Pre-vocational (low)	Grade 8	37.80%	30.04%
	Grade 9	33.86%	34.75%
Pre-vocational (medium)	Grade 8	40.40%	42.06%
	Grade 9	47.79%	48.39%
Pre-vocational (high)	Grade 8	48.60%	48.66%
	Grade 9	54.52%	53.40%
General	Grade 8	57.87%	55.43%
	Grade 9	61.26%	61.82%
	Grade 10	68.02%	63.96%
Pre-university	Grade 8	67.24%	65.25%
	Grade 9	69.89%	69.66%
	Grade 10	72.67%	70.49%

Table 5: Model text level analysis (including non-significant factor Coherence marking)

Random effects	**Estimates**	**SE**	***Z***	***p***	
School	0.015	0.010	1.500	.067[a]	
School: Student	0.212	0.013	16.308	<.001[a]	
Text	0.038	0.007	5.429	<.001[a]	
Fixed effects	**Estimates**	**SE**	***Z***	***p***	**Odds ratio**
Intercept	-0.409	0.125	-3.272	<.001	0.664
Coherence marking: High-CM	0[b]				
Coherence marking: Low-CM	-0.020	0.014	-1.429	.153	0.980
Education level: pre-voc. low	0[b]				
Education level: pre-voc. medium	0.324	0.093	3.484	<.001	1.383
Education level: pre-voc. high	0.578	0.093	6.215	<.001	1.782
Education level: general	0.801	0.112	7.152	<.001	2.228
Education level: pre-uni.	0.864	0.125	6.912	<.001	2.373
Grade 8	0[b]				
Grade 9	0.109	0.045	2.422	.015	1.115
Grade 10[c]	0.212	0.096	2.208	.027	1.236
Reading ability	0.308	0.029	10.621	<.001	1.361
Reading test: R	0[b]				
Reading test: V	-0.535	0.091	-5.879	<.001	0.586
Genre: Public information	0[b]				
Genre: Educational textbook	0.343	0.026	13.192	<.001	1.409

[a] One-sided. [b] Set as reference level. [c] Unbalanced, no 10th grade pre-vocational students were present in the sample.

3.2 Relation level analysis

After the text level analysis, a relation level analysis was conducted to find out whether effects of coherence marking can be found if we restrict the analysis to the manipulated sentences. This analysis also enables us to investigate the possible interaction between linguistic marking of coherence relations and the type of coherence relation. Cloze gaps were selected from sentences to which a connective was added or from which a connective was removed. Five sentences were excluded from this analysis because they contained not one but two manipulations that marked different types of coherence relations (e.g., first connective marked relation with previous sentence; second connective marked relation between two clauses within the sentence). 183 items remained with a total of 38214 observations. Table 6 shows the mean percentage of gaps that was answered correctly per text version, Education level and Grade for the remaining sentences.

The analysis revealed main effects for Coherence Marking, Education level, Grade, Reading ability, Reading test and Genre. Performance was better in the High-CM version and increased with Education level, Grade and Reading ability. Odds of a correct answer were also higher for educational texts than for public information texts. Students performed better on High-CM host sentences than on Low-CM host sentences. However, when Coherence type was added to the model

the overall effect of Coherence marking disappeared and a significant interaction between Coherence marking and Coherence type was found. This final model is presented in Table 7. Causal and High-CM are selected as reference levels for Coherence type and Coherence marking respectively. The estimate given after 'Low-CM' therefore denotes the estimated effect for causal relations when they are not marked with a connective. This effect is negative though only marginally significant. For the other coherence types the 'Low-CM' estimate is adjusted and can be found by adding up the 'Low-CM' estimate and the estimate for the interaction 'Low-CM*Contrast', 'Low-CM*Temporal' or 'Low-CM*Additive' respectively. For temporal relations, this adjustment is not significant, but for contrast relations and additive relations it is. The coherence marking effect for contrast relations is in the same direction as for causal relations (i.e., Low-CM is more difficult), but the effect for contrast relations is significantly larger than for causal relations. Conversely, for additive relations the effect is also stronger than for causal relations, but in the opposite direction. Additive relations were easier when they were not marked by a connective. In sum, coherence marking significantly improved cloze performance for contrast relations and marginally improved performance for causal relations. It had no effect on temporal relations[7] and an opposite effect on additive relations (see Figure 1). No interactions with Education level, Grade, Reading ability or Genre were found.

Table 6: Mean probability of a correctly answered cloze gaps in host sentences per text version, Education level and Grade

		Coherence marking	
Education level	**Grade**	**High-CM**	**Low-CM**
Pre-vocational (low)	Grade 8	37.40%	28.49%
	Grade 9	32.45%	32.02%
Pre-vocational (medium)	Grade 8	39.44%	40.68%
	Grade 9	48.07%	45.83%
Pre-vocational (high)	Grade 8	48.12%	47.34%
	Grade 9	53.27%	51.35%
General	Grade 8	57.19%	53.56%
	Grade 9	59.90%	59.20%
	Grade 10	67.42%	63.05%
Pre-university	Grade 8	67.96%	64.11%
	Grade 9	67.97%	69.03%
	Grade 10	71.52%	71.72%

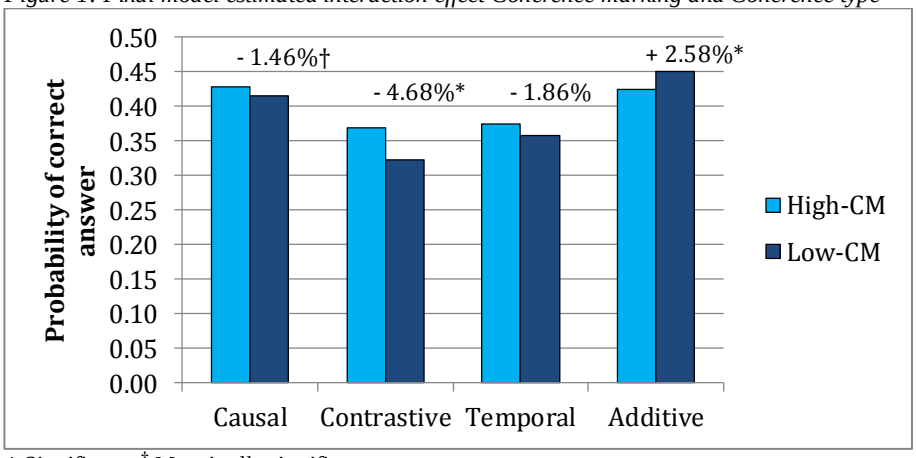
[7] Temporal relations were underrepresented with only 7 cases. The non-significant result may be caused by a lack of power.

Table 7: Final model relation level analysis

Random effects	**Estimates**	**SE**	***Z***	***p***	
School	0.032	0.013	2.462	<.001[a]	
School: Student	0.189	0.015	12.600	<.001[a]	
Sentence	0.039	0.007	5.571	<.001[a]	
Fixed effects	**Estimates**	**SE**	***Z***	***p***	**Odds ratio**
Intercept	-0.284	0.140	-2.029	.042	0.753
Coherence marking: High-CM	0[b]				
Coherence marking: Low-CM	-0.060	0.031	-1.935	.053	0.942
Coherence type: Causal	0[b]				
Coherence type: Contrast	-0.249	0.038	-6.553	<.001	0.780
Coherence type: Temporal	-0.224	0.078	-2.872	.004	0.799
Coherence type: Additive	-0.022	0.040	-0.550	.582	0.978
Low-CM * Contrast	-0.147	0.054	-2.722	.006	0.863
Low-CM * Temporal	-0.018	0.110	-0.164	.870	0.982
Low-CM * Additive	0.165	0.057	2.895	.004	1.179
Education level: pre-voc. low	0[b]				
Education level: pre-voc. medium	0.356	0.101	3.525	<.001	1.428
Education level: pre-voc. high	0.578	0.102	5.667	<.001	1.782
Education level: general	0.802	0.122	6.574	<.001	2.230
Education level: pre-uni.	0.903	0.137	6.591	<.001	2.467
Grade 8	0[b]				
Grade 9	0.099	0.048	2.063	.039	1.104
Grade 10[c]	0.183	0.105	1.743	.081	1.201
Reading ability	0.296	0.031	9.548	<.001	1.344
Reading test: R	0[b]				
Reading test: V	-0.541	0.104	-5.202	<.001	0.582
Genre: Public information	0[b]				
Genre: Educational textbook	0.141	0.025	5.640	<.001	1.151

[a] One-sided. [b] Set as reference level. [c] Unbalanced, no 10th grade pre-vocational students were present in the sample.

Figure 1: Final model estimated interaction effect Coherence marking and Coherence type[8]



** Significant. † Marginally significant.*

3.3 Exploration of subcategories of coherence types

The relation level analysis showed different effects of coherence marking for causal, contrast, temporal and additive relations. In order to further explore these differences we performed four additional analyses in which we subcategorized these coherence relations. The following subcategorizations were used:

1. Additive relations were split into elaboration and list relations;
2. Causal relations were split into objective and subjective relations;
3. Causal relations were split into forward and backward relations;
4. Causal relations were split into objective-forward, objective-backward, subjective-forward and subjective-backward relations.

The final models of these analyses can be found in Appendix 9. The subcategorization of additive relations indicated differences between the subcategories. Coherence marking seemed not to affect elaboration relations and to negatively affect list relations. However, a comparison of the effect of Coherence marking for lists and elaborations did not reach statistical significance (β = -0.171, SE = 0.108, p = .113).

Subcategorizations within causal relations did not reveal any significant differences or trends towards significance, suggesting that all subcategories of causal relations tend to benefit somewhat from coherence marking.

[8] Estimates are set on the reference levels for Education level, Grade, Reading ability, Reading test and Genre (see Table 7). They do not reflect overall mean scores.

4. Discussion

We investigated the effect of coherence marking in twenty randomly selected authentic Dutch texts. Coherence marking was manipulated by adding or removing causal, temporal, contrast or additive connectives. Texts were presented to secondary education students differing in reading proficiency. Their comprehension of the texts was measured with HyTeC-cloze tests. Our results show that adding connectives to a text only affects comprehension on a local level. The presence of connectives did not influence comprehension on a global text level. Cloze scores were higher for sentences to which a connective was added than for sentences where a connective was removed. These results are in line with previous findings that show that connectives mainly function at a local level in the text (Sanders et al., 2007). Contrary to findings of O'Reilly and McNamara (2007) and Ozuru et al. (2009), our coherence marking effect was independent from reading proficiency. Both high-proficiency and low-proficiency readers performed better when relations were marked with a connective. Coherence marking also did not interact with Education level. Students enrolled in the highest levels of Dutch secondary education benefitted as much as students enrolled in the lower levels.

Coherence marking did interact with the type of coherence relation. Connectives facilitated comprehension of contrast and causal relations, but decreased cloze scores in additive relations. The results for contrast and causal connectives are in line with our expectations. These relations represent the strongest connections and as such are the most informative with regard to the upcoming segment. Readers benefitted the most from the linguistic marking of contrast relations. This effect occurred despite of the other contrast cues that were necessarily still present in the sequence to keep it interpretable as a contrast relation (see Section 1.3). Causal connectives only showed a trend towards facilitation. This may seem surprising given the relatively large amount of literature showing an effect of causal connectives. However, most of this literature shows an effect on on-line processing and is not focused on comprehension effects. One way to explain this limited effect of causal connectives in comprehension is that readers will try to connect segments in a causal way by default (Causality-by-default hypothesis; Sanders, 2005), and therefore readers already set out to connect the segments causally. Adding a connective only confirms their expectation, which explains the only marginally significant effect for causals.

Another possibility is that causal relations are too heterogeneous and that different subtypes have opposite or reduced effects. However, the explorative analyses we conducted – in which causal relations were divided depending on

subjectivity and linear order – did not give any indication that effects differed for different subtypes of causal relations.

With regard to additive relations, connectives decreased comprehension. Adding a connective resulted in lower cloze scores. It seems readers where thrown off track rather than guided by these connectives. It may be that highlighting such relations drew unnecessary focus towards the integration of the segment. Rather than taking two consecutive segments as sharing a (sub)topical focus – which is a given in an implicit relation by simple juxtaposition – additive connectives forced readers to look at this relation more closely. For list relations, segments 'A' and 'B' are parallel elements connected under a global topic. Marking this relation will force the reader to identify this topic. It thereby blocks the interpretation that B is a continuation of A, elaborating on the same subtopic. In addition, it forces readers to identify the shared topic and reinterpret A as part of a list as well. This may force the reader to restructure his mental representation (see Kintsch & Vipond, 1979). In addition, because A and B are separate points, their relation offers little information regarding the content of either segment (compared to contrast relations for instance, in which one of the segments contradicts an expectation based on the other segment). In contrast, elaborations flow continuously from the first segment and therefore do not require restructuring of the mental representation. An elaboration connective confirms the natural continuation of the text, but offers little information on how the text will continue. This would be in line with the absence of an effect for elaboration relations. Although these are mere speculations based on our results, at the very least our results indicate the need for more theoretical and experimental research into different types of coherence relations and their linguistic expression.

In contrast to mixed signal studies, coherence marking was manipulated in the present study with only one type of signal: connectives. Although connectives are prototypical cues for relational coherence, relational coherence can be marked with many other cohesive devices including cue phrases, verbs and syntax. Our results show that connectives can facilitate or disrupt comprehension depending on the coherence relation that they mark. Whether this holds for other coherence markers is yet to be determined. By teasing apart different cohesive devices as well as different types of coherence relations, we have gained more insight into the complex mechanisms that influence how coherence is established by the reader.

6 Predicting the readability of texts for Dutch adolescents

In the previous three chapters we discussed how lexical, syntactic and cohesive factors can affect readability. Using an experimental setup, we have examined the contribution of these factors to comprehension. We have seen that in these controlled experiments – where only stylistic difficulty is manipulated and conceptual difficulty is the same across text versions – lexical, syntactic and cohesive factors affect comprehension. In this final chapter we shift our attention towards readability prediction. We combine the data collected in the previous chapters and investigate how well linguistic features explain comprehension differences between different texts.

In this chapter we focus on the comprehension aspect of readability rather than the processing aspect of readability since this has been the main focus of most readability research. By focusing on comprehension we can compare our findings with those of (traditional) studies. Also, our processing data is less extensive than our comprehension data, both in number of texts as in number of participants. The larger numbers improve the generalizability of our findings. Still, we do not want to completely disregard the processing aspect of readability since so many studies have already ignored it. We conduct a small exploratory study in which we compare how well predictors of comprehension predict processing ease. This exploration is meant to guide further investigations and to illustrate the need for such research into the processing aspect of readability. For the most part however, this chapter is focused on finding the linguistic features that best predict the comprehension scores of Dutch adolescents.

1. Readability in the twenty-first century

Readability instruments have been around since the 1920s and have been popular among the general public for many years. Among scholars, the popularity of these instruments quickly dropped once extensive research in the seventies and eighties showed that these instruments were invalid and did not reflect true text difficulty (Anderson & Davison, 1988; Bruce, Rubin & Starr, 1981; Davison & Green, 1988; Davison & Kantor, 1982; Duffy & Kabance, 1982; Kintsch & Vipond, 1979; Redish & Selzer, 1985; among others). Most critiques focus on the fact that these readability instruments employed shallow predictors which might correlate with text

difficulty but are not causally related. Some scholars seemed to lose all hope, but others believed that with a more extensive theoretical foundation and with discourse inspired indices, we could do better (Kintsch & Vipond, 1979; Noordman & Vonk, 1994).

In the last two decades, scholars are looking at readability with renewed interest (Benjamin, 2012; Collins-Thompson, 2014). Inspired by new insights from experimental research and empowered by advances in computational linguistics, a new era for readability has begun. While in the twentieth century readability instruments were rather simple (A * word length + B * sentence length = text difficulty; e.g., Flesch, 1948), now readability instruments are evolving into sophisticated, automated, analytical tools. These tools can automatically analyze text and they return a wide range of indices that represent linguistic and/or discourse features. Some tools also provide an actual readability assessment in the form of a grade level or score. These tools provide an important service in a world where written communication is a vital part of society. But important questions that are often overlooked by users are: how do these tools get to their assessments and what is their scientific basis?

2. Differences in research goals and designs

Studies presenting and testing new indices of text difficulty are numerous, but there are important differences between them with regard to research goals and designs.[1] While some studies are interested in the 'why-question' (i.e., why is a text difficult), others are mainly focused on finding the best model to predict text difficulty.[2] For prediction studies it is not important that predictors are theoretically valid (i.e., causally related to readability) or transparent. It is less important what the exact model looks like, as long as it works best. Studies can employ shallow linguistic features like sentence length or even predictors that are completely unrelated to text difficulty. Conversely, when the 'why-question' is important, it is vital that the predictors are theoretically and practically relevant. These different research goals influence three important design choices: 1) the linguistic features that are included, 2) the data used to calibrate these features, and 3) the statistical method(s) used to perform the calibration. We will discuss these in more detail below.

[1] It is beyond the scope of this chapter to give a complete review of these studies. We refer the interested reader to Benjamin (2012), Collins-Thompson (2014) and proceedings of conferences like PITR (http://mcs.open.ac.uk/nlg/pitr2014/), IEEE (http://ieeexplore.ieee.org), ACL (https://www.aclweb.org/website/), and EMNLP (http://emnlp2017.net/).

[2] This issue is related to the readability prediction versus readability improvement distinction discussed in Chapter 1).

2.1 Linguistic features

In the early days of readability research, all linguistic features had to be calculated by hand. Features were therefore limited to 'things that are easy to count', which often resulted in shallow predictors like word and sentence length. To date, computer programs can extract more sophisticated and meaningful linguistic features directly from texts. Of course the computer does not get it right all the time (e.g., Pander Maat et al., 2014), but the overall performance of these systems is remarkable, especially given the high level of ambiguity intrinsic to language use (Bailin & Grafstein, 2016; Jurafsky & Martin, 2009).

There are two types of linguistic features: traditional-style features and features that are the result of language modeling techniques. Traditional-style features are transparent proxies of linguistic features, like PoS-distributions and grammatical structures. Language model features are the result of statistical language modeling, like n-gram probabilities (Collins-Thompson, 2014). These features are trained on example data, after which the language model is applied to new materials. These features are harder to interpret, which is why researchers that are interested in the 'why'-question generally opt for traditional-style features. However, they work really well for prediction purposes (Collins-Thompson, 2014).

Features can also be divided into feature classes, like lexical, syntactic and discourse features. All studies include features from the lexical and syntactic classes, which traditionally have always provided the strongest predictors and in many cases they still do. In addition, with the creation of Coh-Metrix (Graesser, McNamara, Louwerse & Cai, 2004) and a rise in discourse annotated corpora like the Penn Discourse Treebank (Prasad et al., 2008), more and more discourse features are included in readability research (Collins-Thompson, 2014; Crossley, Skalicky, Dascalu, McNamara & Kyle, 2017; Feng, Elhadad & Huenerfauth, 2009; Feng Jansche, Huenerfauth & Elhadad, 2010; Pitler & Nenkova, 2008).

Although many studies use indices that are very similar, the exact computation is often not the same. For instance, a simple feature like word length can be calculated in letters, syllables or morphemes. Some analytical tools provide different indices and let the user choose, others provide just a narrow selection. How features are scaled or computed affects their interpretation. Preferably, computations are based on theoretical considerations. For example, as readers learn to read they progress from focusing on individual letters to combination of letters (e.g., Ehri, 1991). For beginning readers, it thus makes sense to use a letter based feature, but for more advanced readers this feature loses validity. This raises the more general issue that features must be appropriate for the target population. Using a word frequency metric that is based on a frequency list from the seventies may be a good strategy when testing 60 year olds, but not for younger or older readers. It is thus important to look critically at how features are exactly computed.

2.2 Calibration data

On their own, the values of extracted features do not say much; they are purely descriptive. These values need to be 'mapped' onto text difficulty. For this purpose, scholars need a set of calibration data: texts that are ranked by text difficulty.

Studies employ different kinds of calibration data. For one, their data differ with regard to the type of texts that are included. Some studies are limited to one text genre (e.g., news articles, Pitler & Nenkova, 2008; Crossley et al., 2017; health content; Zeng-Treitler et al., 2012; classroom magazines, Feng et al., 2010) others include texts from multiple genres (McNamara, Graesser & Louwerse, 2012; De Clercq et al., 2014; De Clercq & Hoste, 2016; Feng et al., 2009). Although most of these studies target the English language, work is also conducted on other languages like French (François & Miltsakaki, 2012), Russian (Mikk & Elts, 1999) and Dutch (De Clercq et al., 2014). Another difference is the population they implicitly or explicitly target[3] (e.g., adults, Zeng-Treitler et al., 2012; L2-learners, François & Miltsakaki, 2012; K-12 students; McNamara et al., 2012). However, even when studies claim they target a certain population, in practice there are many that do not actually use data from real subjects to calibrate their linguistic features.

It is common practice to use expert judgments to assign a difficulty level to texts (Collins-Thompson, 2014; François, 2015). Nowadays, scholars also have easy access to graded corpora. Graded corpora contain texts that are already ranked by difficulty, for instance by U.S. grade level or L2-level. These labels are often determined by experts but sometimes these corpora themselves have been calibrated with existing readability formulae, which naturally causes methodological problems when these corpora are used to calibrate new formulae. Expert-judgments may also be problematic, because experts are not always capable to predict the difficulties that readers experience and they often disagree on specific rankings of texts (Klare, 1976b; De Jong & Lentz, 1996; Lentz & De Jong, 1997).

It is of course preferable to determine text difficulty by collecting behavioral data of the target population.[4] Naturally, this method is time-consuming and has its own methodological caveats. But if it is done correctly, it provides the optimal basis for calibration and the creation of a valid readability instrument. Mechanical cloze tests have been an all-time favorite for collecting comprehension data in readability research (Bormuth, 1969; Jansen & Boersma, 2013; Klare, 1976a; Staphorsius, 1994; Taylor, 1953). Cloze tests are easy to construct, can be applied to a wide range of texts, and because items are distributed across the entire text they sample the overall difficulty of the text in an objective way. However, according to

[3] Some studies do not state their target population and the target population must be inferred from the data they use.

[4] With behavioral data we refer to any objective measurement of comprehension or processing ease. This does not include self-reports (i.e., judgments) of the reader.

some critics cloze is not a valid measure of text comprehension (e.g., Klein-Braley & Raatz, 1984; Pearson & Hamm, 2005; Shanahan, Kamil &Webb Tobin, 1982). They claim that cloze is only sensitive to local and not to discourse level constraints. Advocates of the cloze have challenged this view and have proven that a large percentage of cloze gaps does require a discourse level representation (Brown, 1983; Jonz, 1994; Kobayashi, 2002a; Trace, Brown, Janssen, & Kozhevnikova, 2017).

A compromise between expert-judgments and behavioral data from the target population is to let readers from the target population judge the difficulty level of a text. Readers are presented with two different texts and have to judge which text they find easier to understand (Crossley et al., 2017; De Clercq et al., 2014; Pitler & Nenkova, 2008). This method has become increasingly popular with the availability of crowdsourcing technology in platforms like Amazon Mechanical Turk. These platforms enable researchers to easily collect data from large numbers of heterogeneous participants (cf. Keuleers, Stevens, Mandera, & Brysbaert, 2015). However, the question remains whether such judgments reflect true understanding or reading ease.

2.3 Statistical analysis

Once the calibration data are collected, they can be used to calibrate the linguistic features. The data are handled differently across studies, both in terms of the applied statistical methods as well as the use of aggregated versus individual-level data. It depends on the type of collected data which statistical options are available.

2.3.1 Aggregated versus individual-level data

Depending on how the calibration data is collected, original datasets may include one or multiple observations per text. For graded corpora, each text has received a 'gold-standard' difficulty label and only occurs once in the dataset.[5] For most behavioral data, however, each text has been read by more than one participant so there are multiple observations per text. In addition, each participant usually reads more than one text, perhaps even all texts depending on the number of texts in the study. This means that for each participant multiple observations are included in the dataset as well.

Traditionally, observations are aggregated over participants (Anderson & Davison, 1988). This means that each text receives an average score based on the scores of all the participants that read the text. The result is a dataset similar to a graded corpus: one 'observation' per text. Individual-level observations no longer

[5] Note that graded corpora may be based on judgments of multiple experts, in which case the original dataset that predates the corpus included multiple observations per text. Therefore, most graded corpora are actually aggregated datasets.

exist. One problem with such aggregated datasets is that all variance that was related to the text is removed. The variance is therefore systematically underestimated which leads to an increased risk of a Type I error: a possible overestimation of the significance of the predictors and a wrongful rejection of the H_0 (Quené & Van den Bergh, 2004; 2008).[6] In addition, aggregation over participants eliminates the possibility to investigate reader-text interactions, and disregards the importance of the reader (see also Chapter 1).

2.3.2 Statistical methods

There are roughly three ways to map the linguistic features onto the calibration data. One way is to use multiple regression. Linguistic features are entered into a regression model to see whether they improve the prediction of the model or not. It is a fairly simple, transparent method of mapping linguistic features to an outcome variable (e.g., comprehension score, label). Regression models can be build using unilevel or multilevel modeling. Unilevel regression models assume that all observations are independent. Multilevel models assume that the data is hierarchically structured and that some observations are not independent. For instance, when participants read multiple texts, these observations are related. They are not independent and the observations often correlate. Multilevel models take into account the data structure and model the shared variance of such observations. Unilevel models do not take into account these sources of variance. This can lead to an overestimation of statistical significance. Multilevel modeling is therefore preferred if the data is hierarchically structured. However, unilevel regression was and still is a very popular method in readability research (e.g., Bormuth, 1969; Crossley, Dufty, McCarty & McNamara, 2007; Crossley et al., 2017; Flesch, 1948; Staphorsius, 1994), partly because of its ease and transparency. For aggregated datasets multilevel modeling is generally not even an option since the variance normally modeled in the hierarchical structure no longer exists.[7]

Regression modeling also has its drawbacks. One issue is multicollinearity. If predictors correlate, regression models have trouble producing reliable regression coefficients for the individual predictors. This is a considerable problem in readability research, because many linguistic features correlate highly. A practical solution is to investigate the inter-correlations of predictors before adding them to the regression model and if predictors correlate, to select one over the other.

---

[6] Observations for one text naturally vary. Linguistic features cannot explain this variance because their values are linked to the text. When the variance is removed by aggregation, the effect of a linguistic feature no longer needs to 'handle' the variance of the observations and therefore effects can be inflated.

[7] Of course, if the data consists of more than two levels it is still possible to aggregate the lowest observation level and model the second, third etc. using a hierarchical structure.

Another way to deal with the high correlations between linguistic features is to use Principal Component Analysis or PCA (e.g., Graesser & McNamara, 2011; Staphorsius, 1994; Van Oosten, Tanghe & Hoste, 2010). PCA is "*a way of identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences.*" (Smith, 2002, p. 12). PCA reduces a large number of features to functional dimensions by combining features that are similar into one component. However, interpreting these components can be complicated depending on the combination of predictors that load onto each component.

The final way to map features onto text difficulty is by using state-of-the-art computational algorithms, usually referred to as machine learning. Most machine learning techniques are black-boxes. It is a data-driven approach, focused on predicting the best output (e.g., Collins-Thompson & Callan, 2005; Schwarm & Ostendorf, 2005). What goes on in the black-box is irrelevant for many applications of machine learning. Texts are put in, something happens and an output is given. Depending on the specific technique, it may be less observable or completely unobservable what determines text difficulty. "*...unlike traditional methods, advanced machine learning frameworks use dozens or even thousands of features and can express sophisticated 'decision spaces' that are better at capturing the complex interactions between many variables that may characterize document difficulty for different reading levels and readers."* (Collins-Thompson, 2014, p. 102). The benefit of using such models is that they are more accurate and reliable than other models, as long as the input is similar to the data that was used to train them, and as long as the training set was large enough. The drawback is that they generally do not answer the 'why'-question.

2.4 Present study

We have seen that there are big differences between readability studies and that these differences are influenced by research goals. In the present study, we have two objectives: 1. to find valid and transparent linguistic features that best predict the difficulty of Dutch texts; 2. to examine to what extent the use of aggregated datasets inflates the predictive power of linguistic features.

We focus on Dutch adolescent readers since there is currently no readability assessment tool available that is designed or validated for this group. Old and new linguistic features are put to the test to find out which combination of features best predicts how difficult a text is for these adolescents. We base our predictive model ('the Utrecht Readability model' or in short 'U-Read') on real comprehension data of more than 2300 readers. To make sure our sample includes readers varying in reading proficiency and intelligence, we include readers enrolled in different education levels and grades. By comparing unilevel and multilevel regression models we show how important it is to use individual level data. To test

the performance of the U-Read model, we compare it to predictions of two traditional readability formulae (*Flesch-Douma*; Douma, 1960; *CLIB*; Staphorsius, 1994).

In the final results section, we step away from comprehension and focus on another important aspect of readability: processing ease. Linguistic features may benefit comprehension but can affect processing ease quite differently. We test our predictors against collected reading time data to see whether our predictors are also indicative of the processing ease aspect of readability.

3. Method

3.1 Feature extraction

The linguistic features were extracted with T-Scan (Kraf & Pander Maat, 2009; Pander Maat et al., 2014). T-Scan is a text complexity tool for Dutch. At present, T-Scan computes more than 400 features at text, paragraph, sentence and word level that are theoretically and practically relevant (Pander Maat et al., 2014). It combines various tools and resources developed by the Dutch computational linguistics community, such as *Alpino* (for dependency parsing; Bouma, Van Noord, & Malouf, 2001), *Frog* (for tokenization, lemmatization, PoS-tagging, named entity recognition; Van den Bosch, Busser, Canisius & Daelemans, 2007), and frequency lists (*SUBTLEX-NL*; Keuleers, Brysbaert & New, 2010; *SoNaR*; Oostdijk, Reynaert, Hoste & Van den Heuvel, 2013). Table 1 gives an overview of the feature classes currently included in T-Scan and some examples of each class.

Within a feature class many indices are closely related. That is: they largely measure the same construct but with small differences in the computation. These differences are prompted by theoretical considerations and experimental findings. For instance, T-Scan computes word frequencies with and without names because there is evidence that names are processed differently than other content words (Camblin et al., 2007; Gordon, Groz & Gilliom, 1993). While names are not very frequent, they are generally not difficult for readers to understand. In addition, T-Scan provides related indices computed at different scales (occurrences per 1000 words; occurrences per sentence; occurrences per clause; proportions).

We include all T-Scan indices in the analysis, with the exception of the language model features (i.e., probability measures). Although language model features are interesting, we prefer insightful features that can easily be understood interpret with regards to their effect on readability.

Table 1: T-Scan overview (adapted from Pander Maat et al., 2014, p.55)

Feature class	Examples
Lexical complexity	Word and lemma frequencies for two corpora Frequency rank class membership Word prevalence Nominalizations
Sentence complexity	(Subordinate) clauses per sentence Passives Negations Syntactic dependency lengths (e.g., subject-verb, object-verb) NP modifiers (number; kind)
Referential cohesion and lexical diversity	Type-token-ratio Measure of Lexical Diversity in Text Repeated arguments from sentence n-1 Repeated arguments from the last X words Anaphoric pronouns
Coherence	Connectives Situation model dimensions (e.g., spatial, temporal and causal words)
Concreteness	Semantic type for nouns / adjectives / verbs Universal nouns Geographical, organization and product names
Personal style	Personal pronouns Personal nouns and person names
Verbs and time	Tense Aspect Action / process / state verbs
Part-of-speech	Densities for 10 PoS-tags
Probability features	Forward trigram probability Backward trigram probability Perplexities

3.2 Comprehension data

Our readability models are calibrated using comprehension data of the target population. Below we describe how the data was collected.

3.2.1 Participants

Forty-one Dutch secondary schools participated with in total 2339 students from grade 8, 9 and 10 (age: 13 - 16). Testing was introduced as part of the regular school curriculum. The students were enrolled in different levels of secondary education: in pre-university education (‘vwo’), general secondary education (‘havo’) or pre-vocational education (‘vmbo-gt’, ‘vmbo-kb’ or ‘vmbo-bb’). The distribution of

participants over Grades and Level of education is given in Table 2. Grade 10 students of pre-vocational education were not included in the study because pre-vocational students are graduating in grade 10.

Table 2: Distribution of participants over Grade and Level of education

	Pre-voc. (low)	**Pre-voc. (medium)**	**Pre-voc. (high)**	**General education**	**Pre-uni.**	**Total**
Grade 8	74	239	447	139	169	1068
Grade 9	154	246	196	156	211	963
Grade 10	-	-	-	193	115	308
Total	228	485	643	488	495	2339

3.2.2 Materials

Sixty texts were quasi-randomly selected from a set of 120 Dutch public information texts and 146 Dutch educational textbook texts (see Chapter 1). All texts were 300 to 420 words long and did not contain figures or tables. Thirty texts came from educational textbooks on history, geography, Dutch language and economics. These texts were written especially for students in secondary education. The other thirty texts were public information texts which discussed matters related to health (e.g., diabetes, donor registration), the environment (e.g., pest control), public safety (e.g., fire safety), and other socially important matters. These texts were written for the general public, but were also relevant for the participant population. Each text was manipulated on one stylistic feature: lexical complexity (see Chapter 3), syntactic complexity (see Chapter 4) or coherence marking (see Chapter 5). As a result, the sample included a total of 120 texts, consisting of 60 text pairs. All texts were automatically analyzed using T-Scan.

3.2.3 Measures

Comprehension assessment. The texts were transformed into cloze tests following the Hybrid Text Comprehension cloze procedure (HyTeC-cloze; see Chapter 2). This hybrid cloze procedure combines the strengths of mechanical and rational cloze tests into a valid and reliable measure of text comprehension. The rational strategy is used to exclude words that do not rely on text level comprehension from becoming cloze gaps. This includes words that can be reconstructed using only grammatical knowledge or knowledge of usage conventions (e.g., articles and multi-word expressions). Also excluded are words that can only be guessed, such as names and numbers. All other words in the text are candidates for deletion, except for the words that were altered as part of the manipulations. The remaining candidates are divided over different cloze versions via mechanical selection. Two cloze versions were randomly selected to serve in the study. In total 10% of the words were deleted. Depending on the text length, the cloze tests contained 30 to 42 cloze gaps.

Reading ability. Standardized reading ability scores were made available for all students. Two different tests were used to measure Reading ability. Although scores of both tests were mapped to the same scale, analyses showed that scores for one of the tests were consistently higher. To control for this, the factor Reading test was included in the analyses.

3.2.4 Design

The data was collected using a matrix sampling design (e.g., Gonzalez & Rutkowski, 2010). Each participant was given four different cloze tests: two of educational texts and two of public information texts. To balance out possible order effects, each combination of cloze tests was presented in two orders. Participants were semi-randomly assigned to a test package; they were divided over packages based on their education level and grade.

3.2.5 Procedure

All testing took place at the participating schools. The tests were administered by the school teachers in classroom settings. Cloze tests were presented digitally on computers. Participants filled in the cloze gaps on the screen. To fill in all four cloze tests, participants took part in two sessions of 45 minutes. Schools scheduled all sessions themselves over the course of a couple weeks.

3.2.6 Scoring procedure and data-clean up

The answers to each cloze gap were dichotomously scored (1 = correct; 0 = incorrect) according to the acceptable word scoring procedure (see Chapter 2). Following the acceptable word scoring procedure, not just originally deleted words were scored as correct but semantically correct alternatives were given the same score (including spelling errors and typos). The acceptability of alternative answers was judged by the global appropriateness criterion which means that the answer had to fulfill "*all the contextual requirements of the entire discourse context in which it appears*" (Oller & Jonz, 1994, p.416). Each answer was scored by two independent judges from a pool of 16 judges. When the judges disagreed, a third judge made the final decision. All judges received a short training to familiarize them with the scoring procedure.

Cloze scores were summed for each participant-text combination. The resulting summed cloze score was normalized to correct for differences in number of cloze gaps between texts. This number ranged from 30 to 42. The normalized scores represent a score on a 30 gap test. Internal reliability of the cloze tests was high (Cronbach's α between .707 and .899) and the summed scores were normally distributed (see Chapter 2).

10% of the data was removed because students repeatedly gave non-serious answers or did not answer the cloze gaps at all. Cases where students occasionally failed to fill in a gap were regarded as incorrect answers rather than missing answers and were included in the sum-scores. The final dataset contained scores for 8640 cloze tests. The dataset with cloze scores was augmented with the values of the linguistic features extracted by T-Scan. These values were centered on their grand mean before they were added to the dataset.

3.3 Processing data

Processing data was collected for a sample of the texts from the cloze study (see Chapters 3 and 4). These data were used for an exploratory study.

3.3.1 Participants

In total 181 Dutch ninth-grade students participated in the study (119 female; 62 male). 80 participants were enrolled in pre-university education, 54 in general education and 47 in pre-vocational medium education. Eye movement data of 20 participants was removed from the analyses because either the calibration procedure failed or because the registration proved to be unstable. All remaining participants had normal or corrected to normal vision. Standardized reading ability scores were available for all but one student.

3.3.2 Materials

Eight texts were selected from the public information texts used in the cloze study (easy and difficult text versions). Stimuli were presented with black letters on a white computer screen. Because the texts were too long to present on one screen, they were split up into three to five screens. To avoid unnatural text breaks, breaks either coincided with paragraph endings or other natural break points.

3.3.3 Design

The texts were divided over two lists, as a Latin-Square. As a result, participants read every text but only in one condition. Half of the participants read the texts in the reversed order.

3.3.4 Apparatus

The eye movements of the participants were recorded with a desktop eye-tracker: the SR Research EyeLink 1000. The eye-tracker recorded the position of the right pupil via a Logitech QuickCam Pro 5000 webcam at a rate of 500Hz. A remote setup with target sticker was used, allowing participants to move their head slightly. Accuracy of this eye-tracker is 0.5 degrees. Stimuli were presented on a 17 inch computer screen (1280x1024px).

3.3.5 Procedure

Participants took part in two sessions of approximately half an hour each spread over two days. Recording took place in a private room at the participating schools. Each session started with an oral instruction during which the equipment and procedure were explained. The participants were instructed to read each text at their own pace, but to make sure that they could answer comprehension questions at the end of the text.

The instruction was followed by a 13-point calibration and validation procedure. Participants fixated on a sequence of dots which appeared on various locations on the computer screen. After a successful calibration and validation sequence the testing started with a practice text and three practice questions to familiarize the participant with the procedure.

Each text fragment started with a single dot on the screen which indicated the location of the first word of the fragment. When the participant fixated on the dot, the dot vanished and the fragment appeared. To progress to the next text fragment participants pressed the 'next' button on the button-box. After the last text fragment, participants answered eight multiple-choice questions before continuing to the next text. There was no time limit.

3.3.6 Data clean-up and calculation of reading times

13 cases (0.90%) were removed because the student did not read the entire text, either intentionally or accidentally (i.e., student pressed the button too early). For the remaining 1435 cases we calculated the total time a participant spent on a text. Because texts were divided over multiple screens, reading times for each screen were added up to find the total time spent on the text.[8] We adjusted these times for text length because text varied from 300 to 420 words. Reading times are therefore presented as milliseconds per character.

3.4 Analyses

3.4.1 Comprehension data

Two analyses were performed on the cloze data to show the different results of traditional and modern regression analysis. In the first analysis, aggregated data was modelled using unilevel regression. In the second analysis the individual level, unaggregated data was modelled using multilevel regression.

[8] Note that this measure necessarily includes noise related to the participant having to press a button and moving the eyes back to the beginning of the page for each new text fragment. However, since we only use these data in an exploratory analysis, this provisional measure was deemed good enough.

The features were selected and calibrated using the following procedure. First, we investigated the correlations between the linguistic features extracted by T-Scan and the cloze scores. Due to the reduced variance in the aggregated dataset, correlations were much higher for the aggregated data compared to the unaggregated data. For the aggregated dataset, we selected features that correlated .40 or higher with the scores. For the unaggregated dataset, we lowered the threshold to .20. Within groups of related indices (see Section 3.1), we only selected the feature that correlated the highest with the cloze scores. The remaining indices were checked for multicollinearity by examining their inter-correlations. If features correlated above .70, once again the feature that correlated the highest with the cloze scores was selected. Features that survived the selection procedure were added one-by-one to the unilevel or multilevel regression model.

For the unilevel analysis, features were entered into the model using a stepwise procedure. The performance of the model was checked by conducting a 5-fold-cross-validation. In a 5-fold-cross-validation the data is randomly divided into five partitions. The model is trained on four partitions and its performance is tested on the fifth. This procedure is repeated 4 more times, each time with a different partition left out of the training set and used as the test set. If the model is accurate, it will have no problem predicting the text cloze scores in the test sets.

For the multilevel analysis, we added the predictors to a base model which included the hierarchical structure of the data as well as known reader characteristics. Observations were crossed between students and texts, and with students nested in schools. The reader characteristics Education level, Grade, Reading ability and Reading test were included as fixed predictors (see Table 3 for descriptions). Linguistic features were added to the model following a stepwise procedure. In order to keep the model transparent, we implemented an R^2-change threshold: we do not accept predictors that do not result in an increase of predictive power below .001. The resulting model is in this sense not statistically optimal but restrained (i.e., highly parsimonious). We will refer to this model as Utrecht Readability model (U-Read). For the interested reader, a statistically optimal multilevel model is presented in Appendix 11. The robustness of the estimates was checked with a bootstrap procedure.

Table 3: Descriptions of factors in base model text comprehension

Factors	**Description**	**Levels**
Education level	Level of education in which the student is enrolled	5 levels
Grade	Grade in which the student is enrolled	3 levels
Reading ability	Reading ability score on Dutch reading ability test (centered and standardized)	Continuous
Reading test	Reading test used to test reading ability	2 levels

3.4.2 Processing data

Reading times were analyzed using multilevel regression to see whether the predictors of the comprehension data were also indicative of processing difficulty. The procedure was similar to the multilevel analysis for the cloze data, with the exception that only predictors from the U-Read model were selected. These predictors were added one-by-one to the multilevel base model which included the hierarchical structure of the data as well as known reader characteristics (Education level and Reading ability). Observations were crossed between students and texts.

4. Results

4.1 Traditional unilevel regression

From the extracted T-Scan features, features were selected that correlated .40 or more with the individual cloze scores and which did not have inter-correlations above .70. This resulted in a list of 16 predictors. Only five predictors proved to be significant in the stepwise unilevel regression:

1. Word frequency[9]
2. Content words per clause
3. Adjectival past participles
4. Proportion of concrete nouns
5. Maximum syntactic dependency length (SDL)

The model is presented in Table 4 and a description of the features can be found in Appendix 10.

[9] Content words only, no names and corrected for compound nouns (see Appendix 10).

Table 4: Unilevel model text comprehension

Predictor	B	SE	β	t	p	Multiple R^2	R^2-change
Intercept	16.623	0.156					
Word frequency	5.260	0.727	0.425	7.236	<.001	.575	.575
Content words per clause	-1.241	0.268	-0.295	-4.633	<.001	.693	.118
Adjectival past participles	-16.345	6.208	-0.139	-2.633	.010	.733	.039
Concrete nouns	3.114	1.054	0.158	2.954	.004	.752	.020
Max SDL	-0.241	0.104	-0.130	-2.323	.022	.764	.011

Word frequency and the proportion of concrete nouns had a positive relationship with the cloze scores. Texts with highly frequent words or concrete nouns had higher cloze scores than texts with less frequent words or more abstract nouns. The other predictors had a negative relationship with the cloze scores. When the number of content words per clause, the number of adjectival pas participles per clause, or the maximum syntactic dependency length was high, cloze scores were lower.

A 5-fold-cross-validation showed that the model performed well on the test sets. R-squared for the test sets ranged from .724 to .783 with a mean value of .753 (see Table 5). These values are almost identical to the values for the training set, as well as the model trained on the entire dataset. For the entire dataset, observed R-squared was .764, meaning that three-quarters of the observed variance between cloze scores is accounted for by the five predictors.

Table 5: Results cross-validation

Partition	R (R^2) training set	R (R^2) test set
1	.886 (.785)	.851 (.724)
2	.874 (.764)	.877 (.769)
3	.880 (.774)	.852 (.726)
4	.871 (.759)	.873 (.762)
5	.871 (.759)	.885 (.783)
Mean	.876 (.767)	.868 (.753)

4.2 Multilevel regression

4.2.1 Base model text comprehension

A multilevel base model was built which included the hierarchically structured random effects and fixed effects for known reader characteristics (Education level, Grade and Reading ability). Main effects were found for Education level, Grade, Reading ability and Reading test (see Table 6). As expected, performance increased

with rising education levels, grades[10] and reading ability. The prediction of the fixed factors correlated .583 (R^2 = .340) with the observed individual cloze scores. Reader characteristics thus accounted for 34% of the observed variance within cloze scores. After adding the reader characteristics, the random variance at student level was reduced by 51%.

Table 6: Multilevel base model text comprehension

Random effects	**Estimates**	**SE**	***t***	***p***
School	0.864	0.327	2.642	.004[a]
School: Student	5.214	0.323	16.142	<.001[a]
Text	0.638	0.152	4.197	<.001[a]
Residual	18.844	0.335		
Fixed effects	**Estimates**	**SE**	***t***	***p***
Intercept	13.603	0.494	27.536	<.001
Education level: pre-voc. low	0[b]			
Education level: pre-voc. medium	1.625	0.292	5.565	<.001
Education level: pre-voc. high	3.400	0.314	10.828	<.001
Education level: general	5.277	0.391	13.496	<.001
Education level: pre-uni.	6.095	0.406	15.012	<.001
Grade 8	0[b]			
Grade 9	0.685	0.170	4.029	<.001
Grade 10[c]	0.555	0.354	1.568	.117
Reading ability	1.988	0.111	17.910	<.001
Reading test: R	0[b]			
Reading test: V	-3.053	0.452	-6.754	<.001

[a] One-sided. [b] Set as reference level. [c] Unbalanced, no 10th grade pre-vocational students were present in the sample.

4.2.2 U-Read multilevel model

From the 400+ features extracted by T-Scan, predictors were selected that correlated .20 or more with the individual cloze scores and which did not have inter-correlations above .70. This resulted in a list of 15 predictors. Predictors were added to the base model using a stepwise procedure. The final model (i.e., the Utrecht Readability model) is presented in Table 7.

Five predictors were statistically significant and resulted in an R^2-change above our .001 threshold: *Word frequency*, *Content words per clause*, *Concrete nouns*, *Maximum syntactic dependency length* and *Adjectival past participles* (see Table 8).[11] The U-Read model explained more than 57% of the observed variance in individual cloze scores (R^2 = .572). Given that the reader characteristics explained

[10] No effect was found for Grade 10. Once linguistic features were included, however, the effect was significant in the expected direction.

[11] Correlations between these linguistic features and the individual cloze scores can be found in Appendix 11 (Table A11.2).

34%, this means that the five linguistic features accounted for roughly 23% of the variance. Word frequency and the proportion of concrete nouns had a positive relation with the cloze scores. As text included more frequent words and a high proportion of concrete nouns, cloze scores increased. The number of content words per clause, the maximum syntactic dependency length and the number of adjectival past participles all had a negative relation with the cloze scores, suggesting a negative effect on comprehension when these features increased.

Additional predictors did not reach the threshold. We also tested the model for interactions between the top linguistic predictors and reader characteristics. Only word frequency showed a significant interaction. Pre-vocational medium and pre-vocational high students were more strongly affected by word frequency than pre-vocational low, general and pre-university students. However, the interaction did not improve the explanatory power of our model beyond the threshold of .001 in R^2-change and for that reason did not make the final model.

Table 7: U-Read multilevel model

Random effects	**Estimates**	**SE**	***t***	***p***
School	1.004	0.308	3.260	<.001[a]
School: Student	4.423	0.234	18.902	<.001[a]
Text	0.123	0.046	2.674	.004[a]
Residual	11.330	0.201		
Fixed effects	**Estimates**	**SE**	***t***	***p***
Intercept	13.544	0.454	29.833	<.001
Education level: pre-voc. low	0[b]			
Education level: pre-voc. medium	1.464	0.250	5.856	<.001
Education level: pre-voc. high	3.205	0.269	11.914	<.001
Education level: general	5.118	0.339	15.097	<.001
Education level: pre-uni.	6.053	0.352	17.196	<.001
Grade 8	0[b]			
Grade 9	0.662	0.146	4.534	<.001
Grade 10[c]	0.897	0.304	2.951	.003
Reading ability	1.935	0.095	20.368	<.001
Reading test: R	0[b]			
Reading test: V	-2.740	0.430	-6.372	<.001
Word frequency	5.073	0.182	27.874	<.001
Content words per clause	-1.172	0.069	-16.986	<.001
Concrete nouns	3.670	0.268	13.694	<.001
Max SDL	-0.278	0.026	-10.692	<.001
Adjectival past participles	-11.361	1.574	-7.218	<.001

[a] One-sided. [b] Set as reference level. [c] Unbalanced, no 10th grade pre-vocational students were present in the sample.

Table 8: Model comparisons

Model	Features	R	Multiple R^2	R^2-change
Base model:	Education level + Grade + Reading ability + Reading test	.583	.340	.340
1:	+ Word frequency	.716	.513	.173
2:	+ Content words per clause	.742	.550	.037
3:	+ Concrete nouns	.750	.562	.012
4:	+ Max SDL	.754	.568	.006
5:	+ Adjectival past participles	.756	.572	.003

4.3 Performance compared to traditional readability formulae

The performance of our multilevel model was compared to the performance of two classical Dutch readability formulae: the Flesch-Douma (1) and the CLIB (2). Although these two formulae were not especially developed for adolescent readers, they are the most popular formulae for Dutch. In addition, CLIB was the result of extensive empirical research and offers the best available comparison of readability modelling based on real Dutch comprehension data.

(1) Flesch-Douma = 206.835 - 0.77 * word length [syllables/words] - 0.93 * sentence length [words/sentences]

(Douma, 1960)

(2) CLIB = 46 - 6.603 * word length [letters/words] + 0.474 * percentage highly frequent words [Freq77] - 0.365 * Type/Token ratio + 1.425 * inverse sentence length [sentences/words]

(Staphorsius, 1994)

The predictors used in the Flesch-Douma and CLIB formulae were added to the base model presented in Section 4.2.1.[12] All predictors of these formulae were available within T-Scan, except for word length in syllables per word (which is used in the Flesch-Douma). Therefore, we included two alternative indices that were available in T-Scan: letters per word and morphemes per word. Letters per word together with sentence length resulted in the best alternative Flesch-Douma model. This model explained 53% of the observed variance (see Table 9). Of course, 34% of the variance was already explained by the reader characteristics as can be seen from the base model. This means that the fixed linguistic predictors of the Flesch-Douma

[12] We only added the predictors. We did not use the coefficients as they are given in the formulae.

formula explained 19% of the observed variance. Predictors used in the CLIB formula did not outperform the Flesch-Douma's and also accounted for 19% of the variance. In comparison, our own model explained 23% of the observed variance. Even when we limited the number to our respective best two or four predictors – thereby equaling the number of predictors of the Flesch-Douma and CLIB – the predictive power of our model was higher than that of the older models. Our predictors clearly outperform the old predictors used in the Flesch-Douma and CLIB formulae.

Table 9: Model comparisons

Model	**Features**	**R**	**Multiple R^2**	**R^2-change***
Base model	Education level + Grade + Reading ability + Reading test	.583	.340	
Flesch-Douma	+ Word length + Sentence length	.727	.529	.189
CLIB	+ Word length + Inverse sentence length + Percentage of frequent words + TTR	.728	.530	.190
U-Read	+ Word frequency + Content words per clause + Concrete nouns + Max SDL + Adj. past participles	.756	.572	.232
U-Read Top 2 predictors	+ Word frequency + Content words per clause	.742	.550	.210
U-Read Top 4 predictors	+ Word frequency + Content words per clause + Concrete nouns + Max SDL	.754	.568	.228

**Compared to the base model*

4.4 Exploratory analysis of reading times

An exploratory analysis was conducted to see whether the best predictors of the comprehension data would also be indicative of processing difficulty. A multilevel analysis was run with reading time in milliseconds per character as a dependent variable and with participants and texts as crossed random intercepts. Education level and Reading ability were included as known reader characteristics. A log-transformation was carried out to normalize the distributions.

The five predictors of the U-Read model (i.e., Word frequency, Content words per clause, Concrete nouns, Max SDL, Adj. past participles) were added separately to the base model and ranked according to strength. Only Word frequency had a significant effect on reading times (see Table 10). A forced entry procedure, in which all predictors are forced into the model at the same time, revealed the same pattern. Reading times decreased as Word frequency increased. No interactions between Word frequency and the reader characteristics reached significance.

Table 10: Final model reading times

Random effects	**Estimates**	**SE**		
Student	0.0084	0.0914		
Text	0.0002	0.0125		
Residual	0.0026	0.0514		
Fixed effects	**Estimates**	**SE**	***t***	***p***
Intercept	1.723	0.014	121.197	<.001
Education level: pre-voc. medium	0[b]			
Education level: general	-0.039	0.019	-2.120	.035
Education level: pre-uni.	-0.047	0.018	-2.640	.009
Reading ability	-0.029	0.007	-3.933	<.001
Word frequency	-0.040	0.012	-3.473	.004

[a] *One-sided.* [b] *Set as reference level.*

We compared the correlations of the five comprehension predictors and the reading times to correlations of other T-Scan features with the reading times. This exploration showed that our five predictors did not correlate highly with reading times at all, and that other features showed stronger correlations with the reading times. For Word frequency and Concrete nouns other related measures outperformed the ones we included. Interestingly, indices of lexical diversity (e.g., TTR and MLTD) were highly represented among the highest correlating features. These indices were not represented in the comprehension study because they did not reach the .20 correlation threshold.

5. Discussion

In this study we investigated the predictive power of linguistic features on comprehension data of Dutch adolescents. 60 pairs of texts were transformed into cloze tests using the HyTeC-cloze procedure and presented to more than 2300 adolescents enrolled in Dutch secondary education. We analyzed these data using traditional unilevel regression and modern multilevel regression techniques which revealed the 'Utrecht Readability model (U-Read)'. U-Read is a restrained multilevel model that uses five linguistic features: *Word frequency*, *Content words per clause*, *Concrete nouns*, *Maximum syntactic dependency length* and *Adjectival past participles*. These features are available in the Dutch automatic analytical tool T-Scan. We compared the U-Read predictors against those of two known Dutch readability formulae: Flesch-Douma and CLIB. In terms of explained variance our U-Read model improved both models by approximately 20%.

5.1 U-Read predictors

Of the five predictors in the U-Read model, Word frequency was the strongest. Texts with high word frequency were easier to comprehend than texts with low word frequency. This finding is in line with results of experimental studies, including our own lexical experiment presented in Chapter 3 (Freebody & Anderson, 1983ab; Radach, Huestegge & Reilly, 2008; Stahl, Jacobson, Davis & Davis, 1989; Williams & Morris, 2004). Highly frequent words are easier to understand and to process than low frequent words. Interestingly, our best frequency measure did not include names and the frequency of compound nouns was corrected by taking the frequency of the head noun. It was calculated using the subtitle corpus SUBTLEX-NL (Keuleers et al., 2010). Frequencies based on subtitles are thought to be more representative of everyday language use than frequencies based on edited texts (Brysbaert & New, 2009). Our corrected word frequency measure outperformed related indices, including word frequencies from other corpora, frequency rankings and word prevalence measures.

In addition to word frequency, concreteness also contributed positively to comprehension. Texts with a high proportion of nouns that referred to persons, organisms, artifacts, places, times and/or units of measurements received higher cloze scores than texts with a low proportion of such concrete nouns. Similar results were found previously by Sadoski and colleagues with regard to recall (Sadoski, Goetz & Fritz, 1993; Sadoski, Goetz & Rodriguez, 2000). According to Sadoski: "*...concrete language promotes referential processing, the evocation of mental images related to that language. The consequent dual encoding of language in verbal and imaginal forms promotes elaboration, comprehension, and memory.*" (Sadoski et al., 2001, p.93).

Maximum syntactic dependency length (SDL) was negatively related to comprehension. Texts where the maximal dependencies in sentences were long were more difficult than texts with shorter maximum SDLs. Since SDL is expressed as the number of words between a head and its dependent (e.g., verb-subject), it naturally correlates with the traditional shallow predictor sentence length. However, this relationship is not fully bidirectional. In order for the SDLs to be high, the sentence must be relatively long since it requires a length that is at least the SDL number plus the words of the dependency. Yet, long sentences do not necessarily have a high SDL and sentences of the same length can have different SDLs. Our experiments in Chapter 4 showed that SDL can negatively affect comprehension and processing even when sentence length and the content of the sentence are kept equal.

The number of content words per clause was also negatively related to comprehension. When the number of content words per clause was high, texts were harder to understand. Content words per clause is closely related to the measure propositional density that has long been shown to affect comprehension (Kemper,

Jackson, Cheung & Anagnopoulos; 1993; Kintsch & Keenan, 1973; Kintsch & Vipond, 1979). One key difference, however, is that propositional densities are standardized on words while our content word measure is standardized on a grammatical unit, the clause. For our data the number of content words per clause correlated much higher with individual comprehension scores than the density of content words throughout the text (r = -.404 vs. r = -.218). Content words per clause was a stronger predictor than words per clause. Thus, it is the accumulation of larger numbers of propositions within a particular grammatical unit that affects comprehension most, not the number of propositions or content words per 1000 words.

Perhaps the least expected predictor is Adjectival past participles (APP). This predictor was also negatively related to comprehension. Texts with a high number of APPs (e.g., *the skinned potatoes*) received lower cloze scores than texts with no or a low number of adjectival past participles. APP-constructions are very dense propositions. They do not just denote a property of the following noun; they also denote a resultant state: an action or event that had to transpire in order to produce that specific referent (Ackerman & Goldberg, 1996; Parsons, 1990). This is different from other descriptive adjectives (cf. *the brown potato*, *the small potato*). An exploration of the texts with high numbers of adjectival past participles revealed that in many instances the adjectival past participle construction was part of an even more complex noun phrase (see (3)-(5)) and that these constructions were mostly used in rather formal texts.

(3) *het in 1994 opgerichte Internationaal Monetair Fonds*

'the in 1994 constituted International Monetary Fund'

(4) *de door het Centraal Bureau voor de Statistiek (CBS) verzamelde statistische gegevens*

'the by the Central Bureau for Statistics (CBS) collected statistical data'

(5) *de in de Troonrede genoemde beleidsplannen*

'the in the King's speech mentioned policies'

Although our predictors are attained from different feature classes, some classes are not represented. Most notably, U-Read does not include discourse features. None of the predictors is related to relational or referential coherence. Only one measure correlated highly enough with the individual cloze scores to be entered into the analysis in the first place: lemma overlap in the last 50 words. This feature correlated .304 with the cloze scores. However, even in the optimal multilevel model lemma overlap dropped out once other predictors were added to the model.

Interestingly, overlap measures were also the only cohesion features that survived the initial predictor selection of Crossley and colleagues (2017). In their analysis, however, the feature 'Content word overlap in adjacent paragraphs' did end up in the final model together with age of acquisition, lexical decision accuracy and average number of verbs. Our findings are more in line with those of Feng and colleagues (2009; 2010). They found that for predicting readability, discourse features are not that helpful. When they are combined with other features, discourse features lose their significance in predictive models. In contrast, Pitler and Nenkova (2008) did find that discourse features benefitted their model. However, Pitler and Nenkova used gold-standard discourse features which were based on an annotated discourse corpus. In Feng's study, as well as in our own, all features were computed automatically. Since the automatic computation of discourse features still relies on relatively shallow proxies, it makes sense that gold-standard discourse features would provide better results.[13] However, for practical applications annotating the texts before they are entered into the prediction model is not really an option.

In our analysis we did not include language model features like probability features. We left these features out in favor of transparent features. Perplexity and Entropy correlated highly with the individual cloze scores (>.30) and could potentially have been used in the analysis. However, it is unclear what these features reflect. The probabilistic features correlate highly with a number of syntactic features: backward entropy even correlated .944 with the number of words per sentence. It is unlikely that it is just sentence length that increases the probability of a sentence. Therefore, we advise some caution when using and interpreting probability features in readability research.

5.1.1 Influence of the study's design on predictor selection

We included 60 pairs of texts in this study, totaling 120 texts. Each text was manipulated on one stylistic feature while the content of the text remained the same (see 3.2.2). Therefore, the 120 texts in the sample are not completely independent. There is a possibility that this design created a biased predictor selection. That is: favoring linguistic features that were manipulated above features that were not manipulated. The lexical and syntactic manipulations were reflected in Word frequency and Maximum SDL. We investigated whether our design resulted in a biased predictor selection and affected the consequent U-Read model.

The dependencies within the sample were eliminated by splitting the data into two partitions. Each partition only included the easy or the difficult version of a text. We repeated our predictor selection procedure for the two partitions separately. The results were nearly the same as when the predictor selection was performed on

[13] For syntactic and semantic features, De Clercq and Hoste (2016) showed that gold-standard features do not improve predictions compared to automatically computed features.

the full dataset with only some minor changes that did not involve the U-Read predictors. Next, we checked the estimates and explained variance of the U-Read predictors when they were added to multilevel models for the two partitions. The dependencies between texts did not bias the results. Both the estimates and the explained variance of the models were similar for the two partitions and in line with the U-Read model (see Appendix 12).

5.2 Unilevel versus multilevel analysis

This study also investigated the differences between traditional unilevel analysis and modern multilevel analysis. Both analyses showed that *Word frequency*, *Content words per clause*, *Concrete nouns*, *Maximum syntactic dependency length* and *Adjectival past participles* were the five strongest predictors for our cloze data. Although the predictors were the same, there were important differences. First of all, the order of the predictors was not identical. Adjectival past participles entered the unilevel model as the third predictor, but entered the multilevel model as the fifth predictor. Secondly, the unilevel regression tapped out at five predictors. Adding more predictors did not statistically improve the model. The multilevel models on the other hand could accommodate much more predictors, including interactions. Although these interactions did not make the R^2-threshold implemented in the U-read model, the optimal multilevel model showed that interactions are present (see Appendix 11). Not all readers were equally affected by the linguistic features, but these interactions only help to explain a very small portion of the observed variance.

The most important difference between the unilevel and multilevel models is that the unilevel model overestimated the variance that can be explained by linguistic features. The unilevel model accounted for 76% of the observed variance of average text scores. The same predictors only explained 23% of the variance observed in individual scores. This means that whether a reader understands a given text is lot more uncertain than unilevel models indicate (see Anderson & Davison, 1988). Besides being more realistic about the predictive power of linguistic features, multilevel models offer a window on the interplay between reader and text factors, regardless of whether there are interactions between these variable groups.

Although we did not aggregate our cloze data over participants, we did aggregate our data within texts. We summed the scores of the cloze gaps in a text to create a text level score. This approach made it possible to align cloze scores with the values of all our linguistic features. However, our text level analysis ignores variance within the text, both with regard to the distribution of the cloze scores as well as values of linguistic features. The difficulty of a text can vary from one part to the next and the difference between such parts may be small or big. Similarly, a linguistic feature like word frequency could be very stable or fluctuate from high to

low values from one sentence to the next. At text level we only look at the text's average difficulty given the average value of the linguistic features. Although this might be okay for readability prediction purposes, for readability improvement purposes it would be interesting to see if we could make more localized predictions and predict where in a text problems are likely to occur. To build such a model, it is necessary to descend to a lower analytical level, for instance sentence level or even word level (see Bormuth, 1966; 1969). Although our data allows for such lower level investigations, it was beyond the scope of the current study.

5.3 Comprehension versus processing ease

We explored whether the predictors of text comprehension would also be indicative of text processing by using the predictors of the U-Read model to predict reading times. Only our word frequency measure proved to be a significant predictor. Texts with high word frequencies were read faster compared to texts with low word frequencies. No effects were found for the features Content words per clause, Concrete nouns, Maximum syntactic dependency length and Adjectival past participles, not even when they were added to the model separately.

Of course our findings can be due to our measure of reading time. Because of the differences in text length we computed reading time per character. This is a rather rough measure of processing ease compared to localized eye-tracking measures like *First pass gaze duration* or *Regression path duration*. Processing studies have previously found that at least concreteness and syntactic dependency length can affect on-line processing at the word level (e.g., Bartek, Lewis, Vasishth & Smith, 2011; Demberg & Keller, 2008; Juhasz & Rayner, 2003). However, our results indicate that they are not that indicative of processing ease at the text level. Furthermore, other linguistic features show higher correlations with our reading time measure than the U-Read model predictors. Together, these results indicate that comprehension and processing are different aspects of readability and should be studied separately and preferably also in tandem.

5.4 Conclusion

In this study we combined insights from prior readability studies and experimental work with recent developments in computational linguistics to improve readability prediction for Dutch adolescents. In contrast to most recent studies, we based our predictive model on real comprehension data of our target population. Using improved statistical methods we were able to make more realistic predictions of how much variance linguistic features can explain. The resulting U-Read model uses the features *Word frequency*, *Content words per clause*, *Concrete nouns*, *Maximum syntactic dependency length* and *Adjectival past participles* to predict text difficulty for Dutch adolescents. These five predictors outperformed the predictors used in

popular Dutch readability formulae. In contrast to traditional shallow predictors (e.g., word and sentence length) support for the causal relevance of our predictors can be found in prior theoretical and experimental research. Our findings are insightful for researchers interested in predicting readability, but also for researchers interested in answering the question why texts are difficult.

7 Conclusion

In this final chapter we summarize our main findings and reflect on four key notions that have been important for this study's design: the distinction between conceptual and stylistic difficulty, the comprehension and processing ease aspects of readability, the role of the reader, and the importance of causally relevant predictors. Subsequently, we will discuss the study's limitations and directions for further research. The chapter will end with a general conclusion.

1. Summary of results

The aim of this dissertation was to investigate the effects of linguistic features on how Dutch adolescents understand and process texts, and to collect data necessary for the construction of a new valid readability assessment tool for Dutch ('LIN'). Our first step was to design a reliable method to measure text comprehension across different texts. This method had to be easily applicable to a large number of texts and had to be sensitive to differences between texts, text versions and readers. We developed a new cloze procedure: the Hybrid Text Comprehension cloze (HyTeC-cloze) which was presented in Chapter 2. Our hybrid cloze procedure combines the strengths of mechanical and rational cloze tests into a valid and reliable measure of text comprehension. First, the rational deletion strategy excludes words that are locally predictable and do not rely on discourse level comprehension (e.g., articles, copula & multi-word expressions), as well as words that can only be guessed (e.g., numbers, first mention of names, geographical locations). All remaining words are candidates for deletion. Next, the mechanical deletion strategy guarantees that an unbiased, distributed sample of the gap candidates ends up in the final cloze test. The procedure was used to collect comprehension data for the dissertation chapters as well as data for the development of the LIN-tool.

Sixty texts were quasi-randomly selected from a compiled corpus of 146 educational texts and 120 public information texts, and subsequently assigned to one of three experimental studies: a study on lexical complexity, a study on syntactic complexity and a study on coherence marking. Goal of the experimental studies was to test the effects of stylistic features on readability in a controlled setting. That is: without confounding effects of content or other stylistic features. The second objective was to see whether effects could be generalized across a large number of

texts taken from two different genres. The results of these studies were presented in Chapters 3, 4 and 5.

In Chapter 3 ten educational texts and ten public information texts were manipulated to create a text version with high lexical complexity and a version with low lexical complexity. 20% of the content words were substituted for an intuitively less familiar or more familiar alternative. Next, all texts were transformed into cloze tests using the HyTeC-cloze procedure and administered to students enrolled in grades 8 through 10 of Dutch secondary education. Results of the cloze tests showed that lexical complexity significantly affected comprehension. Reducing the lexical complexity of the texts improved comprehension especially for students enrolled in pre-vocational medium and pre-vocational high education.

Reducing the lexical complexity also positively affected on-line processing. An eye-tracking study on four of the public information texts showed that reading times and the number of immediate refixations on manipulated words were reduced in the Low-lexical complexity version. Once again, pre-vocational medium students were most affected by the manipulation.

In Chapter 4 we investigated the influence of syntactic complexity on readability. We increased and decreased the distance between syntactical heads and their dependents to create two text versions of ten educational and ten public information texts. Four public information texts were used in an eye-tracking study and all twenty texts were used in a cloze study. The syntactic dependency length (SDL) negatively affected processing ease and comprehension. Reading times were longer for sentences with increased SDLs. Increasing the syntactic dependency length negatively affected comprehension as measured by the cloze tests. However, this effect depended on the text genre and the size of the increase. Comprehension of public information was affected even when the increase in SDL was small, but comprehension of educational texts was only affected when the increase was large. We believe that students were able to overcome small increases in the educational texts because they were easier than the public information texts. They could no longer compensate once the increase in SDL was large.

In Chapter 5 we investigated whether making coherence relations explicit enhances comprehension. While many processing studies show benefits of coherence marking, effects on comprehension are less consistent. Twenty texts were assigned to this study. Coherence marking was manipulated by adding and removing connectives in one-third of the coherence relations. The effect of coherence marking was restricted to the immediate context of the connective and depended on the type of relation. While adding a contrast or causal connective increased comprehension, adding an additive connective did not and comprehension was even negatively affected by the presence of the connective.

Finally, in Chapter 6 we focused on readability prediction. We presented the Utrecht Readability model (U-Read): a multilevel model that can predict text

difficulty for Dutch adolescent readers. Using the predictors *Word frequency*, *Content words per clause*, *Concrete nouns*, *Maximum syntactic dependency length* and *Adjectival past participles* the model explained 23% of the observed variance on top of the 34% variance explained by known reader characteristics (i.e., Education level, Grade and Reading ability). The U-Read model performed roughly 20% better than two classical Dutch readability formulae with shallow predictors. This suggests that the U-Read's semantically and syntactically informative features are to be preferred over shallow features such as word and sentence length.

2. Closing in on readability

In Chapter 1 we proposed an integrated approach to fill some of the gaps that have become apparent in 100 years of readability research. In this section we highlight these gaps and show how our research contributed to the field and helped us to close in on readability.

2.1 Stylistic and conceptual difficulty

A key notion in this dissertation was the difference between conceptual difficulty and stylistic difficulty, conceptual difficulty being the message (i.e., 'content') and stylistic difficulty being the manner in which the message is relayed (e.g., Leech & Short, 2007). Such a distinction is not often made in traditional readability research (Bailin & Grafstein, 2016; Gray & Leary, 1935; Johnson & Otto, 1982; Klare, 1976a). Readability predictions are based on differences between texts. These texts differ in content and in style. In such a design, conceptual and stylistic difficulty are confounded. The effect of a linguistic predictor may largely reflect differences in conceptual difficulty, not in stylistic difficulty. As a result, stylistic interventions will not yield the predicted effects and the effects of linguistic predictors are often overestimated. We suspect that 'between-text' effects of predictors will be larger than the 'between-text-version' effects.

In order to disentangle conceptual and stylistic difficulty, we have incorporated three controlled experiments within our overall design. In these experiments we created two text versions of every text by manipulating a stylistic feature. While texts differed in conceptual and stylistic difficulty, text versions of the same text only differed in stylistic difficulty. From the perspective of the manipulated features, this means that their values varied between texts and between text versions. This design enables us to assess their performance in predicting both text differences and text version differences. Such a direct comparison will also show how much the effects of stylistic features are overestimated when style and content are confounded.

Additional analyses were performed on the cloze data to compare the between-texts and between-text-versions effects of our manipulated features. First, three linguistic features were selected that directly reflected our manipulations. Lexical complexity was reflected in Word frequency[1], Syntactic complexity was reflected in Maximum syntactic dependency length (Max SDL) and Coherence marking was reflected in Connectives per clause. The mean difference between the easy and difficult text versions was calculated as well as the standard deviation of this difference (Table 1). We also calculated the mean difference and its standard deviation between all texts (incl. all text versions). For the texts whose lexical complexity was manipulated, difficult text versions had on average a 0.25 lower word frequency than easy text versions. Our syntactic manipulation resulted in an average difference in SDL of 1.54 words. The Coherence marking manipulation led to a mean difference between text versions of 0.21 connectives per clause. For all three linguistic features, the variance between the easy and difficult text versions was much smaller than the variance between all texts (see also Chapter 3: Figure 1, Appendix 6 and Appendix 7). This was to be expected because we manipulated the stylistic difficulty within the realm of possibilities provided by the texts without changing the content or other stylistic properties and while keeping the text natural. As a result, differences between versions are relatively small compared to differences between texts.

Table 1: Means and standard deviations of linguistic features that reflect manipulations

Manipulation	**Feature**	**Easy text version (N=20)**	**Difficult text version (N=20)**	**Mean Diff. (N=20)[a]**	**SD Diff. (N=20)[a]**	**Mean Diff. (N=40)[b]**	**SD Diff. (N=40)[b]**
Lexical complexity	Word frequency[c]	4.57	4.32	0.25	0.05	0.32	0.24
Syntactic complexity	Max SDL	5.45	6.99	1.54	0.61	2.49	2.02
Coherence marking	Connectives per clause	0.49	0.28	0.21	0.04	0.19	0.14

[a] Between text pairs. [b] Between all texts. [c] Per billion words (log-transformed).

Next, multilevel analyses were run for each manipulation separately. Each analysis started out with a base model similar to the base model presented in Chapter 6 (Section 4.2.1) with a random structure and reader characteristics (Education level, Grade and Reading ability). Two predictors were added to the base models, first separately and later combined. The first predictor was the dichotomous predictor

[1] Based on SUBTLEX-NL (Keuleers, Brysbaert & New, 2010) without names and corrected for compound nouns (see Appendix 10).

Text version, with the easy text version as reference level and the estimate showing what happens to the summed cloze score in case of a difficult text version. The second predictor was the respective linguistic feature: Word frequency, Max SDL or Connectives per clause. While the predictor Text version only explains variance between the easy and difficult text versions, the linguistic feature reflects both version and text differences. In the combined model, we can see how these predictors interact. A summary of the results is presented in Table 2.

Table 2: Explained variance and predictor estimates for the multilevel models

Manipulation	**Model**	**R^2**	**Estimate Difficult version**	**Estimate Linguistic feature**
Lexical complexity	Base model	.363		
	Difficult version	.373	-1.213	
	Word frequency	.527		8.492
	Difficult version + Word frequency	.538	1.403	9.840
Syntactic complexity	Base model	.338		
	Difficult version	.339	-0.311	
	Max SDL	.440		-0.835
	Difficult version + Max SDL	.449	1.146	-0.932
Coherence marking	Base model	.307		
	Difficult version	.307	-0.148[a]	
	Connectives per clause	.310		-1.746
	Difficult version + Connectives per clause	.314	-0.878	-3.513

[a] *Not significant.*

Overall we see that the predictor Text version only explains a very small portion of the variance on top of the linguistic features (max 1%), with the lexical manipulation explaining most, syntactic complexity a little less, and coherence marking explaining the least amount of variance. These results are not surprising since the linguistic features include differences between texts while the predictor 'Difficult version' does not. Still, the sizes of the differences are telling and show that text features used in readability models are much better in predicting differences between texts than between text versions.

When we look at the estimates for the single predictor models, we see that for all manipulations the estimate of the difficult text version is negative. This means that the difficult text versions were more difficult to comprehend than the easy text

versions.[2] The estimates of the linguistic features are also in the expected direction. Word frequency had a positive effect. Texts with frequent words were easier to comprehend than texts with less frequent words. The estimate for Maximum SDL was negative. Texts with long dependencies were more difficult to understand. Similarly, texts were more difficult when they had a high number of connectives per clause. Although this last finding may seem counterintuitive, it was expected since complex relations are more often explicitly marked than simple relations (Asr & Demberg, 2012).

Adding both the Text version and the linguistic feature predictor to the model resulted in the best model for all three manipulations. When we look at the estimates of these combined models, we see that the estimates change. For the lexical and syntactic manipulations, the estimate of the linguistic feature increases and the estimate of the Difficult version changes direction. At first glance it seems that in the combined model the difficult version is easier to comprehend than the easy text version (i.e., the estimate is positive). However, part of the effect of the difficult version is captured by the linguistic feature and so if we want to know what prediction the model would make for a difficult text version, we must also add the value of the linguistic feature and look at the net effect rather than the separate predictors. For instance, difficult texts in the lexical condition had a word frequency of -0.25 compared to easy texts. If we add this value into our combined model we find that the net effect for a difficult text version is still negative: -1.057 = 1.403 + (9.840 * -0.25). The model thus correctly predicts that difficult text versions are harder to comprehend than easy text versions, but it is important to note that the Text version predictor is now correcting the Word frequency predictor. It shows that between text versions the word frequency effect is smaller than between texts. The same pattern is found for the syntactic manipulation: the estimate for the difficult version changes direction to compensate the effect of the predictor Maximum SDL.

A different pattern emerges for the coherence marking manipulation. The estimates become larger when both predictors are in the model, but the estimates do not change direction. The number of connectives per clause remains negatively related to comprehension. The estimate of the difficult version also remains negative. On its own the predictor Text version was not significant, but that changes when both predictors are in the model. Text version is now also significant and correcting Connectives per clause. The net effect for a difficult text is therefore negative: -0.140 = -0.878 + (-3.513 * -0.21). This means that although texts that have a large number of connectives are generally hard to understand removing connectives within a text makes the text more difficult.

[2] Except in the coherence marking manipulation where the Text version predictor did not reach significance. This is in line with our results presented in Chapter 5. We only found local effects of coherence marking and no global effects on text level.

Thus, we find that the version differences cannot be correctly predicted on the basis of the estimate of the linguistic feature alone. The predictors Word frequency and Syntactic dependency length overestimate the version differences, as they are driven by text differences. A way to show how much correlational features overestimate text version differences is by comparing the actual version difference to the predicted version difference. We apply the size of the difference between the easy and difficult text versions to the estimate of the linguistic feature. For the lexical manipulation, Word frequency yielded an estimate of +8.492 and the difficult version yielded an estimate of -1.213. In Table 1 we see that the difference in Word frequency between easy and difficult text versions is on average 0.25. Thus, according to the model with Word frequency comprehension scores should be -2.123 (= 8.492 * -0.25) lower in the difficult version. In reality, comprehension scores for difficult text versions were only 1.213 lower, about 57% of the Word frequency estimate. For the syntactic manipulation we see an even larger overestimation. An increase of 1.54 in Maximum SDL, should reduce comprehension scores by 1.286 (= 0.835 * 1.54), but the Text version estimate only shows a reduction of 0.311.

For coherence marking, the linguistic feature prediction is not simply an overestimation of the text version difference, it points in the wrong direction. According to the predictor Connectives per clause the reduction of 0.21 in connectives per clause in the difficult version should yield an *increase* in comprehension scores of 0.367 (= 1.746 * 0.21). However, the Text version estimate indicates that comprehension scores are *reduced* by 0.148 in the difficult version.

These findings have important consequences for theories of linguistic complexity and for the field of readability improvement. For any linguistic feature it is crucial to differentiate between how it predicts conceptual difficulty and how it predicts stylistic difficulty. In our data, this comparison has revealed two different outcomes.

In the first scenario, the linguistic feature's relations with conceptual and stylistic difficulty go in the same direction, but differ in strength. This was the case for Word frequency and for Syntactic dependency length. Texts with high word frequencies are easier than texts with low word frequencies and increasing a text's word frequency makes the text a little easier to comprehend. Similarly, texts with short syntactic dependencies are easier to comprehend than texts with long dependencies and reducing the dependency length increases comprehension. The strength of the relationships differs though: linguistic features vary more in response to content differences than in response to stylistic differences. As a result, a given difference in the value of the feature yields stronger comprehension differences when texts with different contents are being examined as opposed to different versions of the same text. This means that the results of stylistic modifications will

be overestimated when they are based on prediction models that confound content and style.

In the second scenario, a linguistic feature's relationships with conceptual and stylistic difficulty are in opposite directions. This was the case for Connectives per clause. Texts with a high number of connectives are more difficult than texts with a low number of connectives, but adding connectives to a text generally improves comprehension. Such a scenario is perhaps more troublesome than the first scenario because on the basis of between-text comparisons, we could believe that removing connectives improves text comprehension, whereas the opposite is the case.

2.2 Comprehension and processing ease

In the final chapters we mostly focused on the comprehension aspect of readability, but we also included work on processing ease in our study. Both aspects are important for the readability of a text (Dale & Chall, 1949; Johnson & Otto, 1982; Klare, 1963; Miller & Kintsch, 1980). A text has a high level of readability when it is easy to understand with a limited amount of effort. In Chapters 3 and 4 we have seen that lexical and syntactic complexity can affect both comprehension and processing. In both cases reducing the complexity increased comprehension and reduced processing effort. In Chapter 5 we did not perform our own processing experiment, but we can draw from a large amount of work by others. Connectives reduce processing effort for the upcoming segment and the more informative a connective is (i.e., contrastive or causal), the larger the reduction. Our own findings on comprehension are not unidirectional. Some connectives have a positive influence on comprehension while others reduce comprehension. These results do not completely match previous findings in on-line processing. Thus, processing and comprehension effects do not always align. Furthermore, in Chapter 6 we found that strong predictors of comprehension are not necessarily indicative of processing ease. Only Word frequency was strongly predictive of both comprehension and processing ease. Our findings indicate that processing and comprehension are different concepts and that their relationship merits further study.

2.3 The importance of the reader

For reasons of convenience or statistical limitations, reader characteristics are often ignored in readability research (Anderson & Davison, 1988; Collins-Thompson, 2014; Duffy & Kabance, 1982). A clear example is the reliance on expert ratings rather than behavioral data[3] of the target population. But even when behavioral data

[3] With behavioral data we refer to any objective measurement of comprehension or processing ease. This does not include self-reports (i.e., judgements) of the reader.

is used, results are often aggregated over participants and reader characteristics are not included in the analysis of the data.

We took a different approach in which we acknowledged the importance of the reader. Comprehension and processing data were collected from our target population and analyzed with multilevel statistics. Our results show that a large amount of variance can be ascribed to the reader. In the U-Read model roughly 34% of the total variance was explained by the known reader characteristics Education level, Grade and Reading ability. An additional 23% was explained by text features.

Although our approach made it possible to test for reader-text interactions, we did not find many. Only in our lexical manipulation study (Chapter 3) and in the optimal multilevel model (Chapter 6; Appendix 11) we found some interactions which indicated that not all readers were equally affected by linguistic features. However, these interactions were of limited importance when compared to the main effects of reader characteristics.

2.4 Causal relevance

The lack of causally relevant predictors has been one of the classical critiques on readability research (Anderson & Davison, 1988; Bailin & Grafstein, 2001; Davison & Kantor, 1982; Johnson & Otto, 1982). Although causal relevance is not a prerequisite for readability prediction, even in prediction it would be preferable when features are supported by theoretical and experimental evidence before they are included in readability prediction. For readability improvement, causality is vital. Readability can only be improved by adapting stylistic features that actually reduce the difficulty of the text.

We designed our study accordingly, from the way features were calculated within T-Scan to the integration of experiments within our overall correlational study. Our experiments provide explicit evidence that two of the predictors in the U-Read model are indeed causally related to comprehension. Although the effects are smaller than the correlational model suggests (see Section 2.1), changing the word frequency or syntactic dependency length of a text will affect its readability. In contrast, reducing the complexity of shallow predictors like word and sentence length generally has no or little effect on readability (Davison & Kantor, 1982; Duffy & Kabance, 1982; Klare, 1976a). In addition, in our correlational analyses the U-Read predictors outperformed these shallow predictors. The traditional predictors have made way for causally related, informative predictors. For lexical complexity, Word length is surpassed by Word frequency and Concreteness. For syntactic

complexity, Sentence length is captured in Syntactic dependency length, Content words per clause and Adjectival past participle constructions.[4]

Another issue is the practice of building predictive models on graded text corpora or expert judgments instead of actual behavioral data. As a result, predictors that supposedly correlate with readability are not even really correlated with readability as experienced by the intended readers. They correlate with expert judgments which in turn supposedly correlate with actual readers. Therefore, even the correlations these models are based on are questionable. We believe that behavioral data like we have collected should be the gold standard rather than expert judgments and graded corpora.

3. Limitations and future research

In this section we discuss some of the limitations of our research and present avenues for further research.

3.1 Measuring readability

Many procedures have been used to measure readability and to validate readability predictors. Among these are expert and subject ratings, graded corpora, recall, reading times, eye-tracking, question answering, cloze, and comparing new formulae with old formulae (see Bailin & Grafstein, 2016; Collins-Thompson, 2014; Kintsch & Vipond, 1979). The readability of our texts was measured in two ways: with HyTeC-cloze tests and with eye-tracking. We used this two-pronged approach to differentiate between the comprehension and processing aspects of readability. In contrast to classical studies of Bormuth (1969), Dale and Chall (1948; 1995), and Flesch (1948) we did not limit our investigation to the comprehension aspect of readability.

3.1.1 Comprehension

Comprehension was measured with the specially developed HyTeC-cloze test. While cloze tests have a strong history in readability research, not all scholars believe they provide valid measurements of text comprehension. We already addressed these critiques in Chapter 2 and showed that our HyTeC-cloze was

[4] The term syntactic complexity must be taken in a broad sense here. The dependency length factor might be considered as pointing to syntactic complexity in a classical sense, but the other two factors reflect the informational load in the clause. As all three factors correlate substantially with sentence length, we suggest that they indicate the difficulties for which sentence length has long been used as a proxy.

sensitive to differences between text versions, texts and readers without being overly sensitive to local predictable items unrelated to text comprehension. We do agree that cloze tests, including the HyTeC-cloze, might capture some aspects of comprehension less clearly as some other methods may do. This may be the reason why our coherence marking manipulation did not appear to influence comprehension of the entire text. However, at the level of local coherence the HyTeC-cloze was sensitive to coherence marking and even detected differences between relation types. While some other methods, like asking inference or situation model questions (McNamara, 2001; Kamalski, 2007; Van Silfhout, Evers-Vermeul & Sanders, 2014), might be more adaptable to measure higher-order levels of comprehension, their results are also less comparable across texts. Item difficulty will be confounded with text difficulty, making it impossible to compare different texts. In contrast, cloze tests do not confound question with text difficulty since item difficulty is a direct reflection of text difficulty (Klare, 1976a).

Finally, it is important to note that even widely accepted standardized reading tests do not measure the exact same construct. Each test slightly favors other components of reading and has reduced sensitivity to measuring others. Studies that have compared standardized reading tests show correlations between test scores ranging from .3 to .8 (Cutting & Scarborough, 2006; Keenan, Betjemann & Olson, 2008). Our cloze scores showed stable correlations centering around .6 with standardized reading and vocabulary tests. In addition, our scores correlated .5 with the multiple-choice questions which were used in the eye-tracking experiments. In Chapter 3 we found effects of equal sizes for lexical complexity on cloze and multiple-choice questions. In Chapter 4 we were unable to capture syntactic complexity with the multiple-choice questions, but the HyTeC-cloze was sensitive enough to capture this subtle manipulation. It would be interesting to repeat such comparisons with other comprehension methods to see how our cloze holds up, but we are confident that the HyTeC-cloze has been a valid choice for the current investigation. It proved to be a stable, reliable method that provides a fair comparison between texts.

3.1.2 Processing ease

As is the case for text comprehension, there are different methods to measure processing ease. We chose eye-tracking for its precision, speed and the fact that it is fairly undisruptive to the reading process (compared to self-paced-reading or moving-window paradigms). Eye-tracking allowed us to look at processing ease at different levels and locations in the texts. For instance, our lexical manipulation was analyzed at word level, our syntactic manipulation was analyzed at sentence level and our exploration in Chapter 6 was run on text level data. However, to get precise measurements we needed to divide texts across multiple screens. Although these

screen breaks co-occurred with natural breaks, they limited the possibility to regress to all locations in the text. Therefore, our data may not be entirely representative of text level processing.

We also note that within this dissertation, the processing aspect of readability has received somewhat less attention than the comprehension aspect. Given the timeframe of our project, we could not investigate both aspects to the same extent. For instance, we would have liked to include a full correlational analysis of the eye-tracking data similar to our analysis of the cloze data that we presented in Chapter 6. We are planning to run this and other analyses in the near future. In addition, we have collected eye-tracking data of 167 Dutch adolescents reading multiple texts. These data open up possibilities for a wide range of research questions. We have seen similar opportunities arise from other eye-tracking corpora like the Dundee corpus (Kennedy, 2003) and the Potsdam sentence corpus (Kliegl, Nuthmann, & Engbert, 2006). We believe our eye-tracking corpus will prove to be a valuable resource for further reading research.

3.1.3 Affective aspects of readability

Comprehension and processing are both cognitive aspects of readability. It has been suggested that there is also an affective aspect to readability (Anderson & Davison, 1988; Dale & Chall, 1949; Hidi, Baird & Hidyard, 1982; Hidi, 2001; Klare, 1963; 1976a). This concerns the interestingness of the text and to which extend the reader is being motivated to read on. It may help the reading process when the text is interesting, but adding unnecessary information in order to make the text more interesting can decrease comprehension (Harp & Mayer, 1998; Van Silfhout, 2014). We did not address the affective aspect in our investigation of readability. Although we do believe that motivation and interest can have important effects on the reading process, we focused our investigation on the cognitive aspects of readability.

3.2 Texts and target population

Like all readability studies, our findings are only valid for text and readers that are similar to the ones we included in this study (Bruce, Rubin & Starr, 1981; Redish, 2000). Our prediction models do not automatically extend to Dutch adults or other text genres like literary works or blogs. However, within these parameters we are confident that our results are robust. We took a cross section of Dutch adolescents by including secondary school students from multiple grades and students enrolled in different levels of the Dutch educational system. This included students with low reading proficiency and students with high reading proficiency, but also students with dyslexia or attentional deficits, since they are all part of this population of Dutch adolescents attending secondary schools in the Netherlands. Since most

schools participated with entire classes, it is likely that the distribution of these subgroups is representative of the distribution of the entire population.

We also included a fair amount of texts in our study which were relevant for the target population. In the experimental studies our number of texts far exceeded the usual 2 to 4 texts per experiment. Moreover, the texts were randomly selected and not selected based on their potential for manipulation success.[5] Out of all features that we could manipulate we chose a lexical, a syntactic and a coherence marking feature based on their relative importance in the literature and in prior readability research. The results of Chapter 6 indicate that we chose well. The lexical complexity feature Word frequency and the syntactic complexity feature Maximum SDL were found to be causally relevant for comprehension ánd strong predictors of predicting differences between texts. Both features made it into the U-Read model. Our coherence marking manipulation was in that sense less successful. It had the least effect on comprehension and did not make it into the U-Read model. However, we limited our manipulation of coherence marking to the addition and removal of connectives. This is a very narrow working definition of coherence marking. It ignores other explicit signals of coherence (including cue phrases and advanced organizers, but also signals of referential coherence). In addition, any coherence marker is only a proxy for actual coherence since coherence is essentially a property of mental representations (Graesser, McNamara, Louwerse & Cai, 2004; Sanford, 2006; Sanders & Pander Maat, 2006, Zwaan & Rapp, 2006). Coherence is therefore difficult to capture in an objective measure, let alone in an automatically computed measure.

We included texts from two different genres in our study: educational textbook texts and public information texts. Whether our findings hold for other genres, is debatable. Genres are characterized by different linguistic features (Biber & Conrad, 2009; Graesser & McNamara, 2011; Pander Maat & Dekker, 2016; Pander Maat, 2017). We suspect that some features, like word frequency, will be important for most if not all genres. However, genre-specific features that were not present or uncommon in the genres we included may end up to be strong predictors for other genres.

Although our findings may be limited to our reader population, text genres and manipulations, our approach can be easily applied to explore other avenues. In spin-offs of this study we are already looking at the generalizability of our findings to different populations, in particular to Dutch adults and Flemish speakers of Dutch. Similar spin-offs are possible for new genres and manipulations.

A more fundamental question is whether we should even focus on text difficulty when image-based media seem to infringe on the territory of (traditional)

[5] We did impose a minimum number of manipulations, but no limitation was placed on the strength of the manipulation (see Chapter 1).

texts. As texts lose some terrain and are supplanted by or combined with graphics and video, we may need to redefine our definition of 'readability'. It may be worthwhile to draw from insights of dual coding theory (Paivio, 1991) and multimedia-learning (Mayer, 2005) to extend our definition to include combinations of text and images. However, we do not believe that texts will disappear. Text will remain an important medium. If anything, people are asked to be more self-reliant and to actively seek out and process information at work and in their private life. The bulk of this information is communicated via text.

3.3 Technological limitations

Language is ambiguous, infinite and subject to change. Capturing such a system is a difficult task (Manning & Schütze, 1999). Any study that uses automated text analysis is limited to automatically computable indices and their reliability (De Clercq & Hoste, 2016; Klare, 1976a). We could only include features in our study that did not require human intervention. Some text features, like coherence and figurative language, are by their very nature difficult to measure and so for now we can only use proxies of these features (Graesser & McNamara, 2011; Loenneker-Rodman & Narayanan, 2012; Manning & Schütze, 1999). In addition, we are bound to the upper reliability limits of indices. Parsers make mistakes, lists are vulnerable to polysemy and limited coverage, and spelling or punctuation errors pose an external threat to the reliability of most indices. Even manual computation is not perfect; human annotators do not always agree 100%. So determining the gold standard that the automated system should emulate is difficult in itself.

Another issue concerns the way indices are computed. The exact calculation of indices is important for how to interpret the results. For instance, our results showed that word frequencies without names and corrected for compound nouns was a stronger predictor than 'all-inclusive', uncorrected measures of frequency. Unfortunately, not all studies and readability tools offer detailed information on how they have computed their indices. This also makes it difficult to replicate their results. We used the analytical tool T-Scan to compute our features. T-Scan comes with an extensive manual in which the indices are explained in detail (Pander Maat, Kraf & Dekker, 2017). T-Scan provides alternative computations for a wide range of features which allows users to choose the metric best suited for them. This includes multiple indices that differ on what is and is not included in their calculation (like our frequency measure), but also the unit of measurement (occurrences per 1000 words, per sentence, proportion).

But of course, technology keeps progressing and new indices are still being developed. One big advantage of this study is that we now have behavioral data that can be used to test these new indices. The collected cloze and eye-tracking data were not only necessary for the studies presented in this dissertation, but they provide the

basis for further research and validation of linguistic features in future endeavors. As technological or theoretical advances create new and improved features, these features can also be validated using the data collected within this project. As a result, this work will not quickly become obsolete.

4. Final remarks

This study combined insights from readability research and discourse processing with current language technologies to investigate the relationship between linguistic features and two aspects of readability: comprehension and processing ease. We used an integrated methodological design in which we combined experimental with correlational work to disentangle causal effects on readability from correlational relationships. The results reported in the previous chapters allow the conclusion that our approach successfully dealt with many critiques on readability research which have often been voiced but have hardly ever been (simultaneously) addressed. These include the lack of causal relevance, disregard for the role of the reader and the use of non-behavioral data for validation and calibration of readability tools.

Our investigation revealed five linguistic features that best predict the text comprehension of Dutch adolescents: *Word frequency*, *Content words per clause*, *Concrete nouns*, *Maximum syntactic dependency length* and *Adjectival past participles*. These semantic- and syntactic-based features outperformed the shallow features used in traditional Dutch readability formulae. For two features - Word frequency and Maximum syntactic dependency length - we were able to show that they were also causally related to comprehension and processing ease. However, these causal effects were much smaller than correlational models would predict. Combining correlational and experimental studies allowed us to disentangle the way linguistic features reflect conceptual difficulty and stylistic difficulty. Our findings are especially important for the fields of readability improvement and discourse processing as they show the extent to which the reading process is influenced by stylistic interventions and provide a realistic (and sobering) view on the possible reduction of text difficulty when the content of a text cannot be altered. Due to our design, we are able to generalize these results across a large number of texts and across readers differing in reading proficiency.

Our study was driven by a practical need for a reliable, valid readability tool for Dutch. It lies on the intersection between readability improvement and prediction, between refining cognitive models of reading and practical applications. We hope that this study has helped to solidify the interconnections between these areas, and that insights from this study will benefit researchers but also - in the very near future - the general public.

References

Abraham, R. G., & Chapelle, C. A. (1992). The meaning of cloze test scores: An item difficulty perspective. *The Modern Language Journal, 76*(4), 468-479. doi:10.1111/j.1540-4781.1992.tb05394.x

Ackerman, F., & Goldberg, A. E. (1996). Constraints on adjectival past participles. In. A.E. Goldberg (Ed.), *Conceptual structure, discourse and language* (pp.17-30). Stanford: CSLI.

Alderson, J. C. (1979a). The cloze procedure and proficiency in English as a foreign language. *TESOL Quarterly, 13*(2), 219-227.

Alderson, J. C. (1979b). The effect on the cloze test of changes in deletion frequency. *Journal of Research in Reading, 2*(2), 108-119. doi:10.1111/j.1467-9817.1979.tb00198.x

Alderson, J. C. (2000). *Assessing Reading*. Cambridge: Cambridge University Press.

Anderson, R. C., & Davison, A. (1988). Conceptual and empirical bases of readability formulas. In A. Davison & G. M. Green (Eds.), *Linguistic complexity and text comprehension. Readability issues reconsidered* (pp. 23-53). Hillsdale: Lawrence Erlbaum Associates.

Asr, F. T., & Demberg, V. (2012). Implicitness of discourse relations. Paper presented at the *Proceedings of COLING 2012*, Mumbai. 2669-2684.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Bachman, L. F. (1982). The trait structure of cloze test scores. *TESOL Quarterly, 16*(1), 61-70. doi:10.2307/3586563

Bachman, L. F. (1985). Performance on cloze tests with fixed-ratio and rational deletions. *TESOL Quarterly, 19*(3), 535-556. doi:10.2307/3586277

Bailin, A., & Grafstein, A. (2001). The linguistic assumptions underlying readability formulae: A critique. *Language & Communication, 21*(3), 285-301. doi:10.1016/S0271-5309(01)00005-2

Bailin, A., & Grafstein, A. (2016). *Readability: Text and context.* New York, NY: Palgrave Macmillan.

Bartek, B., Lewis, R. L., Vasishth, S., & Smith, M. R. (2011). In search of on-line locality effects in sentence comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(5), 1178-1198. doi:10.1037/a0024194

Beck, I. L., McKeown, M. G., Sinatra, G. M., & Loxterman, J. A. (1991). Revising social studies text from a text-processing perspective: Evidence of improved comprehensibility. *Reading Research Quarterly, 26*(3), 251-276. doi:10.2307/747763

Benjamin, R. G. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review, 24*(1), 63-88.

Bertram, R. (2011). Eye movements and morphological processing in reading. *The Mental Lexicon, 6*(1), 83-109.

Biber, D., & Conrad, S. (2009). *Register, genre, and style*. Cambridge: Cambridge University Press.

Bloom, B. S. (1956). *Taxonomy of educational objectives: The classification of educational goals.* London: Longman Group.

Bormuth, J. R. (1966). Readability: A new approach. *Reading Research Quarterly, 1*(3), 79-132. doi:10.2307/747021

Bormuth, J. R. (1969). *Development of readability analyses.* (Final Report, Project No. 7-0052, Contract No. 1, OEC-3-7-070052-0326). Office of Education, U.S. Department of Health, Education, and Welfare, Office of Education, Bureau of Research.

Bouma, G., Van Noord, G., & Malouf, R. (2001). Alpino: Wide-coverage computational analysis of Dutch. *Language and Computers, 37*(1), 45-59.

Breland, H. M. (1996). Word frequency and word difficulty: A comparison of counts in four corpora. *Psychological Science, 7*(2), 96-99.

Britton, B. K., & Gülgöz, S. (1991). Using Kintsch's computational model to improve instructional text: Effects of repairing inference calls on recall and cognitive structures. *Journal of Educational Psychology, 83*(3), 329-345. doi:10.1037/0022-0663.83.3.329

Brown, J. D. (1980). Relative merits of four methods for scoring cloze tests. *The Modern Language Journal, 64*(3), 311-317. doi:10.1111/j.1540-4781.1980.tb05198.x

Brown, J. D. (1983/1994). A closer look at cloze validity. In J. W. Oller & J. Jonz (Eds.), *Cloze and Coherence* (pp. 189-196). Lewisburg: Bucknell University Press. (Reprinted from J.W. Oller (Ed.), Issues in Language Testing (pp.237-250). Rowley, Massachusetts: Newbury House, 1983).

Brown, J. D. (1993). What are the characteristics of natural cloze tests? *Language Testing, 10*(2), 93-116. doi:10.1177/026553229301000201

Brown, J. D. (2002). Do cloze tests work? Or, is it just an illusion? *Second Language Studies, 21*(1), 79-125.

Brown, J. D. (2013). My twenty-five years of cloze testing research: So what. *International Journal of Language Studies, 7*(1), 1-32.

Bruce, B., & Rubin, A. (1988). Readability formulas: Matching tool and task. In A. Davison & G. M. Green (Eds.), *Linguistic complexity and text comprehension. Readability issues reconsidered* (pp. 5-22). Hillsdale: Lawrence Erlbaum Associates.

Bruce, B., Rubin, A., & Starr, K. (1981). Why readability formulas fail. *IEEE Transactions on Professional Communication, PC-24*(1), 50-52. doi:10.1109/TPC.1981.6447826

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior research methods, 41*(4), 977-990.

Cain, K., & Nash, H. M. (2011). The influence of connectives on young readers' processing and comprehension of text. *Journal of Educational Psychology, 103*(2), 429-441. doi:10.1037/a0022824

Camblin, C., Ledoux, K. Boudewyn, M., Gordon, P.C., & Swaab, T.Y. (2007). Processing new and repeated names: Effects of coreference on repetition priming with speech and fast RSVP. *Brain Research, 1146*, 172-184.

Canestrelli, A. R., Mak, W. M., & Sanders, T. J. M. (2013). Causal connectives in discourse processing: How differences in subjectivity are reflected in eye movements. *Language and Cognitive Processes, 28*(9), 1394-1413.

Chaffin, R., Morris, R. K., & Seely, R. E. (2001). Learning new word meanings from context: A study of eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(1), 225-235. doi:10.1037/0278-7393.27.1.225

Chall, J. S., & Dale, E. (1995). *Readability revisited: The new Dale-Chall readability formula*. Cambridge, MA: Brookline Books.

Chávez-Oller, M. A., Chihara, T., Weaver, K. A., & Oller, J. W. (1985/1994). When are cloze items sensitive to constraints across sentences? In J. W. Oller & J. Jonz (Eds.), *Cloze and coherence* (pp. 229-245). Lewisburg: Bucknell University Press. (Revised from Language Learning, 35(2), 181-206, 1985).

Chen, L. (2004). On text structure, language proficiency, and reading comprehension test format interactions: a reply to Kobayashi, 2002. *Language Testing, 21*(2), 228-234. doi:10.1191/0265532202lt227oa

Chihara, T., Oller, J. W., Weaver, K. A., & Chávez-Oller, M. A. (1977/1994). Are cloze items sensitive to constraints across sentences? In J. W. Oller & J. Jonz (Eds.), *Cloze and coherence* (pp. 135-147). Lewisburg: Bucknell University Press. (Revised from Language Learning, 27(1), 63-73, 1977).

Clifton Jr, C., & Frazier, L. (1989). Comprehending sentences with long-distance dependencies. In G. N. Carlson & M. K. Tanenhaus (Eds.), *Linguistic structure in language processing* (pp. 273-317) Dordrecht, The Netherlands: Kluwer academic publishers. doi:10.1007/978-94-009-2729-2_8

Cohen, A. D. (1984). On taking language tests: What the students report. *Language Testing, 1*(1), 70-81. doi:10.1177/026553228400100106

Coleman, C., Lindstrom, J., Nelson, J., Lindstrom, W., & Gregg, K. N. (2010). Passageless comprehension on the Nelson-Denny Reading Test: Well above chance for university students. *Journal of Learning Disabilities, 43*(3), 244-249. doi: 10.1177/0022219409345017

Coleman, E. B. (1962). Improving comprehensibility by shortening sentences. *Journal of Applied Psychology, 46*(2), 131-134. doi:10.1037/h0039740

Collins-Thompson, K. (2014). Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics, 165*(2), 97-135.

Collins-Thompson, K., & Callan, J. (2005). Predicting reading difficulty with statistical language models. *Journal of the Association for Information Science and Technology, 56*(13), 1448-1462. doi:10.1002/asi.20243

Cozijn, R. (1994). *Manual for the interactive analysis program of eye movement recordings: FIXATION*. (Internal paper). Nijmegen, The Netherlands: Max Planck Institute for Psycholinguistics.

Cozijn, R., Noordman, L. G. M., & Vonk, W. (2011). Propositional integration and world-knowledge inference: Processes in understanding because sentences. *Discourse Processes, 48*(7), 475-500. doi:10.1080/0163853X.2011.594421

Crain, S., & Steedman, M. (1985). On not being led up the garden path: The use of context by the psychological syntax processor. In D. R. Dowty, L. Karttunen & A. M. Zwicky (Eds.), *Natural language parsing: Psychological, computational, and theoretical perspectives* (pp. 320-358). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511597855.011

Cromley, J. G., & Azevedo, R. (2006). Self-report of reading comprehension strategies: What are we measuring? *Metacognition and Learning, 1*(3), 229-247.

Crossley, S. A., Allen, D., & McNamara, D. S. (2012). Text simplification and comprehensible input: A case for an intuitive approach. *Language Testing Research, 16*(1), 89-108. doi:10.1177/1362168811423456

Crossley, S. A., Dufty, D. F., McCarthy, P. M., & McNamara, D. S. (2007). Toward a new readability: A mixed model approach. In *Proceedings of the Cognitive Science Society 29* (pp. 197-202).

Crossley, S. A., Louwerse, M. M., McCarthy, P. M., & McNamara, D. S. (2007). A linguistic analysis of simplified and authentic texts. *The Modern Language Journal, 91*(1), 15-30. doi:10.1111/j.1540-4781.2007.00507.x

Crossley, S. A., Skalicky, S., Dascalu, M., McNamara, D. S., & Kyle, K. (2017). Predicting text comprehension, processing, and familiarity in adult readers: New approaches to readability formulas. *Discourse Processes, 54*(5-6), 340-359. doi:10.1080/0163853X.2017.1296264

Cutler, A. (1983). Lexical complexity and sentence processing. In G. B. Flores d'Arcais & R. J. Jarvella (Eds.), *The process of language understanding* (pp. 43-79). New York: Wiley.

Cutting, L. E., & Scarborough, H. S. (2006). Prediction of reading comprehension: Relative contributions of word recognition, language proficiency, and other cognitive skills can depend on how comprehension is measured. *Scientific Studies of Reading, 10*(3), 277-299. doi:10.1207/s1532799xssr1003_5

Cziko, G. A. (1983). Another response to Shanahan, Kamil, and Tobin: Further reasons to keep the cloze case open. *Reading Research Quarterly, 18*(3), 361-365.

Dale, E., & Chall, J. S. (1948). A formula for predicting readability. *Educational Research Bulletin, 27*(1), 11-28.

Dale, E., & Chall, J. S. (1948). A formula for predicting readability: Instructions. *Educational Research Bulletin, 27*(2), 37-54.

Dale, E., & Chall, J. S. (1949). The concept of readability. *Elementary English, 26*(1), 19-26.

Dascalu, M., Stavarache, L. L., Trausan-Matu, S., Dessus, P., & Bianco, M. (2014). Reflecting comprehension through French textual complexity factors. Paper presented at the *IEEE 26th International Conference on Tools with Artificial Intelligence (ICTAI)*, 615-619.

Davies, A., & Widdowson, H. G. (1974). Reading and writing. In J. P. B. Allen & S. Pit Corder (Eds.), *The Edinburgh course in applied linguistics: Vol. 3. Techniques in applied linguistics* (pp. 155-201). Oxford University Press: London.

Davison, A., & Green, G. M. (Eds.). (1988). *Linguistic complexity and text comprehension. Readability issues reconsidered.* Hillsdale: Lawrence Erlbaum Associates.

Davison, A., & Kantor, R. N. (1982). On the failure of readability formulas to define readable texts: A case study from adaptations. *Reading Research Quarterly, 17*(2), 187-209. doi:10.2307/747483

Davison, A., & Lutz, R. (1985). Measuring syntactic complexity relative to discourse context. In D. R. Dowty, L. Karttunen & A. M. Zwicky (Eds.), *Natural language parsing: Psychological, computational, and theoretical perspectives* (pp. 26-66). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511597855.002

Davison, A., Kantor, R. N., Hannah, J., Hermon, G., Lutz, R., & Salzillo, R. (1980). *Limitations of readability formulas in guiding adaptation of texts.* (Tech. Rep. No. 162). Urbana, Illinois: University of Illinois, Center for the Study of Reading.

De Clercq, O., & Hoste, V. (2016). All mixed up? Finding the optimal feature set for general readability prediction and its application to English and Dutch. *Computational Linguistics, 42*(3), 457-490. doi:10.1162/COLI_a_00255

De Clercq, O., Hoste, V., Desmet, B., van Oosten, P., De Cock, M., & Macken, L. (2014). Using the crowd for readability prediction. *Natural Language Engineering, 20*(3), 293-325. doi:10.1017/S1351324912000344

De Jong, M. D. T., & Lentz, L. R. (1996) Expert judgments versus reader feedback. A comparison of text evaluation techniques. *Journal of Technical Writing and Communication, 26*(4), 507–519.

De Jong, M. D. T., & Schellens, P. J. M. C. (1995). *Met het oog op de lezer: Pretestmethoden voor schriftelijk voorlichtingsmateriaal*. Amsterdam: Thesis Publishers.

Degand, L., & Sanders, T. J. M. (2002). The impact of relational markers on expository text comprehension in L1 and L2. *Reading and Writing, 15*(7), 739-757.

Degand, L., Lefèvre, N., & Bestgen, Y. (1999). The impact of connectives and anaphoric expressions on expository discourse comprehension. *Document Design, 1*, 39-51.

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition, 109*(2), 193-210.

Douma, W. H. (1960). *De leesbaarheid van landbouwbladen*. Wageningen, The Netherlands: Afdelingen voor Sociale Wetenschappen aan de Landbouwhogeschool.

DuBay, W. (2006). *The classic readability studies*. Costa Mesa, CA: Impact Information.

Duffy, T. M., & Kabance, P. (1982). Testing a readable writing approach to text revision. *Journal of Educational Psychology, 74*(5), 733-748.

Ehri, L. C. (1991). Development of the ability to read words. In R. Barr, M. L. Kamil, P. B. Mosenthal & P. D. Pearson (Eds.), *Handbook of reading research. Volume 2* (pp. 383-417). New York, NY: Lawrence Erlbaum.

EP-Nuffic. (2015). *Education system the Netherlands: The Dutch education system described*. Retrieved from https://www.epnuffic.nl/en/publications/find-a-publication/education-system-the-netherlands.pdf.

Evers-Vermeul, J., & Sanders, T. (2009). The emergence of Dutch connectives: How cumulative cognitive complexity explains the order of acquisition. *Journal of Child Language, 36*(4), 829-854.

Evers-Vermeul, J., & Sanders, T. (2011). Discovering domains: On the acquisition of causal connectives. *Journal of Pragmatics, 43*(6), 1645-1662.

Farmer, T. A., Misyak, J. B., Christiansen, M. H., Spivey, M., Joannisse, M., & McRae, K. (2012). Individual differences in sentence processing. In M. J. Spivey, K. McRae & M. Joannisse (Eds.), *The Cambridge handbook of psycholinguistics* (pp. 353-364). Cambridge, UK: Cambridge University Press.

Fedorenko, E., Gibson, E., & Rohde, D. (2006). The nature of working memory capacity in sentence comprehension: Evidence against domain-specific working memory resources. *Journal of Memory and Language, 54*(4), 541-553. doi:10.1016/j.jml.2005.12.006

Fedorenko, E., Piantadosi, S., & Gibson, E. (2012). Processing relative clauses in supportive contexts. *Cognitive Science, 36*(3), 471-497. doi:10.1111/j.1551-6709.2011.01217.x

Feng, L. (2010). *Automatic readability assessment*. (Doctoral dissertation). New York: The City University of New York.

Feng, L., Elhadad, N., & Huenerfauth, M. (2009). Cognitively motivated features for readability assessment. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2009)*, Athens, Greece. 229-237.

Feng, L., Jansche, M., Huenerfauth, M., & Elhadad, N. (2010). A comparison of features for automatic readability assessment. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters* (pp. 276-284). Beijing, China.

Ferreira, F., Bailey, K. G., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science, 11*(1), 11-15. doi:10.1111/1467-8721.00158

Flesch, R. F. (1943). *Marks of readable style: A study in adult education*. New York: Teachers College, Columbia University.

Flesch, R. F. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*(3), 221-233. doi:10.1037/h0057532

François, T. (2015). When readability meets computational linguistics: A new paradigm in readability. *Revue Française De Linguistique Appliquée, 20*(2), 79-97.

François, T., & Miltsakaki, E. (2012). Do NLP and machine learning improve traditional readability formulas? Paper presented at the *Proceedings of the First Workshop on Predicting and Improving Text Readability for Target Reader Populations,* 49-57.

Freebody, P., & Anderson, R. C. (1983a). Effects of vocabulary difficulty, text cohesion, and schema availability on reading comprehension. *Reading Research Quarterly, 18*(3), 277-294.

Freebody, P., & Anderson, R. C. (1983b). Effects on text comprehension of differing proportions and locations of difficult vocabulary. *Journal of Reading Behavior, 15*(3), 19-39.

Futrell, R., & Levy, R. (2017). Noisy-context surprisal as a human sentence processing cost model. Paper presented at the *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics,* Valencia, Spain. 688-698. Retrieved from https://pdfs.semanticscholar.org/9ce4/c4ab20328738429fe4f01e7f082684046ee2.pdf

Futrell, R., Mahowald, K., & Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences of the United States of America, 112*(33), 10336-10341. doi:10.1073/pnas.1502134112

Gellert, A. S., & Elbro, C. (2013). Cloze tests may be quick, but are they dirty? Development and preliminary validation of a cloze test of reading comprehension. *Journal of Psychoeducational Assessment, 31*(1), 16-28.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition, 68*(1), 1-76. doi:10.1016/S0010-0277(98)00034-1

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In Y. Miyashita, A. Marantz & W. O'Neil (Eds.), *Image, language, brain* (pp. 96-126). Cambridge: MIT Press.

Glenberg, A. M., Wilkinson, A. C., & Epstein, W. (1982). The illusion of knowing: Failure in the self-assessment of comprehension. *Memory & Cognition, 10*(6), 597-602.

Gonzalez, E., & Rutkowski, L. (2010). Principles of multiple matrix booklet designs and parameter recovery in large-scale assessments. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments, 3*, 125-156.

Gordon, P. C., Grosz, B. J., & Gilliom, L. A. (1993). Pronouns, names, and the centering of attention in discourse. *Cognitive science, 17*(3), 311-347.

Gordon, P. C., Hendrick, R., Johnson, M., & Lee, Y. (2006). Similarity-based interference during language comprehension: Evidence from eye tracking during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(6), 1304-1321. doi:10.1037/0278-7393.32.6.1304

Gough, P. B. (1966). The verification of sentences: The effects of delay of evidence and sentence length. *Journal of Verbal Learning and Verbal Behavior, 5*(5), 492-496. doi:10.1016/S0022-5371(66)80067-1

Graesser, A. C., & McNamara, D. S. (2011). Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science, 3*(2), 371-398.

Graesser, A. C., McNamara, D. S., Louwerse, M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers, 36*(2), 193-202.

Graesser, A., Ozuru, Y., & Sullins, J. (2010). What is a good question? In M. G. McKeown & L. Kucan (Eds.), *Bringing reading research to life* (pp. 112-141). New York: Guilford Press.

Gray, W. S., & Leary, B. E. (1935). *What makes a book readable.* Chicago: University of Chicago Press.

Greene, B. B. (2001). Testing reading comprehension of theoretical discourse with cloze. *Journal of Research in Reading, 24*(1), 82-98.

Grodner, D., & Gibson, E. (2005). Consequences of the serial nature of linguistic input for sentenial complexity. *Cognitive Science, 29*(2), 261-290. doi:10.1207/s15516709cog0000_7

Grodner, D., Gibson, E., & Watson, D. (2005). The influence of contextual contrast on syntactic processing: Evidence for strong-interaction in sentence comprehension. *Cognition, 95*(3), 275-296. doi:10.1016/j.cognition.2004.01.007

Haberlandt, K. (1982). Reader expectations in text comprehension. *Advances in Psychology, 9*, 239-249.

Hacquebord, H., & Lenting-Haan, K. (2012). Kunnen we de moeilijkheid van teksten meten? Naar concrete maten voor de referentieniveaus. *Levende Talen Tijdschrift, 13*(2), 14-23.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics* (pp. 1–8). Pittsburgh, PA: Carnegie Mellon University.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. London: Longman.

Hanna, G. S., & Oaster, T. R. (1980). Studies of the seriousness of three threats to passage dependence. *Educational and Psychological Measurement, 40*(2), 405-411. doi:10.1177/001316448004000218

Harp, S. F., & Mayer, R. E. (1998). How seductive details do their damage: A theory of cognitive interest in science learning. *Journal of educational psychology, 90*(3), 414-434.

Hashkes, B., & Koffman, N. (1982). *Strategies used in a cloze test.* (Course paper). School of Education, Hebrew University of Jerusalem.

Henk, W. A. (1982). A response to Shanahan, Kamil, and Tobin: The case is not yet clozed. *Reading Research Quarterly, 17*(4), 591-595.

Hidi, S. (2001). Interest, reading, and learning: Theoretical and practical considerations. *Educational Psychology Review, 13*(3), 191-209. doi:10.1023/A:1016667621114

Hidi, S., Baird, W., & Hildyard, A. (1982). That's important but is it interesting? Two factors in text processing. *Advances in Psychology, 8*, 63-75.

Holmes, V. M., & O'Regan, J. K. (1981). Eye fixation patterns during the reading of relative-clause sentences. *Journal of Verbal Learning and Verbal Behavior, 20*(4), 417-430. doi:10.1016/S0022-5371(81)90533-8

Honeyfield, J. (1977). Simplification. *TESOL Quarterly, 11*(4), 431-440.

Jansen, C. (2005). Kwaliteit overheidscommunicatie meetbaar gemaakt? Pretenties van BureauTaal tegen het licht gehouden. *Tekst[Blad], 11*(4), 9-11.

Jansen, C., & Boersma, N. (2013). Meten is weten? Over de waarde van de leesbaarheidsvoorspellingen van drie geautomatiseerde Nederlandse meetinstrumenten. *Tijdschrift voor Taalbeheersing, 35*(1), 47-62.

Jansen, C., & Lentz, L. (2008). Hoe begrijpelijk is mijn tekst? De opkomst, neergang en terugkeer van de leesbaarheidsformules. *Onze Taal, 77*(1), 4-7.

Jansen, C., & Woudstra, E. (1979). Theorie en praktijk van het Nederlandse leesbaarheidsonderzoek. *Tijdschrift voor Taalbeheersing, 1*(1), 43-60.

Johnson, L. L., & Otto, W. (1982). Effect of alterations in prose style on the readability of college texts. *Journal of Educational Research, 75*(4), 222-229.

Jonz, J. (1994). Cloze item types and constraint on response. In J. W. Oller & J. Jonz (Eds.), *Cloze and Coherence* (pp. 317-344). Lewisburg: Bucknell University Press.

Juhasz, B. J., & Rayner, K. (2003). Investigating the effects of a set of intercorrelated variables on eye fixation durations in reading. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 29*(6), 1312-1318.

Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (2nd ed.). Upper Saddle River, New Jersey: Prentice Hall.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review, 99*(1), 122-149. doi:10.1037/0033-295X.99.1.122

Kaiser, E., & Trueswell, J. C. (2004). The role of discourse context in the processing of a flexible word-order language. *Cognition, 94*(2), 113-147. doi:10.1016/j.cognition.2004.01.002

Kamalski, J. (2007). *Coherence marking, comprehension and persuasion on the processing and representations of discourse.* (Doctoral dissertation). Utrecht, The Netherlands: LOT.

Kamalski, J., Lentz, L., Sanders, T., & Zwaan, R. A. (2008). The forewarning effect of coherence markers in persuasive discourse: Evidence from persuasion and processing. *Discourse Processes, 45*(6), 545-579.

Kamalski, J., Sanders, T., & Lentz, L. (2008). Coherence marking, prior knowledge, and comprehension of informative and persuasive texts: Sorting things out. *Discourse Processes, 45*(4-5), 323-345.

Katz, S., Lautenschlager, G. J., Blackburn, A. B., & Harris, F. H. (1990). Answering reading comprehension items without passages on the SAT. *Psychological Science, 1*(2), 122-127.

Keenan, J. M., & Betjemann, R. S. (2006). Comprehending the gray oral reading test without reading it: Why comprehension tests should not include passage-independent items. *Scientific Studies of Reading, 10*(4), 363-380. doi:10.1207/s1532799xssr1004_2

Keenan, J. M., & Meenan, C. E. (2014). Test Differences in Diagnosing Reading Comprehension Deficits. *Journal of Learning Disabilities, 47*(2), 125-135. doi:10.1177/0022219412439326

Keenan, J. M., Betjemann, R. S., & Olson, R. K. (2008). Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension. *Scientific Studies of Reading, 12*(3), 281-300.

Kemper, S., & Kemtes, K. A. (2002). Limitations on syntactic processing. In S. Kemper & R. Kliegl (Eds.), *Constraints on language: Aging, grammar, and memory* (pp. 79-105). Boston: Kluwer academic publishers. doi:10.1007/0-306-46902-2_4

Kemper, S., Jackson, J. D., Cheung, H., & Anagnopoulos, C. A. (1993). Enhancing older adults' reading comprehension. *Discourse Processes, 16*(4), 405-428.

Kennedy (2003). *The Dundee Corpus* [CD-ROM]. School of Psychology, The University of Dundee.

Kennedy, A., Pynte, J., Murray, W. S., & Paul, S. (2013). Frequency and predictability effects in the Dundee corpus: An eye movement analysis. *The Quarterly Journal of Experimental Psychology, 66*(3), 601-618. doi:10.1080/17470218.2012.676054

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods, 42*(3), 643-650. doi:10.3758/BRM.42.3.643

Keuleers, E., Stevens, M., Mandera, P., & Brysbaert, M. (2015). Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. *The Quarterly Journal of Experimental Psychology, 68*(8), 1665-1692. doi:10.1080/17470218.2015.1022560

King, J., & Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language, 30*(5), 580-602. doi:10.1016/0749-596X(91)90027-H

Kintsch, W. (1992). How readers construct situation models for stories: The role of syntactic cues and causal inferences. In A. F. Healy, S. M. Kosslyn & R. M. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes* (pp. 261-278). Hillsdale, NJ: Erlbaum.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. New York: Cambridge university press.

Kintsch, W. (2012). Psychological models of reading comprehension and their implications for assessment. In J. P. Sabatini, E. R. Albro & T. O'Reilly (Eds.), *Measuring up: Advances in how we assess reading ability* (pp. 21-38). Lanham, MD: Rowman & Littlefield Education.

Kintsch, W., & Keenan, J. (1973). Reading rate and retention as a function of the number of propositions in the base structure of sentences. *Cognitive Psychology, 5*(3), 257-274.

Kintsch, W., & Vipond, D. (1979). Reading comprehension and readability in educational practice and psychological theory. In L. Nilsson (Ed.), *Perspectives on memory research: Essays in honor of Uppsala university's 500th anniversary* (pp. 329-365). Hillsdale, NY: Lawrence Erlbaum Associates.

Kintsch, W., & Yarbrough, J. C. (1982). Role of rhetorical structure in text comprehension. *Journal of Educational Psychology, 74*(6), 828-834.

Klare, G. R. (1963). *The measurement of readability*. Ames: Iowa State University.

Klare, G. R. (1974). Assessing readability. *Reading research quarterly, 10*(1), 62-102.

Klare, G. R. (1976a). A second look at the validity of readability formulas. *Journal of Literacy Research, 8*(2), 129-152. doi:10.1080/10862967609547171

Klare, G. R. (1976b). Judging readability. *Instructional Science, 5*(1), 55-61.

Klare, G. R. (1984). Readability. In P. Pearson, R. Barr, M. Kamil & P. B. Mosenthal (Eds.), *Handbook of reading research* (pp. 681-744). New York: Longman.

Kleijn, S., Mak, P., & Sanders, T. (2011). Daardoor dus! Effecten van specificiteit en subjectiviteit op de verwerking van connectieven. [Effects of specificity and subjectivity on on-line processing of connectives]. *Toegepaste Taalwetenschap in Artikelen, 85*(1), 103-112.

Klein-Braley, C., & Raatz, U. (1984). A survey of research on the C-test. *Language Testing, 1*(2), 134-146.

Kliegl, R., Nuthmann, A., & Engbert, R. (2006). Tracking the mind during reading: The influence of past, present, and future words on fixation durations. *Journal of Experimental Psychology: General, 135*(1), 12-35.

Knott, A., Oberlander, J., O'Donnell, M., & Mellish, C. (2001). Beyond elaboration: The interaction of relations and focus in coherent text. In T. Sanders, J. Schilperoord & W. Spooren (Eds.), *Text representation: Linguistic and psycholinguistic aspects* (pp. 181-196). Amsterdam: John Benjamins.

Kobayashi, M. (2002a). Cloze tests revisited: Exploring item characteristics with special attention to scoring methods. *The Modern Language Journal, 86*(4), 571-586.

Kobayashi, M. (2002b). Method effects on reading comprehension test performance: text organization and response format. *Language Testing, 19*(2), 193-220.

Kobayashi, M. (2004). Investigation of test method effects: text organization and response format: a response to Chen, 2004. *Language Testing, 21*(2), 235-244.

Konieczny, L. (2000). Locality and parsing complexity. *Journal of Psycholinguistic Research, 29*(6), 627-645. doi:10.1023/A:1026528912821

Koornneef, A., Dotlačil, J., Van den Broek, P., & Sanders, T. (2016). The influence of linguistic and cognitive factors on the time course of verb-based implicit causality. *The Quarterly Journal of Experimental Psychology, 69*(3), 455-481. doi:10.1080/17470218.2015.1055282

Kraf, R., & Pander Maat, H. (2009). Leesbaarheidsonderzoek: oude problemen, nieuwe kansen. *Tijdschrift voor Taalbeheersing, 31*(2), 97-123.

Kraf, R., Lentz, L., & Pander Maat, H. (2011). Drie Nederlandse instrumenten voor het automatisch voorspellen van begrijpelijkheid. Een klein onderzoek. *Tijdschrift voor Taalbeheersing, 33*(3), 249-265.

Kuperman, V., & Van Dyke, J. A. (2011). Effects of individual differences in verbal skills on eye-movement patterns during sentence reading. *Journal of Memory and Language, 65*(1), 42-73. doi:10.1016/j.jml.2011.03.002

Kuperman, V., Drieghe, D., Keuleers, E., & Brysbaert, M. (2013). How strongly do word reading times and lexical decision times correlate? Combining data from eye movement corpora and megastudies. *The Quarterly Journal of Experimental Psychology, 66*(3), 563-580. doi:10.1080/17470218.2012.658820

Lakoff, R. (1971). If's, and's and but's about conjunction. In C. J. Fillmore & D. T. Langendoen (Eds.), *Studies in linguistic semantics* (pp. 114-149). New York: Holt, Rinehart, & Winston.

Land, J. (2009). *Zwakke lezers, sterke teksten? Effecten van tekst- en lezerskenmerken op het tekstbegrip en de tekstwaardering van vmbo-leerlingen*. (Doctoral dissertation). Delft: Eburon.

Land, J., Sanders, T., & Van den Bergh, H. (2008). Effectieve tekststructuur voor het vmbo een corpus-analytisch en experimenteel onderzoek naar tekstbegrip en tekstwaardering van vmbo-leerlingen voor studieteksten. *Pedagogische Studiën, 85*(2), 76-94.

Leech, G. N., & Short, M. (2007). *Style in fiction: A linguistic introduction to English fictional prose* (2nd ed.). Harlow: Pearson Education Limited.

Lentz, L., & De Jong, M. (1997). The evaluation of text quality: Expert-focused and reader-focused methods compared. *IEEE Transactions on Professional Communication, 40*, 224-234.

Levenston, E. A., Nir, R., & Blum-Kulka, S. (1984). Discourse analysis and the testing of reading comprehension by cloze techniques. In A. J. Pugh & J. M. Ulijn (Eds.), *Reading for professional purposes. Studies and practices in native and foreign languages* (pp. 202-212). London: Heinemann Educational Books.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition, 106*(3), 1126-1177. doi:10.1016/j.cognition.2007.05.006

Levy, R. P., & Keller, F. (2013). Expectation and locality effects in German verb-final structures. *Journal of Memory and Language, 68*(2), 199-222. doi:10.1016/j.jml.2012.02.005

Levy, R., Fedorenko, E., & Gibson, E. (2013). The syntactic complexity of Russian relative clauses. *Journal of Memory and Language, 69*(4), 461-495. doi:10.1016/j.jml.2012.10.005

Levy, R., Fedorenko, E., Breen, M., & Gibson, E. (2012). The processing of extraposed structures in English. *Cognition, 122*(1), 12-36. doi:10.1016/j.cognition.2011.07.012

Leyland, L., Kirkby, J. A., Juhasz, B. J., Pollatsek, A., & Liversedge, S. P. (2013). The influence of word shading and word length on eye movements during reading. *The Quarterly Journal of Experimental Psychology, 66*(3), 471-486. doi:10.1080/17470218.2011.599401

Linderholm, T., Everson, M. G., Van Den Broek, P., Mischinski, M., Crittenden, A., & Samuels, J. (2000). Effects of causal text revisions on more- and less-skilled readers' comprehension of easy and difficult texts. *Cognition and Instruction, 18*(4), 525-556.

Liu, C., Kemper, S., & Bovaird, J. A. (2009). Comprehension of health-related written materials by older adults. *Educational Gerontology, 35*(7), 653-668.

Liu, H., Xu, C., & Liang, J. (2017). Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews, 21*, 171-193. doi:10.1016/j.plrev.2017.03.002

Loenneker-Rodman, B., & Narayanan, S. (2012). Computational approaches to figurative language. In M. Spivey, K. McRae & M. Joanisse (Eds.), *The Cambridge Handbook of Psycholinguistics* (pp. 485-504). New York: Cambridge University press.

Loman, N. L., & Mayer, R. E. (1983). Signaling techniques that increase the understandability of expository prose. *Journal of Educational Psychology, 75*(3), 402-412.

Lorch Jr, R. F., & Lorch, E. P. (1986). On-line processing of summary and importance signals in reading. *Discourse Processes, 9*(4), 489-496.

MacDonald, M. C., & Christiansen, M. H. (2002). Reassessing working memory: Comment on Just and Carpenter (1992) and Waters and Caplan (1996). *Psychological Review, 109*(1), 35-54. doi:10.1037/0033-295X.109.1.35

MacGinitie, W. H. (1961). Contextual constraint in English prose paragraphs. *Journal of Psychology*, 51, 121-130.

Mann, W. C., & Taboada, M. (2017). *RST website*. Retrieved from http://www.sfu.ca/rst/02analyses/index.html

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge: MIT press.

Martin, W., & Maks, I. (2005). *Referentie Bestand Nederlands.* [In collaboration with S. Bopp and M. Groot].

Maury, P., & Teisserenc, A. (2005). The role of connectives in science text comprehension and memory. *Language and Cognitive Processes, 20*(3), 489-512.

Mayer, R. E. (Ed.). (2005). *The Cambridge handbook of multimedia learning*. New York: Cambridge university press.

McNamara, D. S. (2001). Reading both high-coherence and low-coherence texts: Effects of text sequence and prior knowledge. *Canadian Journal of Experimental Psychology, 55*(1), 51-62.

McNamara, D. S., Graesser, A. C., & Louwerse, M. M. (2012). Sources of text difficulty: Across the ages and genres. In J. P. Sabatini, E. Albro & T. O'Reilly (Eds.), *Measuring up: Advances in how to assess reading ability* (pp. 89-116). Lanham, MD: Rowman & Littlefield Education.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. New York, NY: Cambridge University Press.

McNamara, D. S., Kintsch, E., Butler Songer, N., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction, 14*(1), 1-43.

Meyer, B. J. F. (1975). *The organization of prose and its effects on memory*. Amsterdam, The Netherlands: North-Holland Publishing.

Meyer, B. J. F. (2003). Text coherence and readability. *Topics in Language Disorders, 23*(3), 204-224.

Meyer, B. J. F., & Freedle, R. O. (1984). Effects of discourse type on recall. *American Educational Research Journal, 21*(1), 121-143.

Mikk, J., & Elts, J. (1999). A reading comprehension formula of reader and text characteristics. *Journal of Quantitative Linguistics, 6*(3), 214-221.

Miller, G. R., & Coleman, E. B. (1967). A set of thirty-six prose passages calibrated for complexity. *Journal of Verbal Learning and Verbal Behavior, 6*(6), 851-854. doi:10.1016/S0022-5371(67)80148-8

Miller, J. R., & Kintsch, W. (1980). Readability and recall of short prose passages: A theoretical analysis. *Journal of Experimental Psychology: Human Learning and Memory, 6*(4), 335-354.

Millis, K. K., & Just, M. A. (1994). The influence of connectives on sentence comprehension. *Journal of Memory and Language, 33*(1), 128-147.

Millis, K. K., Graesser, A. C., & Haberlandt, K. (1993). The impact of connectives on the memory for expository texts. *Applied Cognitive Psychology, 7*(4), 317-339.

Millis, K., Magliano, J., & Todaro, S. (2006). Measuring discourse-level processes with verbal protocols and latent semantic analysis. *Scientific Studies of Reading, 10*(3), 225-240. doi:10.1207/s1532799xssr1003_2

Mosenthal, P. B. (1996). Understanding the strategies of document literacy and their conditions of use. *Journal of Educational Psychology, 88*(2), 314-332. doi:10.1037/0022-0663.88.2.314

Mulder, G. (2008). *Understanding causal coherence relations*. (Doctoral dissertation). Utrecht, The Netherlands: LOT.

Murphy, E. (2013). *The effect of working memory and syntactic complexity on sentence comprehension*. (Master thesis). Brandeis University, Waltham, Massachusetts. Retrieved from http://bir.brandeis.edu/bitstream/handle/10192/24970/MurphyThesis2013.pdf?sequence=1&isAllowed=y

Murray, J. D. (1995). Logical connectives and local coherence. In R. F. Lorch Jr. & E. J. O'Brien (Eds.), *Sources of coherence in reading* (pp. 107-125). Hillsdale, NJ: Lawrence Erlbaum Associates.

Murray, J. D. (1997). Connectives and narrative text: The role of continuity. *Memory & Cognition, 25*(2), 227-236.

Nagy, W. E., & Hiebert, E. H. (2010). Toward a theory of word selection. In M. L. Kamil, P. D. Pearson, E. B. Moje & P. P. Afflerbach (Eds.), *Handbook of reading research* (4th ed., pp. 388-404). New York, NY: Routledge.

Nagy, W. E., Anderson, R. C., & Herman, P. A. (1986). *The influence of word and text properties on learning from context.* (Technical Report No. 369). Champaign, Ill.: University of Illinois Center for the Study of Reading.

Nagy, W. E., Herman, P. A., & Anderson, R. C. (1985). Learning words from context. *Reading Research Quarterly, 20*(2), 233-253. doi:10.2307/747758

Nelson, J., Perfetti, C., Liben, D., & Liben, M. (2012). *Measures of text difficulty: Testing their predictive value for grade levels and student performance.* Washington, DC: Council of Chief State School Officers.

Nicenboim, B., Logačev, P., Gattei, C., & Vasishth, S. (2016). When high-capacity readers slow down and low-capacity readers speed up: Working memory and locality effects. *Frontiers in Psychology, 7*(280). doi:10.3389/fpsyg.2016.00280

Nicenboim, B., Vasishth, S., Gattei, C., Sigman, M., & Kliegl, R. (2015). Working memory differences in long-distance dependency resolution. *Frontiers in Psychology, 6*(312). doi:10.3389/fpsyg.2015.00312

Noordman, L., & Vonk, W. (1994). Text processing and its relevance for literacy. In L. Verhoeven (Ed.), *Functional literacy: Theoretical issues and educational implications* (pp. 75-92). Amsterdam: John Benjamins.

Norman, S., Kemper, S., & Kynette, D. (1992). Adults' reading comprehension: Effects of syntactic complexity and working memory. *Journal of Gerontology, 47*(4), 258-265. doi:10.1093/geronj/47.4.P258

Oakland, T., & Lane, H. B. (2004). Language, reading, and readability formulas: Implications for developing and adapting tests. *International Journal of Testing, 4*(3), 239-252.

Oller, J. W., & Jonz, J. (1994). Why cloze procedure? In J. W. Oller & J. Jonz (Eds.), *Cloze and Coherence* (pp. 1-20). Lewisburg: Bucknell University Press.

Oller, J. W., & Jonz, J. (Eds.). (1994). *Cloze and coherence*. Lewisburg: Bucknell University Press.

Oostdijk, N., Reynaert, M., Hoste, V., & Van den Heuvel, H. (2013). *SoNaR user documentation. Version 1.0.4*. Retrieved from https://ticclops.uvt.nl/SoNaR_end-user_documentation_v.1.0.4.pdf

O'Reilly, T., & McNamara, D. S. (2007). Reversing the reverse cohesion effect: Good texts can be better for strategic, high-knowledge readers. *Discourse Processes, 43*(2), 121-152.

O'Toole, J. M., & King, R. A. R. (2010). A matter of significance: Can sampling error invalidate cloze estimates of text readability? *Language Assessment Quarterly, 7*(4), 303-316.

O'Toole, J. M., & King, R. A. R. (2011). The deceptive mean: Conceptual scoring of cloze entries differentially advantages more able readers. *Language Testing, 28*(1), 127-144.

Ozuru, Y., Best, R., Bell, C., Witherspoon, A., & McNamara, D. S. (2007). Influence of question format and text availability on the assessment of expository text comprehension. *Cognition and Instruction, 25*(4), 399-438. doi:10.1080/07370000701632371

Ozuru, Y., Dempsey, K., & McNamara, D. S. (2009). Prior knowledge, reading skill, and text cohesion in the comprehension of science texts. *Learning and Instruction, 19*(3), 228-242.

Paivio, A. (1991). Dual coding theory: Retrospect and current status. *Canadian journal of psychology, 45*(3), 255-287.

Pander Maat, H. L. W. (2001). Unstressed en/and as a marker of joint relevance. In T. Sanders, J. Schilperoord & W. Spooren (Eds.), *Text representation: Linguistic and psycholinguistic aspects* (pp. 197-230). Amsterdam: John Benjamins.

Pander Maat, H. L. W. (2002). *Tekstanalyse. Wat teksten tot teksten maakt.* Bussum: Coutinho.

Pander Maat, H. L. W. (2017). Zinslengte en zinscomplexiteit: Een corpusbenadering. *Tijdschrift voor Taalbeheersing, 39*(3), 297-328. doi: 10.5117/TVT2017.3.PAND

Pander Maat, H. L. W., & Dekker, N. (2016). Tekstgenres analyseren op lexicale complexiteit met T-scan. [Using T-Scan to analyse the lexical complexity of text genres]. *Tijdschrift voor Taalbeheersing, 38*(3), 263-304. doi:10.5117/TVT2016.3.PAND

Pander Maat, H. L. W., & Sanders, T. J. M. (2006). Connectives in text. In K. Brown (Ed.), *Encyclopedia of language and linguistics* (2nd edition ed., pp. 597-607). London: Elsevier.

Pander Maat, H. L. W., Kraf, R. L., & Dekker, N. (2017). *Handleiding T-Scan. Versie 1 april 2017.* Retrieved from https://github.com/proycon/tscan/blob/master/docs/tscanhandleiding.pdf

Pander Maat, H. L. W., Kraf, R. L., Van den Bosch, A., Van Gompel, M., Kleijn, S., Sanders, T. J. M., & Van der Sloot, K. (2014). T-scan: A new tool for analyzing Dutch text. *Computational Linguistics in the Netherlands Journal, 4*, 53-74.

Parsons, T. (1990) *Events in the semantics of English: A study in subatomic semantics.* Cambridge: MIT Press.

Pearson, P. D., & Hamm, D. N. (2005). The assessment of reading comprehension: A review of practices—Past, present, and future. In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 13-69). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Perfetti, C. A. (1997). Sentences, individual differences, and multiple texts: Three issues in text comprehension. *Discourse Processes, 23*(3), 337-355. doi:10.1080/01638539709544996

Perfetti, C., & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific Studies of Reading, 18*(1), 22-37. doi:10.1080/10888438.2013.827687

Pitler, E., & Nenkova, A. (2008). Revisiting readability: A united framework for predicting text quality. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (pp.186-195). Honolulu, HI.

Pollatsek A., & Hyönä J. (2006). Processing of morphologically complex words in context: What can be learned from eye movements. In S. Andrews (Ed.), *From Inkmarks to Ideas: Current Issues in Lexical Processing* (pp. 275-298). Hove: Psychology Press.

Porter, D. (1978). Cloze procedure and equivalence. *Language Learning, 28*(2), 333-341. doi:10.1111/j.1467-1770.1978.tb00138.x

Poulsen, M., & Gravgaard, A. K. D. (2016). Who did what to whom? The relationship between syntactic aspects of sentence comprehension and text comprehension. *Scientific Studies of Reading, 20*(4), 325-338. doi:10.1080/10888438.2016.1180695

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., & Webber, B. (2008). The Penn discourse treebank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)* (pp. 2961-2968). Marrakech, Morocco.

Pressley, M., & Allington, R. L. (2015). *Reading instruction that works: The case for balanced teaching* (4th ed.). New York: Guilford Publications.

Quené, H., & Van den Bergh, H. (2004). On multi-level modeling of data from repeated measures designs: A tutorial. *Speech Communication, 43*(1), 103-121.

Quené, H., & Van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language, 59*(4), 413-425. doi:10.1016/j.jml.2008.02.002

Radach, R., Huestegge, L., & Reilly, R. (2008). The role of global top-down factors in local eye-movement control in reading. *Psychological Research, 72*(6), 675-688. doi:10.1007/s00426-008-0173-3

Rankin, E. F., & Thomas, S. (1980/1994). Contextual constraints and the construct validity of the cloze procedure. In J. W. Oller & J. Jonz (Eds.), *Cloze and Coherence* (pp. 165-175). Lewisburg: Bucknell University Press. (Reprinted from M.L. Kamil & A.J. Moe (Eds.), Perspectives on Reading: Research and Instruction (pp. 47-55). Washington, DC: National Reading Conference, 1980).

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological bulletin, 124*(3), 372-422.

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *The Quarterly Journal of Experimental Psychology, 62*(8), 1457-1506.

Rayner, K., Kambe, G., & Duffy, S. A. (2000). The effect of clause wrap-up on eye movements during reading. *The Quarterly Journal of Experimental Psychology: Section A, 53*(4), 1061-1080. doi:10.1080/713755934

Redish, J. (2000). Readability formulas have even more limitations than Klare discusses. *ACM Journal of Computer Documentation, 24*(3), 132 – 137.

Redish, J. C., & Selzer, J. (1985). The place of readability formulas in technical communication. *Technical Communication, 32*(3), 46-52.

Renkema, J. (1982). Leesbaarheidsformules. Een overzichtsartikel van de werkgroep begrijpelijkheidsonderzoek. *Massacommunicatie, 10*, 115-123.

Renkema, J. (1989). Tangconstructies, experimenteel onderzoek naar leesbaarheid en attentiewaarde. *Spektator, 18*(6), 444-462.

Robinson, C. G. (1981). Cloze procedure: A review. *Educational Research, 23*(2), 128-133. doi:10.1080/0013188810230206

Rodriguez, N., & Hansen, L. H. (1975). Performance of readability formulas under conditions of restricted ability level and restricted difficulty of materials. *The Journal of Experimental Education, 44*(1), 8-14.

Ross, S., Long, M. H., & Yano, Y. (1991). Simplification or elaboration? The effects of two types of text modifications on foreign language reading comprehension. *University of Hawaï Working Papers in English as a Second Language, 10*(2), 1-32.

Sadoski, M., Goetz, E. T., & Fritz, J. B. (1993). Impact of concreteness on comprehensibility, interest, and memory for text: Implications for dual coding theory and text design. *Journal of Educational Psychology, 85*(2), 291.

Sadoski, M., Goetz, E. T., & Rodriguez, M. (2000). Engaging texts: Effects of concreteness on comprehensibility, interest, and recall in four text types. *Journal of Educational Psychology, 92*(1), 85.

Sanders, T. J. M. (2001). Structuursignalen in informerende teksten. Over leesonderzoek en tekstadviezen. *Tijdschrift voor Taalbeheersing, 23*(1), 1-22.

Sanders, T. J. M. (2005). Coherence, causality and cognitive complexity in discourse. Paper presented at the *Proceedings/Actes SEM-05, First International Symposium on the Exploration and Modelling of Meaning*, 105-114.

Sanders, T. J. M., & Canestrelli, A. R. (2012). The processing of pragmatic information in discourse. In H. Schmid (Ed.), *Cognitive pragmatics [handbook of pragmatics, vol.4]* (pp. 201-232). Berlin: Mouton de Gruyter.

Sanders, T. J. M., & Noordman, L. (1988). Tekststructuur en begrijpelijkheid. Begrijpelijkheidscriteria van ervaren tekstbeoordelaars. *Tijdschrift voor Taalbeheersing, 10*(2), 81-92.

Sanders, T. J. M., & Noordman, L. G. M. (2000). The role of coherence relations and their linguistic markers in text processing. *Discourse Processes, 29*(1), 37-60.

Sanders, T. J. M., & Pander Maat, H. L. W. (2006). Cohesion and coherence: Linguistic approaches. In K. Brown (Ed.), *Encyclopedia of language and linguistics[vol.2]* (2nd edition ed., pp. 591-595). London: Elsevier.

Sanders, T. J. M., & Spooren, W. P. M. (2007). Discourse and text structure. In D. Geeraerts & J. Cuykens (Eds.), *Handbook of cognitive linguistics* (pp. 916-941). Oxford, UK: Oxford University Press.

Sanders, T. J. M., & Spooren, W. P. M. (2009). The cognition of discourse coherence. In J. Renkema (Ed.), *Discourse, of course: An overview of research in discourse studies* (pp. 197-212). Amsterdam: John Benjamins.

Sanders, T. J. M., Land, J., & Mulder, G. (2007). Linguistics markers of coherence improve text comprehension in functional contexts. *Information Design Journal, 15*(3), 219-235.

Sanders, T. J. M., Spooren, W. P. M., & Noordman, L. G. M. (1992). Toward a taxonomy of coherence relations. *Discourse Processes, 15*(1), 1-35.

Sanders, T. J. M., Spooren, W. P. M., & Noordman, L. G. M. (1993). Coherence relations in a cognitive theory of discourse representation. *Cognitive Linguistics, 4*(2), 93-133.

Sanford, A. J. (2006). Coherence: Psycholinguistic approach. In K. Brown (Ed.), *Encyclopedia of language and linguistics[vol.2]* (2nd edition ed., pp. 585-591). London: Elsevier.

Schilling, H. E. H., Rayner, K., & Chumbley, J. I. (1998). Comparing naming, lexical decision, and eye fixation times: Word frequency effects and individual differences. *Memory & Cognition, 26*(6), 1270-1281. doi:10.3758/BF03201199

Schwarm, S. E., & Ostendorf, M. (2005). Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics* (pp. 523-530). Association for Computational Linguistics.

Segal, E. M., Duchan, J. F., & Scott, P. J. (1991). The role of interclausal connectives in narrative structuring: Evidence from adults' interpretations of simple stories. *Discourse Processes, 14*(1), 27-54.

Shain, C., Van Schijndel, M., Futrell, R., Gibson, E., & Schuler, W. (2016). Memory access during incremental sentence processing causes reading time latency. Paper presented at the *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)* (pp.49-58). Osaka, Japan.

Shanahan, T., Kamil, M. L., & Webb Tobin, A. (1982). Cloze as a measure of intersentential comprehension. *Reading Research Quarterly, 17*(2), 229-255.

Smith, L. I. (2012). A tutorial on Principal Components Analysis. Retrieved from http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf

Spooren, W., & Sanders, T. (2008). The acquisition order of coherence relations: On cognitive complexity in discourse. *Journal of Pragmatics, 40*(12), 2003-2026.

Spyridakis, J. H. (1989a). Signaling effects: A review of the research—Part I. *Journal of Technical Writing and Communication, 19*(3), 227-240.

Spyridakis, J. H. (1989b). Signaling effects: Increased content retention and new answers—Part II. *Journal of Technical Writing and Communication, 19*(4), 395-415.

Spyridakis, J. H., & Standal, T. C. (1987). Signals in expository prose: Effects on reading comprehension. *Reading Research Quarterly, 22*(3), 285-298. doi:10.2307/747969

Stahl, S. A. (1991). Beyond the instrumentalist hypothesis: Some relationships between word meanings and comprehension. In P. Schwanenflugel (Ed.), *The psychology of word meanings* (pp. 157-178). Hillsdale, N.H.: Lawrence Erlbaum associates.

Stahl, S. A. (2003a). How words are learned incrementally over multiple exposures. *American Educator, 27*(1), 18-22.

Stahl, S. A. (2003b). Vocabulary and readability: How knowing word meanings affects comprehension. *Topics in Language Disorders, 23*(3), 241-247.

Stahl, S. A., & Nagy, W. E. (2006). *Teaching word meanings*. Mahwah, NJ: Lawrence Erlbaum associates.

Stahl, S. A., Jacobson, M. G., Davis, C. E., & Davis, R. L. (1989). Prior knowledge and difficult vocabulary in the comprehension of unfamiliar text. *Reading Research Quarterly, 24*(1), 27-43. doi:10.2307/748009

Staphorsius, G. (1994). *Leesbaarheid en leesvaardigheid. De ontwikkeling van een domeingericht meetinstrument.* (Doctoral dissertation). Arnhem: Cito.

Stine-Morrow, E. A. L., Ryan, S., & Leonard, J. S. (2000). Age differences in on-line syntactic processing. *Experimental Aging Research, 26*(4), 315-322. doi:10.1080/036107300750015714

Street, J. A., & Dąbrowska, E. (2010). More individual differences in language attainment: How much do adult native speakers of English know about passives and quantifiers? *Lingua, 120*(8), 2080-2094.

Taboada, M. (2006). Discourse markers as signals (or not) of rhetorical relations. *Journal of Pragmatics, 38*(4), 567-592.

Taboada, M. (2009). Implicit and explicit coherence relations. In J. Renkema (Ed.), *Discourse, of course: An overview of research in discourse studies* (pp. 125-138). Amsterdam: John Benjamins.

Taylor, W. L. (1953). Cloze procedure. A new tool for measuring readability. *Journalism Quarterly, 30*(4), 415-433.

Taylor, W. L. (1956/1994). Recent developments in the use of cloze procedure. In J. W. Oller & J. Jonz (Eds.), *Cloze and Coherence* (pp. 81-90). Lewisburg: Bucknell University Press. (Reprinted from Journalism Quarterly, 33, 42-48, 1956).

Temperley, D. (2007). Minimization of dependency length in written English. *Cognition, 105*(2), 300-333. doi:10.1016/j.cognition.2006.09.011

Trace, J., Brown, J. D., Janssen, G., & Kozhevnikova, L. (2017). Determining cloze item difficulty from item and passage characteristics across learner backgrounds. *Language Testing, 34*(2), 151-174. doi:10.1177/0265532215623581

Traxler, M. J., Bybee, M. D., & Pickering, M. J. (1997). Influence of connectives on language comprehension: Eye tracking evidence for incremental interpretation. *The Quarterly Journal of Experimental Psychology: Section A, 50*(3), 481-497.

Traxler, M. J., Long, D. L., Tooley, K. M., Johns, C. L., Zirnstein, M., & Jonathan, E. (2012). Individual differences in eye-movements during reading: Working memory and speed-of-processing effects. *Journal of Eye Movement Research, 5*(1), 1-16. doi:10.16910/jemr.5.1.5

Ulijn, J. M., & Strother, J. B. (1990). The effect of syntactic simplification on reading EST texts as L1 and L2. *Journal of Research in Reading, 13*(1), 38-54. doi:10.1111/j.1467-9817.1990.tb00321.x

Unsworth, S. (2005). *Child L1, child L2 and adult L2 acquisition: Differences and similarities. A study on the acquisition of direct object scrambling in Dutch.* (Doctoral Dissertation). Utrecht, The Netherlands: LOT.

Van den Bosch, A., & Berck, P. (2009). Memory-based machine translation and language modeling. *The Prague Bulletin of Mathematical Linguistics*, 91, 17-26.

Van den Bosch, A., Busser, B., Canisius, S., & Daelemans, W. (2007). An efficient memory-based morphosyntactic tagger and parser for Dutch. In P. Dirix, I. Schuurman, V. Vandeghinste & F. Van Eynde (Eds.), *Computational linguistics in the Netherlands 2006: Selected papers of the seventeenth CLIN meeting* (pp. 191-206). Utrecht, The Netherlands: LOT.

Van Dyke, J. A., Johns, C. L., & Kukona, A. (2014). Low working memory capacity is only spuriously related to poor reading comprehension. *Cognition, 131*(3), 373-403. doi:10.1016/j.cognition.2014.01.007

Van Oosten, P., Tanghe, D., & Hoste, V. (2010). Towards an improved methodology for automated readability prediction. Paper presented at the *7th Conference on International Language Resources and Evaluation (LREC 2010)*, 775-782.

Van Silfhout, G. (2014). *Fun to read or easy to understand? Establishing effective text features for educational texts on the basis of processing and comprehension research.* (Doctoral dissertation). Utrecht, The Netherlands: LOT.

Van Silfhout, G., Evers-Vermeul, J., & Sanders, T. J. M. (2014). Establishing coherence in schoolbook texts: How connectives and layout affect students' text comprehension. *Dutch Journal of Applied Linguistics, 3*(1), 1-29.

Van Silfhout, G., Evers-Vermeul, J., & Sanders, T. J. M. (2015). Connectives as processing signals: How students benefit in processing narrative and expository texts. *Discourse Processes, 52*(1), 47-76. doi:10.1080/0163853X.2014.905237

Van Silfhout, G., Evers-Vermeul, J., Mak, W. M., & Sanders, T. J. M. (2014). Connectives and layout as processing signals: How textual features affect students' processing and text representation. *Journal of Educational Psychology, 106*(4), 1036-1048. doi:10.1037/a0036293

Van Veen, R. (2011). *The acquisition of causal connectives: The role of parental input and cognitive complexity*. (Doctoral dissertation). Utrecht, The Netherlands: LOT.

Vasishth, S., & Drenhaus, H. (2011). Locality in German. *Dialogue & Discourse, 2*(1), 59-82. doi:10.5087/dad.2011.104

Vasishth, S., & Lewis, R. L. (2006). Argument-head distance and processing complexity: Explaining both locality and antilocality effects. *Language, 82*(4), 767-794.

Vennink, M. (2014). Meer begrip van tekstbegrip: Een geslaagd experiment met een activerende examentraining tekstbegrip in vwo 6. *Levende Talen Magazine, 101*(3), 10-13.

Vidal-Abarca, E., & Sanjose, V. (1998). Levels of comprehension of scientific prose: The role of text variables. *Learning and Instruction, 8*(3), 215-233.

Warren, T., Reichle, E. D., & Patson, N. D. (2011). Lexical and post-lexical complexity effects on eye movements in reading. *Journal of Eye Movement Research, 4*(1), 1-10.

Warren, T., White, S. J., & Reichle, E. D. (2009). Investigating the causes of wrap-up effects: Evidence from eye movements and E–Z reader. *Cognition, 111*(1), 132-137. doi:10.1016/j.cognition.2008.12.011

Watanabe, Y., & Koyama, D. (2008). A meta-analysis of second language cloze testing research. *Second Language Studies, 26*(2), 103-133.

White, S. J., Drieghe, D., Liversedge, S. P., & Staub, A. (2016). The word frequency effect during sentence reading: A linear or nonlinear effect of log frequency? *The Quarterly Journal of Experimental Psychology*. doi:10.1080/17470218.2016.1240813

Williams, R., & Morris, R. (2004). Eye movements, word familiarity, and vocabulary acquisition. *European Journal of Cognitive Psychology, 16*(1-2), 312-339. doi:10.1080/09541440340000196

Zakaluk, B. L., & Samuels, S. J. (1988). Toward a new approach to predicting text comprehensibility. In B. L. Zakaluk & S. J. Samuels (Eds.), *Readability: Its Past, Present and Future* (pp. 121-144). Newark, DE: International Reading Association.

Zeng-Treitler, Q., Ngo, L., Kandula, S., Rosemblat, G., Kim, H., & Hill, B. (2012). A method to estimate readability of health content. *HI-KDD '12*, Beijing, China.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological bulletin, 123*(2), 162.

Zwaan, R. A., & Rapp, D. N. (2006). Discourse comprehension. In M. J. Traxler & M. A. Grensbacher (Eds.), *Handbook of psycholinguistics* (2nd Edition ed., pp. 725-764). London: Elsevier.

Zwaan, R. A., Langston, M. C., & Graesser, A. C. (1995). The construction of situation models in narrative comprehension: An event-indexing model. *Psychological science, 6*(5), 292-297.

Appendix 1: Source materials

Table A1: List of original materials used in the study

Text	Title	Genre	Condition	Source
1	Internationaal Monetair Fonds	Edu	Coherence	Haak, J.K. van den (2002). Economie in Balans, 2e fase totaalvak havo, theorieboek 1. Amersfoort: Nijgh Versluys.
2	Produceren	Edu	Coherence	Haak, J.K. van den (2002). Economie in Balans, 2e fase totaalvak havo, theorieboek 1. Amersfoort: Nijgh Versluys.
3	Steden in de late middeleeuwen	Edu	Coherence	Memo (2012). Memo. Geschiedenis voor de tweede fase. 's-Hertogenbosch: Malmberg.
4	Wisselkoers van valuta	Edu	Coherence	Scholte, P., Janssen, K., Kuijpers, M., Voorend, P., Wevers, F. (2003). Index vmbo kgt voor de basisvorming (eerste druk). Amersfoort: ThiemeMeulenhoff.
5	Leugendetectors	Edu	Coherence	Kraaijeveld, R. (2003). Op Niveau 2 vmbo kgt basisboek. Amersfoort: ThiemeMeulenhoff.
6	Grenzen in Europa	Edu	Coherence	Kruis, M. (2006). Pincode onderbouw vmbo kgt leerboek. Groningen: Noordhoff.
7	Europese Unie	Edu	Coherence	Kunnen, L., Nonnekes, H., Reichard, A., & J. Remmers-Kamp (Redactie: Nonnekes, H., & Reichard, A. (2003). Terra vmbo kgt 1 informatieboek (tweede editie). Groningen: Wolters-Noordhoff.
8	Regiomarketing	Edu	Coherence	J.H. Bulthuis (redactie: J. Bos, J. Hofker) (1999). De Geo havo bovenbouw. Informatieboek: Regionale beeldvorming. Amsterdam: Meulenhoff.
9	Inkomen en sociale zekerheid	Edu	Coherence	Huitema, J., Peters, L., Vaart, I. van der (2004). Economisch bekeken, basisvorming vmbo-kgt handbook (zesde druk). 's-Hertogenbosch: Malmberg.
10	Vakantie	Edu	Coherence	Hooghuis, F., Nijnatten, H., Peenstra, T., Schouten, M., & Wanrooij, B. van (Eindredactie: Hooghuis, F., Nijnatten, H., & Peenstra, T.) (2003). Atlantis vmbo kgt, deel 1. Amersfoort: Thieme-Meulenhof.
11	Verhuizen	Edu	Syntax	Gerits, G. (Eindredactie: I. Hendriks) (2004). Atlantis Havo tweede fase (tweede druk). Amersfoort: Thieme-Meulenhoff.

Text	Title	Genre	Condition	Source
12	Cultuur	Edu	Syntax	Kunnen, L., Nonnekes, H., Reichard, A., & J. Remmers-Kamp (Redactie: Nonnekes, H., & Reichard, A. (2003). Terra vmbo kgt 1 informatieboek (tweede editie). Groningen: Wolters-Noordhoff.
13	De eerste Kruistocht	Edu	Syntax	Buskop, H., Dalhuisen, L., van der Geest, R., Steegh, F., & van der Waal, C. (2004). Sprekend verleden handboek 1 (druk 4). Amersfoort: Nijgh Versluys.
14	Een beroep kiezen	Edu	Syntax	Kruis, M. (2006). Pincode onderbouw vmbo kgt leerboek. Groningen: Noordhoff.
15	Magna Carta	Edu	Syntax	Memo (2012). Memo. Geschiedenis voor de tweede fase. 's-Hertogenbosch: Malmberg.
16	Egypte	Edu	Syntax	Buskop, H., Dalhuisen, L., van der Geest, R., & Steegh, F., Bastiaans, C., & de Waal, C. (1998). Sprekend verleden 2e fase Handboek A (druk 2, oplage 1). Amersfoort: Nijgh Versluys.
17	Michiel de Ruyter	Edu	Syntax	Schrover, W., & Boxtel, C. (2012). Memo handboek KGT deel 2. Groningen: Noordhoff.
18	Export in ontwikkelings-landen	Edu	Syntax	Haak, J.K. van den (2002). Economie in Balans, 2e fase totaalvak havo, theorieboek 1. Amersfoort: Nijgh Versluys.
19	Lichamelijk contact	Edu	Syntax	Mulder, E. (n.d.). Talent 4/5 havo tekstenboek. 's-Hertogenbosch: Malmberg.
20	Ontwikkelings-landen	Edu	Syntax	Van der Berg. G., Bloothoofd, T., de Boer, M., Botter, H., & van Oorschot, F. (red.) (2007). Wereldwijs vmbo kgt, deel 1 (vierde druk). 's-Hertogenbosch: Malmberg.
21	Begin van de Tweede Wereldoorlog	Edu	Lexical	Kreek, R. de, Verberne, L., Veldkamp, M., Venner, J., Bosch, A. J., Voorst, A. van, Oudheusden, J. van, Boonstra, R., Heijden C. van der, Haperen M. van (2007). Feniks, Havo tweede fase, overzicht v/d geschiedenis. Amersfoort: Thieme-Meulenhoff.
22	Beleidsplannen	Edu	Lexical	Duijm, H., Gorter, G.F. (2002). Percent havo, Totaalvak 1. Theorieboek. Amersfoort: Thieme-Meulenhoff.

Text	Title	Genre	Condition	Source
23	Huwelijken in de middeleeuwen	Edu	Lexical	Hageraats, B., van der Heyden, C., van Oudheusden, J., van de Pol, L., Raaijmakers, J., Rongen, W., Salman, J., Schuitemaker, P., van Voorst, A., & van der Kaap, A. (1998). Pharos themaboek voor de tweede fase (druk 1). Amsterdam: Meulenhoff Educatief.
24	Achteruitgang van de cultuur	Edu	Lexical	Kraaijeveld, R. (2007). Op Niveau 4/5 havo verwerkingsboek. Amersfoort: ThiemeMeulenhoff.
25	Oost- en West-Duitsland	Edu	Lexical	Ten Brinke, W.B., Broeke, J.L., Groen, H., De Jong, C., & Klauw, van der, E. (2006). De Geo VMBO KGT lesboek 1. Amersfoort: Thieme-Meulenhof.
26	Geld	Edu	Lexical	Scholte, P., Janssen, K., Kuijpers, M., Voorend, P., Wevers, F. (2003). Index vmbo kgt voor de basisvorming (eerste druk). Amersfoort: ThiemeMeulenhoff.
27	Goede en slechte economische tijden	Edu	Lexical	Hinloopen, J., Adriaansen, P., Zuiderwijk, A. (2009). Praktische economie, havo voor de 2e fase (5e druk). 's-Hertogenbosch: Malmberg.
28	Arbeidsovereen-komst	Edu	Lexical	Kruis, M. (2006). Pincode onderbouw vmbo kgt leerboek. Groningen: Noordhoff.
29	Droogte in Kenia	Edu	Lexical	Ten Brinke, W.B., Broeke, J.L., Groen, H., De Jong, C., & Klauw, van der, E. (2006). De Geo VMBO KGT lesboek 1. Amersfoort: Thieme-Meulenhof.
30	Centrum en periferie	Edu	Lexical	Van der Berg. G., Bloothoofd, T., de Boer, M., Botter, H., & van Oorschot, F. (red.) (2007). Wereldwijs vmbo kgt, deel 1 (vierde druk). 's-Hertogenbosch: Malmberg.
31	Ongedierte in huis bestrijden	Pub	Coherence	Milieu centraal (n.d.) Ongedierte in huis bestrijden [Website]. Retrieved from http://www.milieucentraal.nl/themas/ongedierte-in-huis-bestrijden
32	Veiligheid sector gas en elektriciteit	Pub	Coherence	AIVD (n.d.) Gas & elektriciteit [Website]. Retrieved from https://www.aivd.nl/onderwerpen/veiligheidsbevordering/inhoud/bedrijven/gas-en-elektriciteit
33	Nodeloze brandmeldingen	Pub	Coherence	Brandweer (n.d.) Nodeloze brandmeldingen [Brochure]. Retrieved from http://www.brandweer.nl/publish/pages/2460/121_nodelozebrandmeldingen.pdf

Text	Title	Genre	Condition	Source
34	Donorregistratie	Pub	Coherence	Donorregister (Ministerie van VWS) (n.d.) Over donorregistratie [Website]. Retrieved from http://www.donorregister.nl/over_donorregistratie/
35	Examen bromfietsrijbewijs	Pub	Coherence	CBR (n.d.) Examen doen voor het bromfietsrijbewijs [Brochure]. Retrieved from https://www.cbr.nl/downloadbrochure.pp?id=9
36	Rijden onder invloed	Pub	Coherence	Veilig verkeer Nederland (n.d.) Rijden onder invloed [Website]. Retrieved from http://www.veiligverkeernederland.nl/bob/node/10186
37	De wereld achter je eten	Pub	Coherence	Voedingscentrum (n.d.) De wereld achter je eten [schoolkrant]. Retrieved from http://www.voedingscentrum.nl/Assets/Uploads/Documents/Voedingscentrum/Professionals - Onderwijs/Onderwijs/Schoolkranttekst_Verspilling_VO.doc
38	Het landelijk fietsdiefstalregister	Pub	Coherence	RDW (n.d.) Het landelijke fietsendiefstalregister [Website]. Retrieved from http://www.rdw.nl/Particulier/Paginas/Fiets.aspx
39	Ontstaan van mist	Pub	Coherence	KNMI (n.d.) Ontstaan van mist [Brochure]. Retrieved from http://www.knmi.nl/bibliotheek/weerbrochures/FS_Zicht.pdf
40	Eenzaamheid	Pub	Coherence	GGD Midden Nederland (n.d.) Eenzaamheid [Brochure]. Retrieved from http://www.ggdmn.nl/gezond-leven/brochure-eenzaamheid.pdf
41	Brandwonden	Pub	Syntax	Brandweer (n.d.) Brandwonden [Brochure]. Retrieved from http://www.brandweer.nl/publish/pages/534/3_brandwonden_voorkomeneneersteh ulp.pdf
42	Burgerarrest en eigenhandig optreden	Pub	Syntax	Openbaar ministerie (n.d.) Burgerarrest en eigenhandig optreden [Website]. Retrieved from http://www.om.nl/onderwerpen/burgerarrest_en/
43	Snelrecht en supersnelrecht	Pub	Syntax	Openbaar ministerie (n.d.) Snelrecht en supersnelrecht [Website]. Retrieved from http://www.om.nl/onderwerpen/snelrecht_en/

Text	Title	Genre	Condition	Source
44	Digitalisering	Pub	Syntax	KNAW (n.d.) Een digitaal doordrenkte wereld [Rapport]. Retrieved from https://www.knaw.nl/nl/actueel/publicaties/digitale-geletterdheid-in-het-voortgezet-onderwijs/@@download/pdf_file/20121027.pdf
45	Drinkwater	Pub	Syntax	Rijksoverheid (n.d.) Wat is drinkwater [Website]. Retrieved from http://www.rijksoverheid.nl/onderwerpen/drinkwater/vraag-en-antwoord/wat-is-drinkwater.html
46	Diabetes	Pub	Syntax	Sugarkids (n.d.) Wat is diabetes [Website]. Retrieved from http://www.sugarkids.nl/diabetes/zorgwijzer-
47	Bronnen van energie	Pub	Syntax	Milieu centraal (n.d.) Bronnen van energie [Website]. Retrieved from http://www.milieucentraal.nl/themas/bronnen-van-energie
48	Kleine blusmiddelen	Pub	Syntax	Brandweer (n.d.) Kleine blusmiddelen [Brochure]. Retrieved from http://www.brandweer.nl/publish/pages/13/1_rookmeldersenbrandblussers.pdf
49	Milieubewust eten	Pub	Syntax	Milieu centraal (n.d.) Milieubewust eten [Website]. Retrieved from http://www.milieucentraal.nl/themas/milieubewust-eten
50	Rookmelders	Pub	Syntax	Brandweer (n.d.) Rookmelders [Brochure]. Retrieved from http://www.brandweer.nl/publish/pages/13/1_rookmeldersenbrandblussers.pdf
51	Het Koninklijk Nederlands Meteorologisch Instituut (KNMI)	Pub	Lexical	KNMI (n.d.) Het Koninklijk Nederlands Meteorologisch instituut [Website]. Retrieved from http://www.knmi.nl/over_het_knmi/
52	Moord en doodslag	Pub	Lexical	Politie (n.d.) Moord en doodslag [Website]. Retrieved from http://www.politie.nl/onderwerpen/moord-doodslag.html
53	Rabiës bij vleermuizen	Pub	Lexical	RIVM (n.d.) Rabiës bij vleermuizen [Brochure]. Retrieved from http://toolkits.loketgezondleven.nl/site_files/php/download.php?location=rabies&file=005228-vleermuis-rabies-folder-a5-v10.pdf

Text	Title	Genre	Condition	Source
54	Tolerantie homoseksualiteit	Pub	Lexical	Inspectie van het Onderwijs (n.d.) Homoseksualiteit [Brochure]. Retrieved from http://www.onderwijsinspectie.nl/binaries/content/assets/Actueel_publicaties/2009/Anders+zijn+is+van+iedereen+%28printversie%29.pdf
55	Werkende jongeren	Pub	Lexical	Rijksoverheid (n.d.) Werkende jongeren [Website]. Retrieved from http://www.rijksoverheid.nl/onderwerpen/jongeren-en-werk/werkende-jongeren
56	Vandalisme	Pub	Lexical	Politie (n.d.) Vandalisme [Website]. Retrieved from http://www.politie.nl/onderwerpen/vandalisme.html
57	Staatsbosbeheer en jeugd	Pub	Lexical	Staatsbosbeheer (n.d.) Staatsbosbeheer en jeugd [Website]. Retrieved from http://www.staatsbosbeheer.nl/Nieuws en achtergronden/Themas/Jeugd en natuur/Staatsbosbeheer en jeugd.aspx
58	Griep en verkoudheid	Pub	Lexical	RIVM (n.d.) Griep en verkoudheid [Brochure]. Retrieved from http://toolkits.loketgezondleven.nl/site_files/php/download.php?location=griep&file=005927-toolkit-griep-folder-kleur-a5-v3def-lr-web.pdf
59	Bankrekening-fraude	Pub	Lexical	Politie (n.d.) Bankrekeningfraude [Website]. Retrieved from http://www.politie.nl/onderwerpen/bankrekeningfraude.html
60	KNMI waarschuwing zicht	Pub	Lexical	KNMI (n.d.) KNMI risicosignalering zicht [Brochure]. Retrieved from http://www.knmi.nl/bibliotheek/weerbrochures/FS_Zicht.pdf

** Edu = Educational textbook text; Pub = Public information text*

Appendix 2: Construction manual HyTeC-cloze

The following procedure can be used to create cloze tests to assess text comprehension. The procedure is developed especially for Dutch. The procedure should be transferable to other languages, though some conditions that are tailored to Dutch might be irrelevant and other language dependent conditions might be missing. Scholars should check whether all conditions are appropriate for their target language, given the rationale presented in Chapter 2.

The HyTeC-cloze procedure is a 3-step-procedure. Step I consists of selecting all words that are gap candidates. The selection of candidates is done via a reversed procedure, meaning all words are candidates as long as they are not excluded based on the heuristics presented in Chapter 2, Section 3.3. For transparency reasons, the restrictions are divided in three categories: 1. General; 2. Heuristic 1; 3. Heuristic 2. Step II consists of determining the number of possible cloze versions and distributing the candidates among them. Finally, Step III consists of selecting two cloze versions out of all possible versions.

Step I: Selection of candidates

- The following words are <u>not</u> candidates:

  General
  - Manipulated words (words that are not identical over text versions)
  - Title and first sentence of the text

  Heuristic 1 (predictable)
  - Articles
  - Prepositions
  - Relative and interrogative pronouns
  - Interjections
  - Copula, auxiliary verbs and modal verbs
  - The conjunct '*en*'[and]
  - Non-infinitive phrasal verbs
  - Parts of multi-word-units/common expressions
  - Antonym pairs

  Heuristic 2 (unpredictable)
  - First use of a name or technical term
  - Numbers (incl. dates)

  - Units of measurement
  - Cardinal directions
  - Hyphenated words (except for words with suspended hyphens)
  - Abbreviations (excl. names)

- All remaining words are candidates. Make these words **boldfaced**.

Step II: Creating cloze versions

1. Count the number of words of the text and determine how many gaps are needed ('number of gaps' = 'number of words' * 0.1). If the number of words differs between versions, use the highest number.
2. Put an asterisk (*) after lemmas that occur multiple times as a candidate.
3. If the lemma makes up X% of the candidates, it can maximally make up X% of the gaps within a version as well (see Point 10 below).
4. Highlight all words that are the only candidate in their sentence with yellow. These words will become gaps in all cloze versions.
5. Subtract the number of candidates highlighted with yellow from the necessary gaps to calculate how many gaps are still to be determined.
6. Count the remaining candidates.
7. Divide the number of remaining candidates by the number of gaps to be chosen to see how many versions can be made. (e.g., if there are 155 candidates and 36 remaining gaps, 155/36 = 4.30. So we can either create four versions, but they will have too many gaps (155/4 = 38.75) or we can create five versions which will have too few gaps (155/5 = 31). We choose to create four versions because then only 2 or 3 gaps have to be deselected, while with five versions five additional gaps have to be selected.)
8. Start with one text version[1] and highlight the different cloze versions in the text mechanically. (e.g., With 4 versions: the first, fifth, ninth word are highlighted green, the second, sixth, tenth word are highlighted red, etc.)
9. Highlight in the other text version the same words with the same colors as were used for Point 8. (i.e., for each cloze version the same words are deleted in the easy and difficult text version).
10. Check whether Point 3 above holds for each version. If not, deselect lemmas in such a way that the remaining few are distributed evenly over the text. Remove the asterisks.
11. → When there are exactly enough gaps per cloze version:
    - Proceed directly to Step III.

[1] Alternate between texts whether to start with the easy or the difficult text version.

→ When there are too few gaps per version (i.e., additional selection):
- Locate in each cloze version where the distance between gaps is the largest. Select candidates from other versions that maximally decrease these differences. These candidates thus become gaps in multiple versions.

→ When there are too many gaps per version (i.e., deselection):
- Locate in each cloze version the clauses that contain multiple gaps. Deselect one of the gaps by determining the distance between the previous gap and the next and by choosing the gap that is in the middle of the previous and next gap. If there are still too many, do the same for sentences with multiple gaps.

Step III: Choosing cloze versions

1. If there are two possible cloze versions, both versions will be used.
2. If there are more than two possible versions, use a random number generator (e.g., www.random.org) to select two versions.

Appendix 3: Construction manual HyTeC-cloze (Dutch)

De volgende procedure kan gebruikt worden om clozetoetsen te produceren die tekstbegrip meten. De procedure is ontwikkeld voor het Nederlands en volgt drie stappen. In Stap I worden alle woorden geselecteerd die als clozegat kunnen dienen: de 'kandidaten'. Deze selectie vindt via een omgekeerde procedure plaats. Alle woorden zijn kandidaat zolang ze niet op basis van de restricties in hoofdstuk 2 (§3.3) uitgesloten worden. De uitzondering zijn verdeeld in drie categorieën: 1. Algemeen; 2. Heuristiek 1; 3. Heuristiek 2. In Stap II worden het aantal mogelijke clozeversies per tekst bepaald en worden de kandidaten verspreid over deze versies. Tot slot worden in Stap III twee clozeversies geselecteerd voor het onderzoek.

Stap I: Kandidaatselectie

- De volgende woorden zijn geen kandidaten:

  Algemeen
  - Gemanipuleerde woorden (woorden die verschillen tussen tekstversies)
  - Titel en eerste zin van de tekst

  Heuristiek 1 (voorspelbaarheid)
  - Lidwoorden
  - Voorzetsels
  - Betrekkelijke en vragende voornaamwoorden
  - Tussenwerpsels
  - Koppelwerkwoorden, hulpwerkwoorden en modale werkwoorden
  - Het additieve connectief 'en'
  - Meerdelige werkwoorden (m.u.v. infinitieven)
  - Vaste combinaties en uitdrukkingen ('een mening hebben, trouw blijven')
  - Antoniemkoppels ('goed en kwaad')

  Heuristiek 2 (onvoorspelbaarheid)
  - Eerste keer dat een naam of technische term gebruikt wordt
  - Getallen (incl. data)
  - Maten
  - Wind(richtingen)
  - Woorden met *verbindings*streepjes (m.u.v. woorden met weglatingsstreepjes)

- Afkortingen (m.u.v. namen)

- Alle overgebleven woorden zijn kandidaten. Maak deze woorden **vet**.

Stap II: Creatie van clozeversies

1. Tel het aantal woorden in de tekst en bepaal hoeveel clozegaten er nodig zijn. Het aantal gewenste gaten is 10% van het aantal woorden. Wanneer het aantal woorden per tekstversie verschillend is, houd dan het hoogste woordaantal aan.
2. Plaats een asterisk (*) achter lemma's die vaker dan één keer voorkomen als kandidaat.
3. Wanneer X% van de kandidaten hetzelfde lemma betreft, mag dat lemma ook maar maximaal X% van de gaten representeren binnen een versie (zie Punt 10).
4. Markeer met de gele 'highlight' van Word alle woorden die de enige kandidaat zijn binnen hun zin. Deze woorden zullen een gat worden in elke clozeversie.
5. Trek het aantal gemarkeerde gaten af van het aantal benodigde gaten uit punt 1. Dit is het aantal gaten wat nog nodig is.
6. Tel het aantal overgebleven kandidaten.
7. Deel het aantal overgebleven kandidaten door het nog te bepalen aantal gaten om zo te zien hoeveel clozeversies er gemaakt kunnen worden. Kies het meest efficiënte aantal. (Bijv. als er 155 kandidaten zijn en 36 overgebleven gaten, 155/36 = 4.30. We kunnen dan of vier versies kunnen maken, maar dan hebben we iets te veel gaten per versie (155/4 = 38.75 gaten) óf we kunnen vijf versies maken maar dan hebben we te weinig gaten (155/5 = 31). We kiezen voor vier versies omdat we dan slecht 2 of 3 gaten moeten deselecteren, terwijl we bij vijf versies 5 extra gaten moeten selecteren.)
8. Start met een van de tekstversies[1] en markeer de verschillende clozeversies met verschillende kleuren door op een mechanische manier de overgebleven kandidaten te verdelen. (D.w.z. met 4 clozeversies worden de 1^e, 5^e en 9^e kandidaat groen (= versie 1), de 2^e, 6^e en 10^e kandidaat worden blauw, etc.)
9. Markeer in de andere tekstversie dezelfde woorden met dezelfde kleur als je in de andere tekstversie hebt gedaan bij Punt 8. Voor iedere clozeversie geldt dus dat dezelfde woorden gaten worden in de moeilijke en makkelijke tekstversie.
10. Controleer of elke versie voldoet aan Punt 3. Zo niet, deselecteer deze lemma's op zo'n manier dat de overbleven paar regelmatig verspreid zijn over de tekst. Verwijder de asterisken.
11. → Wanneer er precies genoeg gaten zijn per clozeversie:
    - Ga door naar Stap III.

[1] Wissel per tekst af of je met de makkelijke of moeilijke tekstversie begint.

→ Wanneer er te weinig gaten zijn per clozeversie:
- Dan moeten er extra gaten worden geselecteerd uit kandidaten die al aan een andere versie zijn toegewezen.
  Bepaal voor elke versie waar de afstand tussen twee gaten het grootst is. Selecteer kandidaten van andere versies die deze afstand het meest verminderen. Deze woorden worden dus een gat in meer dan één versie.

→ Wanneer er te veel gaten zijn per clozeversie:
- Dan moeten er gaten worden gedeselecteerd.
  Bepaal voor elke versie waar er meerdere gaten in dezelfde deelzin staan. Deselecteer er één door de afstand tussen het vorige en het volgende gat te bepalen. Deselecteer het gat wat het meest in het midden ligt. Selecteer kandidaten van andere versies die deze afstand het meest verminderen. Als er nog steeds te veel gaten zijn, doe dan hetzelfde voor hele zinnen.

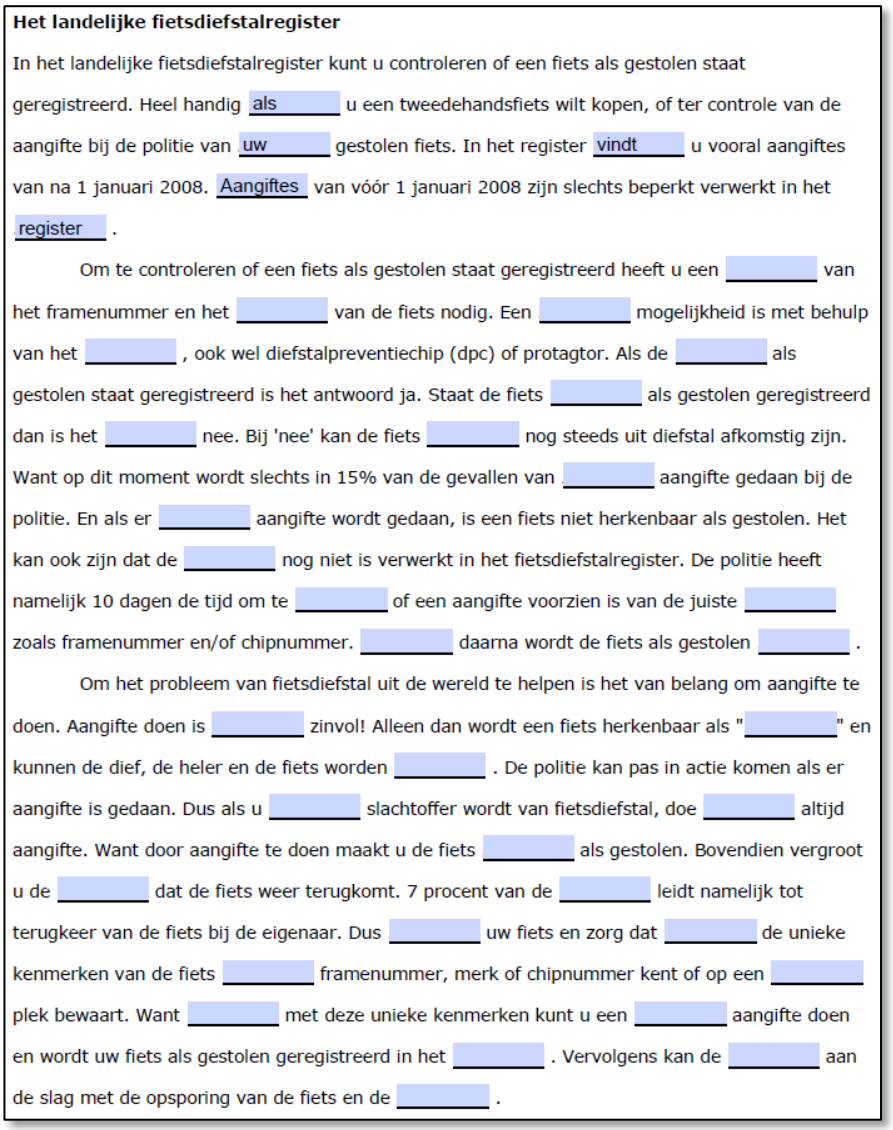
Stap III: Selectie clozeversie

1. Wanneer er slechts twee clozeversies mogelijk zijn, worden beide versies gebruikt.
2. Wanneer er meer dan twee clozeversies mogelijk zijn, worden er willekeurig twee versies geselecteerd met behulp van een randomisatie tool (bijv. www.random.org).

Appendix 4: Example HyTeC-cloze test

Figure A4 shows a HyTeC-cloze test as presented on screen. Fields were given a lavender background color and were underlined. All fields would be blank when the participant started.

Figure A4: Example of HyTeC-cloze test



Het landelijke fietsdiefstalregister

In het landelijke fietsdiefstalregister kunt u controleren of een fiets als gestolen staat geregistreerd. Heel handig als u een tweedehandsfiets wilt kopen, of ter controle van de aangifte bij de politie van uw gestolen fiets. In het register vindt u vooral aangiftes van na 1 januari 2008. Aangiftes van vóór 1 januari 2008 zijn slechts beperkt verwerkt in het register .

Om te controleren of een fiets als gestolen staat geregistreerd heeft u een _____ van het framenummer en het _____ van de fiets nodig. Een _____ mogelijkheid is met behulp van het _____ , ook wel diefstalpreventiechip (dpc) of protagtor. Als de _____ als gestolen staat geregistreerd is het antwoord ja. Staat de fiets _____ als gestolen geregistreerd dan is het _____ nee. Bij 'nee' kan de fiets _____ nog steeds uit diefstal afkomstig zijn. Want op dit moment wordt slechts in 15% van de gevallen van _____ aangifte gedaan bij de politie. En als er _____ aangifte wordt gedaan, is een fiets niet herkenbaar als gestolen. Het kan ook zijn dat de _____ nog niet is verwerkt in het fietsdiefstalregister. De politie heeft namelijk 10 dagen de tijd om te _____ of een aangifte voorzien is van de juiste _____ zoals framenummer en/of chipnummer. _____ daarna wordt de fiets als gestolen _____ .

Om het probleem van fietsdiefstal uit de wereld te helpen is het van belang om aangifte te doen. Aangifte doen is _____ zinvol! Alleen dan wordt een fiets herkenbaar als "_____" en kunnen de dief, de heler en de fiets worden _____ . De politie kan pas in actie komen als er aangifte is gedaan. Dus als u _____ slachtoffer wordt van fietsdiefstal, doe _____ altijd aangifte. Want door aangifte te doen maakt u de fiets _____ als gestolen. Bovendien vergroot u de _____ dat de fiets weer terugkomt. 7 procent van de _____ leidt namelijk tot terugkeer van de fiets bij de eigenaar. Dus _____ uw fiets en zorg dat _____ de unieke kenmerken van de fiets _____ framenummer, merk of chipnummer kent of op een _____ plek bewaart. Want _____ met deze unieke kenmerken kunt u een _____ aangifte doen en wordt uw fiets als gestolen geregistreerd in het _____ . Vervolgens kan de _____ aan de slag met de opsporing van de fiets en de _____ .

Appendix 5: Quantitative checks lexical manipulation

The lexical manipulation was quantitatively checked to unsure that only the intended linguistic features differed between texts versions. Anovas were performed using text level data.

Table A5: Anova results

Feature	**Anova**
Word frequency (corrected for compound nouns and names)	$F(1,38) = 9.658$, $p = .004$
Frequency Top 1000	$F(1,38) = 7.119$, $p = .011$
Frequency Top 2000	$F(1,38) = 8.693$, $p = .005$
Frequency Top 3000	$F(1,38) = 9.593$, $p = .004$
Frequency Top 5000	$F(1,38) = 7.136$, $p = .011$
Frequency Top 10000	$F(1,38) = 5.662$, $p = .022$
Frequency Top 20000	$F(1,38) = 2.833$, $p = .101$
Word prevalence (z-score)	$F(1,38) = 4.942$, $p = .032$
Word length (in letters)	$F(1,38) = 1.606$, $p = .213$
Morphemes per word	$F < 1$
Nominalizations	$F < 1$
Type-token-ratio	$F < 1$
Concrete nouns (strict)	$F < 1$
Universal nouns	$F < 1$

Appendix 6: SDL per text and text version

The maximum SDL for each text was calculated by summing the maximum SDL of each sentence in the text and dividing it through the number of sentences. Calculations were performed automatically by T-Scan. All texts were used in the cloze experiments; texts 12, 15, 16 and 20 were used in the eye-tracking experiment.

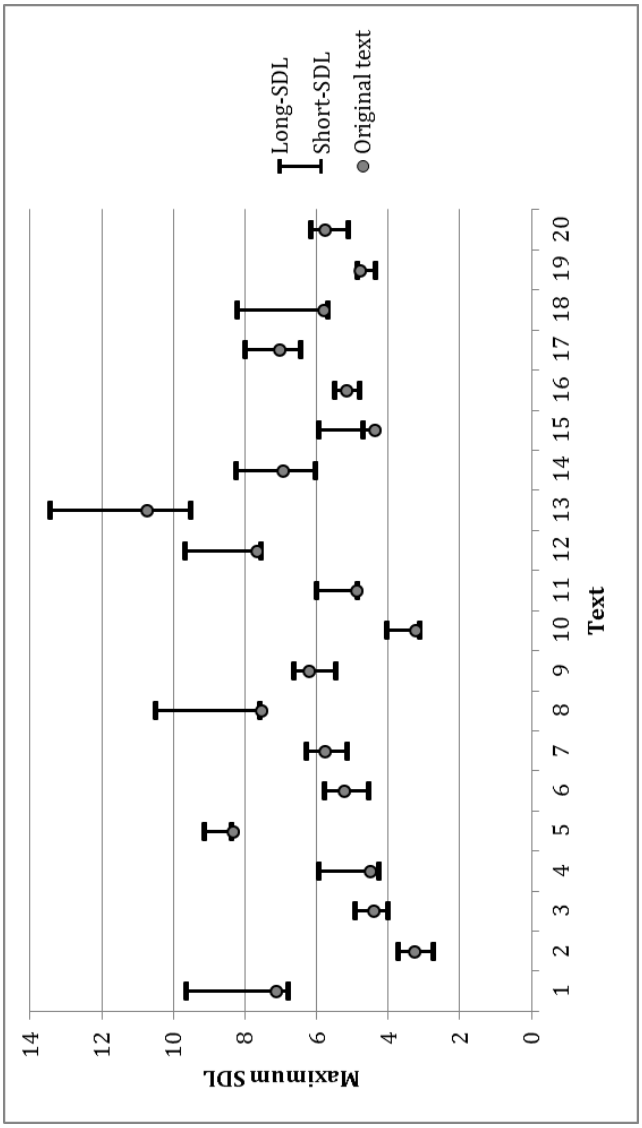


Figure A6: Average maximum SDL per text and text version[a]

[a] *Calculated over the whole text, including sentences that were not manipulated.*

Appendix 7: Connectives per text and text version

The number of connectives in each text was determined using T-Scan. The number was standardized to connectives per clause.

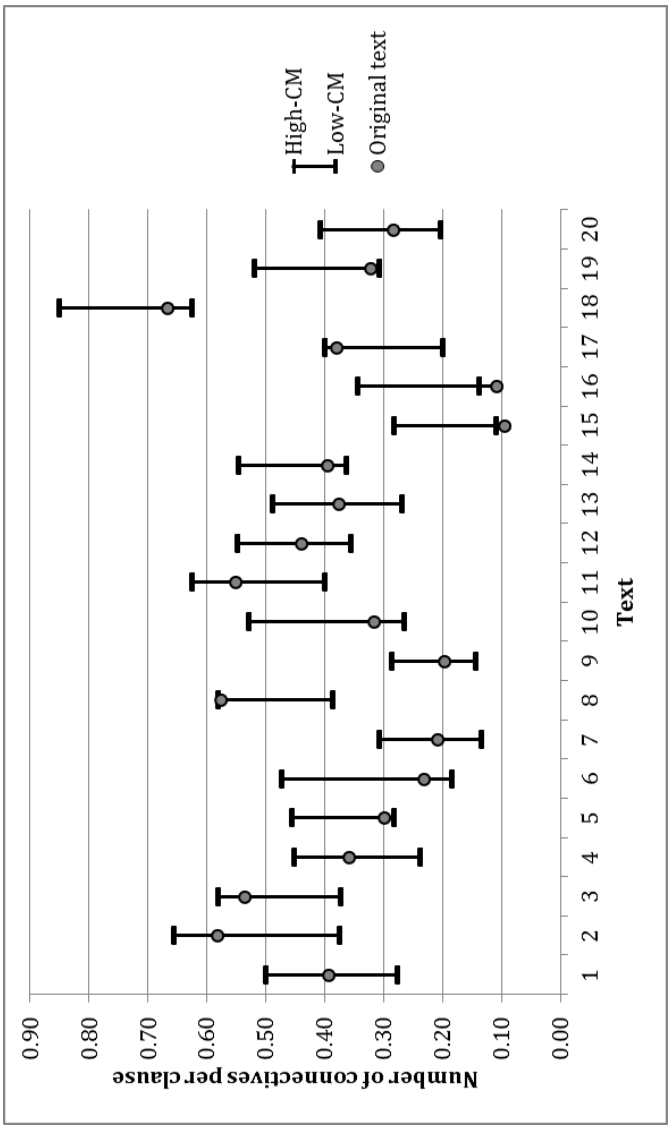


Figure A7: Average number of connectives per text and text version[a]

[a] *Calculated over the whole text, including connectives that were not manipulated.*

Appendix 8: Example text coherence marking

Dutch:
Donorregistratie

In het Donorregister kunt u laten vastleggen of u uw organen na uw overlijden wel of niet beschikbaar stelt voor transplantatie. U kunt er ook voor kiezen uw nabestaanden of één specifieke persoon te laten beslissen na uw overlijden. U kunt uw beslissing alleen vastleggen door het invullen en versturen van het donorformulier. Dit kan online of per post.

Het laten registreren van uw keuze in het Donorregister geeft duidelijkheid en zekerheid voor iedereen die bij orgaan- en weefseldonatie betrokken is, zoals potentiële donoren, uw naasten, maar ook artsen en verpleegkundigen. Registratie in het Donorregister is niet verplicht. Maar als uw keuze niet geregistreerd staat, betekent dit dat uw nabestaanden na uw overlijden moeten beslissen of u donor bent of niet.

Niet iedereen kan zich inschrijven in het Donorregister: u moet minstens twaalf jaar zijn; en daarnaast moet u ingeschreven staan bij een Nederlandse gemeente. Op dit moment staan ruim vijf miljoen keuzes in het Donorregister vastgelegd. De meeste geregistreerden geven toestemming voor donatie, al dan niet met uitsluitingen. Als u uw keuze heeft laten vastleggen, kunt u deze altijd wijzigen. Daarvoor moet u een nieuw donorformulier invullen.

Minderjarigen kunnen vanaf hun twaalfde hun wens in het Donorregister laten opnemen. Ouders of voogden hoeven hiervoor géén toestemming te verlenen. Maar als minderjarigen instemmen met donatie en voor hun zestiende overlijden, kunnen ouders of voogden alsnog weigeren. Ouders hebben namelijk een vetorecht en zonder hun goedkeuring worden er dus geen organen of weefsels uitgenomen. Als de ouders of voogden in zo'n geval niet bereikbaar zijn, wordt de geregistreerde wens van de minderjarige uitgevoerd. Maar het vetorecht geldt andersom niet. Ouders of voogden kunnen dus géén toestemming verlenen voor donatie als de minderjarige zelf heeft laten vastleggen juist geen donor te willen zijn. Vanaf zestien jaar heeft iemand volledige beslissingsbevoegdheid en dan geldt de eigen wilsverklaring.

English:
Donor registration

In the Donor Register you can record whether you want your organs to be available for transplantation after your death. You can also choose to have your relatives or one specific person decide after you die. You can record your decision by filling out and sending in the donor form. This can be done online or by regular mail.

The registration of your choice in the Donor Register provides clarity and certainty for all those involved in organ and tissue donation, such as potential donors, your loved ones, but also doctors and nurses. Registration in the Donor Register is not mandatory. However, if your choice is not registered, it means that after your death your relatives must decide whether to donate your organs or not.

Not everyone can register in the Donor Register: you must be at least twelve years old and in addition you must be a registered citizen of a Dutch municipality. At present, more than five million choices are recorded in the Donor Register. Most of the registered people give permission for donation, with or without exclusions. If you register your choice, you can always change it. To do so you need to fill out a new donor form.

Minors can record their preference in the Donor Registers from age twelve. Parents or guardians do not have to consent for this. However, if minors agree to become a donor and die before they are sixteen, parents or guardians can still refuse. That is: parents have the veto and therefore without their permission the organs or tissues are not harvested. If the parents or guardians cannot be reached, the registered choice of the minor is executed. There is, however, no veto for the reversed situation. So, parents or guardians cannot consent to donation if the minor has registered against donation. From age sixteen onward, people have full decision-making authority and then their own declaration of intention applies.

Appendix 9: Additional relational analyses coherence marking

Table A9.1: Final model relation level analysis with list and elaboration subcategorization

Random effects	**Estimates**	**SE**	***Z***	***p***	
School	0.033	0.013	2.538	<.001[a]	
School: Student	0.184	0.015	12.267	<.001[a]	
Sentence	0.045	0.007	6.429	<.001[a]	
Fixed effects	**Estimates**	**SE**	***Z***	***p***	**Odds ratio**
Intercept	-0.245	0.146	-1.678	.093	0.783
Coherence marking: High-CM	0[b]				
Coherence marking: Low-CM	0.175	0.065	2.692	.007	1.191
Coherence type: List	0[b]				
Coherence type: Elaboration	-0.228	0.076	-3.000	.003	0.796
Coherence type: Causal	-0.035	0.050	-0.700	.484	0.966
Coherence type: Contrast	-0.287	0.055	-5.218	<.001	0.751
Coherence type: Temporal	-0.268	0.088	-3.045	.002	0.765
Low-CM * Elaboration	-0.171	0.108	-1.583	.113	0.843
Low-CM * Causal	-0.232	0.072	-3.222	.001	0.793
Low-CM * Contrast	-0.382	0.079	-4.835	<.001	0.682
Low-CM * Temporal	-0.250	0.125	-2.000	.046	0.779
Education level: pre-voc. low	0[b]				
Education level: pre-voc. medium	0.357	0.101	3.535	<.001	1.429
Education level: pre-voc. high	0.586	0.102	5.745	<.001	1.797
Education level: general	0.805	0.122	6.598	<.001	2.237
Education level: pre-uni.	0.901	0.138	6.529	<.001	2.462
Grade 8	0[b]				
Grade 9	0.096	0.048	2.000	.046	1.101
Grade 10[c]	0.187	0.105	1.781	.075	1.206
Reading ability	0.296	0.031	9.548	<.001	1.344
Reading test: R	0[b]				
Reading test: V	-0.530	0.104	-5.096	<.001	0.589
Genre: Public information	0[b]				
Genre: Educational textbook	0.119	0.026	4.577	<.001	1.126

[a] One-sided. [b] Set as reference level. [c] Unbalanced, no 10th grade pre-vocational students were present in the sample.

Table A9.2: Final model relation level analysis with causal objective and subjective subcategorization

Random effects	**Estimates**	**SE**	***Z***	***p***	
School	0.031	0.013	2.385	<.001[a]	
School: Student	0.188	0.015	12.533	<.001[a]	
Sentence	0.038	0.007	5.429	<.001[a]	
Fixed effects	**Estimates**	**SE**	***Z***	***p***	**Odds ratio**
Intercept	-0.368	0.141	-2.610	.009	0.692
Coherence marking: High-CM	0[b]				
Coherence marking: Low-CM	-0.074	0.044	-1.682	.093	0.929
Coherence type: Causal-objective	0[b]				
Coherence type: Causal-subjective	0.164	0.044	3.727	<.001	1.178
Coherence type: Contrast	-0.159	0.044	-3.614	<.001	0.853
Coherence type: Temporal	-0.131	0.081	-1.617	.106	0.877
Coherence type: Additive	0.069	0.045	1.533	.125	1.071
Low-CM * Causal-subjective	0.026	0.062	0.419	.675	1.026
Low-CM * Contrast	-0.135	0.063	-2.143	.032	0.874
Low-CM * Temporal	-0.004	0.114	-0.035	.972	0.996
Low-CM * Additive	0.177	0.065	2.723	.006	1.194
Education level: pre-voc. low	0[b]				
Education level: pre-voc. medium	0.349	0.101	3.455	<.001	1.418
Education level: pre-voc. high	0.566	0.101	5.604	<.001	1.761
Education level: general	0.791	0.122	6.484	<.001	2.206
Education level: pre-uni.	0.890	0.137	6.496	<.001	2.435
Grade 8	0[b]				
Grade 9	0.099	0.048	2.063	.039	1.104
Grade 10[c]	0.180	0.104	1.731	.083	1.197
Reading ability	0.298	0.031	9.613	<.001	1.347
Reading test: R	0[b]				
Reading test: V	-0.546	0.103	-5.301	<.001	0.579
Genre: Public information	0[b]				
Genre: Educational textbook	0.160	0.025	6.400	<.001	1.174

[a] One-sided. [b] Set as reference level. [c] Unbalanced, no 10th grade pre-vocational students were present in the sample.

Table A9.3: Final model relation level analysis with causal forward and backward order subcategorization

Random effects	**Estimates**	**SE**	***Z***	***p***	
School	0.031	0.013	2.385	<.001[a]	
School: Student	0.187	0.015	12.467	<.001[a]	
Sentence	0.041	0.007	5.857	<.001[a]	
Fixed effects	**Estimates**	**SE**	***Z***	***p***	**Odds ratio**
Intercept	-0.397	0.141	-2.816	.005	0.672
Coherence marking: High-CM	0[b]				
Coherence marking: Low-CM	-0.075	0.041	-1.829	.067	0.928
Coherence type: Causal-forward	0[b]				
Coherence type: Causal-backward	0.261	0.043	6.070	<.001	1.298
Coherence type: Contrast	-0.127	0.043	-2.953	.003	0.881
Coherence type: Temporal	-0.113	0.080	-1.413	.158	0.893
Coherence type: Additive	0.103	0.044	2.341	.019	1.108
Low-CM * Causal-backward	0.038	0.061	0.623	.533	1.039
Low-CM * Contrast	-0.133	0.061	-2.180	.029	0.875
Low-CM * Temporal	-0.002	0.113	-0.018	.986	0.998
Low-CM * Additive	0.179	0.063	2.841	.004	1.196
Education level: pre-voc. low	0[b]				
Education level: pre-voc. medium	0.353	0.101	3.495	<.001	1.423
Education level: pre-voc. high	0.572	0.101	5.663	<.001	1.772
Education level: general	0.798	0.121	6.595	<.001	2.221
Education level: pre-uni.	0.897	0.137	6.547	<.001	2.452
Grade 8	0[b]				
Grade 9	0.100	0.048	2.083	.037	1.105
Grade 10[c]	0.181	0.104	1.740	.082	1.198
Reading ability	0.296	0.031	9.548	<.001	1.344
Reading test: R	0[b]				
Reading test: V	-0.543	0.103	-5.272	<.001	0.581
Genre: Public information	0[b]				
Genre: Educational textbook	0.135	0.025	5.400	<.001	1.145

[a] One-sided. [b] Set as reference level. [c] Unbalanced, no 10th grade pre-vocational students were present in the sample.

Table A9.4: Final model relation level analysis with causal objective-forward, objective-backward, subjective-forward and subjective-backward subcategorization

Random effects	**Estimates**	**SE**	***Z***	***p***	
School	0.031	0.013	2.385	<.001[a]	
School: Student	0.188	0.015	12.533	<.001[a]	
Sentence	0.040	0.007	5.714	<.001[a]	
Fixed effects	**Estimates**	**SE**	***Z***	***p***	**Odds ratio**
Intercept	-0.430	0.143	-3.007	.003	0.651
Coherence marking: High-CM	0[b]				
Coherence marking: Low-CM	-0.071	0.054	-1.315	.106	0.931
Coherence type: Causal-obj.-forward	0[b]				
Coherence type: Causal-obj.-backward	0.216	0.068	3.176	.001	1.241
Coherence type: Causal-subj.-forward	0.085	0.061	1.393	.164	1.089
Coherence type: Causal-subj.-backward	0.333	0.055	6.055	<.001	1.395
Coherence type: Contrast	-0.093	0.049	-1.898	.058	0.911
Coherence type: Temporal	-0.077	0.083	-0.928	.353	0.926
Coherence type: Additive	0.136	0.050	2.720	.007	1.146
Low-CM * Causal-obj.-backward	-0.002	0.096	-0.021	.983	0.998
Low-CM * Causal-subj.-forward	-0.012	0.085	-0.141	.888	0.988
Low-CM * Causal-subj.-backward	0.048	0.077	0.623	.533	1.049
Low-CM * Contrast	-0.138	0.070	-1.971	.049	0.871
Low-CM * Temporal	-0.007	0.118	-0.059	.953	0.993
Low-CM * Additive	0.175	0.071	2.465	.014	1.191
Education level: pre-voc. low	0[b]				
Education level: pre-voc. medium	0.351	0.101	3.475	<.001	1.420
Education level: pre-voc. high	0.569	0.101	5.634	<.001	1.766
Education level: general	0.795	0.122	6.516	<.001	2.214
Education level: pre-uni.	0.894	0.137	6.526	<.001	2.445
Grade 8	0[b]				
Grade 9	0.100	0.048	2.083	.037	1.105
Grade 10[c]	0.181	0.104	1.740	.082	1.198
Reading ability	0.297	0.031	9.581	<.001	1.346
Reading test: R	0[b]				
Reading test: V	-0.545	0.103	-5.291	<.001	0.580
Genre: Public information	0[b]				
Genre: Educational textbook	0.143	0.026	5.500	<.001	1.154

[a] One-sided. [b] Set as reference level. [c] Unbalanced, no 10th grade pre-vocational students were present in the sample.

Appendix 10: Descriptions of linguistic features

Table A10: Descriptions of the linguistic features in the U-Read model and the optimal model

Predictor	T-Scan label	Description
2nd person pronouns	pers_vnw2_d	Number of second person pronouns per 1000 words
Adjectival past participle	vd_bv_dz	Number of adjectival past participles per clause (e.g., The *overbooked* flight)
Concrete nouns	conc_nw_ruim_p	Proportion of concrete nouns (broad definition, includes nouns referring to: persons, organisms, artifacts, places, times and measures)
Content words per clause	inhwrd_dz_zonder_abw	Content words per clause (without adverbs)
Max SDL	al_max	Average maximum syntactic dependency length within sentences. SDL is calculated as the number of words between a syntactic head and its dependent (e.g., verb-subject). The length of the longest dependency in each sentence is taken and averaged over all sentences in the text.
Nominalizations	nom_d	Number of nominalizations per 1000 words
Small conjuncts	extra_kconj_per_zin	Number of small, verbless, conjuncts per sentence (e.g., *Dick and Harry* went to the supermarket)
State verbs	Toestww_d	Number of state verbs (i.e. non-dynamic, non-control verbs; Martin & Maks, 2005) per 1000 words
Universal nouns	alg_nw_d	Number of universal nouns per 1000 words (e.g., *method, idea*)
Word frequency	wrd_freq_log_zn_corr	Word frequency of all content words based on the SUBTLEX-NL corpus (per billion words log-transformed), without names and the frequency of compound nouns was corrected by taking the frequency of the head morpheme(s).

Appendix 11: Optimal multilevel readability model

The U-Read model presented in Section 4.2 includes the five strongest predictors. Adding additional predictors does not really improve the explanatory power of the model (i.e., R^2-change < .001). However, adding more predictors does improve the model significantly, due to our large number of cases. Below we present a statistically optimal multilevel model. For this model we did not include an R^2-change threshold. All predictors that significantly affected the -2LogLikelihood of the model were included.

Like for the U-Read model presented in the results section, for the optimal model predictors were selected that correlated .20 or more with the individual cloze scores and which did not have inter-correlations above .70. This resulted in 15 linguistic features. These features were first added separately to the base model and ranked according to the reduction in -2LogLikelihood. All predictors significantly improved the model ($p < .001$). Next, the model was built up by adding the predictors one-by-one in order of rank. Predictors that did not improve the model were dropped. The final optimal model included 10 of the 15 linguistic features and 3 interactions between educational level and linguistic features (see Table A11.1). A description of these features can be found in Appendix 10. Their effects are discussed in tandem below.

Lexical complexity & concreteness

The word frequency of a text and the use of concrete nouns were positively related to comprehension. When a text contained more frequent words, cloze scores increased. These effects were a little stronger for the pre-vocational medium and pre-vocational high students compared to other students. Texts with a high proportion of concrete nouns had higher cloze scores. In addition, universal nouns were negatively related to comprehension. Universal words are abstract as they refer to universal concepts rather than concrete objects/phenomena. A high number of universal nouns thus indicates that the text is abstract.

Contrary to expectation, nominalizations were also positively related to comprehension in the optimal model. When nominalizations are added as a single predictor however, nominalizations do have a negative relation with the cloze scores (see also inter-correlations in Table A11.2). Further investigation showed that the coefficient for nominalizations was influenced by the predictors Word frequency and Concrete nouns. Without these predictors, the coefficient turned negative instead of positive. We therefore interpret this effect as a correction of the estimates of other predictors.

Informational load & information structure

A number of predictors indicate that the quantity and structure of information within clauses and sentences is related to text difficulty. As the number of content words per clause, the maximum syntactic dependency length, the number of small conjuncts and the number of adjectival past participles increased, cloze scores decreased. The effect of Small conjuncts was moderated by an interaction with Education level. High level students were more affected by the number of small conjuncts than low level students which might indicate that they are more sensitive to this type of structure. State verbs were positively related, which indicates that texts that described states rather than changes to states (i.e., actions or processes) were easier to comprehend.

Personal style

The number of second person pronouns had a positive relationship with comprehension. Texts with second person pronouns received higher cloze scores. However, the effect was moderated by the Education level of the student. Only the scores of pre-vocational students were positively related to the use of second person pronouns. No significant relation was found for general and pre-university students.

Model performance

Together, reader characteristics and linguistic features predicted 58% (r =.762) of the observed variance. 34% was accounted for by the reader characteristics, which means that the linguistic features (including interactions) predicted an additional 24% of the variance. Compared to the U-Read model presented in the result section, this means a 1% gain in predictive power. This gain comes at the cost of five more predictors and three interactions.

Table A11.1: Optimal multilevel model

Random effects	**Estimates**	**SE**	***t***	***p***
School	0.971	0.302	3.215	<.001[a]
School: Student	4.414	0.230	19.191	<.001[a]
Text	0.159	0.052	3.058	<.001[a]
Residual	11.007	0.196		

Fixed effects	**Estimates**	**SE**	***t***	***p***
Intercept	13.530	0.450	30.067	<.001
Education level: pre-voc. low	0[b]			
Education level: pre-voc. medium	1.479	0.248	5.964	<.001
Education level: pre-voc. high	3.207	0.268	11.966	<.001
Education level: general	5.136	0.337	15.240	<.001
Education level: pre-uni.	6.029	0.350	17.226	<.001
Grade 8	0[b]			
Grade 9	0.660	0.145	4.552	<.001
Grade 10[c]	0.937	0.303	3.092	.002
Reading ability	1.952	0.094	20.766	<.001
Reading test: R	0[b]			
Reading test: V	-2.740	0.425	-6.447	<.001
Word frequency	4.772	0.538	8.870	<.001
Word frequency * Pre-voc. medium	1.435	0.644	2.228	.026
Word frequency * Pre-voc. high	0.720	0.623	1.156	.248
Word frequency * General	-0.383	0.653	-0.587	.558
Word frequency * Pre-uni.	-0.471	0.614	-0.767	.443
Content words per clause	-0.585	0.107	-5.467	<.001
Concrete nouns	4.103	0.328	12.509	<.001
Max SDL	-0.267	0.026	-10.269	<.001
Adjectival past participles	-17.061	1.645	-10.371	<.001
Small conjuncts	-0.675	0.440	-1.534	.125
Small conjuncts * Pre-voc. medium	-0.432	0.508	-0.850	.395
Small conjuncts * Pre-voc. high	-0.628	0.501	-1.253	.210
Small conjuncts * General	-1.850	0.524	-3.531	<.001
Small conjuncts * Pre-uni.	-1.503	0.490	-3.067	.002
Universal nouns	-0.011	0.003	-3.667	<.001
Nominalizations	0.021	0.002	10.500	<.001
State verbs	0.007	0.003	2.333	.020
2nd person pronouns	0.017	0.008	2.125	.034
2nd Pers. pronouns * Pre-voc. medium	0.004	0.009	0.444	.657
2nd Pers. pronouns * Pre-voc. high	-0.004	0.009	-0.444	.657
2nd Pers. pronouns * General	-0.009	0.009	-1.000	.317
2nd Pers. pronouns * Pre-uni.	-0.013	0.009	-1.444	.149

[a] One-sided. [b] Set as reference level. [c] Unbalanced, no 10th grade pre-vocational students were present in the sample.

Table A11.2: Correlations predictors with individual cloze scores

	Max SDL	Uni-versal nouns	State verbs	Con-crete nouns	Small con-juncts	Con-tent words	Nomi-naliza-tions	2nd Pers. pro-nouns	Adj. past parti-ciples	Word freq-uency
Cloze score	-.310	-.305	.247	.299	-.290	-.404	-.263	.203	-.269	.419
Max SDL		.352	-.243	-.322	.241	.545	.233	-.290	.372	-.383
Uni-versal nouns	.352		-.105	-.620	.581	.549	.605	-.158	.294	-.428
State verbs	-.243	-.105		-.001	-.321	-.616	-.184	.337	-.065	.488
Con-crete nouns	-.322	-.620	-.001		-.226	-.385	-.462	.081	-.426	.391
Small con-juncts	.241	.581	-.321	-.226		.619	.563	-.084	.083	-.502
Con-tent words	.545	.549	-.616	-.385	.619		.507	-.483	.341	-.601
Nomi-naliza-tions	.233	.605	-.184	-.462	.563	.507		-.124	.394	-.539
2nd Pers. pro-nouns	-.290	-.158	.337	.081	-.084	-.483	-.124		-.177	.303
Adj. past parti-ciples	.372	.294	-.065	-.426	.083	.341	.394	-.177		-.317
Word freq-uency	-.383	-.428	.488	.391	-.502	-.601	-.539	.303	-.317	

Appendix 12: Results 2-fold-cross-validation

Table A12.1: Estimates and standard errors for partition 1 and 2 models

	Partition 1	Partition 2
Random effects	**Estimate (SE)**	**Estimate (SE)**
School	1.011 (0.342)	0.715 (0.290)
School: Student	3.879 (0.345)	3.998 (0.351)
Text	0.153 (0.088)	0.295 (0.131)
Residual	11.723 (0.350)	11.634 (0.352)
Fixed effects	**Estimate (SE)**	**Estimate (SE)**
Intercept	13.541 (0.500)	13.662 (0.480)
Education level: pre-voc. low	0[a]	0[a]
Education level: pre-voc. medium	1.528 (0.292)	1.307 (0.292)
Education level: pre-voc. high	3.347 (0.315)	2.991 (0.315)
Education level: general	5.157 (0.390)	5.034 (0.390)
Education level: pre-uni.	6.12 (0.403)	5.875 (0.405)
Grade 8	0[a]	0[a]
Grade 9	0.705 (0.170)	0.678 (0.171)
Grade 10[b]	0.935 (0.351)	1.048 (0.356)
Reading ability	1.903 (0.111)	1.986 (0.113)
Reading test: R	0[a]	0[a]
Reading test: V	-2.894 (0.460)	-2.549 (0.430)
Word frequency	5.079 (0.284)	5.611 (0.277)
Content words per clause	-1.323 (0.104)	-1.057 (0.100)
Concrete nouns	3.257 (0.405)	3.522 (0.392)
Max SDL	-0.299 (0.042)	-0.298 (0.037)
Adjectival past participles	-10.768 (2.299)	-13.316 (2.311)

[a] Set as reference level. [b] Unbalanced, no 10th grade pre-vocational students were present in the sample.

Table A12.2: Results cross-validation

Partition	R	Multiple R^2
1	.750	.563
2	.761	.579
Full dataset	.756	.572

Samenvatting in het Nederlands

Leesbaarheid ontrafeld: Hoe linguïstische kenmerken tekstbegrip en tekstverwerking beïnvloeden en voorspellen

Sinds de publicatie van de eerste leesbaarheidsformules, zo'n 100 jaar geleden, is de behoefte aan objectieve meetinstrumenten van leesbaarheid alleen maar toegenomen. Naast uitgevers en docenten houden ook de Nederlandse overheid, diverse bedrijven en organisaties zich bezig met de vraag of hun teksten geschikt zijn voor de lezers die zij voor ogen hebben. Helaas zijn de meeste bestaande Nederlandse instrumenten die zij tot hun beschikking hebben niet gevalideerd op basis van empirisch onderzoek. Ook is zelden bekend waar de instrumenten op gebaseerd zijn (Jansen, 2005; Jansen & Boersma, 2013; Kraf, Lentz & Pander Maat, 2011). De validiteit van deze instrumenten is dus twijfelachtig en met hun uitkomsten moet behoedzaam worden omgegaan. Alleen de CLIB-formule ('Cito LeesIndex voor het Basis- en speciaal onderwijs'; Staphorsius, 1994) is uitgebreid gevalideerd en gebaseerd op empirisch onderzoek. Maar dit instrument is alleen geschikt voor lezers in het basisonderwijs. Daarom startte in 2012 het LIN-project ('LeesbaarheidsIndex voor het Nederlands') waarin onderzoekers, programmeurs en experts van Universiteit Utrecht, Radboud Universiteit, Cito en de Nederlandse Taalunie samenwerken om een gevalideerd, automatisch leesbaarheidsinstrument te ontwikkelen voor het Nederlands. Dit instrument is de in eerste plaats gericht op het voorspellen van leesbaarheid voor middelbare scholieren met het uiteindelijke doel om naderhand ook leesbaarheid voor volwassenen te gaan voorspellen. Voordat dit instrument gebouwd kan worden, moeten een aantal belangrijke stappen worden gezet. Ten eerste moeten teksten automatisch geanalyseerd kunnen worden. Linguïstische kenmerken moeten met een druk op de knop uit de tekst worden geëxtraheerd. Vervolgens zijn er empirische data nodig waaruit blijkt welke tekstkenmerken de leesprestaties van middelbare scholieren beïnvloeden. Met andere woorden, de kenmerken moeten op basis van deze data gevalideerd en gekalibreerd worden. Uiteindelijk kunnen we die resultaten gebruiken voor het bouwen van een gevalideerd leesbaarheidsinstrument. Als onderdeel van het LIN-project heeft dit proefschrift als doel de effecten van linguïstische kenmerken op de leesbaarheid van teksten te onderzoeken en empirische data te verzamelen op basis waarvan het nieuwe leesbaarheidsinstrument gebouwd kan worden.

Problemen rondom (traditioneel) leesbaarheidsonderzoek

Ondanks de populariteit van leesbaarheidsformules en -instrumenten is er al sinds de jaren zeventig veel kritiek op leesbaarheidsonderzoek (o.a. Anderson & Davison, 1988; Bailin & Grafstein, 2016; Bruce, Rubin & Starr, 1981; Duffy & Kabance, 1982; Klare, 1976a). De meest gehoorde kritiek is dat de kenmerken die gebruikt worden om de leesbaarheid van teksten te voorspellen niet causaal relevant zijn. Het zijn slechts symptomen die correleren met tekstmoeilijkheid, zonder dat ze daar de oorzaak van zijn (Kintsch & Vipond, 1979). De geldigheid van die kritiek hangt wel af van het doel dat wordt nagestreefd met het leesbaarheidsonderzoek: tekstmoeilijkheid voorspellen of teksten verbeteren (zie ook Klare, 1984). Voorspellingen worden gebruikt om bijvoorbeeld teksten te selecteren voor specifieke lezers, bijvoorbeeld 3e klas middelbare scholieren. Deze voorspellingen zijn gebaseerd op correlaties tussen linguïstische kenmerken en tekstmoeilijkheid. Daarvoor is een causaal verband misschien wenselijk, maar niet noodzakelijk. Wanneer het doel is teksten te verbeteren, is causaliteit wel een vereiste. Moeilijke teksten bevatten misschien vaak lange zinnen, maar dat betekent niet automatisch dat een tekst makkelijker wordt wanneer de zinnen worden opgesplitst in korte zinnen. Wanneer er geen causaal verband bestaat tussen tekstkenmerken en tekstmoeilijkheid, zal het aanpassen van deze kenmerken niet het gewenste resultaat hebben en kan het zelfs leiden tot verslechtering van de leesbaarheid (Davison & Kantor, 1982).

Zelfs als er wel een causaal verband bestaat tussen tekstkenmerken en leesbaarheid, kan dit verband veel zwakker zijn dan de correlatie doet vermoeden. Dit komt doordat er twee componenten zijn die de moeilijkheid van een tekst bepalen: de conceptuele complexiteit en de stilistische complexiteit (zie ook Leech & Short, 2007). De conceptuele complexiteit wordt bepaald door de inhoud van de tekst, de boodschap die de tekst probeert over te brengen. De stilistische complexiteit betreft de vorm, de manier waarop deze boodschap wordt verteld. Wanneer een tekst te moeilijk is voor het beoogde publiek, kan alleen de stilistische complexiteit worden aangepast. De boodschap zelf moet immers onveranderd blijven; dit is wat de lezer volgens de schrijver moet weten. Vanuit dit perspectief kan de potentiele ‘winst’ van tekstverbetering uitsluitend voortkomen uit stilistische aanpassingen. Dit onderscheid tussen stilistische en conceptuele complexiteit wordt in veel leesbaarheidsonderzoeken niet gemaakt.

Een ander probleem met eerder onderzoek is dat de lezer vaak wordt vergeten. Leesbaarheidsformules worden gekalibreerd op basis van de tekstbeoordelingen van experts in plaats van begripsscores van de beoogde lezers. Dit is problematisch omdat deze experts het veelal oneens zijn. Daarnaast is gebleken dat zij vrij slecht zijn in het aanwijzen van de problemen die daadwerkelijk door lezers worden ervaren (De Jong & Lentz, 1996; Lentz & De Jong, 1997; Feng,

2010). Hoewel er ook onderzoek is waarin beoordelingen van de lezers zelf worden gebruikt (Crossley, Skalicky, Dascalu, McNamara & Kyle, 2017; De Clercq & Hoste, 2016), is ook hier de vraag in hoeverre deze beoordelingen overeenkomen met het werkelijke tekstbegrip van de lezer.

Een andere manier waarop de lezer wordt vergeten is door het gebruik van zogenaamde geaggregeerde begripsscores. Er worden dan wel begripsscores verzameld, maar deze worden samengenomen tot een gemiddelde score per tekst. Het gevolg is dat variantie tussen lezers wordt verwijderd en hierdoor worden verschillen tussen lezers genegeerd (Anderson & Davison, 1988).

Opzet van de huidige studie

In de opzet van de huidige studie hebben we bovenstaande problemen aangepakt. Ten eerste maken we gebruik van de huidige taaltechnologische middelen die geavanceerde linguïstische kenmerken automatisch uit teksten extraheren. We gebruiken daarvoor de applicatie 'T-Scan' (Kraf & Pander Maat, 2009; Pander Maat et al., 2014). Deze tekstanalyse software berekent ruim 400 linguïstische maten op het gebied van lexicale complexiteit, zinscomplexiteit, referentiële en relationele coherentie, concreetheid, persoonlijkheid en woordwaarschijnlijkheden van Nederlandse teksten. De kenmerken die T-Scan berekent zijn geïnspireerd door experimenteel en theoretisch onderzoek (Pander Maat et al., 2014). Ze zijn dan ook veel preciezer dan de standaardmaten 'zinslengte' en 'woordlengte' die wetenschappers in de vorige eeuw tot hun beschikking hadden

Ten tweede gebruiken we data van echte lezers uit onze doelgroep die we niet aggregeren maar analyseren met behulp van multilevel statistische methodes. We kijken hoe middelbare scholieren van verschillende leesvaardigheidsniveaus teksten lezen en begrijpen. We baseren ons dus niet op beoordelingen van experts of lezers maar op objectieve maten van leesbaarheid. De focus ligt daarbij niet alleen op het tekstbegrip maar ook op het leesproces. Dit zijn twee belangrijke aspecten van leesbaarheid. Een tekst is alleen goed leesbaar wanneer het tekstbegrip hoog is en dit niveau zonder al te veel moeite wordt bereikt.

Tot slot combineren we experimenteel en correlationeel onderzoek in één design waardoor we zowel de causale als correlationele effecten van linguïstische kenmerken kunnen onderzoeken. In drie grootschalige experimenten manipuleren we steeds één stilistisch kenmerk van 20 teksten: de lexicale complexiteit, de syntactische complexiteit of de hoeveelheid coherentiemarkeringen in de tekst. De inhoud en andere stilistische kenmerken worden gelijk gehouden. Op die manier kunnen we zien of deze drie kenmerken alleen correleren met leesbaarheid of dat er ook sprake is van een causaal verband. Hiervoor gebruiken we in totaal 60 teksten (3 x 20) van 300 tot 400 woorden. 30 teksten zijn afkomstig uit schoolboeken. Deze teksten zijn geschreven voor middelbare scholieren van verschillende niveaus en

komen uit boeken voor aardrijkskunde, geschiedenis, Nederlands en economie. De voorlichtingsteksten zijn afkomstig van websites en brochures van Nederlandse overheidsinstanties of met de overheid geaffilieerde organisaties. Hoewel deze teksten niet specifiek geschreven zijn voor middelbare scholieren, bespreken ze thema's die vanuit maatschappelijk oogpunt ook relevant zijn voor jongeren (bijvoorbeeld diabetes, donorregistratie en criminaliteit).

Deze teksten zijn voorgelegd aan 3107 middelbare scholieren van verschillende leerjaren en onderwijsniveaus. Aan het onderzoek naar tekstbegrip namen in totaal 2926 scholieren deel afkomstig uit: 2e en 3e klas vmbo; 2e, 3e en 4e klas havo; en 2e, 3e en 4e klas vwo. Vmbo-leerlingen in dit deel van de studie kwamen uit drie verschillende leerwegen: basisberoepsgerichte leerweg ('vmbo-bb'), kaderberoepsgerichte leerweg ('vmbo-kb') en gemengde theoretische leerweg ('vmbo-gt'). Aan het onderzoek naar het leesproces namen in totaal 181 3e klas scholieren uit vmbo-kb, havo en vwo deel.

Tekstbegrip meten met de HyTeC-cloze (Hoofdstuk 2)

Er zijn vele verschillende methodes om tekstbegrip te meten, maar wat de juiste methode is hangt af van de praktische en theoretische eisen waar de test aan moet voldoen. In het geval van het huidige onderzoek moest dat een methode zijn die:

- betrouwbaar en valide tekstbegrip meet;
- de moeilijkheid van de tekst weerspiegelt zonder dat deze gecontamineerd wordt met eventuele vraagmoeilijkheid;
- gevoelig is voor verschillen tussen teksten en tussen tekstversies;
- gevoelig is voor verschillen tussen lezers van verschillende leesvaardigheidsniveaus;
- toepasbaar is op een groot aantal teksten;
- geschikt is voor goede en zwakke lezers;
- makkelijk en snel uitvoerbaar is.

Clozetoetsen kunnen aan al deze eisen voldoen, al is er in de vakliteratuur nog wel discussie over (zie o.a. Brown, 2013; Chen, 2004; Gellert & Elbro, 2013; Greene, 2001; Kobayashi, 2002a, 2002b, 2004; Oller & Jonz (Eds.), 1994; O'Toole & King, 2010, 2011; Trace, Brown, Janssen, & Kozhevnikova, 2017). Tegenstanders menen dat clozetoetsen hoofdzakelijk vaardigheden van lagere orde meten – zoals grammatica en taalkundige kennis – en geen begrip op tekstniveau. Daarom ontwikkelden we een nieuwe clozeprocedure: de 'Hybrid Text Comprehension cloze' (HyTeC-cloze). Deze hybride clozeprocedure combineert de sterke punten van mechanische en rationele clozeprocedures. De rationele procedure wordt gebruikt om te voorkomen dat woorden die niet op tekstniveau begrip meten

clozegaten worden, zoals lidwoorden, koppelwerkwoorden en delen van uitdrukkingen. Ook woorden die de lezer alleen kan 'gokken', zoals getallen en namen, zijn niet toegestaan. Alle woorden die overblijven zijn potentiële clozegaten. Zij worden door middel van een mechanische procedure over verschillende clozeversies verdeeld. Er wordt een deletieratio aangehouden van 1 op 10, wat betekent dat in een tekst van 300 woorden 30 clozegaten worden geselecteerd. De toetsen worden na afname semantisch gescoord.

The HyTeC-clozeprocedure werd geëvalueerd aan de hand van data van 120 unieke clozetoetsen (60 teksten in twee clozeversies). De resultaten toonden aan dat de HyTeC-clozeprocedure een betrouwbare en valide methode is voor het meten van tekstbegrip in experimentele en correlationele studies.

Effecten van lexicale complexiteit (Hoofdstuk 3)

Lexicale kenmerken vormen de sterkste voorspellers van leesbaarheid (Bailin & Grafstein, 2016). Dat is ook niet zo verrassend aangezien het kennen van de woorden in de tekst een logische voorwaarde vormt voor het begrijpen van de tekst (Perfetti & Stafura, 2014). De sterke correlationele relatie tussen leesbaarheid en woordmoeilijkheid doet vermoeden dat wanneer moeilijke woorden worden vervangen door makkelijkere woorden, de leesbaarheid sterk toeneemt. Echter, uit gecontroleerde experimenten blijkt dat de effecten van lexicale manipulaties sterk afhangen van het aantal moeilijke woorden, hun relevantie binnen de tekst en de moeilijkheidsgraad van het woord, gegeven de woordkennis van een lezer. Lexicale simplificatie leidt niet altijd tot beter tekstbegrip (Freebody & Anderson, 1983ab; Stahl, 1991; 2003b; Stahl, Jacobson, Davis & Davis, 1989). Maar het aantal teksten en lezers dat in deze experimenten wordt meegenomen is zeer beperkt en de vraag is of de resultaten te generaliseren zijn naar meer uiteenlopende lezers en teksten. Twee experimenten moesten hier meer duidelijkheid in geven.

In Experiment 1 zijn van 10 schoolboekteksten en 10 voorlichtingsteksten een lexicaal-makkelijke en lexicaal-moeilijke tekstversie gemaakt door 20% van de inhoudswoorden te vervangen (zie (1) en (2)). De manipulaties waren subtiel en in lijn met de inhoud en de stijl van de tekst. Daarnaast werden andere stilistische kenmerken – zoals syntax, woordlengte, argumentoverlap en type-token-ratio – tussen tekstversies gelijk gehouden.

(1) *Rabiës is een infectieziekte die de hersenen <u>beschadigt/aantast</u>.*

(2) *Iemand heeft <u>genoeg geld/voldoende middelen</u> om er een tijdje tussenuit te kunnen.*

De teksten werden omgezet in HyTeC-clozetoetsen en voorgelegd aan tweede tot vierde klas scholieren in het Nederlands voortgezet onderwijs. Multilevel-analyses toonden aan dat lexicale complexiteit een significant effect had op tekstbegrip. Het verminderen van de lexicale complexiteit van een tekst leidde tot hogere scores op de clozetoetsen, met name voor vmbo-kb en vmbo-gt leerlingen.

In Experiment 2 werden vier van de voorlichtingsteksten gebruikt in een oogbewegingsstudie bij scholieren in 3-vmbo-kb, 3-havo en 3-vwo. Gemanipuleerde woorden in de lexicaal-makkelijke versie werden sneller gelezen dan gemanipuleerde woorden in de lexicaal-moeilijke versie. Ook het aantal directe herhaalde fixaties op een woord was lager in de lexicaal-makkelijke versie. Net als in Experiment 1 hadden vmbo-kb leerlingen het meeste profijt van een lexicaal gesimplificeerde tekst. Het vereenvoudigen van de lexicale complexiteit leidt dus zowel tot beter tekstbegrip als tot een snellere verwerking van teksten.

Effecten van afhankelijkheidslengte (Hoofdstuk 4)

Na lexicale complexiteit komt syntactische complexiteit vaak als tweede voorspellende factor van leesbaarheid uit de bus, waarbij syntactische complexiteit meestal wordt geoperationaliseerd als zinslengte (Bailin & Grafstein, 2016; Dale & Chall, 1948; Flesch, 1948; Klare, 1963; 1974; Staphorsius, 1994). Hoewel lange zinnen vaak ook complexer zijn, is lengte niet hetzelfde als complexiteit (Bailin & Grafstein, 2016; Davison et al, 1980; Gough, 1966). Het opsplitsen van lange zinnen leidt daarom meestal niet tot een verbetering van de leesbaarheid.

Een maat die wel de syntactische structuur van zinnen meeneemt is afhankelijkheidslengte (‘AL’). Dat is de afstand tussen een syntactisch hoofd en zijn dependens, zoals de persoonsvorm en het onderwerp, of een zelfstandig naamwoord en een bijbehorend adjectief. Als de afstand tussen een hoofd en zijn dependens toeneemt – bijvoorbeeld omdat er andere zinsdelen tussen hen in staan zoals in een tangconstructie – zou het moeilijker worden om het hoofd en zijn dependens te integreren. Lange afstanden zouden meer mentale capaciteit kosten omdat informatie over een langere periode actief gehouden moet worden en omdat tussenliggende elementen kunnen infereren bij de integratie. Effecten van AL zijn vooral nog gevonden in oogbewegingsstudies waarbij geconstrueerde zinnen in isolatie worden gelezen (Bartek, Lewis, Vasishth & Smith, 2011; Grodner & Gibson, 2005).

In Experiment 3 zijn vier voorlichtingsteksten gebruikt om te onderzoeken of AL in natuurlijke zinnen die onderdeel zijn van authentieke teksten het leesproces beïnvloeden. Van iedere tekst werd een versie met relatief korte afhankelijkheidslengtes (AL-kort) en een versie met relatief lange afhankelijkheidslengtes (AL-lang) gemaakt door de woordvolgorde in 1/3 van de zinnen te veranderen. Hierdoor veranderden de afstanden tussen sommige hoofden

en dependenten in de zin (zie (3)). De grootte van de verandering was afhankelijk van de zinnen en tekst zelf en varieerde van 2 tot 10 woorden.

(3) a. De aangehoudene dient onverwijld overgedragen te worden aan een opsporingsambtenaar.

b. De aangehoudene dient onverwijld aan een opsporingsambtenaar overgedragen te worden.

Resultaten van het oogbewegingsexperiment toonden aan dat leestijden langer waren voor gemanipuleerde zinnen in de AL-lang conditie dan in de AL-kort conditie. Dit effect was onafhankelijk van de mate waarmee de afstand toenam.

In Experiment 4 werden de vier voorlichtingsteksten aangevuld met zes andere voorlichtingsteksten en 10 schoolboekteksten. Van de nieuwe teksten werden op dezelfde manier als in het oogbewegingsexperiment een AL-lang en AL-kort versie gemaakt. Vervolgens werden de 20 teksten omgezet in HyTeC-clozetoetsen en voorgelegd aan middelbare scholieren. De langere afhankelijkheidslengtes zorgden voor lagere scores op de clozetoetsen, maar het effect hing samen met het tekstgenre en de mate waarmee de AL toenam. Voor voorlichtingsteksten was er een effect zelfs wanneer de afstand waarmee de AL toenam klein was (toename van minder dan 5 woorden), maar voor schoolboekteksten was er alleen een significant negatief effect wanneer de afstand groot was (toename van 5 of meer woorden). Omdat de schoolboekteksten over het algemeen ook makkelijker waren dan de voorlichtingsteksten, lijkt het dat scholieren kleine toenames in AL aankunnen. Alleen zodra de tekst te moeilijk wordt of zodra de toename in AL te groot wordt, kunnen ze niet meer voldoende compenseren.

Effecten van coherentiemarkeringen (Hoofdstuk 5)

Lexicale en syntactische kenmerken kennen een lange traditie in leesbaarheidsonderzoek. Voor kenmerken die betrekking hebben op tekstsamenhang (‘coherentie’) geldt dat niet. Hoewel Kintsch en Vipond al in 1979 wezen op het ontbreken van deze factoren in leesbaarheidsonderzoek, neemt pas de laatste 15 jaar het gebruik van dit soort kenmerken in leesbaarheidsonderzoek toe. Eén van de redenen hiervoor is dat coherentie deels in het hoofd van de lezer geconstrueerd wordt (Graesser, McNamara, Louwerse & Cai, 2004; Sanders, Spooren & Noordman, 1992; Sanford, 2006; Zwaan & Rapp, 2006). Daardoor is coherentie per definitie moeilijk vast te stellen met een objectieve maat. Samenhang wordt dan ook vaak afgeleid aan de hand van lexicale signalen, of te wel ‘cohesie’ (o.a. Crossley, Skalicky, Dascalu, McNamara & Kyle, 2017; Feng, Elhadad & Huenerfauth, 2009).

Bekende en veel gebruikte coherentie-signalen zijn connectieven, zoals *dus*, *maar*, en *daarnaast*. Connectieven maken relaties tussen zinnen of tekstdelen expliciet zodat de lezer niet zelf de coherentierelatie hoeft af te leiden. De aanwezigheid van connectieven leidt in het algemeen tot een snellere verwerking van de informatie die direct volgt op het connectief ('het integratie-effect'; o.a. Cozijn, Noordman & Vonk, 2011; Maury & Teisserenc, 2005; Van Silfhout, Evers-Vermeul, Mak & Sanders, 2014; Van Silfhout, Evers-Vermeul & Sanders, 2015). Maar wanneer het gaat om de invloed van connectieven op tekstbegrip zijn de resultaten minder consistent (Degand & Sanders, 2002; McNamara, Kintsch, Butler Songer & Kintsch, 1996; Murray, 1995; Sanders & Noordman, 2000; Van Silfhout, Evers-Vermeul & Sanders, 2014). Connectieven lijken soms tekstbegrip te verbeteren, soms geen effect te hebben en in sommige situaties zelfs een negatief effect te hebben op tekstbegrip.

Om meer duidelijkheid te krijgen over de invloed van connectieven op tekstbegrip zijn van 10 schoolboekteksten en 10 voorlichtingsteksten twee tekstversies gemaakt. Ten opzichte van de andere versie werd in een derde van de coherentierelaties een causaal (4), contrastief (5), additief (6) of temporeel (7) connectief weggehaald dan wel toegevoegd. Zo ontstond een versie met een laag aantal coherentiemarkeringen (laag-CM) en een versie met een relatief hoog aantal coherentiemarkeringen (hoog-CM). Alleen optionele connectieven werden verwijderd; de relatie moest interpreteerbaar blijven, ook zonder connectief.

Causale relatie

(4) Als de productiestructuur inderdaad verbetert, kan het land op de wereldmarkt beter concurreren met andere landen. Daardoor kan de export van het land stijgen en de import van het land afnemen.

Contrastieve relatie

(5) Minderjarigen kunnen vanaf hun twaalfde hun wens in het Donorregister laten opnemen. Ouders of voogden hoeven hiervoor géén toestemming te verlenen. Maar als minderjarigen instemmen met donatie en voor hun zestiende overlijden, kunnen ouders of voogden alsnog weigeren.

Additieve relatie

(6) Als de brandweer wordt gealarmeerd door een brandmelder heeft dit veel consequenties. Ten eerste rukt de brandweer met spoed uit naar het meldadres. Dat brengt verkeersrisico's met zich mee. Ten tweede, als de brandweer uitrukt voor een nodeloze alarmering, is ze op dat moment niet beschikbaar voor andere wel noodzakelijke, hulpverlening.

Temporele relatie

(7) In de late middeleeuwen zien we steeds meer steden in Europa. Handelaren vormden handelsgemeenschappen op plaatsen waar relatief rijke afnemers zaten, zoals een adellijk hof, een militaire vesting of een klooster. Deze handelsgemeenschappen trokken vervolgens ambachtslieden aan.

De teksten werden omgezet in HyTeC-clozetoetsen en voorgelegd aan tweede tot vierde klas scholieren in het Nederlands voortgezet onderwijs.

Het effect van de connectieven bleek zich te beperken tot hun directe omgeving en hing af van het type connectief. Het toevoegen van een contrastief of causaal connectief leidde tot hogere clozescores. Het toevoegen van een additief connectief had echter een negatief effect op tekstbegrip en resulteerde in lagere clozescores. Uit de resultaten blijkt dat connectieven tekstbegrip zowel kunnen faciliteren als hinderen afhankelijk van het type coherentierelatie.

Leesbaarheid voorspellen voor middelbare scholieren (Hoofdstuk 6)

In totaal werden in de drie experimentele studies 60 teksten onderzocht. Van iedere tekst zijn twee versies gemaakt. Dat brengt het totaal op 120 teksten. Voor al die 120 teksten zijn clozescores verzameld. Deze data zijn vervolgens ook gebruikt om een model te bouwen dat de leesbaarheid van teksten voorspelt voor scholieren in het voortgezet onderwijs. De data zijn op twee manieren geanalyseerd: 1. Volgens de traditionele unilevel-regressiemethode, waarbij individuele scores worden geaggregeerd tot een gemiddelde tekstscore; 2. Met multilevel-regressie, waarbij de datastructuur behouden blijft en individuele scores worden gebruikt.

Beide analyses identificeerden dezelfde vijf factoren die samen de leesbaarheid van de teksten het beste voorspelden:

1. Woordfrequentie (gecorrigeerd voor samenstellingen en zonder namen)
2. Inhoudswoorden per deelzin
3. Concrete zelfstandige naamwoorden
4. Maximale afhankelijkheidslengte
5. Bijvoeglijk voltooide deelwoorden (‘De *geschilde* aardappel’)

Deze factoren verklaarden 76% van de variantie van de geaggregeerde tekstscores, maar uit de multilevel-analyse bleek dat zij slechts 23% van de variantie verklaarden in individuele scores. Veruit de meeste variantie (34%) werd in dit laatste model verklaard door lezerskenmerken (onderwijsniveau, leerjaar en leesvaardigheidsscore). Desalniettemin verklaarde dit model (het ‘*Utrecht readability model*’, of te wel U-Read) zo’n 20% meer variantie dan vergelijkbaar modellen met daarin de factoren die gebruikt worden in de Flesch-Douma en CLIB-

formules. Als het gaat om het voorspellen van leesbaarheid voor middelbare scholieren, is het U-Read model dus beduidend beter dan de oude, populaire formules.

Tussen tekst-effecten versus tussen tekstversie-effecten (Hoofdstuk 7)

Een belangrijk onderscheid in dit onderzoek was het verschil tussen conceptuele complexiteit (de boodschap) en de stilistische complexiteit (de vorm) van een tekst. De stilistische complexiteit kan worden aangepast, maar de conceptuele complexiteit staat vast; dit is de boodschap die de lezer volgens de schrijver moet weten. In het meeste leesbaarheidsonderzoek wordt dit onderscheid niet gemaakt. Leesbaarheidsvoorspellingen worden gebaseerd op verschillen tussen teksten. Deze teksten verschillen zowel qua inhoud als qua stijl. In zulke correlationele studies zijn effecten van conceptuele en stilistische complexiteit dus niet uit elkaar te houden. Dat heeft tot gevolg dat stilistische aanpassingen die op basis van deze voorspellingen gedaan worden veelal niet het gewenste effect hebben en dat de effecten van tekstkenmerken worden overschat.

Dankzij ons geïntegreerde onderzoeksdesign waarin experimenteel en correlationeel onderzoek wordt gecombineerd, kunnen we voor drie tekstkenmerken wel kijken hoe ze samenhangen met stilistische en conceptuele complexiteit. In drie experimenten werden tekstversies met elkaar vergeleken die alleen verschilden in stilistische complexiteit. De inhoud werd gelijk gehouden en alleen de lexicale complexiteit, syntactische complexiteit of het aantal coherentiemarkeringen verschilde tussen twee tekstversies. Zodoende kunnen we voor deze drie kenmerken zien hoe goed ze zijn in het voorspellen van verschillen tussen teksten (die verschillen in stilistische en conceptuele complexiteit) en hoe goed ze zijn in het voorspellen van verschillen tussen tekstversies (die alleen verschillen in stilistische complexiteit).

De eerste vraag is dan hoeveel variantie in tekstbegripsscores de kenmerken kunnen verklaren tussen teksten en hoeveel variantie tussen versies van dezelfde tekst. Aanvullende analyses lieten zien dat tussen tekstversies maximaal 1% wordt verklaard, en tussen teksten 3% tot 16%. Dat is op zich niet zo vreemd, omdat tekstkenmerken tussen teksten veel meer verschillen dan tussen tekstversies. Zoveel ‘rek’ zit er niet in het stilistisch aanpassen van teksten wanneer de inhoud gelijk moet blijven.

Een tweede vraag is: wanneer de voorspellende waarde van een tekstkenmerk bepaald wordt op basis van verschillen tussen teksten, kun je dan die waarde ook gebruiken om verschillen in tekstbegripsscore tussen twee tekstversies te voorspellen? Dit onderzochten we voor de drie gemanipuleerde kenmerken. Voor lexicale en syntactische complexiteit blijkt dat modellen op basis van verschillende teksten tot een overschatting leiden van de kracht van de voorspellers als we die

willen gebruiken voor versieverschillen. De richting van de twee voorspellingen is wel hetzelfde.

Voor coherentiemarkeringen ligt dat anders. Het model gebaseerd op tussen-tekstverschillen voorspelt dat connectieven teksten moeilijker maken. Dit komt waarschijnlijk omdat complexe relaties vaker gemarkeerd worden met connectieven, en dus is een tekst met veel connectieven over het algemeen moeilijker dan een tekst met weinig connectieven. Wanneer we kijken naar tekstversieverschillen, zien we echter dat het toevoegen van connectieven aan een gegeven tekst (over het algemeen) een positief effect heeft op het tekstbegrip. Het weghalen van de connectieven zal de tekst niet makkelijker maken en waarschijnlijk de leesbaarheid zelfs verminderen.

Deze resultaten illustreren dat het 'toeschrijven naar formules' die gebaseerd zijn op onderzoek naar verschillende teksten gevaarlijk kan zijn en dat zulke formules niet altijd causaal geïnterpreteerd mogen worden. Tenminste, wanneer de factoren die zij gebruiken niet ook gestaafd worden door experimenteel onderzoek.

Conclusie

In dit proefschrift zijn inzichten vanuit leesbaarheids- en tekstverwerkingsonderzoek gecombineerd met hedendaagse taaltechnologieën om de relatie tussen linguïstische kenmerken en twee aspecten van leesbaarheid te onderzoeken: tekstbegrip en leesproces. Het resultaat is een gevalideerde leesbaarheidsformule voor het voortgezet onderwijs die een verbetering is ten opzichte van de oude formules. Door middel van een geïntegreerd empirisch design met experimentele en correlationele studies konden ook causale effecten van linguïstische kenmerken op leesbaarheid van correlationele relaties worden onderscheiden. De bevindingen zijn met name belangrijk voor leesbaarheidsverbeterings- en tekstverwerkingsonderzoek omdat ze laten zien in hoeverre het lezen en begrijpen van een tekst wordt beïnvloed door linguïstische kenmerken. De resultaten schetsen een realistisch, en enigszins ontnuchterend beeld van de mate waarin teksten verbeterd kunnen worden wanneer de inhoud niet kan veranderen. Door de schaal van het onderzoek kunnen we er zeker van zijn dat de resultaten robuust zijn en gegeneraliseerd kunnen worden over een groot aantal teksten en lezers die verschillen in leesvaardigheid.

Curriculum Vitae

Suzanne Kleijn was born on the 16th of October 1988 in Nieuwegein, The Netherlands. After obtaining her VWO diploma (pre-university education) in 2006, she studied Communication and Information sciences at Utrecht University. She obtained her bachelor's degree in 2009 and her master's degree in 2010 (cum laude). She went on to study linguistics at Utrecht University obtaining her second master's degree in 2012 (cum laude). In November of that same year she started her PhD-research at the Utrecht Institute of Linguistics – OTS. Suzanne Kleijn currently works as a researcher within the department of applied linguistics.