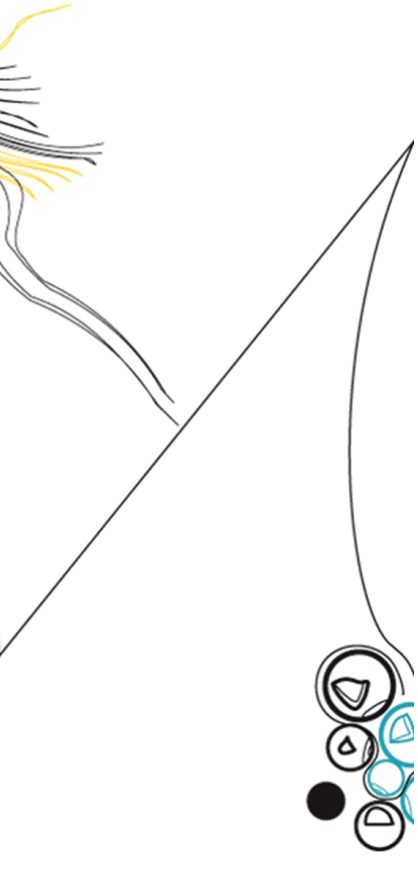
UNIVERSITY OF TWENTE.

Faculty of Electrical Engineering, Mathematics & Computer Science

Large Scale Online Readability Assessment

Rutger Varkevisser

M.Sc. Thesis
November 2016



Supervisors:
Prof. Dr. T. Huibers
Dr. D. Hiemstra
Dr. T. Westerveld

Human Media Interaction Group
Faculty of Electrical Engineering,
Mathematics and Computer Science
University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

Summary

The internet is an incredible resource for information and learning. By using search engines like Google, information is usually just a click away. Unless you are a child, in which case most of the information on the web is either (way) too difficult to read and/or understand, or impossible to find. This research aims to successfully combine the areas of readability assessment and gamification in order to provide a technical and theoretical foundation for the creation of an automatic large scale child feedback readability assessment system. In which correctly assessing the readability level of online (textual) content for children is the central focus. The importance of having correct readability scores for online content, is that it provides children with a guideline on the difficulty level of textual content on the web. It also allows for external programs i.e. search engines, to potentially take readability scores into account based on the known age/proficiency of the user. Having children actively participate in the process of determining readability levels should improve any current systems which usually rely on fully automated systems/algorithms or human (adult) perception.

The first step in the creation of the aforementioned tool is to make sure the underlying process is scientific valid. This research has adapted the Cloze-test as a method of determining the readability of a text. The Cloze-test is an already established and researched method of readability assessment, which works by omitting certain words from a text and tasking the user with filling in the open spots with the correct words. The resulting overall score determining the readability level.

For this research we want to digitize and automate this process. However, while the validity of the Cloze-test and its results in an offline (paper) environment have been proven, this is not the case for any digital adaptation. Therefore the first part of this research focusses on this central issue. By combining the areas of readability assessment (the Cloze-test), gamification (the creation of a digital online adaptation of the Cloze-test) and child computer interaction (a user-test on the target audience with the developed tool) this validity was examined and tested. In the user-test the participants completed several different Cloze-test texts, half of them offline (on paper) and the other half in a recreated online environment. This was done to measure the correlation between the online scores and the offline scores, which we already

know are valid. Results of the user-test confirmed the validity of the online version by showing significant correlations between the offline and online versions via both a Pearson correlation coefficient and Spearman's rank-order analysis.

With the knowledge that the online adaptation of the Cloze-test is valid for determining readability scores, the next step was to automate the process of creating Cloze-tests from texts. Given that the goal of the project was to provide the basis of a scalable gamified approach, and scalable in this context means automated. Several methods were developed to mimic the human process of creating a Cloze-test (i.e. looking at the text and selecting which words to omit given a set of general guidelines). Included in these methods were TF.IDF and NLP approaches in order to find suitable extraction words for the purposes of a Cloze-test. These were tested by comparing the classification performance of each method with a 'baseline' of manually classified/marked set of texts. The final versions of the aforementioned methods were tested, and resulted performance scores of around 50%, i.e. how well they emulated human performance in the creation of Cloze-tests. A combination of automated methods resulted in an even bigger performance score of 63%. The best performing individual method was put to the test in a small Turing-test style user-test which showed promising results. Presented with 2 manually- and 1 automatically created Cloze-test participants attained similar scores across all tests. Participants also gave contradicting responses when asked which of the 3 Cloze-tests was automated.

This research concludes the following:

1. Results of offline- and online Cloze-tests are highly correlated.
2. Automated methods are able to correctly identify 63% of suitable Cloze-test words as marked by humans.
3. Users gave conflicting reports when asked to identify the automated test in a mix of both automated- and human-made Cloze-tests.

In order for a final large scale online gamified readability assessment system to be completed, the automated methods detailed in this report need to be further developed and tested in order to attain (near) human performance levels. Additionally, a lot of work and thought has to go into the gamification aspect, as this is crucial in the success or failure of the final system, which has fallen outside the scope of this project.

Preface

I would like to start by sincerely thanking all my supervisors; Theo, Djoerd and Thijs who have assisted- and provided me with feedback throughout this project. I could not have done this without you.

While doing parts of my research at the WizeNoze offices I was able to ask questions and receive feedback from some of the team members located there. These included experts in Information Retrieval (IR), Machine Learning (ML), Natural Language Processing (NLP), Child Computer Interaction (CCI) and others. I would like to thank all of the WizeNoze team, especially Hanna, Gerben and Rosa for helping me out during my time there.

Another big thank you goes to everyone who participated in either user-test for this research. Including Elsbeth, Diana and Remco, the teachers of the classes of children which participated in one of those user-tests. Your help and input were invaluable.

Finally I would like to sincerely thank my parents who have supported me throughout this phase of my life.

Contents

List of acronyms

AI	Artificial Intelligence
AVI	Analyse Voor Individualiseringsvormen
CCI	Child Computer Interaction
GWAP	Games With A Purpose
HCI	Human Computer Interaction
IDF	Inverse Document Frequency
IR	Information Retrieval
ML	Machine Learning
NARA	Neale Analysis of Reading Ability
NLP	Natural Language Processing
POS	Part-of-speech
TF	Term Frequency

Chapter 1

Introduction

As the years go by, more and more people all around the world now have access to the internet. In the last 8 recorded years alone, the percentage of users who have access to the internet via any means (computer, phone, tv, etc. . .), has gone up by more than 23% to a global internet penetration level of 40.7% in 2014 [1]. These numbers, of course, are much higher when only looking at the westernised world. This is exemplified in Figure 1.1 where the internet penetration is shown at a global- and a European scale in the last 8 recorded years.

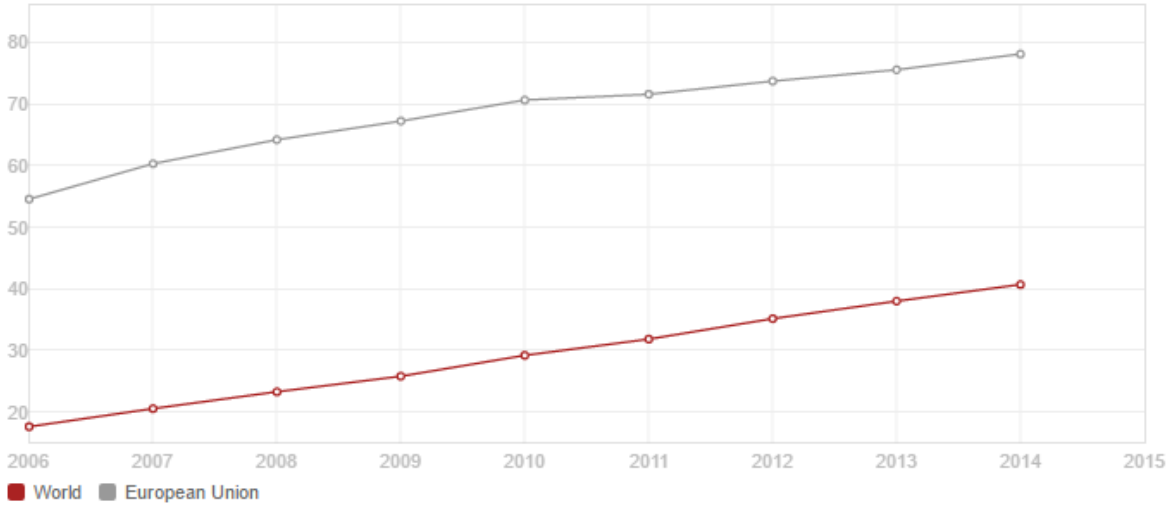


Figure 1.1: Internet users (per 100 people), graph sourced from [1]

This increase in internet penetration across the world now also allows for more people to access the wealth of information which can be found online. This is especially useful for children who can use the web to help them with school work or in order to teach themselves all sorts of things. Another result of the omnipresent nature of the internet is that children nowadays grow up using the internet and the various devices which are connected to it, *Digital Natives*, as some people describe them [2].

However, the rise of the internet has also enormously changed the way we deal with information. Children grow up in a society where (digital) information is more accessible than ever. This accessibility also introduces a (major) problem. Everyone can post anything they want online without verification. This means that people are now easily confronted with partial, wrong or potentially damaging information. This is particularly true for children who (generally) do not yet possess the skill of judging information for its reliability, usability and readability.

And it is that last feature, *readability*, that is the focus of this particular research. And for a good reason. As mentioned in the paragraphs before, the internet is absolutely filled with useful information, if you know where to look. Now there are already various methods and tools out there which can help you, as a user, to sort through the abundance of online content to find good, verified and factual information. The problem of readability comes into play when talking about the difficulty of those pieces of content. Research has shown that in order to read and have a good understanding of a text, around 95% of the words have to be known by the reader [3]. Naturally children have smaller vocabulary sizes, depending on their age, which mean they generally have a hard time reading and understanding various texts which are not specifically made for them.

Therefore readability is an important feature to know for pieces of (textual) content in order to assess the difficulty, and potentially inform the readers, the content creators, or other internet services (e.g. search engines), that a particular piece of content is suited to people of certain age- or reading comprehension levels. The addition of such a feature would greatly benefit children, whom currently often have to look through pages and pages of online content before they find something they can actually understand.

This research investigates methods to automatically ascertain readability levels of textual documents by doing research and user-testing, culminating in the development of readability assessment tool. Which provide a theoretical and technological foundation for a potential future gamified approach to this problem of large scale readability assessment.

1.1 WizeNoze

This research was done in cooperation with WizeNoze, an Amsterdam based start-up company which aims to make the internet a useful resource for children of any age and proficiency. By collecting, indexing and classifying online content for their suitability towards children of certain age- and grade levels WizeNoze is able to present understandable content from across the web to children via their own search

engine JouwZoekMachine[1]. While this classification process for content in their search-engine is working well, the goal is to always keep striving to make it better for the children viewing the content. And since the ideal method of assessing readability by asking children directly about the difficulty of textual content is not viable, the idea of automatic readability assessment via child-feedback with potentially similar results was well received.

1.2 Motivation

Giving children the ability to quickly ascertain whether (online) content is easy enough for them to understand is a goal of this research. With this research we hope to successfully combine two areas of research, readability assessment and gamification, and provide a technical and theoretical foundation for the creation of an automatic large scale child feedback readability assessment tool. The purpose of this tool would be to include children directly in the process of determining readability levels of texts, instead of purely relying on automated algorithms or human (adult) perception.

Previous work in the field of readability assessment (see Section 2.1) has largely remained focused on traditional (paper) methods and formulas, and has widely stayed away from (partial) automation with the current technological advances. One of the goals for this research is to take the current knowledge/information from traditional readability assessment methods and attempt to translate those findings into a digital system.

The gamification of traditionally non-gaming interactions/actions is another part of this research, as the ultimate goal of this research (if proven to be effective) is to have children play and interact with an online system. While the gamification aspect is not the main focus of this research, it certainly is something that needs to be accounted for, as any future system using these, or similar methods of collecting feedback on readability levels on a large scale, have to be interesting and compelling enough for children to use.

1.3 Research Goals & Questions

There are two themes to this research. Readability assessment and gamification. These two themes will be combined in order for this research to be a success.

Concerning readability assessment (see Section 2.1), we have chosen to utilize the *Cloze-test* method because of its widespread use, the potential for automation

[1] http://www.jouwzoekmachine.nl.

and the possibility of gamification. This choice was made based on the preliminary work and research done for this project. In this preliminary work, a prototype system was developed wherein multiple digitized and gamified approaches of the Cloze-test readability assessment method were tested and evaluated. The results of which showed promise for using such a digitized adaptation of a 'standard' readability assessment method, and was therefore adapted for this project as well.

By using a modified/enhanced version of the prototype system, this research hopes to determine whether that system, a combination of a digitized and gamified traditional readability assessment method, can accurately measure readability. If not, what changes can be made to make it work as intended?

Another aspect of this research, once the readability assessments accuracy has been established, is scalability. Scalability in the context of this research means that we want to design the system in such a way that it allows for hundreds to thousands of users (children) to be able to play and interact with the tool, while the system is continually processing and handling the actions and results and also adding and updating content without major human involvement. While the development of a completely functioning scalable system is outside the scope of this research, scalability itself is heavily emphasized in the development of a digital readability assessment tool.

The main research goal is as follows:

- Delivering the technological and theoretical foundation of a scientifically substantiated readability assessment tool which provides the basis for a future gamified approach of large scale readability assessment.

In addition to the main research goal the following items are the various subgoals which together with the main research goal determine the successful outcome of this research.

- Any online or digitized readability result must reflect/correspond to that of (traditional) offline Cloze-test results. I.e. the measurements have to be remain valid.
- The final readability assessment tool must be developed in such a way in that it emphasizes scalability and that it provides a basis for any future gamified adaptations of the tool.

Based on the research goals as stated above, the following are the research questions which this research attempts to answer.

1. Is the developed (prototype) system capable of correctly measuring the readability of a text given the users proficiency level?

2. Assuming we are able to assess the readability correctly, how can the developed system and its results be made scalable?

1.4 Methodology

In order to answer the research questions and attain the research goals as stated in the previous section, this research project has divided itself into two main stages. A research stage and a developmental stage. With the results of the first stage defining the foundation of the second stage.

The focus of the first stage, Stage A (see Chapter 3), is on answering the question whether or not the digitization of the traditional (paper) Cloze-tests has any influence on the validity of the results. To answer this question a user-test was set up with children of similar ages, and tasked them with filling in a couple of Cloze-test texts, of different difficulty levels, with everything being identical in both the paper as the online version apart from the method of interaction.

The second stage, Stage B (see Chapter 4), is focussed on researching whether or not the process of converting a 'regular' text into a Cloze-test text can be automatized. This is done by emulating the (regular) manual process in creating these Cloze-test texts into potential automated systems. Finally the result of these automated Cloze-test texts are tested by performing a Turing-test style test wherein the automatically generated Cloze-tests are compared to manually created ones. This developed tool should then provide a technological and theoretical basis for future readability assessment applications.

Each of the previously mentioned chapters contain a comprehensive description and in-depth analysis of the topics therein and a discussion and analyses of the results. The final result of these chapters, and therefore this research, should be a technological and theoretical blueprint for the continued research and development into a system capable of detecting readability levels and presenting them in a gamified manner.

1.5 Thesis Structure

In the following chapter (Chapter 2) the reader will be introduced to relevant research into fields relating to this study. This will include topics such as readability assessment, gamification and child computer interaction.

Chapter 3 (stage A) contains a comprehensive description of the research done into the validity of the digitization process of the traditional (paper) Cloze-test, centered around a user-test.

Chapter 4 (stage B) looks at the method of automating the process of extracting words from a text for the purposes of a Cloze-test, and whether the automated systems can provide a similar level of outcome compared to a manually constructed test.

To close of this thesis, Chapter 5 will discuss the results of the previously mentioned chapters (Stages A&B) and concludes the work done for this research project including looking at the research goals and answering the research questions stated in Section 1.3. Included in Chapter 5 is a discussion on the potential future work derived from the results of this research.

Chapter 2

Literature Review

This research aims to take elements from various fields of research and attempt to successfully combine them for use in this project. The following sections will outline some of major points gained from research into the different fields.

2.1 Readability Assessment

When talking about readability you have to look at your audience and their abilities when it comes to reading and comprehending textual content. For this particular project the targeted audience are children, and what differentiates children from adults is their rapid rate of acquired (textual) knowledge. Children learn words in different ways, most of which are learned through daily social interactions with others without any (special) effort required and is therefore learned implicitly [4].

When a child starts to read, newly acquired words are stored together with current knowledge. An explicit growth of a child's vocabulary size begins from the moment they go to school. The reason why this is important to this research is that vocabulary size is highly correlated with reading comprehension [5]. A good example of this rapid growth can be seen in Figure 2.1.

The implications of a low vocabulary size in a child or person in general, whether due to age or proficiency, is that it can affect the person's desire to read or lead to avoiding reading altogether. This severely hinders their progress and development of reading comprehension skills. The children who do have the ability to read well tend to read more and improve their readability and reading comprehension abilities [7]. This is corroborated by research done by Biemiller (2003) which shows a high correlation (.91) between vocabulary size and reading comprehension/readability skills [8].

The reason why a small vocabulary size affects the ability to read a text well has also been researched. Research by Appel & Vermeer (1994) showed that in order to

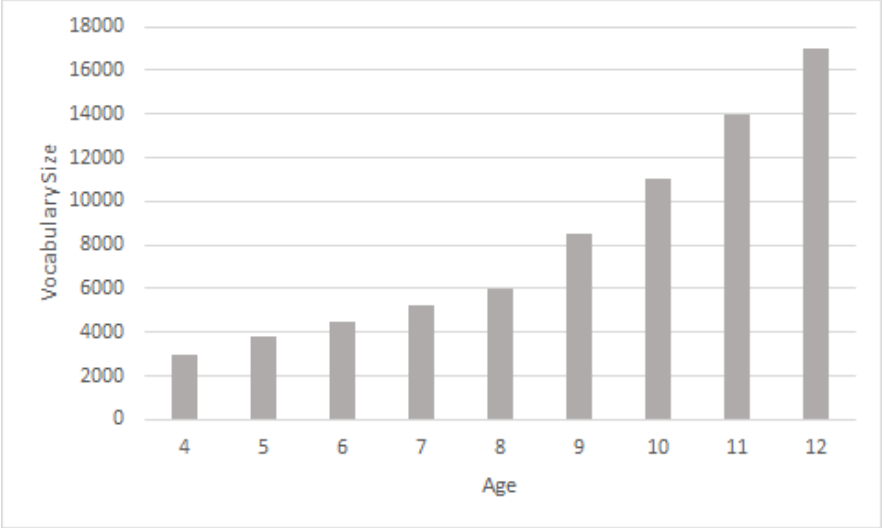


Figure 2.1: Vocabulary size of Dutch children in primary education [6]

read and have a good understanding of a text, 95% of the words have to be known by the reader. If readers do not know or understand $>$ 5% of the words in a text, they will have a problem understanding and reading it. The implication of this is that in order to reach that 95% text coverage, a minimum vocabulary size of around 5000 words is required for being able to seamlessly read everyday texts [3].

Almost every country in the world have their own method of standardized assessment of reading ability in children. Examples of these are the Neale Analysis of Reading Ability (NARA) system in the UK, and the Analyse Voor Individualiseringsvormen (AVI) system in the Netherlands. These assessment systems provide a grading scale on which the reading abilities of a child are measured. For example, the (Dutch) AVI system is made out of 12 different levels which signify various stages of a child's progress through the school system[1].

- **AVI start**; Entry level; prior to groep 3
- **AVI M3/E3**; Mid groep 3 / End of groep 3
- **AVI M4/E4**; Mid groep 4 / End of groep 4
- **AVI M5/E5**; Mid groep 5 / End of groep 5
- **AVI M6/E6**; Mid groep 6 / End of groep 6
- **AVI M7/E7**; Mid groep 7 / End of groep 7
- **AVI Plus**; Anything above E7

[1] For a comparison between international school systems, see Appendix A.

These 'ratings' signify the reading comprehension of children through the primary education system. In the AVI system, as shown above, a child should theoretically be at the level corresponding to the groep (grade) he/she is in. These levels can be higher or lower depending on the (reading) skill level of the child. But to assess the reading comprehension level itself there has to be some form of testing. Two of the methods which these tests can consist of are listed below [9]:

- **Multiple Choice**; Multiple choice is different from other tasks as it asks the user to select which of the four (or more) responses is the correct answer to a question. This task has a higher demand on the user because it forces them to compare the different options. The multiple choice format is usually suitable for (large scale) group administration, and has the advantage of an easy scoring system. It is also highly flexible in nature and can be converted/included to suit the needs of other tasks.

- **Cloze- and C-tasks**; Cloze- and C-tasks are methods in which words, or part of words are deleted in a certain interval and the reader is tasked to fill in the correct replacement. These tasks can easily be administered to groups, which is helpful for assessing large numbers of students/children and scoring is very simple as it is just a percentage of the total number of correct replacements.

The reason for singling out the previously listed methods is that they are the ones most suited to be adopted for automatization in a digital environment. The multiple choice method is very well known and widely used, but when thinking about a digital adaptation for assessing readability there is one big (foreseeable) issue. The issue is in the 'generation' of meaningful options besides the answer. Whatever the method for asking the question is, for the correct application of a multiple-choice type test it is needed that all the options make sense in some way as to make the reader think about which answer to choose. This becomes immediately more complicated when talking about an adoption into a fully automated system which will have to automatically generate those other options.

The Cloze-test, as previously mentioned in this thesis, was originally developed by W.L. Taylor in 1953 [10] and is a method for determining the readability of a text, by challenging the reader's abilities to deal with the content and structure of the text they're reading. This is done by omitting words at a certain interval and replacing those words with blank lines, see Figure 2.2 for an example paragraph. Depending on the type of Cloze-test you prefer to use, certain words can be preferred (e.g. verbs) or skipped (e.g. names, numbers) for substitution. There are 2 general methods in which the holes of a Cloze-test can be constructed. A random- *(fixed ratio)* or a selective approach *(rational fill)* [4]. In the *fixed ratio* approach the researcher

selects holes by deleting words in a regular frequency, i.e. every third, fourth or sixth word. The advantage of this approach is that it is less sensitive to the (inadequate) choices or assumptions of the researcher. However, the disadvantage of this method is that it can create holes in places where the answer is too difficult to give since the selected words might rely on extra-textual knowledge, or too easy because of their grammatical- or lexical predictability. In the *rational fill* approach the holes are chosen by the researcher i.e. only deleting nouns or verbs. This approach has the advantage that the selected words only call on contextual knowledge, both in- and outside of the context of the sentence [11]. Note that in the Cloze-test example paragraph in Figure 2.2 the omitted words are ambiguous and do not seem to have one clear answer, this would suggest that a rational fill approach would be better suited to score the Cloze-test containing this paragraph.

After selecting an approach it is the reader's task to read the text and fill in the words. Cloze-tests have clear quantifiable results [12] and guidelines [13] have been written to aid in constructing these tests. The Cloze-test is currently widely used in teaching environments (i.e. learning a second language [14]), since it 'forces' students to read more carefully, use contextual clues and become actively involved into what they were reading [15].

```
Today, I went to the ________ and bought some milk and eggs. I
knew it was going to rain, but I forgot to take my ________, and
ended up getting wet on the way.
```

Figure 2.2: Example Cloze-test paragraph[2]

As opposed to straight up readability formulas, this research/project wants to know and take into account (some) factors about the reader. For example, if the text is easy, but the context of it is unknown to children thereby causing them to make mistakes, is something that does affect the overall readability and is information we want to know. Both the C-test, and even more so the Cloze-test, can give us that information (on a large scale) by requiring users/readers to fill in words which contain contextual information about the text.

There are multiple methods to score the reader's performance with these methods. These are *exact*- and *semantic* scoring. In the *exact* scoring method the answer given is only marked as correct if the reader fills in the exact word which was removed from the text. In the *semantic* scoring method both the exact word and words which are correct at a contextual level are marked as correct. This method solves the 'issue' of readers filling in words which are synonyms or contextually correct but not counted as correct in the *exact* scoring method.

[2]Source: https://en.wikipedia.org/wiki/Cloze_test.

The validity of Cloze-test in particular have been examined and are compared to that of the 'regular' assessment methods such as; open questions and multiple-choice questions [4]. A validation experiment from Kamalski (2005) showed that Cloze-tests have a high correlation with questions that measure conceptual understanding. In the area of internal reliability the Cloze-test scores even higher then the 'regular' assessment methods. A recent study by Gellert & Elbro (2012) also found a strong (.84) correlation between the Cloze-test and traditional 'regular' tests [16].

2.2 Gamification

With over 350 million players around the world [17] and revenues in excess of $91.5 billion[3], overtaking both the music- and film industry, the gaming industry is now looked at by both researchers and companies for clues on how to improve their research and/or products.

In both interaction design and digital marketing, the concept of using game design elements in non-game contexts in order to motivate and increase user activity and retention, is on the rise [18]. This is known as *gamification* or *'serious games'*. While the term 'gamification' is relatively recent its application in various forms, areas and fields can be dated back millennia starting with military applications and eventually moving towards education and business in the mid to late 1990s [19].

Within the gamification spectrum there are different areas which applications can focus on. One of these areas is the area in which the computer assists humans in performing and/or completing real-life tasks by making them more fun and interactive. An example of which is the fitness application *Zombies, Run!*[4], which uses interactive storytelling and rewards for the purposes of motivating and increasing the users' real life running and therefore their stamina and fitness levels. This research focusses on another area of gamification, namely Games With A Purpose (GWAP)s, also known as 'serious games'.

GWAPs are games that ask users to perform basic tasks which cannot be automated. In these games the task the users are asked to perform have a side effect where the action(s) they are taking result in a useful computation [20]. What this basically means is that humans 'train' computers with the results of their actions. Machine learning algorithms can take the input of thousands of these users and use it to train themselves in areas which are (usually) incredibly difficult to automate e.g. computer vision problems. Another big part of GWAPs is the ability to make 'work' fun. Human Computer Interaction (HCI) researchers have recognized the

[3]Source: Gamesindustry.biz (http://www.gamesindustry.biz/articles/2015-04-22-gaming-will-hit-usd91-5-billion-this-year-newzoo).

[4]Developed by: Six to Start (https://zombiesrungame.com/).

importance and fun in user interfaces. By creating interfaces which use game-like interaction it is possible to increase the enjoyment and engagement between the user and the software.

Figure 2.3: The ESP game

An example of this type of game is *the ESP Game* developed by Luis von Ahn[5] (2006), the pioneer of GWAPs. In the ESP game (see Figure 2.3), two players are randomly paired for two-and-a-half minutes as they are shown a series of images to label. The game itself does not directly ask them to label the images however, but rather both players must try to enter the same word as their 'partner' for each image on the screen. Neither players can see their partner's words. When the players agree on a word, each is given a new image to label. The goal is then to agree with the partner on words for as many images as possible. The words the players agree on for each image can then be used as labels for images throughout the web, as they are extremely accurate. This in turn greatly improves image search technologies which relies heavily on image data (such as labels) in order to correctly retrieve a set of relevant images. This is a perfect example of a task which humans are (currently) much better at than computers. While labelling images is normally a pretty boring task to do, by *gamifying* the process the players are provided with a form of entertainment, and the system uses the input of the players to further train its algorithms. A win/win situation.

The rules of a GWAP should encourage players to correctly perform the necessary steps to solve the problem and should, if possible, involve a probabilistic guar-

[5]Luis von Ahn is one of the biggest pioneers in crowd-sourcing technologies. He is the founder of ReCAPTCHA (sold to Google in 2009) and is currently the CEO of Duolingo, currently the most popular (free) language learning platform in the world.

antee that the game's output is correct, even if the players do not want it to be correct. In his research, von Ahn also describes how to build a (successful) GWAP [20]. According to him building a GWAP should start with first creating a game so that its structure encourages computation and correctness of the output. The next step is to implement methods which improve player enjoyment and increase the challenge level [21]. Since the success of almost every GWAP rests on its ability to attract and retain the attention of players, it is important to keep in mind certain factors which have been shown to improve the challenge level of the game.

First thing to note is having a **timed response** in the game. By setting time limits for game sessions it introduces challenge into the game. When completing assignments within the time limits extra points may be given. Research has shown that goals that are both well-specified and challenging lead to higher levels of effort and task performance than goals that are too easy or vague [22]. The second method, and maybe most direct method for motivation is **score keeping**. By awarding points for each instance of successful output during the game it increases player motivation and also provides players with performance feedback. Introducing **player skill levels** can also have a positive effect on your game. By earning points players have the ability to increase their skill level, which in turn motivates players to keep playing in order to increase that level. A simple addition of **high score lists** can act as a motivating tool for players to perform better and/or keep playing. By including multiple lists (e.g. hourly/daily/all time) goals of increasing difficulty can be defined. The final factor which is mentioned by von Ahn is to include **randomness** into the game. By including randomness it keeps the game interesting and engaging to players. It also adds a bit of uncertainty whether the task can be completed within the time limit.

While these methods help making the game more challenging and enjoyable for the players, it is very important to know whether or not the output of the actual games can be trusted. Players will always try and circumvent the game's default systems in order to attain a high score or just to mess with the system. Von Ahn describes several mechanisms that have proved to be successful which can be applied to guard against player collusions and guaranteeing the correctness of the computations across all game types and structures.

- **Random Matching**; Random matching vastly lowers the chance of players knowing their partners identity and therefore preventing cheating behavior.
- **Player Testing**; The game may present players with (randomly chosen) inputs for which all possible correct outputs are already known. When the output of these inputs do not match the known outputs, the player can be marked as 'suspicious' and none of their results should be trusted. A 50/50 split in presenting players with testing/new input should theoretically ensure a high

probability of correct answers. I.e. the answers/results of players who are not intended on 'properly' playing the game e.g. giving bogus answers, should now be filtered out.

- **Repetition**; The game should be designed so it does not consider an output correct until a certain number of players have entered it. This strategy for determining correctness enables any GWAP to guarantee correct output with high probability.

- **Taboo Outputs**; By preventing the players to enter certain words, a larger dataset can be acquired to describe the image it is associated with. For example, players are tasked to describe a picture wherein a dog is chasing a ball in the park. Taboo outputs could contain the words: 'dog', 'ball' and 'park' in order to get more additional information on the elements in the picture.

Before concluding this section on gamification, there is one more aspect of gamification that is interesting to examine further. That is the *replayability* aspect of gamification. Especially important for repetitive tasks such as the one proposed in this research.

To get a better understanding of how to tackle this aspect we can look towards the gaming sector where this is a major factor. A study by Frattessi et al. [23] analyzed existing research on play and replayability and condensed this into five different specific aspects that they believe is a driving force for replayability. The following is a condensed summation of the five factors:

1. **Difficulty**; difficulty uses a form of competition and challenge to generate replayability.

2. **Completion**; completion feeds upon the natural curiosity of players and encourages them to try each possible scenario. Designers can enhance this factor by adding achievements and unlockables to the game.

3. **Social Aspects**; players want to be the best, but this threshold is always in flux. Direct competition and high score tables take advantage of natural competitive nature of players.

4. **Randomization**; the most obvious solution to keeping something new and interesting is to keep changing it. A properly randomized game will keep a player's interest as the experience varies ever so slightly.

5. **The Experience**; Games can sometimes bring something unique to a player. This uniqueness brings players in and draws players back because they cannot find other games which offer the same experience.

While these factors might be more specifically aimed at video games, it does provide an insight into elements which can potentially aid this research.

2.3 Child Computer Interaction

Reading this section you might be wondering why we changed the 'standard' HCI with CCI. The reason for this is that children are very different in their interaction with computers compared to adults. And even within this group of children there are large differences. At different ages, children's relation with computers and other interactive technologies vary widely, reflecting on their changing interests. A study by Markopoulos & Bekker (2003) describes four stages of development in children [24] using the model from Acuff & Reiher (1997) [25]. From the four stages two are especially relevant to this research, as it comprises of the targeted user group.

- **Rule/role stage (8-12 years-old)**; In this stage the interest from children generally shift from fantasy towards reality. They start to play in groups/pairs and are more interested in competition. Products targeting this age group can therefore also be more complex and challenging, incorporating elements of variation and competition within their design. A big change for this group is that they shift from *learning to read* towards *reading to learn*. They can understand abstract terms and complex sentences and also develop the ability to (critically) analyse what they are reading. This shift also brings a rapidly expanding vocabulary size (see Figure 2.1 on page 8) and writing and spelling also greatly improve within this stage. The children in this stage also start using laptops, phones and tablets with a more serious design and look, moving away from the colourful design for products targeting the earlier stages. This is the first stage were (a part of) the targeted audience start to emerge. Because of the much improved ability to read, write and interact with digital systems/devices, this is an audience we are interested in targeting.

- **Early- and late adolescence (13+ years-old)**; This stage sees children moving away from being dependent on their parents and peers. The children in this stage become more socially and more goal-oriented. They can handle abstract problems and complexity better and also understand difficult concepts. Products designed for this stage are very similar to products designed for adults which relate to activities that appeal to this group such as sports and social activities. This is secondary age-group that is targeted for this project. The children in this group have an even better grasp of grammar, language and sentence structure which greatly helps in assessing readability.

Keeping the characteristics of the above mentioned user groups in mind, we now turn our attention towards the design process when it comes to these user groups, and the limitations that arise as a result.

One of the skills needed for operating (input) devices such as mobile phones and computers, are fine motor skills [26]. These skills are first learned around the age of 3 years old and rapidly progress afterwards. Older children (7+) see the speed of their movements increase and the variability in their movements decrease. Also important in the interaction of children with computers is something called *bimanual coordination*. Bimanual coordination involves coordinating the use of both hands in time and space. Specifically to the computer, this means that they are able to perform multiple key strokes on the keyboard and a combination between keyboard and mouse. The very basic of bimanual coordination start around 2 years-old, and increases significantly in complexity with every year [27].

The visual complexity of the program should be taken into account when designing for children, since they cannot process visual information as quickly as adults can [28]. One suggested method for children is to start with a very basic visual model, and gradually introducing more actions and objects to the user interface as they become proficient with the system [29].

Within user interfaces it is very important to allow for rapid actions, since children will often be less patient compared to adult software users. Children require quick feedback, and if they do not get it, they are likely to move to another activity. This includes giving feedback on the status of an action (e.g. loading % or progress bars) and should be able to interact with the system and cancel the action if they wish to do so. To encourage the exploration of technologies it is important to allow for the reversibility of actions, depending on the type of application of your program. A final (action) design element which can be very helpful (specifically) to children is making actions incremental. This will help children and avoid the need for them to formulate complex instructions. Combined with timely and informative feedback, this can help children accomplish otherwise (very) complex tasks [30].

Given the type of research of this project, one of the issues that might need to be tackled is the issue of *typing*. While not limited to children alone, typing is something children have a generally hard time with. Most children, especially at younger ages, do not have typing skills and dexterity comparable to that of most adults. This does not only impact the precision and speed of filling in words via keyboard, it also present a visual challenge. Since the focus of children during typing tasks is on the keyboard, they cannot, or have difficulty, focusing on the screen and any elements that might require the users attention [31]. Another issue that arises with requiring typing is the issue of spelling. This can cause problems when entering commands or when the program requires an exact input without applying

any spelling correction [32].

The above mentioned topics cover some of the (visual) design aspects for creating an interactive environment for children. These guidelines help us think and relate our design decisions to a children's perspective but a (big) problem remains in the design process itself. Because the design is made by adults for children there will almost always be some misalignment between what the developer/designer think is best, given their previous experiences and preferences, and what children actually want. To help this process Markopoulos et al. [24] discusses the potential of including children in some way in the design process.

A model introduced by Druin (2002), see Figure 2.4, illustrates the relation of the various roles children can play during the design process. The inner circle represent the traditional role of children as end-users of technology, with no involvement in its design. Moving outwards from the inner circles in the diagram of Figure 2.4, the role of children changes in two ways: first it becomes more active and responsible and second, children get involved in more stages of the design activity. A more common and pragmatic view is to include children as testers of products i.e. usability test. The more radical approach, shown in the outer layers of the diagram shown in Figure 2.4 is that children should act as designers.

For the purposes of this research we incorporate children mainly as the two inner circles of the diagram as shown in Figure 2.4, as a user and a tester.

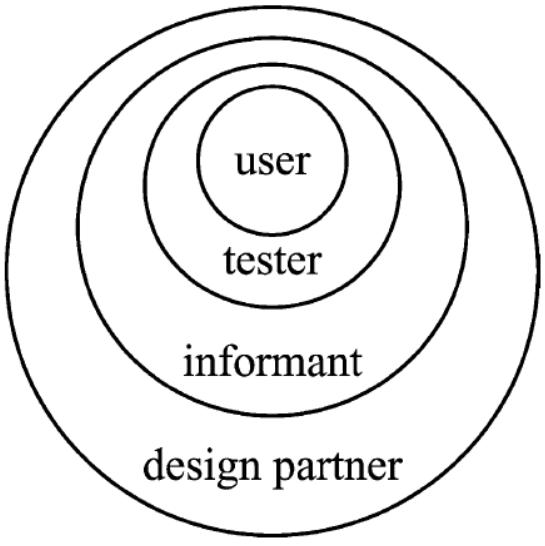


Figure 2.4: The four roles that children may have in the design of new technologies[7]

One research which combined the elements of *gamification* from Section 2.2 and CCI is a study done by van den Bosch et al. (2010). In their study they developed

[7]Source: Druin (2002) [33].

a GWAP-based collaborative approach to common sense resource development, specifically aimed at children [34]. The study contained two distinct phases. In the first phase of the game children were given a concept to describe (e.g. 'house') for which they could choose from a limited number of templates (e.g. 'isA'). In the second phase, to provide a type of in-game human validation typical to GWAPs, children were asked to guess the given concept based on another user's assertion with the subject hidden (e.g. '. . . is a means of transport').

Results of the study by van den Bosch et al. showed that children (aged 10-12) were able to draw upon their common sense knowledge and their language skills to describe concrete and abstract words in the form of valuable assertions. A noteworthy observation made during the study was that the participating children reported to the researchers how much they liked doing the task and had fun doing it. Overall, they concluded that their study indicated that using a child-oriented GWAP could lead to common sense knowledge resource creation. Additionally, they believe that the age range of 10-12 years-old is the right target for the type of GWAP they used in their study.

This study, while not exactly like the one we are proposing to do in this research, shows us a couple of things. First of all, the study was done on a similar age group we attempt to target for this research, and proved that GWAPs or gamified systems targeting that age group can be successful and fun. Secondly results from the study by van den Bosch et al. showed that children can be 'used' to obtain and gain knowledge for several purposes. Lastly, the study showed us that a project such as theirs, using some form of textual assessment, is viable for the targeted age-group and can be successfully done.

2.4 Conclusion

These three topics, **readability**, **gamification** and **Child Computer Interaction** form the basis of the research presented in this document. Because of the broad scope of these fields of research this study looks at certain relevant topics within these field and attempts to apply it where applicable.

One of the major questions surrounding the readability topic for this research is whether the digitization and automation of the Cloze-test is possible while retaining its validity of assessing readability of a text. This has not been done before and is therefore a very interesting part of this research.

For this study we can apply a lot of general lessons gained from previous research into the area of gamification. Specific adaptations however, such as the GWAPs designed by Louis von Ahn differ from the versions proposed for this research, given that the underlying method for this study is an already established test

dating back to the early 1950s. We can certainly apply (some) lessons learned from different gamification research, but the application is restricted in the sense that it is crucial that the validity of the original Cloze-test is retained.

Most of the interaction- and psychological aspects from this study pertaining to the interaction between children and computers are covered in the research done in Section 2.3. However, there are factors which might contribute towards a potential deviation from the 'regular' approach e.g. differences in Dutch upbringing/schooling compared to other countries.

Chapter 3

Stage A: Digital Conversion Assessment

The central theme of this study is whether it is possible to create and sustain a scientifically accurate automated digital readability assessment system. This stage, stage A, is the first and crucial step in the successful creation of such a system. The reason for this stage being the first step is because it pertains to one of the core topics of this research, readability assessment. Before any further steps can be taken into factors such as enjoyment, accessibility and interactivity, we need to know whether or not the readability that is measured via the proposed system is actually scientifically valid. In order to answer this question a user-test was set up.

This chapter will detail the motivation, design, execution, results, analysis and discussion of a user-test into the digitization of the Cloze-test readability assessment method.

3.1 Motivation

The core of this entire study is based on the premise that we are able to extract the readability level of a given text by applying a readability assessment method. The method that was researched and chosen for this research is the 'Cloze-test' method (see Section 2.1 on page 7).

The Cloze-test has been used for a long time as a measurement of readability. In short, the Cloze-test takes a (regular) text for which the readability needs to be assessed and omits a number of words from the text. The user is then tasked with filling in the blank spaces in the text with the correct words, or via a method such as multiple choice. The readability itself is scored by taking the percentage of correct answers from the total number of answers given. The scoring is measured

as follows[1]:

- $>$ 60% correct replacements - independent reading level
- 40-60% correct replacements - instructional reading level
- $<$ 40% correct replacements - frustration levels

This brings us to the central question. This research uses the Cloze-test as a core method for assessing readability, however since we are attempting to build a large scale online system, we have to digitize this process. Now while the validity of the readability assessment of 'regular' (paper) Cloze-tests have already been researched and proven [16], this is not the case for any possible digital adaptation. This means that before we can continue onwards with this study we first need to prove whether or not any differences are introduced by the digitization of the Cloze-test.

3.2 Research Questions and Hypothesis

Given the problem as described in the previous section, the following central research question was formulated for this stage of the research which the user-test will seek to answer.

Does the online adaptation of the Cloze-test measure similar results as the offline (paper) version?

In addition to the main research goal of the user-test as stated above, there are several additional questions we can attempt to answer following the results of the user-test. These might not all be critical for the success or failure of this project but could give us insight into various aspects of readability and the target audience.

- *Given that we perform a user-test on classes from different grades, is there a significant difference in performance results between those classes?*
- *Is there a significant difference in results between boys and girls?*
- *Does the order in which the participant take the test, online or offline first, affect the outcome in a significant way?*
- *Is the amount of recreational reading done by a participant an indicator of improved performance?*

[1] Source: M. Hathcock (2013) [12].

3.3 Research Design

In order to answer the central research question and other questions posed in the previous section a user-test had to be designed. This user-test had to incorporate all the features needed to reach a definitive and scientifically valid answer to the question on whether the online version of the Cloze-test is statistically identical to the offline version.

The following subsections will detail the design and procedure of the user-test. Instead of describing the entire (online) system and all of its functionalities in depth, section 3.3.1 will list a number of features which are core to the user-test and the system. Section 3.3.2 will detail the general outline of the user-test.

3.3.1 Core Features

Cloze-Test

As previously mentioned in this chapter, the core method chosen for this research is the Cloze-test. Much about this particular method has already been mentioned in previous sections of this report, see section 2.1 and section 3.1, but there are a number of things about the Cloze-test method which are pertinent for both the online- and offline portions of this user-test and warrant further explanation.

The Cloze-test method is a method for assessing readability by omitting words from a text and having the user fill them in. As mentioned in section 2.1 there are two approaches which are commonly used when it comes to creating Cloze-test, a fixed-ratio approach and a selective (rational fill) approach. While the fixed-ratio approach would be much easier to implement and test, both online and offline, we wanted to avoid having omitted words in the texts where the answer was either too difficult to give, since they might rely on extra-textual knowledge, or too easy because of their grammatical- or lexical predictability. Therefore we chose to use a selective (rational) approach which has the advantage of being able to only select words which call on contextual knowledge, both in- and outside of the context of the sentence/text.

While this choice of using a selective approach did result in having to spend more time creating the Cloze-tests, there were a few useful aides available to assist in this process. Guidelines have already been written on how to convert a text into a good Cloze-test. For this research we have adopted the guidelines from Kraf, Lentz & Pander Maat [13] as previously used in a research by Raaijmakers [4]. These guidelines list the following:

1. Do not omit words the title and first sentence of a text. This is done to allow the reader to get familiarized with the text.

2. Do not omit words in the text when only little textual knowledge is required in order to fill in the hole:
    - 'function'-words (prepositions, determiners, conjunctions, pronouns).
    - Name or technical term which is used for the first time.
    - Names and numbers.
    - Auxiliary verbs.
    - Copular verbs.
3. Do not omit words where there is local lexical predictability.
    - I.e. words part of a common expression.
4. Try to omit nouns and verbs wherever possible, since these words usually contain the most information.
    - Try to use as little adverbs and adjectives as possible.

One of the constraints in this stage of the research was the maximum allotted time for participants (school children) who took part in this user-test. In order to give each participant enough time to finish 4 tests, we chose to limit the size of each Cloze-test and omit a maximum of 9 words from each text

With this in mind and by using the guidelines as previously mentioned we went through each of the 10 different texts (more on the texts in section 3.3.2) and selected around 7-9 words in a relatively steady interval throughout the text to be omitted.

The final step in the creation of the Cloze-tests was the decision on the method in which the participants give the answers. There were 3 viable options:

1. **Exact fill**; the user is tasked to write down the correct answer on each open spot in the text.
2. **Multiple Choice**; the user is tasked to select the correct answer from a list of several possible options for each open spot in the text.
3. **Drag and Drop**; underneath the 'clozefied' text all words that are removed from that text are listed randomly, the user is then tasked to link each word with the correct open spot in the text.

Since the goal is to scientifically prove whether or not a correlation exists between a traditional offline Cloze-test and a similar online test, it was pertinent to stick as closely to 'proven' methods from literature as possible, therefore the *drag*

and drop feature was removed as a possibility as this particular approach is not very well researched. This left the choice between the exact fill and the multiple choice method. While the exact fill approach is the most common, used and researched Cloze-test approach there were a number of (potential) issues that made me decide on using the multiple choice method instead.

There are several elements introduced when designing a Cloze-test which requires the users to fill in the answer. Both for the online and offline portions of the test. For one thing, it introduces typing into the equation for the online part. This is a skill which not all children possess quite well (yet) and might result in a lot of misspelling of words and increases the time required for the test itself. Additionally it also greatly increases the time spent analysing the results, since (depending on how you do your scoring) multiple answers might be correct in a given open spot in the text. Additionally where do you draw the line then in what answer is correct or not? Also at what point do misspelled or mistyped words become wrong answers?

To avoid these problems the choice was made to use the multiple choice format. This is also a format which has been proven to give accurate results in terms of readability scores [35]. A study by Alderson (1990) showed that providing choices for the deleted words lessens the participants memory load and makes the test taking process easier and faster [36].

While the multiple-choice approach does have its own drawbacks e.g. more time is required in the creation process of the test, for the purposes of this particular research we believe it is the better choice compared to the Exact-fill method. While we were now required to think of (3) possible options for each omitted word in a Cloze-test, the time requirement for a participant taking the test is reduced, scoring is made much easier since there is only 1 single correct answer for each open spot in the text, and the differences between the offline and online test can be kept to a minimum without requiring the participant to possess certain computer skills (like typing).

Offline and Online Tests

Using the previously established blueprint (the multiple-choice Cloze-test), there is now a basis for comparing the offline readability results with the online results. However, this blueprint still had to be implemented.

To start the process of creating multiple Cloze-test, 10 online texts on various subjects were taken that were near, or slightly above or below the targeted user-group reading proficiency level, using the WizeScan program[2] (more on this in sec-

[2]WizeScan is a chrome plug-in developed by WizeNoze which assesses the difficulty of text on a webpage and gives it a score on a scale of 1-5. 1 indicating suitability for children with very basic

tion 3.3.2).

In order to convert each of the 10 text fragments into Cloze-tests suitable for the user-test several steps had to be taken. At first, we went through each text fragment and marked all words suitable for deletion via the guidelines as described in the previous section. Secondly, given the participant time constraints of the user-test a total of around 7-10 words out of all marked words per text fragment were selected in a relatively consistent interval. Finally the last in this process was to think of three substitute answers for each removed word from each text, since we are using the multiple-choice Cloze-test as a blueprint for this user-test. The resulting (offline) stories are added as a appendix to this report and can be found in Appendix B on page 77.

Having created the offline Cloze-test, all the stories plus the omitted words and their three suggestions each now had to be converted for use in the online system. To do this smoothly a few online (admin) tools were created to assist in this process. These tools can be seen in Figures 3.1 and 3.2. To convert the offline text into a Cloze-test format the tool seen in Figure 3.1 was used. This tool allowed the user to paste a text and manually mark the words which you want to extract from a text for Cloze-test purposes. Additionally you can give a title to a text and select the base difficulty level for the text. After completing all these options the text would be saved in a database, along with the extracted words, title and level for use in the online Cloze environment.

Another ability that had to be implemented in the online system which was the ability to add additional options (words) for each word extracted from the text for use in multiple-choice or other potential formats which uses additional choices besides the correct ones. This tool can be seen in Figure 3.2 and shows how for each word there are a maximum of three potential additional answers. Again, the result of changes with this tool were saved in a database which links the additional answers with the actual correct answer together.

Although the goal was to keep the Cloze-test as identical as possible, there were some (small) differences between the online- and the offline test. As explained in the paragraph above for both the on- and offline versions there were now four possible options to choose from for each open position in a Cloze-test. The correct answers plus three additional options. As can be seen in Appendix B, each open spot in the text was labelled with a number indicating the position in the text and each possible answer was labelled with a, b, c or d. To ensure a random distribution on the position of each answer for each word in the Cloze-text, a random number generator was used to generate a number between 1 and 4 which corresponded with a position (a/b/c/d) of the possible answers. This was done differently in the online version. As

reading skills, 5 indicating a text being suitable for children/people with a higher education.

Figure 3.1: Cloze-test creator

Figure 3.2: Word substitutions

opposed to the offline version where the correct position (a/b/c/d) for each extracted word from the text was fixed, this was digitally randomized for each user in the online version. This was done in an effort to prevent the blind copying of answers (e.g. 1=a, 2=c, 3=c, etc. . .) from other participants who might have the same text either on- or offline.

The final difference between the offline and online test is that in the online tests it shows, after having selected an answer for each of the missing words, what the score was and which answers were wrong and which were not, and asks the participants to correct any mistakes made. This, of course, was not possible for the offline tests. This distinction between tests was made in order to safeguard against potential miss clicking of the 'complete' button and otherwise randomized guessing. For every 'attempt' of the participant the system would log the result and would give additional data to analyze.

Tracking Data and Privacy

This user-test was designed in order to answer the central research question as posed in Section 3.2. To do this, the data from the user-test results had contain enough relevant information to answer this question. Wherever possible additional data was collected in order to answer the other research questions as well as providing data for further non-research question/goal related analyses which could provide

an interesting insight into the participants, the system or the user-test itself.

Before listing the various pieces of data that were collected during this user-test we want to emphasize the notion of privacy. For any user-test privacy would be important, but given that the participants in this user-test were children attending primary school, extra measures were taken to ensure anonymity. Firstly, no identifiable information was gathered. We deliberately did not ask for any name-, address-, e-mail-, phone or any other identifiable contact information of any participant. Nor did we ask for any school records and/or results for any of the participants. Additionally no single participant results were shown to any teacher, parent or anyone else not associated with this research.

The following is a list of data that was collected during the course of this user-test. Starting with the information from the offline Cloze-tests[3]:

- **ID**; Every offline Cloze-test form was marked with an ID number (1-100). This was done in order for us to keep track of the tests that were completed and assist with the gathering of data for the final analyses.

- **Code**; A 3 character code was written on the top right of each paper for the participants to use when accessing the online part of the user-test. This code was required to ensure that each participant got the correct combinations of offline and online tests (more on this in Section 3.3.2). It is therefore also used to track and link the results for a single participant for both the offline and online tests. The reason for using a code instead of a number (like the ID) to link the stories is that it would be to easy for the participant to make a mistake and enter the wrong number which would interfere with the distribution of texts between the offline and online versions.

- **Offline Results**; For the offline Cloze-test every participant was asked to circle or otherwise mark one of the multiple-choice options for each removed word from the text. Afterwards the results were digitized and converted to a percentage (%) score for each Cloze-test.

As can be seen relatively little data was gathered from the offline Cloze-tests. A lot more could be, and was, gathered from the online Cloze-tests and its subsequent questionnaire. These include the following[4]:

- **Age**; One of three personal data items asked from each participant. Used for analytical purposes.

[3]The offline Cloze-test forms as shown in Appendix B of this report do not show the data as listed for the offline-tests since these were participant specific.

[4]Questionnaire data items are marked with a (*). The questionnaire itself can be seen in Appendix D.

- **Groep**[5]; Second piece of personal data. The 'groep' data is used to perform the final analysis between the results of different classes.
- **Gender**; The gender identification is used for analytical purposes e.g. analysing whether there is a significant difference in the results of males vs females.
- **Online Results**; This is the same data as previously described for the offline results but in digital form and stored in an online database.
- **Mistakes**; All mistakes made by the participants while performing the online Cloze-test were recorded and stored. Mistakes are made when a participant selects the wrong answer for an omitted word and has clicked the 'Done' button.
- **Attempts**; An attempt is counted every time a participant clicks the 'Done' button after having filled out all the answers on the Online Cloze-test or trying to correct previous mistakes. This data is used to track the participants' progress in a single online Cloze-test and is therefore linked with the 'Online Results'.
- **Time**; While not possible for the offline version of the Cloze-test, we were able to track the time a participant needed to complete each online Cloze-test. The time was recorded in 'attempts' (see previous item) in seconds. The initial measurement (for the first attempt) of each online Cloze-test started when the page loaded and ended the first time the 'Done' button was clicked. For any subsequent attempts the timer recorded the time between the previous- and the next click of the 'Done' button.
- **Reading***; One factor of interest was in seeing whether the amount of casual reading (outside of school) done by a participant is an indication of their performance in the Cloze-test, both off- and online. Potential answers were: *never/daily/weekly/monthly/yearly*.
- **Dyslexia***; Dyslexia can (potentially) be a major influence in the results and overall performance of a participant in any environment where reading is required. Therefore we felt it would be pertinent to include this in the questionnaire and thereby the analysis of the results of this user-test. Potential answers were: *yes/no/I don't know*.
- **AVI-level***; The AVI-level (see Section 2.1) is the Dutch method for standardized assessment of reading ability in children. Potential answers were: *AVI-E5/M6/E6/M7/E7/Plus/I don't know*.

[5]See Appendix A for a comparison between international school year equivalents.

- **Enjoyment***; One of three questions using a Likert-scale response type using smilies emoticons with an accompanying label. It asked what the participant thought of the Online-test. Potential answers were: *very unenjoyable/unenjoyable/average/enjoyable/very enjoyable*.
- **Perceived Difficulty***; This question asked the participants what they thought of the difficulty level. Potential answers were: *very hard/hard/normal/easy/very easy*.
- **Comparison Online/Offline***; This question asked the participants whether they thought the online Cloze-test was easier or harder compared to the offline Cloze-test. Potential answers were: *much harder/harder/similar/easier/much easier*.

3.3.2 Study Design

This section details the general outline and overall design of the user-test and various elements that are a part of that process. The aim of this user-test was to have around 40-60 participating children from groep 7&8 of the Dutch primary school system in this study. Averaging around 20-30 students per class this meant we required around 2 full classes.

This user-test required participants to complete multiple Cloze-tests, both offline and online. Given the nature of this study and the time allowed to perform it, this required finding a good balance between the number of Cloze-tests each student would be able to complete and the total time required. In order to get a good time estimation we contacted an expert in child user-testing and the teachers of the classes where the user-test was schedules to take place. Based on our needs and the experts' input the final number of Cloze-test each child would need to complete was set at 4. Given that 2 tests each would not provide us with enough data since we did not have enough participants to offset this. Also having 3 tests each would cause inconsistency when switching between the off- and online versions of the test.

As mentioned earlier in this chapter the decision was made to use 10 different texts for use in the Cloze-test. In earlier iterations of the user-test the number of texts was set at 4, but having this few texts would not allow for a detailed ranking of the results of each text in order to determine and analyse the correlation between the off- and online Cloze-tests.

Table 3.1 below lists the 10 stories used for this user-test. The individual stories can also be seen (in full) in Appendix B. What Table 3.1 also shows is a *level* column for each text used. This shows the initial difficulty classification level of the text as determined by the WizeScan system (see Section 3.3.1). A text level of 3 denotes a

text as being suitable to children in groep 7/8 (our targeted user-group in this user-test). A text level of 2 is one difficulty level lower and signifies a (slightly) easier text. A text level of 4 signifies a (slightly) more difficult text. By looking at this level feature and including it in the selection process of choosing the stories we made sure that texts were not too difficult, or too easy for the targeted user-group of this study. It also provided us with a basic ranking system between stories (by difficulty) which we would later use to compare to the ranking of the stories by overall results.

Table 3.1: Cloze-test stories

STORY	Level	Title
A	4	Romeinse Geneesmiddelen
B	2	Chinese Muur
C	4	Tjernobyl
D	3	Vikingen
E	3	Duinen
F	2	Schrijven
G	3	Treinen
H	2	Brandweer
I	4	Facebook Onderzoek
J	3	Olympische spelen

Since our goal was to use the participants overall results for both the offline and online Cloze-test to determine a ranking of texts, the 10 texts need to be logically distributed between both the offline and online versions of test. To ensure a (relatively) even distribution of texts between participants and the different versions, the following user-testing schema was set-up, see Table 3.2.

Table 3.2: Sample distribution of texts and versions

	Offline		Online		code
1	A	B	C	D	P6L
2	B	C	D	E	J6S
3	C	D	E	F	V7T
4	D	E	F	G	W2F
5	E	F	G	H	W6Q
6	F	G	H	I	C6Y
7	G	H	I	J	L8G
8	H	I	J	A	X2Z
9	I	J	A	B	K4U
10	J	A	B	C	K1I

Table 3.2 shows a sample distribution of 10 texts (A-J) for 10 users (after which the cycle repeats except for the 'code' which is unique to each user) to ensure a relatively even distribution. Additionally the column 'code' contains the tracking code of that particular user which is used to ensure that every participant receives the correct stories, see also Section 3.3.1.

In the example of Table 3.2 User 1 would complete Cloze-tests **A** and **B** on paper, while completing **C** and **D** online. User 2 would complete Cloze-tests **B** and **C** on paper, while completing **D** and **E** online. This distribution method ensures that every story gets used for both offline and online versions.

Moving on from the distribution of Cloze-tests, Figure 3.3 below shows a flow diagram of the user-test and within its various steps.

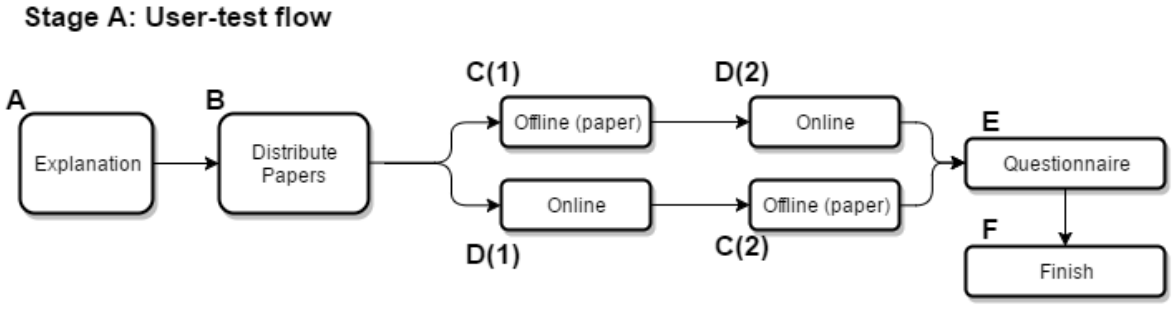


Figure 3.3: Stage A user-test flow diagram

The flow diagram as shown in Figure 3.3 is relatively self-explanatory but there are a few parts which warrant further explanation. After step B in the diagram the path splits into 2 different options, offline (C(1)) and online (D(1)). This split is done to prevent the order effect of always beginning with either version (offline or online). To ensure the validity of the outcome of this user-test a split was made that alternated the starting version of each participant.

Another part which requires additional highlighting is the online portion of the user-test. A flow diagram of which is shown in Figure 3.4[6].

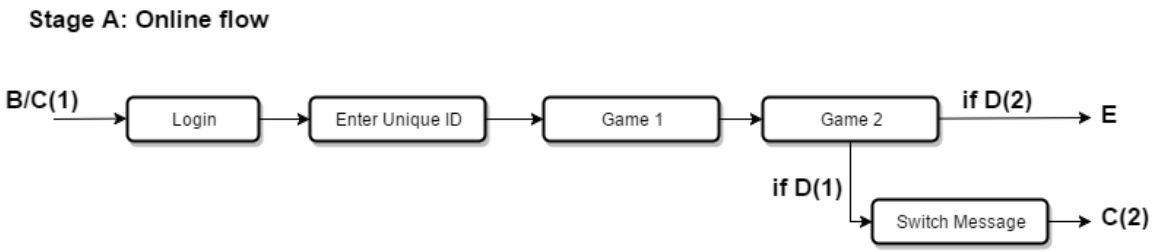


Figure 3.4: Stage A online user-test flow diagram

[6]Examples of the actual pages the participants worked through for the online version are included in Appendix C.

Depending on the previous step in the flow diagram, the participant either completes the user-test with the online version or starts with the online versions. The implication of which is that the participant who got the online version first receives a message upon completion of the online portion to switch to the offline version before continuing with the online part of filling out the questionnaire at the end.

Also shown in Figure 3.4 in the second step, is the inclusion of the participant being required to enter a unique ID. This unique ID (Code) is present on the offline (paper) versions the participants receive and is required/included in order to keep track of the users' results in both the offline and online versions.

Finally, at the end of the user-test the participants are asked to fill in a small digital questionnaire, the results of which are stored in a database alongside the participants results. This questionnaire can be seen in full in Appendix D.

3.4 Results & Analysis

This section details the results of the user-test that was performed at a primary school in Amsterdam. While the primary focus of this user-test was to analyse the correlation between the results of both the offline and online Cloze-tests, data was also collected on various other aspects such as the difference in results between grades, gender differences and others.

3.4.1 General Statistics

These classes that participated in this user-test were groep 7 and groep 8[7]. In total 44 children participated in this study, 22 from groep 7 and 22 from groep 8, see Table 3.3. Combined, these classes consisted of 19 (43.2%) male and 25 (56.8%) female participants. The participants ages ranged from 10 to 12 years old, with a mean age of 11.05 years old (SD=0.61). One participant forgot to fill in and answer one of the two offline tests. Every other participant completed all the required Cloze-tests (4 tests in total each). Out of the total 44 participants 3 reported themselves as being dyslectic on the questionnaire and 5 participants did not know whether they were dyslectic or not.

The results of the user-tests are measured as follows. Every Cloze-test, both offline and online have between 7-9 answers each. The result that is measured and used for this analysis is the percentage of correct answers from the total number of answers. E.g. a participant has correctly marked 5 out of 9 possible answers on one of the offline Cloze-tests, then his score for that test is 5/9 = 0.56 (56%). The

[7] See Appendix A for a comparison between international school year equivalents.

Table 3.3: General participant information

	Male	**Female**	**Dyslectic**	**Avg. Age**
Groep 7	8	14	1	10,77
Groep 8	11	11	2	11,32
Total	19	25	3	11,05

same calculation is done for the online Cloze-test, however in order to make it fair and similar to the offline Cloze-test only the result of the first attempt of each test is included in the result. For statistical purposes results will be measured by taking the average out of all of the Cloze-tests that were completed in this user-test. Whenever this section mentions a 'total score', it refers to the combined average of both the online and offline Cloze-test results.

Earlier in this chapter, in Section 3.2, a number of (additional) research questions were asked. To begin this analysis we start by looking at the first research question which asked whether there was a significant difference between the results of the different grades (groepen) that were tested. The results of these two groups can be seen in Figure 3.5. As can be seen in the graph the average results for groep 8 are higher for both the offline- ($0.83 > 0.72$) and the online ($0.85 > 0.76$) Cloze-test compared to those from groep 7. Mann-Whitney tests confirmed that the difference between the results for group 7 & 8 were significant for both the offline- (U=158.5, Z=-1.970, p=0.049), online- (U=140.0, Z=-2.404, p=0.016) and overall results (U=109.5, Z=-3.112, p=0.002).

For the online portion of the user-test we were able to record the time (in seconds) that was spend on each attempt of every online Cloze-test. For this analysis only the time taken by each participant for the first attempt was used. Figure 3.6 shows the average time (in seconds) required by participants to complete the first attempt on a single online Cloze-test. The difference is quite large, with participants from groep 7 requiring around 5:36 min (336s) and participants from groep 8 (only) around 2:28 min (208s). This is a difference of 2:09 min (128.5s) and is significant according to a Mann-Whitney test (U=118.0, Z=-2.9111, p=0.004).

On the basis of these results we can safely say that there is a significant difference in the performance between the 2 tested groups.

The second research question questioned whether or not a significant difference can be measured between the results of males and females who participated in this user-test. A total of 44 children participated in this user-test, consisting of 19 males (boys) and 25 females (girls).

With the same method used to determine and view the group results, we can also see the results between genders. Figure 3.7 shows the average results for both the offline-, online- and overall Cloze-tests. However, as can be seen in Figure 3.7

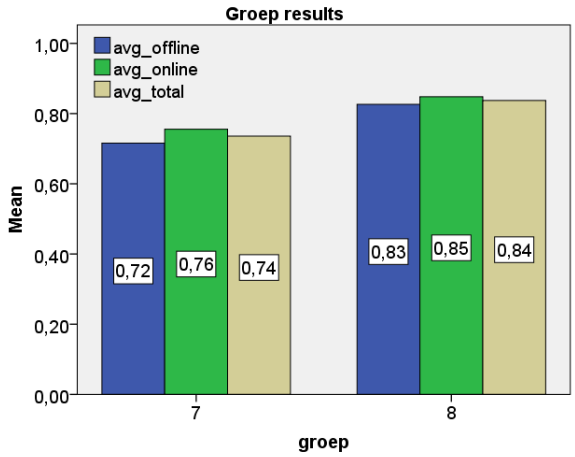


Figure 3.5: Groep results

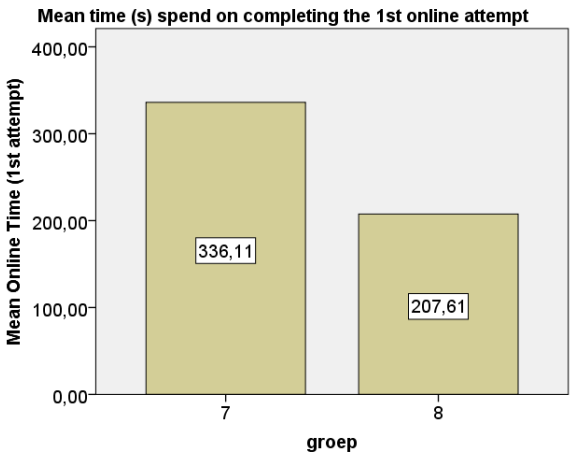


Figure 3.6: Groep time comparison

no gender is clearly above the other in terms of results. In our user-test males score on average slightly higher ($0.79 > 0.76$) in the offline Cloze-test while the females score slightly higher in the online Cloze-test ($0.83 > 0.77$). The overall difference is less than 1% in favour of the females ($0.79 > 0.78$). Mann-Whitney tests confirmed that neither the offline- (U=211.0, Z=-0.631, p=0.528), online- (U=169.5, Z=-1.618, p=0.106) or overall- (U=211.5, Z=-0.616, p=0.538) results are significantly different between genders.

For the analysis of differences between genders we also looked at the average online time required to complete a single attempt. Figure 3.8 shows that the male participants averaged around 4:40 min (280s) to complete a single attempt while the females averaged around 4:25 min (265.5s). This is a difference of about 15 seconds, and is not significant according to a Mann-Whitney test (U=208.0, Z=-0.699, p=0.485).

Both these results show that that there is no significant difference between males and females in this user-test.

The third research question questioned whether the order-effect had any (significant) influence on the outcome. Examples of order-effects include improvement or decline in performance throughout the user-test, which may be due to learning effect, boredom or fatigue[8]. By counterbalancing using a crossover design you can counteract these effects. For this particular user-test the order-effect comes into play when looking at the version of Cloze-test the participant starts with. By alternating the version (offline/online) each participant started with we attempted to avoid the consequences of this effect.

16 participants were asked to start with the offline Cloze-test first and 28 with the online Cloze-test. The results of those groups can be seen in Figure 3.9.

[8]For more information, see: https://en.wikipedia.org/wiki/Repeated_measures_design.

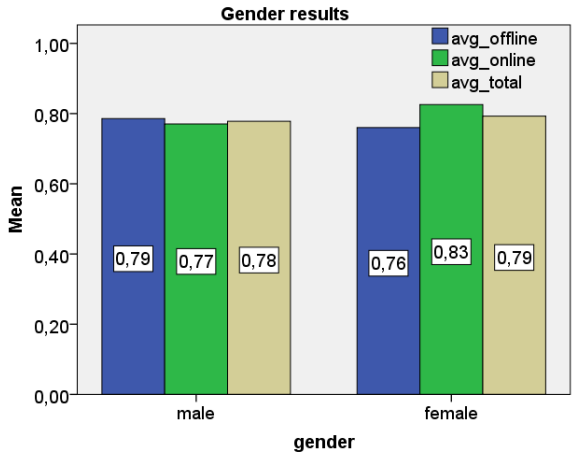


Figure 3.7: Gender results

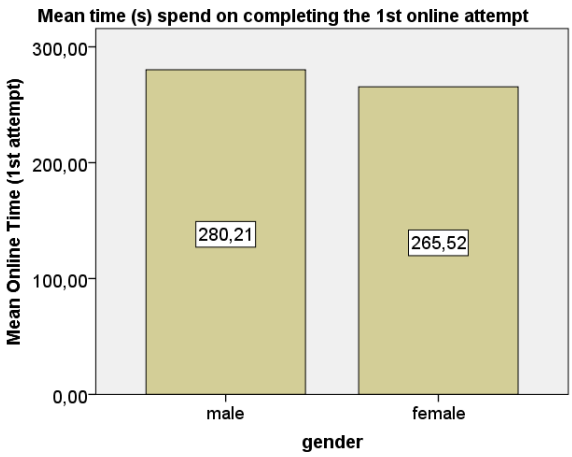


Figure 3.8: Gender time comparison

Further analysis confirmed that there is no significant difference between groups as determined by a one-way ANOVA analysis for either the offline- ($F(1,42)=.754$, $p=0.390$), the online- ($F(1,42)=0.132$, $p=0.718$) or the overall results ($F(1,42)=0.708$, $p=0.405$).

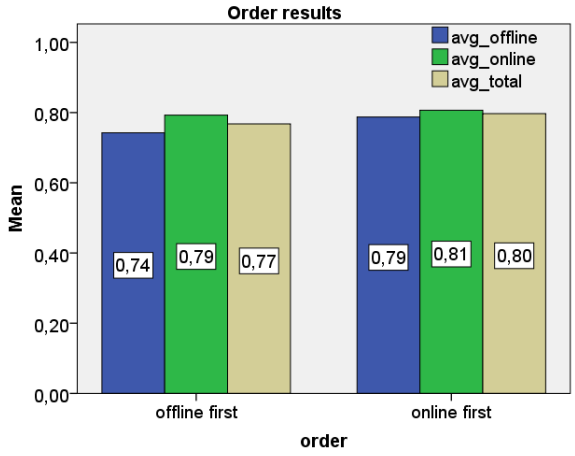


Figure 3.9: Results by order

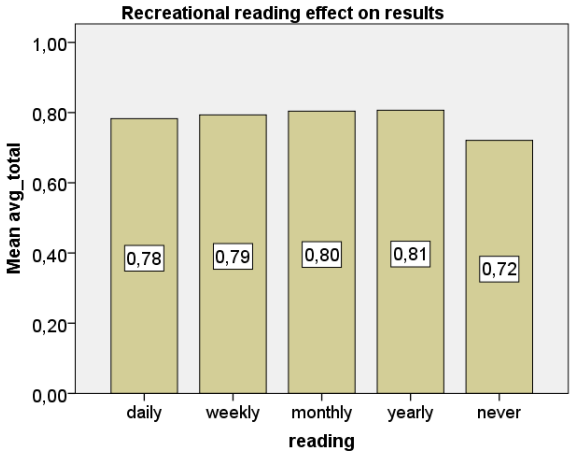


Figure 3.10: Reading effect on results

The final research question concerned recreational reading. This question was included as a personal interest in order to see whether the amount of recreational reading done by the participants affected the overall outcome of the Cloze-test in any way.

Figure 3.10 shows the average overall results distributed across the participants of the user-test. The question on the questionnaire was: *How much do you (the participant) read outside of school?*.

Out of the 44 participants the distribution of answers was as follows: 17 participants said they read daily, 16 weekly, 5 monthly, 3 yearly and 3 never. Looking

at the results in Figure 3.10 there is no real discernible difference (mean=0.7864, sd=0.11), between the overall results of the participants who gave different answers to the questionnaire question.

A Kruskal-Wallis H test confirmed there was no statistically significant difference between the results of participants given their 'reading status', ($\chi^2(4)$=1.965, p=0.742). With a mean rank overall score of 21.71 for daily reading participants, 24.03 for weekly readers, 24.7 for monthly readers, 24.17 for yearly readers and 13.50 for participants who don't read recreationally.

3.4.2 Online and Offline Cloze-test Correlation

The end-goal of this user-test was to find out if a (significant) correlation existed between the offline- and online Cloze-tests. This goal was central to the main research question which was the following:

Does the online adaptation of the Cloze-test measure similar results as the offline (paper) version?

In order to answer this question we needed children (the participants of this user-test) to do complete multiple Cloze-test. As explained in the previous section, we were able to have 44 children complete 4 Cloze-tests each, 2 offline and 2 online. This resulted in the overall 'usage' table, Table 3.4.

Table 3.4: Cloze-text completion table

	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**Total**
Offline	8	10	10	10	9	8	8	8	8	8	**87**
Online	8	8	9	10	10	10	9	8	8	8	**88**

Every Cloze-test text was completed between 8 and 10 times. With one offline result missing due to a participant forgetting to fill in one of the offline Cloze-test forms.

All of these 175 Cloze-tests were scored based as a percentage correct out of all possible answers. Further details on the results based on this data can be read in Section 3.4.1. This section focusses on another feature.

This feature is the ranking of the 10 Cloze-test stories in terms of difficulty, based on the participant scores, for both the online and offline versions of each story. By comparing these lists we are able to see whether or not there is a (significant) correlation between the offline- and the online Cloze-tests.

Having calculated all of the participant scores we can now determine the average

score for each individual Cloze-test story, both offline and online. The result of which is shown in Table 3.5.

Table 3.5: Combined overall results

STORY	LVL	TITLE	OFFLINE RESULTS	ONLINE RESULTS
A	4	Romeinse Geneesmiddelen	0,486	0,607
B	2	Chinese Muur	0,938	0,844
C	4	Tjernobyl	0,790	0,877
D	3	Vikingen	0,728	0,800
E	3	Duinen	0,711	0,700
F	2	Schrijven	0,813	0,825
G	3	Treinen	0,863	0,833
H	2	Brandweer	0,800	0,875
I	4	Facebook Onderzoek	0,831	0,828
J	3	Olympische Spelen	0,738	0,813
		Averages	**0,770**	**0,800**

Looking at Table 3.5, we see that the overall combined average results are higher for the online version compared to those of the offline version ($0.800 > 0.770$). In order to get a better view of the individual story rankings we have included Figures 3.11 & 3.12, which list all Cloze-test stories and display them in order of results, from worst (difficult) to best (easiest).

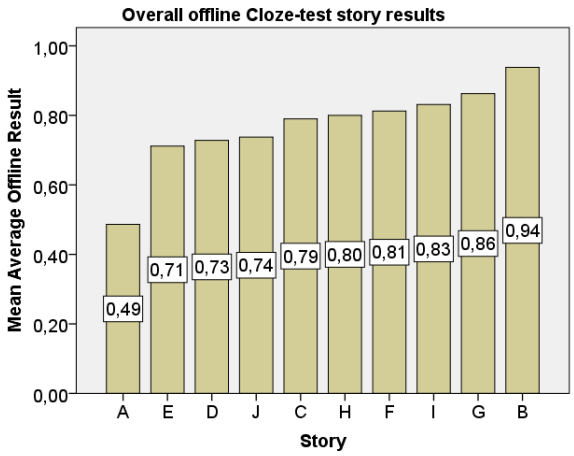


Figure 3.11: Offline Cloze-test story results

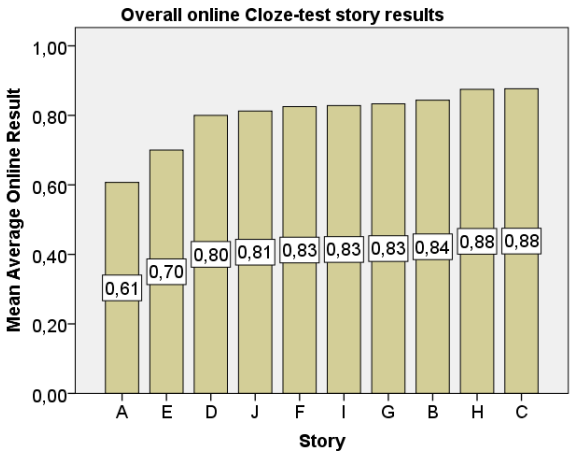


Figure 3.12: Online Cloze-test story results

Using the aforementioned data, we are able to correlate the results and test for significance. The first method of testing the correlation is by performing a correlation test using the Pearson correlation coefficient. By correlating the offline- and online Cloze-test results for each story a very strong and significant correlation was found between both versions **(r(8)=0.849, p=0.002)**.

Table 3.6: Overall story ranking

Story	Offline Rank	Online Rank
A	1	1
B	10	8
C	5	10
D	3	3
E	2	2
F	7	5
G	9	7
H	6	9
I	8	6
J	4	4

Another correlation test is the *'rank-order correlation test'* which looks at the differences in ranks for determining the correlation and its significance. Table 3.6 shows the ranking of the different stories based on their overall scores. A Spearman's rank-order correlation was run on the data from Table 3.6 and results show that there is a statistically significant correlation between the offline- and online Cloze-test results **(rs(8)=0.697, p=0.025)**. Both these analyses answer the central research question, showing that the online adaptation of the Cloze-test measures similar results as the offline (paper) version.

Individual Group Results and Correlations

Before ending this section on the correlation between the offline- and online Cloze-test results, we wanted to include and compare the results and findings from within each class (groep) that participated in this user-test.

When looking at differences and/or correlation within the 2 classes (groep 7/8) an important thing to keep in mind is that the number of completed stories, and therefore data points are halved. This causes (much) greater fluctuation in results i.e. when 1 participants scored exceptionally low/high and is influencing the overall score for that story/version quite heavily. Table 3.7 lists the results for all groups in a single table.

Results from Table 3.7 reiterate the fact that on average groep 7's results are quite a bit lower compared to those of groep 8. This is of course in line with expectations as there is a years worth of education/learning difference between them. From all (average) results groep 7 only score better on the online part of story A compared to groep 8, and even that is by only less than one percent (.003%). The online Cloze-tests (on average) score higher than the offline Cloze-tests. Around

Table 3.7: Overall group results

		Groep 7		Groep 8	
Story	**Title**	**Offline**	**Online**	**Offline**	**Online**
A	Romeinse Geneesmiddelen	.420	.610	.526	.607
B	Chinese Muur	.850	.750	1.000	.938
C	Tjernobyl	.750	.780	.817	.956
D	Vikingen	.470	.780	.898	.813
E	Duinen	.630	.630	.818	.750
F	Schrijven	.810	.750	.813	.875
G	Treinen	.850	.830	.877	.844
H	Brandweer	.790	.850	.815	.938
I	Facebook Onderzoek	.790	.750	.878	.958
J	Olympische Spelen	.690	.750	.785	.875
	Averages	**.705**	**.748**	**.822**	**.855**

4% higher for both groep 7&8 when rounding the overall results.

Looking at the correlation between the offline- and online results within each group, we start to observe something interesting. Using the Pearson correlation coefficient results from groep 7 show a **non-significant** correlation **(rs(8)=0.602, p=0.065)** while the results from groep 8 do show a significant correlation between the offline- and online results for each story **(rs(8)=0.733, p=0.016)**. Additionally we can also run a rank order correlation test and see whether or not those tests result in a significant correlation. Table 3.8 shows the rankings within each group.

Table 3.8: Group story ranking

	Groep 7		Groep 8	
Story	**Offline Rank**	**Online Rank**	**Offline Rank**	**Online Rank**
A	1	1	1	1
B	9	6	10	8
C	5	7	5	9
D	2	8	9	3
E	3	2	6	2
F	8	5	3	6
G	10	9	7	4
H	6	10	4	7
I	7	4	8	10
J	4	3	2	5

As can be seen in Table 3.8, the ranking of each story varies wildly between the

offline- and online versions of the Cloze-test. And between both classes (with the exception of story A, which consistently scores the lowest and is therefore ranked the same).

Spearman rank-order correlation analysis showed **no significant correlation** for either the rank orders of groep 7 **(rs(8)=0.479, p=0.162)** or groep 8 **(rs(8)=0.321, p=0.365)**.

While this is speculation on our part, we believe the lack of correlation is due to a number of factors.

1. The sample size is likely too small to make an accurate analysis on correlation, given that a single participants result has a large influence on the overall score. By combining both the groups' results they average out and become more representative of reality.

2. The rank correlation is based on the ranks of the offline- and online (average) Cloze-test results in ascending order. Looking at the data in Table 3.7 we see that the differences in average scores are so low, often less than 1%, that the ranking of a story could increase or decrease by 4 if the result was 1-4% higher or lower. This causes instability and is a result of the smaller sample size when looking at individual class (groep) results.

3.4.3 Questionnaire Analysis

After completing both the offline- and the online Cloze-tests the participants were asked (on-screen) to fill in a small questionnaire which included a number of questions regarding the test itself, their thoughts and various metrics. The questionnaire itself can be seen in Appendix D. In total there were 7 questions on the questionnaire plus room to leave a comment. We will examine each result in the order in which they were asked.

The first item on the questionnaire was the question of what the participant thought of the test. The answer was a Likert-scale response. Out of the 44 people 3 participants (6.8%) found the Cloze-test *not-enjoyable*, 29 participants (65.9%) said it was *alright*, 9 participants (20.5%) said it was *fun* and 3 participants (6.8%) said it was *super fun*.

The second question on the questionnaire asked whether or not the participant perceived the Cloze-tests they completed as easy or hard. Of the 44 participants 1 (2.3%) said it was *hard*, 20 (45.5%) that it was *normal*, 21 (47.7%) that it was *easy* and 2 participants (4.5%) described the Cloze-tests as *very easy*. There was not enough data to properly test whether any correlation existed between the answer to this question and the Cloze-test results.

The third question asked the participants whether they thought the online Cloze-test was easier or harder than the offline Cloze-test. Of the 44 participants, 7 (15.9%) said it as *harder*, 25 (56.8%) said it was *similar* in difficulty, 10 (22.7%) said it was *easier* and 2 (4.5%) said it was much easier. To see if this actually had any relation to the (overall) results we analysed and graphed the data in Figure 3.13. Figure 3.13 shows the average offline- and online scores of the participants who gave a particular answer. The data does not seem to corroborate the answers given by the participants. When participants answered that the online Cloze-test was *harder* the average online scores were 7% higher than the offline scores, and in the case where the participants said the online Cloze-test was *much easier* the online Cloze-test scored 8% lower compared to the offline results.

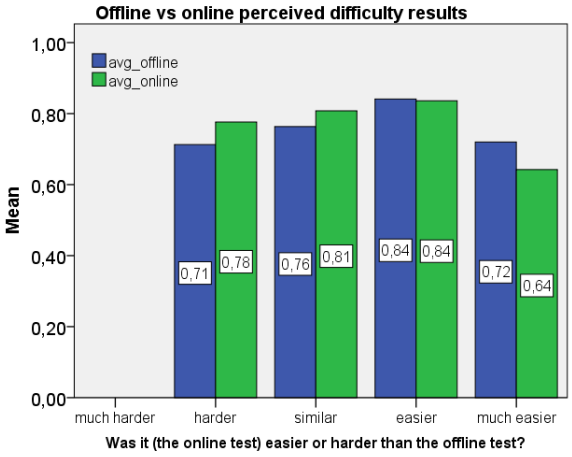


Figure 3.13: Perceived difficulty difference between versions

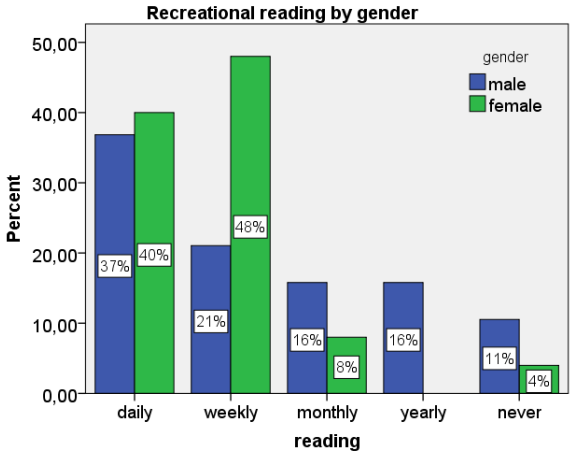


Figure 3.14: Recreational reading by gender breakdown

Earlier in this chapter, we analysed the effect of recreational reading on results. One additional item we wanted to explore in connection with (recreational) reading is the difference in reading habits between males and females. The resulting graph can be seen in Figure 3.14. It contains a breakdown (in %) on the amount of recreational reading for each gender, this was done this way since the females outnumber the males for the particular user-test.

While of course not necessarily representative of today's youth, the results show clear differences between genders. On the female side of the participants almost 90% answered that they read recreationally either daily or weekly. This is much lower in the men, where around 58% say they read daily/weekly. Almost 27% of the male participants answered that they read yearly or never, which is a significant amount.

The final 2 (required) questionnaire questions concerned dyslexia and the AVI reading level. From the 44 participants only 3 reported themselves as being dyslec-

tic and 5 did not know if they were or not. Given the small sample size no real significance should be taken from these results, but the data that we did analyse does not suggest that the participants with dyslexia scored (significantly) lower compared to the children without. The overall scores are actually only 1% apart in favour of the non-dyslectic participants.

Initially it looked obvious to include the AVI reading levels in the questionnaire of this user-test, given that the AVI score directly pertains to the reading level of the child. However, because we performed the user-test on groups 7&8 of the Dutch primary school system, almost everyone, with the exception of a single participant answered either plus (the maximum level) or unknown on the question of their AVI level[9]. The result of which was that we were unable to perform any analysis concerning AVI levels.

Questionnaire Comment Analysis

Lastly, each participant had the option of leaving a comment about the user-test or anything connected with it. The full list of comments (in Dutch) can be seen in Appendix D, but for this section we will list a few general themes within the comment section and translate them accordingly.

One type of comment that recurred multiple times was that the texts were kind of boring. At this particular stage of the research it was not required for the test to be exciting and fun to play/interact with. However, this is an absolutely essential point to address for any future gamified implementation. The issue with these comments about the boring nature of the user-test is that it is very hard to discern whether it is 'boring' because of the texts themselves, or because of the lack of interaction or features. And how does one affect the other? Does having gamified elements decrease the overall 'boringness' level or does a 'boring' text bring down the fun-factor?

Another participant commented that the test was very similar to 'regular' school test. This was something we imagined would come up before even starting with the user-test. Since we intentionally did not want to deviate too much from the original Cloze-formula, which is (regularly) used in teaching environments and school tests We also did not want to introduce too much interactivity in the online section.

3.4.4 General Observations

During the course of the user-test we made several observations of the user-test process and the behavior of the participants in general. These observations might

[9]See Appendix D for the full questionnaire results.

provide some form of explanation to any 'strange' or 'odd' results.

Technically the website worked perfectly without any bugs or problems. Before the second user-test with the children from groep 8 (6^{th} grade), we spoke to the teacher and explained the experience and issues we encountered with the previous user-test group. She (the teacher) was a big help in making sure the second round of user-testing with the children from groep 8 went really well, without major disturbances. Additionally the transition between the offline and online versions went very well, with only a few participants requiring help.

Having little experience in performing user-tests on primary school children the very first class we tested, groep 7 (5^{th} grade) was the most difficult. There were certain unforeseen elements that we were not prepared for including shut-down computers, missing pc equipment (mice, keyboards) and others. We were able to resolve most issues but this process did take about 10-15 minutes. Most of these issues would likely have been prevented by talking to the teacher beforehand. As was done for the next class of participants.

We also miscalculated the total amount of time required for the test, not accounting for the amount of time spend on interruptions and turmoil in the room. Having no expressed authority over the participating children it was difficult to calm them down and point them to the task at hand. We also had to intervene several times when we saw or heard several children communicating or 'cheating' with/off each other. This was of course not intended for this particular user-test. Finally some of the texts, especially for group 7 raised a lot of questions concerning words which were not understood or about the general difficulty of the text. While we foresaw this being a potential issue and having printed out several pieces of paper which listed difficult words from the texts, this did not cover all of it given the amount of questions we were asked.

3.5 Discussion

This section contains a brief discussion on the results, the user-test, and the implications of the results of this user-test. These discussion topics are subdivided by a number of future recommendations based on the user-test and its results.

1. More participants

The outcome of the test positively answered the central research question, showing a significant correlation between the results of the offline- and the online Cloze-tests. However a few reservations must be attached to this outcome. One of the more obvious factors is the relatively small scope and number of participants in this user-test. User-testing with large groups of children is a complex task, both in an organizing

perspective and in the actual execution of the user-test itself. This is something that could have been done better.

2. Re-test on a more diverse group of participants
Another potential factor into the results of this study is that the children from this particular primary school were very similar socio-economically. With a vast majority of them being Caucasian, with presumably (due to the location of the school (central Amsterdam) and comments from colleagues) parents who are highly educated and financially well off. This might therefore not be reflective of the 'average' Dutch classroom.

3. Involve experts in the process
Any (similar) future user-test would certainly have us communicate more with the teacher beforehand about the nature of the user-test and our goals and requirements. Having done this for the second round of user-testing relieved so many potential problems which could have been avoided in the initial user-test group. We would also start a lot earlier in contacting teachers/schools and asking them to perform user-tests on one or more of their classes. This would also potentially alleviate scheduling issues and delays e.g. not being able to do it in a long time period as a result of school vacations and/or exams.

4. Test other variants
The result of this user-test supports the theory that the online Cloze-test correlates with the offline Cloze-test in that they both measure similar outcomes. This would also imply that since the (original) paper Cloze-test is a verified method of readability assessment, this is now also the case for the online Cloze-test. However, some questions remain unanswered. Is this only true for this specific variant of the Cloze-test or do we need to test for correlation for every single variant? And can we make changes to the interaction scheme or visual presentation without compromising the legitimacy of the readability assessment?

Chapter 4

Stage B: Cloze Automatization

In the previous chapter we showed there is a significant correlation between the original offline Cloze-test and the online version. This finding now provides us with a scientific basis to use the Cloze-test in an online environment to determine the readability of a text given the users' performance of the Cloze-test. However, when we start to think of potentially doing these tasks in a large scale (online) environment, certain issues become apparent. Converting regular texts into Cloze-tests is a time consuming process. Going large scale would require an undetermined amount of texts to be converted for use in a Cloze-test. This cannot be done manually, which requires the process to be automated.

However, this automatization process does create additional problems which will have to be solved in order to have a fully automated Cloze-text extractor. Since this concerns a Cloze-test which does not use a fixed-rate word interval scheme, e.g. remove every other fifth word from the text, but a test designed according to the principle of rational deletion, see Sections 2.1 and 3.3.1. Using a rational approach allows the Cloze-test to measure understanding of the Cloze-test instead of grammatical knowledge. Research in this chapter on the automatization process of Cloze-tests uses the same guidelines[1] from Kraf, Lentz & Pander Maat [13] as was done in Chapter 3.

The work in this chapter measures how difficult it is to design and develop a working automated system which can adhere to the aforementioned guidelines as well as a human can. This is due to the nature of the task, which comes down to understanding text at a general- and sentence level and omitting words which are 'important' and suitable for extraction. Due to these challenges, this stage of the research aims to construct a system which can apply the guidelines as best as possible and approximate the performance of a human constructing a Cloze-test from the same source material.

This chapter will detail several methods and principles which all contribute in

[1] For the complete list of guidelines, see page 22.

tackling this issue. Various automatization methods and their performances are examined, analysed and compared. Concluding this chapter is a small Turing-test style user-test measuring differences between manually- and automatically created Cloze-tests.

4.1 The Basics

To analyse the performances of the automated systems we chose to use a classification based method. This allowed us to compare the results of the various methods against each other and against manually created Cloze-tests using the same data. This method works as follows.

Manually, and according to the previously mentioned guidelines (see page 22), we took 10 different texts, and marked every word which was suitable to be omitted from the text with a **(**)** on both sides of the word in order for the computer to recognize the suitable words. A snippet of an example is added below.

> *John Law was een gewiekste man. Hij was econoom, Schot van geboorte en hij was ter dood veroordeeld wegens **moord**. In 1694 wist hij met behulp van **geld** en goede **contacten** uit de Newgate **gevangenis** in Londen te **ontsnappen** en naar Amsterdam te **vluchten**. Hij was in Engeland ter dood veroordeeld, omdat hij in een **duel** om de **gunst** van een **vrouw** zijn **opponent** had **gedood**...*

There are several reasons for using this particular approach. By manually marking every suitable word, the system can be made to operate independent of size requirements of potential resulting Cloze-tests. A Cloze-test may use all, or some of the suitable words, depending on the desired requirements set by the creator. It also provides us with a consistent set of (training) data which is marked following one particular and clear set of guidelines. These are also the reasons why we decided not use pre-existing Cloze-tests as data, because many of the aforementioned factors are unknown.

A Python[2] program was written to process the 10 annotated texts. This program parses each text, and binarily classifies each word based on the presence (or absence) of the previously mentioned markings. This information is then stored, which allows us to compare the classification results of other systems against the manual classification results (created according to the Cloze-test guidelines).

[2]See https://www.python.org/.

Table 4.1: Confusion table

		Predicted Value	
		Not Selected	**Selected**
Actual Value	**Not Selected**	True Negative (TN)	False Positive (FP)
	Selected	False Negative (FN)	True Positive (TP)

Since we are using a binary classification system and are looking to compare one set of results (the predicted value) of various systems against the base (manual) results (the actual value) we can utilize a confusion table as seen in Table 4.1. What this confusion table shows us is that there are four possible outcomes as the results of classification. Which, in the case of this particular research, is when a word is classified by a system as being either **selected** (True) or **not selected** (False) as being suitable for use in a Cloze-test. The four outcomes are as follows:

- **True Negative (TN);** The system classified a word as being not-suitable for use in a Cloze-test. That word was classified the same in the (original) manual classification.
- **True Positive (TP);** The system classified a word as being suitable for use in a Cloze-test. That word was classified the same in the manual classification.
- **False Negative (FN);** The system classified a word as being not-suitable for use in a Cloze-test, while the manual classification indicated that it was.
- **False Positive (FP);** The system classified a word as being suitable for use in a Cloze-test, while the manual classification indicated that it was not.

Using this classification scheme we are able to look at the **precision** and **recall** values. Where the precision value gives us the fraction of retrieved instances that are relevant[3]. And the recall value gives us the fraction of relevant instances which are retrieved. Both the precision and recall values can be calculated using the aforementioned (4) classes as shown in Table 4.1. These calculations are listed below, and include another value, the F-measure (or F1-score), which shows the accuracy of a test by looking at the combination of both the precision- and recall scores.

Precision (P) = $\frac{TP}{TP+FP}$

Recall (R) = $\frac{TP}{TP+FN}$

F-Measure (FM) = $\frac{2}{(\frac{1}{P})+(\frac{1}{R})}$

[3]A precision score of 1.00 (100%) indicates that all of the retrieved instances are relevant (selected). A recall score of 1.00 indicates that all relevant (selected) instances are retrieved.

For this particular research the precision score is more important than the recall score. We want to select as many 'correct' words from the text (via the automated systems) as possible and reduce the amount of 'wrong' words which can be included in the results, since this can have a major effect on the difficulty, feasibility and correctness of the resulting Cloze-test. Still, recall is not unimportant since we do need to extract 'enough' words from the text to create a Cloze-test. As we have chosen to emphasize the precision over the recall scores, we calculate an additional value, the $F_\beta - score$.

A derivative of the F-measure which was previously introduced, the $F_\beta - score$ allows us to assign a value to the β thereby assigning extra weight to either the precision or recall values. Since we chose to emphasize the importance of precision in this particular research over recall we chose a β-value of 0.5. This means that the precision is weighed twice as much as the recall value. The calculation of the $F_\beta - score$ is as follows[4].

$$F_\beta - score = \frac{(1+\beta^2)*recall*precision}{recall+\beta^2*precision}$$

In the following sections various automated systems are described using the calculations and methods as listed above. By comparing their classifications with those of the base method (which was done manually), every word gets a certain classification e.g. TN/FN/TP/FP. From the total list of word classifications of each text the precision, recall and BiasedFM (F_β) scores can be calculated. Which in turn will be used to compare the effectiveness of various systems against each other.

4.2 Interval Classification Method

Interval classification is the first method we applied and is based on a very basic principle. Every xth word from a text is classified as **True** (suitable for use in a Cloze-test) and all of the others as **False**, with x being the number of the interval. This method was never meant as an actual solution or potential method for automating the Cloze-test. It is added to this stage of the research as a method of which the results can be compared and contrasted with by more complicated and specialized methods. Since both the precision- and recall scores are really dependent on the interval that is chosen for this method, we have chosen to test multiple intervals in order to get a picture of the performance of this interval test. Because this is an interval based method, it is impossible to follow the Cloze-test guidelines mentioned earlier in this chapter. It does take into account that no words from the first sentence in a text can be selected. It detects when the first sentence ends after which the

[4]For the remainder of this document the $F_\beta - score$ will be written as BiasedFM (biased F-measure)

interval starts.

Below is an example of the different scores from all (10) texts using an interval of 10.

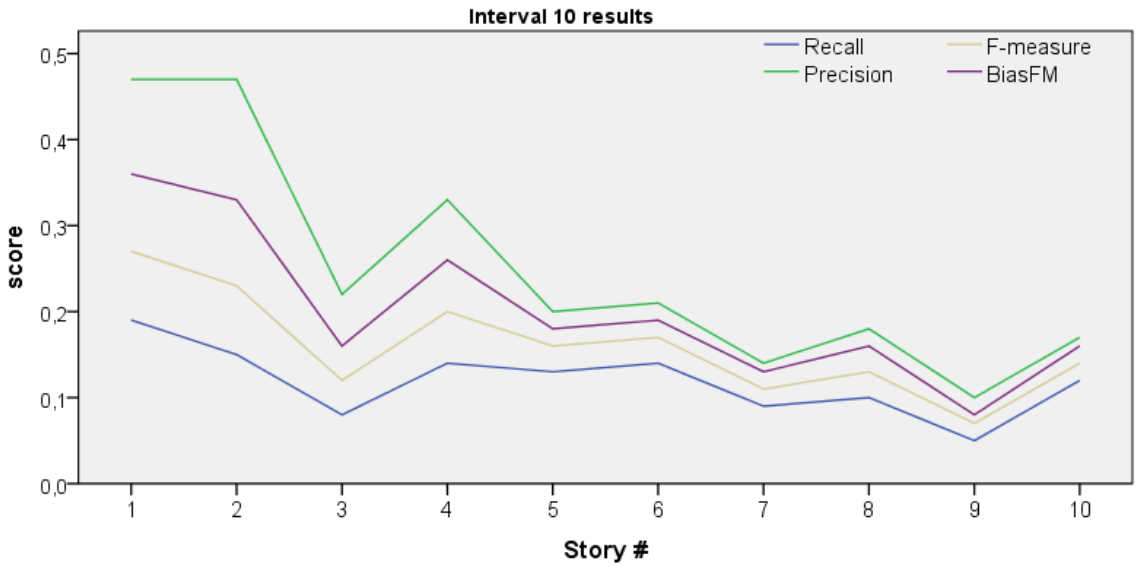


Figure 4.1: Story results using an interval of 10

Table 4.2: Interval parse (10) results

	Story-1	Story-2	Story-3	Story-4	Story-5	Story-6	Story-7	Story-8	Story-9	Story-10	Overall
Recall	0.19	0.15	0.08	0.14	0.13	0.14	0.09	0.10	0.05	0.12	**0.12**
Precision	0.47	0.47	0.22	0.33	0.20	0.21	0.14	0.18	0.10	0.17	**0.25**
F-measure	0.27	0.23	0.12	0.20	0.16	0.17	0.11	0.13	0.07	0.14	**0.16**
BiasedFM	0.36	0.33	0.16	0.26	0.18	0.19	0.13	0.16	0.08	0.16	**0.20**

Visible in Figure 4.1 and Table 4.2 are the differences in scores for each story. All scores (recall, precision, FM, biasedFM) vary between $\pm$0.15-0.40 with an exception for the precision value in the stories 1&2. If we look at the 'important' values from the overall data, we can see that the precision is around 25% and the $F_\beta - score$ (biasedFM) hangs around 20% for an interval of 10. These scores, of course, are highly dependant on the interval that is chosen. Increasing the interval will reduce the overall recall but might increase the precision scores given the random nature of this interval method.

In order to visualize the changes in scores given multiple intervals, the overall (average) results from all 10 texts were taken in an interval range of 1-20. See Figure 4.2.

Clearly visible in Figure 4.2 is the exponential decline in recall scores when increasing the interval. The highest values of both the recall (1.00) (it selects every word therefore the 100% score) and biasedFM (0.24) can be found at an interval of

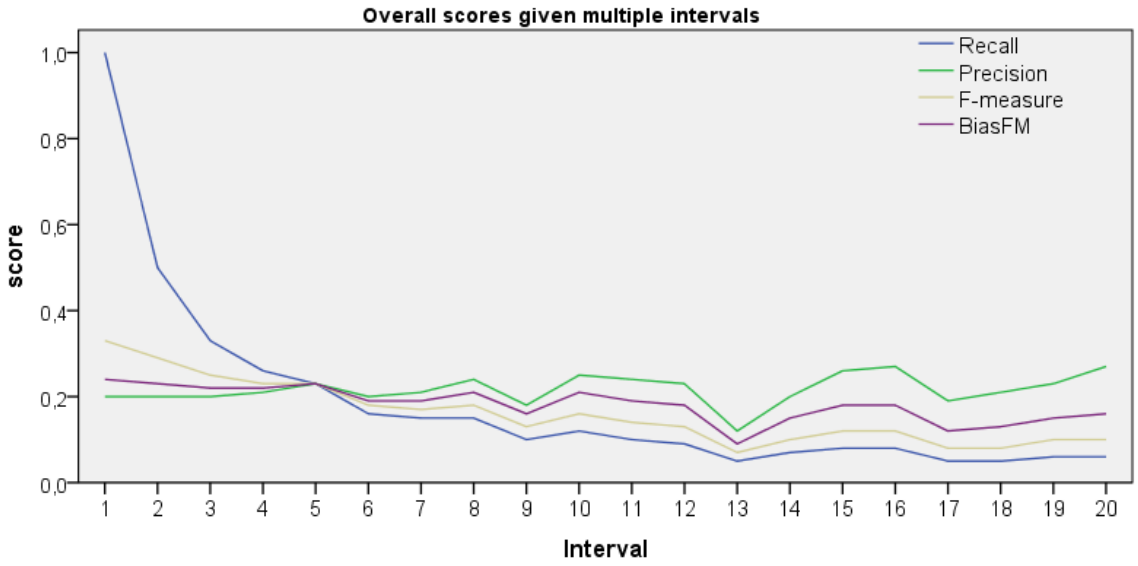


Figure 4.2: Overall scores given multiple intervals

1. The highest precision (0.27) score is found at intervals 16&20. Given the random nature of the interval method it is mostly based on chance that these intervals produce the highest precision values. The large differences in results also shows the unreliability of this method when applied to a rational deletion of words like those in this particular Cloze-test format.

4.3 Custom Classification Method

The 'Custom' classification method of Cloze-texts is not based on a set principle like the ones in the *interval-* or the *NLP* methods (more on this in the next section). It is designed to find 'important' words in the text by looking at various textual features and adhering to the guidelines as best as possible via a coded solution.

Below are the (5) features included in this method which contribute to the eventual score.

1. Start procedure

The 'default' start procedure of the custom classification method is the same as those used in all other methods. According to the guidelines of the Cloze-procedure used for this research, no words in the first sentence of a text can be omitted in order for the reader to get familiarized with the context of the story. The system therefore looks when the first sentence ends and starts applying the custom classification method after that.

2. Word type

Filtering on word type is a relatively straightforward procedure. The system looks if the word contains more than just letters [A-Za-z] and hyphens [-]. If it does, the word is automatically marked as False (not suited for extraction). This is a quick and easy way to filter 'words' from the text which contain numbers or any other special characters, which are usually not suited for extraction (e.g. dates, numbers, scores, years, etc. . .). While this method has the potential of misclassifying words as False Negatives, this effect should rarely occur, resulting in a minimal effect on the overall outcome.

3. Ban-list

As this custom classification method does not contain a NLP/POS-tagger implementation, another method of filtering commonly used words had to be implemented. In an NLP implementation for example you would look at the type of word e.g. *verbs* or *nouns* and make a selection based on that feature. In order to emulate this feature we have chosen to include the feature of a 'ban-list'. A ban-list (in this context) is a list with the x most commonly used words. For this research we have used a ban-list which contains the 100 most common words used in the Dutch Wikipedia version, complemented with a number of 'missing' common words.

4. Word length

Another feature which we started to look at concerning the classification of words is the word-length feature. This is a feature which looks at the length of each individual word and classifies them as *False* if they do not meet a set minimum requirement. This method originated when examining and analysing the words from the base method which were classified as *True*. We noticed that a very large proportion of those words share a similar minimum length. This led to its inclusion as one of the features in this (custom) classification method.

5. Handling capitalized words

One of the issues we came across when testing various methods and analysing the results, was that we did not have a solution for filtering words which were commonly capitalized (i.e. names, locations, countries, etc. . .) which, according to the guidelines, should not be included for extraction. The problem of this particular issue was on how to differentiate between names or countries which start with a capital letter and the word at the start of each sentence. For this particular feature, we made it so the system would mark every capitalized word as False when it was not the first word of a sentence. This did mean that words which were names/locations/countries/etc. . . which also happen to be the start of a sentence,

would not be filtered by this feature. However, the negative effect was measured and found to be minimal.

Potential hazards

The main danger with using these five features as described above is that of over-fitting the features based on the data (the manually marked texts). While this is always a (potential) issue when creating features which are based on data from texts, given the size of the dataset (10 marked texts) and a number of features which are specifically linked to in-text data (e.g. word length), the potential of over-fitting must be kept in mind.

Having described the (5) features and its hazards, we performed an analysis which showed how much these individual features affect the overall performance. The results of which can be seen in Figure 4.3 with the scores measured as an average over all of the data (the 10 stories). The 'default' state in Figure 4.3 represents the base results of the classifier *only* with only using the first feature (start classifying after first sentence ends). For every other feature shown in Figure 4.3 the results indicate what happens to the scores when that feature is added to the default state.

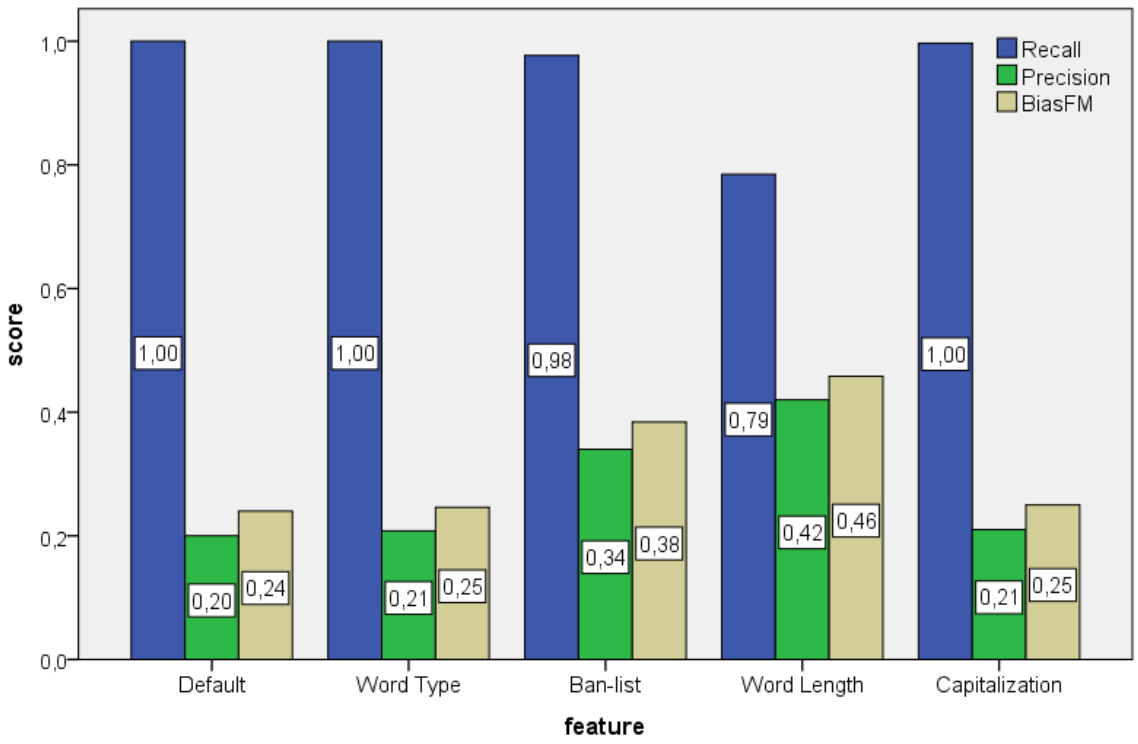


Figure 4.3: Custom classifier feature influence graph

Visible in Figure 4.3 is that some features have a much greater impact than

others. The Word Type feature for example, only shows a 1% increase in precision compared to the default state, indicating that it has very little effect on the overall performance. The effect of the ban-list feature is interesting, as it shows a 13% increase in precision score, while retaining a (very) high recall score of 98%. The biggest change occurs when introducing the Word-Length feature, as it increases the precision score by 22% and decreases the recall score with 21% causing the biasedFM score to get up to 46%. The result of this feature is of course highly variable based on the minimum word length requirement set. The result of the Word-Length feature in Figure 4.3 uses a minimum length requirement of 5. The changes in results of this feature when adjusting the length requirement are examined in detail further on in this section (see Figure 4.4). Finally, as seen in Figure 4.3 the effect of the capitalization feature on its own is almost none (+1% precision).

It is important to reiterate that Figure 4.3 shows the result of each feature individually. The effect of each feature can change when used in combination with the other features. Combining all features result in a biasedFM score of around 45-51% for this classifier depending on certain feature settings, like that of the Word-Length feature.

Figure 4.4 was created to show the change in overall results based on the increasing value of the minimum word length requirement. All other features beside the capitalization feature are included in these results. The capitalization feature itself will be examined separately.

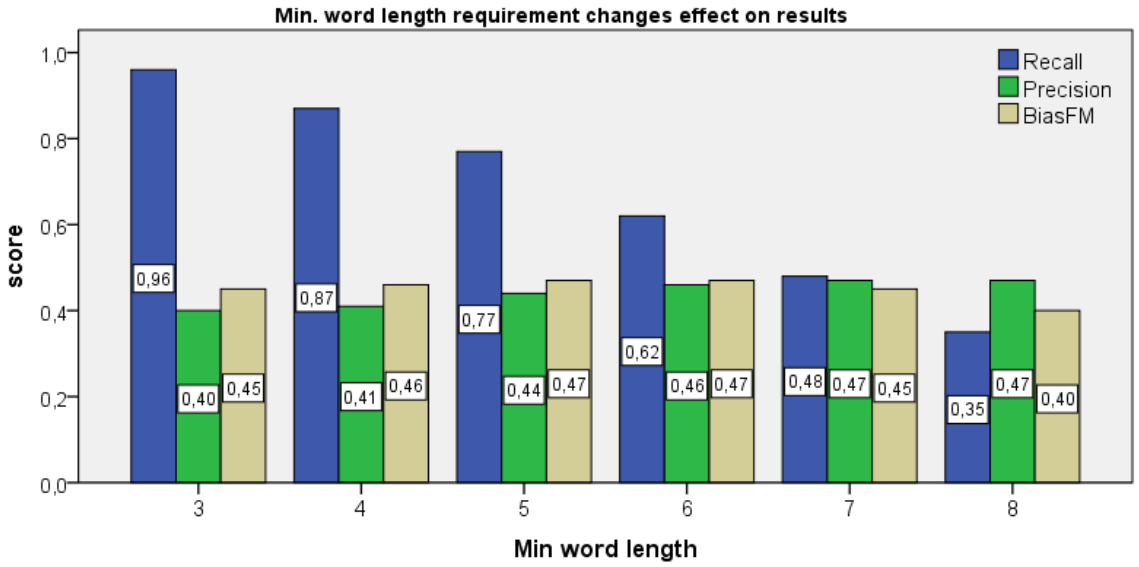


Figure 4.4: Effect of changes in the min. word length req. on overall results

Figure 4.4 clearly shows that by increasing the minimal word length requirement, the recall drops, which is to be expected given that we are looking at an decreas-

ing set of words which match the requirements. However, both the precision and biasedFM increase slightly at higher minimum word length requirements. Peaking around a minimum word length requirement value of around 5-6.

As shown in Figure 4.3, the 5th feature (word capitalization) when viewed on its own has a very tiny effect ($\pm$1%) on overall scores. However, to see whether this is still the case when this feature is introduced after all other features are added, we performed an analysis. Table 4.3 shows the effect of the inclusion of the word capitalization feature combined with different 'settings' of the minimum word length requirement feature as shown in Figure 4.4.

Table 4.3: Word capitalization feature effect on overall results

Min word length	Recall	Precision	BiasedFM	BFM increase
3	0.95	0.43	0.48	0.03
4	0.87	0.44	0.49	0.03
5	0.76	0.47	0.50	0.03
6	**0.61**	**0.50**	**0.51**	**0.04**
7	0.47	0.51	0.48	0.03
8	0.35	0.52	0.43	0.03

What we can see from Table 4.3 is that the inclusion of the capitalization feature has a positive effect on both the precision and biasedFM scores, larger than its individual effect on the scores as shown in Figure 4.3. On average the overall biasedFM score are around 3% higher with the addition of this feature, which also causes the biasedFM value to go above 50% for the first time, at a minimum word length of 5 or 6.

While the presence of potentially over-fitted data is certainly something to keep in mind, and the effect of that is measured of each feature can vary based on the dataset used. This particular method has proven to be relatively effective in classifying the data, showing a 20-25% increase in biasedFM scores compared to that of the top interval results and reaching around a 45-51% overall biasedFM score dependant on certain feature settings.

4.4 NLP Classification Method

Natural Language Processing (NLP) is a field of computer science and computational linguistics which concerns itself with the interactions between computers and human languages[5]. For the purposes of this research the area of NLP which is (potentially) helpful is the area of *Part-of-speech* (POS) tagging which can discern the

[5]See: https://en.wikipedia.org/wiki/Natural_language_processing.

different parts of speech for each word in a sentence. The inclusion of this particular method is primarily done to tackle a specific item(s) from the Cloze-test guidelines which would be otherwise be an extremely difficult problem to solve using 'conventional' code/programming. These are the guidelines we are referring to:

- *Try to omit nouns and verbs wherever possible, since these words usually contain the most information.*
  - *Try to use as little adverbs and adjectives as possible.*

Given the nature of the above mentioned guidelines we need to use the Part-of-speech (POS)-tagger system to differentiate between the different word types. For this research we used the NLTK (Natural Language Toolkit)[6] in conjunction with a POS-tagger trained (using Naive Bayes) on the Dutch Alpino corpus.[7]

In order to measure the performances for this task and the classification of words to either True or False, we specifically look whether a word is a noun or a verb within the context of the text. If a word is either of those types, the word is classified as True (suitable for use in a Cloze-test), otherwise it is set to False. The assignment of word-types to each individual word comes from the POS-tagger. Same as in the other classification methods we skip the first sentence. Figure 4.5 and Table 4.4 show the scores measured for each story.

Table 4.4: Individual and overall story results using the NLP approach

	Story-1	Story-2	Story-3	Story-4	Story-5	Story-6	Story-7	Story-8	Story-9	Story-10	Overall
Recall	0.93	0.74	0.82	0.86	0.74	0.75	0.94	0.77	0.50	0.97	**0.80**
Precision	0.61	0.67	0.57	0.51	0.35	0.30	0.36	0.41	0.32	0.35	**0.45**
F-measure	0.74	0.70	0.67	0.64	0.48	0.43	0.52	0.54	0.39	0.51	**0.56**
BiasedFM	0.66	0.68	0.61	0.56	0.39	0.34	0.41	0.45	0.34	0.40	**0.48**

What we can see from the data in Figure 4.5 and Table 4.4 is that the recall score generally averages between 75-90%, with an exception for the 9th story in the where there is a downward peak. Averaging with a recall score of 80%. The precision data shows a fluctuation between $\pm$30-65%, with an overall average of 45%. The biasedFM score is also similar, averaging at around 48%.

In order to better understand why these scores are what they are, we took the time to further analyse the classification of each individual word types. By looking at each word in a text (from all 10 marked texts) and collecting both the word type and the resulting classification We are able to visualize the overall classification by word

[6] http://www.nltk.org.

[7] The Dutch Alpino corpus is one of the only publicly available Dutch corpora on the web. While its performance is not superb, it is sufficient enough to give us an insight into the performance of using this particular method. Source: http://www.let.rug.nl/vannoord/alp/Alpino/.

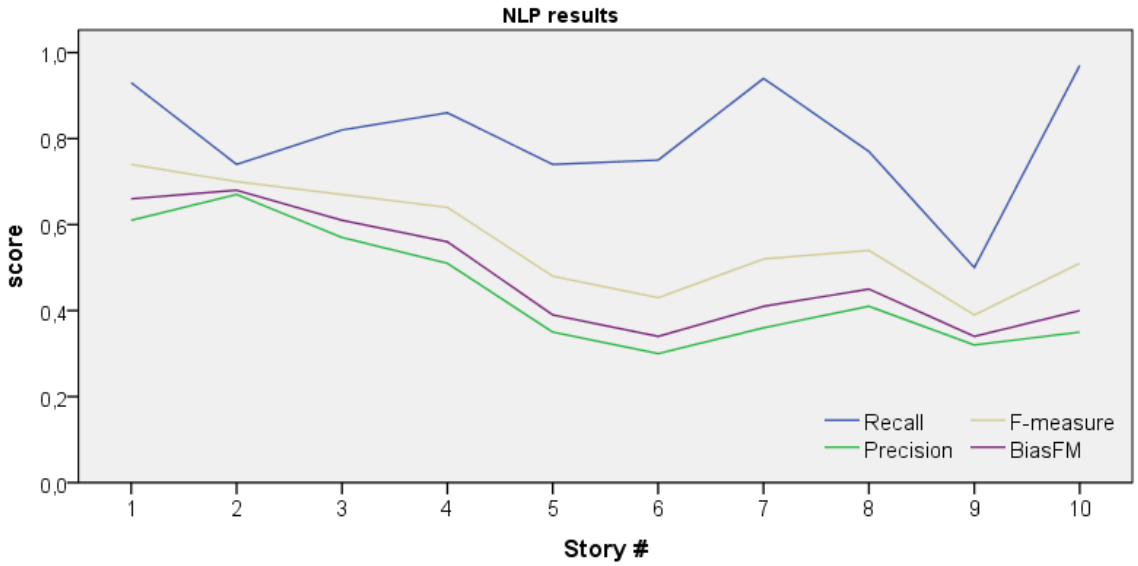


Figure 4.5: Story results using NLP classifier

type i.e. for this specific task/research we could hypothesize that the percentage of words classified as True (suitable for use in a Cloze-test) would heavily favour nouns and verbs. The opposite being true for words classified as False.

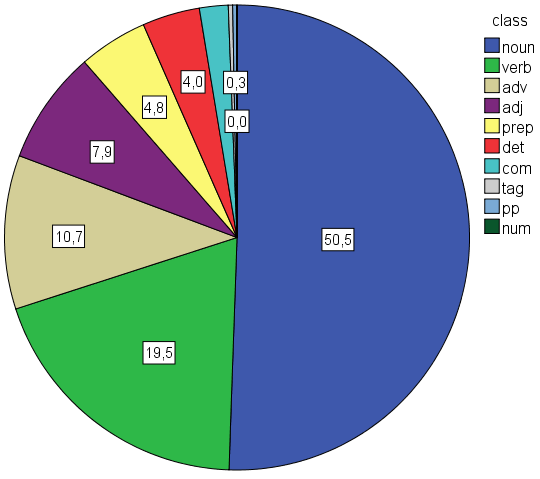


Figure 4.6: Words marked as True

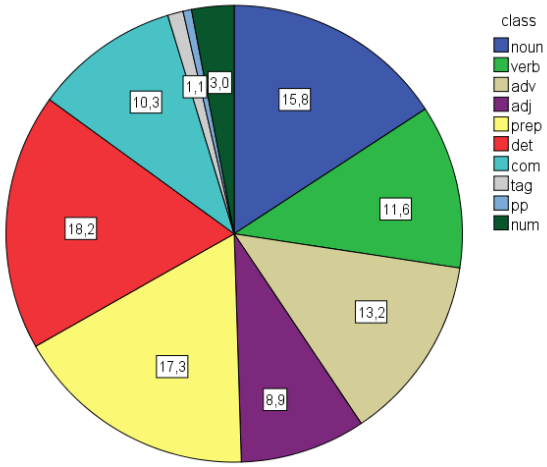


Figure 4.7: Words marked as False

Table 4.5: Overall word-type analysis results

	noun	verb	adv	adj	prep	det	com	tag	pp	num
True	50.6	19.5	10.7	7.9	4.8	4.0	2.0	0.3	0.3	0.0
False	15.8	11.6	13.2	8.9	17.3	18.2	10.3	1.1	0.6	3.0

What we can see in Figures 4.6 & 4.7 and the accompanying data in Table 4.5,

is that around 70% of words which are classified as True are verbs and nouns. For the False classification of verbs and nouns that total combined number is still around 25%. The data and figures show that around 30% of words are missed when purely looking at nouns or verbs. This is something that would warrant further inspection in the future, given the guidelines which state that you should omit nouns and verbs wherever possible. Although 70% is quite a high figure, it might not be as high as expected given the aforementioned guideline rule.

The second part of the guideline stated that you should *"Try to use as little adverbs and adjectives as possible"*. Looking at the data, the adverb and adjectives still account for a combined total of around 18% of the True classification. This appears to be a little more than would be expected.

Of course, as previously already mentioned, not all aberrations in the data are necessarily the 'fault' of the person who marked the texts. The guideline itself says to apply it "wherever possible", which would explain the inclusion of other types of words. Additionally, all these results are based on the Alpino POS-tagger classification, which is likely to cause (some) misclassification's.

4.5 TF.IDF Method

The final method we chose to apply to this problem is a method using Term Frequency (TF).Inverse Document Frequency (IDF) scoring. TF.IDF is a numerical statistic that (theoretically) should reflect the importance of each individual word inside a document (within a collection or corpus)[8]. Simply said, the more often a word occurs inside of a document, the higher the score of that word, offset by the frequency of the word in the entire corpus. The idea behind using this particular method is that words with a high TF.IDF score, i.e. important words inside a text, could (potentially) correlate with the words which are manually marked for extraction, which are often verbs and nouns, and usually key words inside a document.

The first step was to apply the TF.IDF scheme to all the 10 marked texts. This way every word would be scored in addition to the (base) True/False classification it already possesses. After the scoring is done we are able to compare the top TF.IDF scores with the list of words which are classified as True, and hopefully see some correlation.

The scoring for this method is done differently compared to that of the previous methods which used recall/precision/biasedFM scores. This is done because the TF.IDF scheme does not provide binary results (True/False) for each word, but rather a numerical score based on the number of occurrences of the word in the text and

[8]See: https://en.wikipedia.org/wiki/Tf-idf.

corpus. The way scores are calculated for this method is by comparing the TF.IDF scores with the % of the top x. Where x stands for the total number of words in the text classified as True (suitable for use in a Cloze-test). For example, if a text contains 28 words classified as True, then the top 28 TF.IDF scoring words from the text are taken and the resulting score is determined based on the % of correct (True) classifications within that list of the top 28 scoring words[9].

Initially we only used the 10 texts combined as a corpus, but quickly realized that this corpus was way too small and would be unsuited for use as a corpus. At first we were looking for a publicly available (large) corpus containing Dutch texts. Unfortunately we could not find any. However we did manage to find a large Dutch corpus from the University of Leipzig[10] containing 100.000 (100k) to 1.000.000 (1M) individual sentences sourced from a number of Dutch media outlets. A few examples of sentences from that corpus is listed below.

- *"Juist om die ruimte te scheppen."*
- *"Op deze site vindt u meer informatie over onze producten."*
- *"In Zeeland kan de Formatie van Naaldwijk direct bovenop de Formatie van Maassluis liggen."*
- *"Op zoek naar een unieke beleving?"*
- *"Sinds 1984 heeft hij een breed scala aan kennis en ervaring opgebouwd in de administratieve informatievoorziening."*

Even though a large corpus was now available, this corpus was essentially an incredibly large single document containing 100k-1M individual lines or exactly the other way around; a corpus consisting of 100k-1M individual documents all containing a single line. This is an issue given that the IDF score is highly reliant on the number of documents and the size of those documents.

To solve this issue, we elected to split the corpus into a number of documents all containing a similar number of individual lines. After which, we could perform the TF.IDF calculations and analyse the results. We looked at both the TF.IDF and the IDF scores separately as well to see if there are any (major) differences between them, or if one scores better than the other.

After these score calculations are complete for each corpus/document split we could make a list of the individual TF.IDF word scores in each of the 10 marked texts. We were then able to compare this list of top scoring words with the same list of words including each classification (True/False). The results are shown in Table 4.6 for the multiple different corpora size and document splits.

[9]Unless explicitly specified, i.e. the 'top' column in Table 4.6, the score is always based on the accuracy of TF.IDF scores of the total number of words classified as True within a text.

[10]Source: http://corpora.uni-leipzig.de/en.

Table 4.6: TF.IDF results

Corpus (# of lines)	Documents	Lines/Document	Top	TF.IDF	IDF
100k	100k	1	-	0.251	0.208
100k	10k	10	-	0.297	0.310
100k	1k	100	-	0.320	0.309
100k	500	200	-	0.320	0.307
100k	500	200	10	0.380	0.290
100k	500	200	5	0.500	0.220
1M	1M	1	-	0.257	0.212
1M	1k	1k	-	0.322	0.308

What we can see from Table 4.6 is that the results indicate that TF.IDF scores are generally higher than the IDF scores. The split of the corpus into documents also plays a big role in the resulting scores. With a 500/200 (documents/contained sentences) split proven to be the best tested for the 100k corpus and a 1000/1000 split for the 1M corpora.

The resulting accuracy scores seem to hover around the 32% for both the 100k and 1M corpora when taking the scores of all the words in each text into account. Additionally the 1M corpus only very slightly (.2%-.5%) outperforms the 100k corpus in terms of scores. Given these results and the long processing time required for the 1M corpus it seems that the 100k corpus is fine for further/future use. The scores significantly increase when only considering the top 5/10 TF.IDF scores. Reaching a peak score of around 50% when considering the top 5 scoring words from each text.

The approach we took for this method is different than any final solution would be using TF.IDF, mainly due to the (training) data used for all these methods. Because of its non-binary nature, to use TF.IDF in the same way as all of the previous classification methods i.e. providing a binary classification of each word in a text (True/False) and thereby precision- and recall scores, a change has to be made. To do this a threshold has to be set to a certain level of TF.IDF score, which when surpassed will classify the word as either True (suitable for use in a Cloze-test) or False. This threshold level however, has to be based on a lot more training data than available or possible to be made for this research project. Any threshold set using the current data would be highly arbitrary. We therefore decided to use accuracy scores for this method which we believe should provide a decent reflection of the potential of this method.

4.6 Method Comparison

Having examined each classification method and its performance separately, we can now examine how the different methods compare and contrast to each other in terms of performance and suitability.

Beginning with the overall performance, Table 4.7 lists the performance of each of the 4 discussed methods on the data together with an overall score. One additional method was added, but more on this later. The results shown in Table 4.7 are based on the BiasedFM scores of each method, with the exception of the TF.IDF method as explained in the previous section.

As expected given its random nature, the interval (10) method performs the worst with an overall BiasedFM score of 20%. While performing better than the interval (10) method, the TF.IDF method still performs quite bad with a score of 32% compared to the Custom- and NLP method scores which both hover around a $\pm$50%, with the Custom method just edging it with a score of 51%.

To further examine the differences between the methods, specifically the Custom- and NLP results given they are the best performing methods and their close overall performance score, we can look at Figure 4.8 which list both the recall- and precision scores alongside the BiasedFM scores. While the Custom method has a slightly (3%) lower overall recall score, the higher (2%) precision score, which is weighted more in the BiasedFM score, causes it to surpass the NLP BiasedFM score by 3%. No real big differences are found between methods when comparing the standard deviations (SD) of the resulting scores. With the interval (10) method having the lowest SD of 9% and the Custom method having the largest SD of 13%. The limited dataset (10 stories) undoubtedly contributing to a higher SD.

Table 4.7: Method performance comparison table

	Story 1	Story 2	Story 3	Story 4	Story 5	Story 6	Story 7	Story 8	Story 9	Story 10	Overall
Interval(10)	0,36	0,33	0,16	0,26	0,18	0,19	0,13	0,16	0,08	0,16	**0,20**
Custom	0,58	0,74	0,68	0,57	0,47	0,38	0,41	0,37	0,42	0,44	**0,51**
NLP	0,56	0,68	0,61	0,56	0,39	0,34	0,41	0,45	0,34	0,40	**0,48**
TF.IDF	0,48	0,41	0,38	0,38	0,2	0,35	0,12	0,29	0,25	0,35	**0,32**
Custom+NLP	0,67	0,73	0,84	0,70	0,58	0,59	0,61	0,47	0,51	0,64	**0,63**

With the Custom- and NLP methods having such different approaches, but similar resulting scores, it was an interesting prospect of what a combination of these methods would result in. By combining the classification (True/False) results of both methods, and only classifying a word as True when classified as such by each separate method, a new method (Custom+NLP) was formed. The results of which is included in Table 4.7 and Figure 4.8.

Immediately visible in Table 4.7 is that Custom+NLP method scores significantly higher (63%) overall compared to all other methods. Looking at Figure 4.8 we see

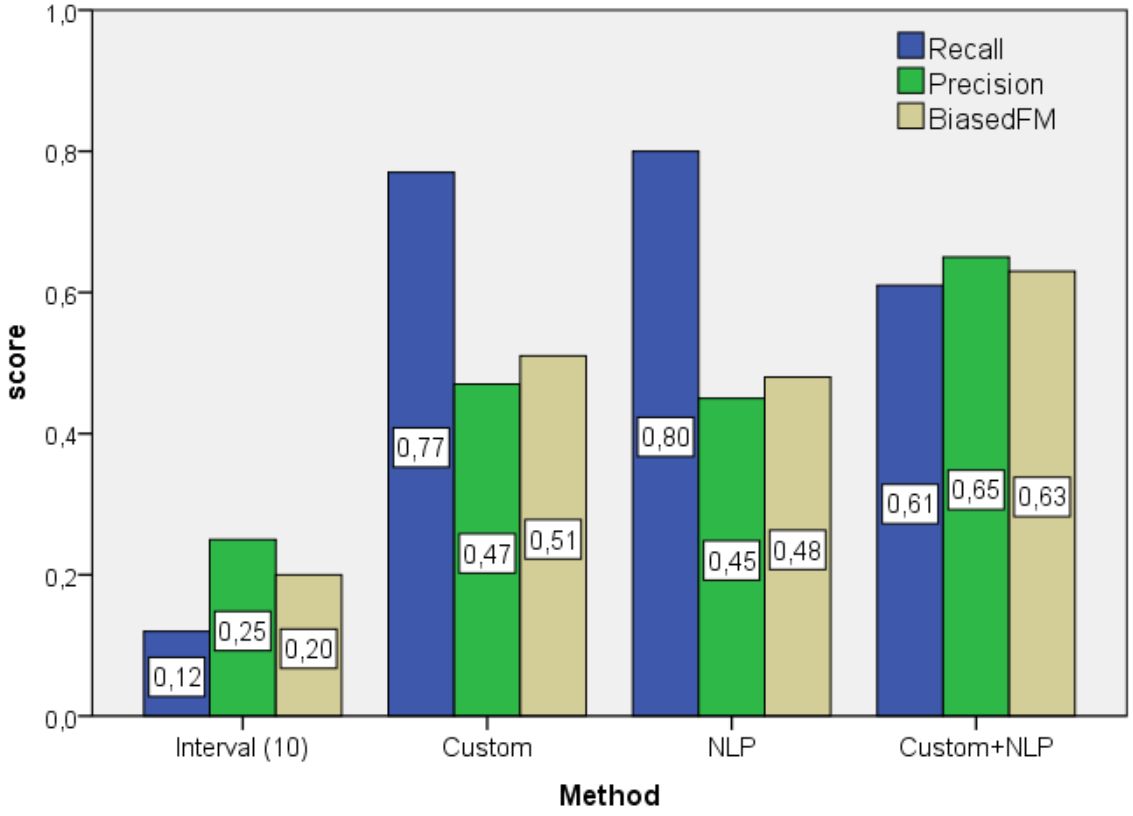


Figure 4.8: Classification methods overall score comparison

that while the recall score is significantly lower at 61%, a drop of around ±18% compared to both the Custom- and NLP recall scores separately, the precision score is also significantly higher at 65%, an increase of almost ±20% over the precision scores of its component methods. As previously mentioned in this chapter, more importance (and weight) is given to the precision score which causes the overall (BiasedFM) score (63%) to increase by 12-15% compared to the Custom- and NLP scores.

Table 4.8: Custom+NLP word length feature influence

	Min. word length				
	2	**3**	**4**	**5**	**6**
Recall	0,78	0,77	0,70	0,61	0,50
precision	0,59	0,61	0,62	0,65	0,67
BiasedFM	0,62	0,63	0,63	0,63	0,62

Tinkering with the settings of the component features does not seem to have a major effect on the overall (BiasedFM) result. As shown in Table 4.8 where the most impactful feature of the Custom Parser, the minimum word length feature, is changed

and the resulting scores are listed. While the recall scores drop significantly based on the increasing minimum word length requirement (as expected), the precision scores do not increase enough to cover this deficit. Therefore the overall (BiasedFM) scores remain steady at around 62-63%.

What these results show us is that future implementations should strongly consider using a mixture of both component classification methods (NLP and Custom) to attain high performance scores.

4.7 Application and Results

In the previous section various methods have been detailed and its (theoretical) performances have been calculated and analysed. However, to gage the performance and get an idea about the effectiveness of these automatic Cloze-test classification methods, a small user-test was set-up. With the focus of the user-test being to provide an insight into the direction of the proposed methods/systems and where they stand compared to a (regular) man-made Cloze-test.

4.7.1 User-Test Details

The user-test itself is modelled as a variation of a Turing test[11], requiring humans to distinguish between manually- and automatically created Cloze-tests.

An online form was created (using Google Forms[12]) to perform this user-test and collect the results. The user-test itself targeted Dutch speaker and was therefore written in Dutch as well. The complete form is included in Appendix F.

The user-test included 3 different Cloze-tests of around $\pm$200 words long containing 8 words omitted each. The amount of Cloze-tests was limited to 3 as to not take up too much time from each participant. 2 of these Cloze-tests were created manually, 1 via an automated system. Each Cloze-test included 2 questions at the end which asked whether they (the participant) thought the Cloze-test was created manually or automatically and how confident they were of that answer. After completing all 3 Cloze-tests the participants were asked to name which of the previous 3 completed Cloze-tests they thought was most likely to have been automated and how confident they were of that answer.

The 2 manually created Cloze-tests are adapted for this user-test from previous use in this study, see Chapter 3. The Cloze-tests concerned the topics 'Duinen'

[11] The original Turing test is a test for intelligence in a computer, requiring that a human being should be unable to distinguish the machine from another human being by using the replies to questions put to both.

[12] See: https://www.google.com/forms/about/.

(dunes) and 'Treinen' (trains).

In order to present participants with a third and automatically generated Cloze-test, the following steps were taken. A short story, about 'Olifanten' (elephants), with a similar source[13] and size (+- 200 words) was taken and converted into a plain text document, without any additional editing. We chose to use the best performing automated classification method as described in the previous section(s) for use in this test[14]. This was the 'Custom Classification Method' as described in Section 4.3. From the 46 words that were classified by the method as 'True' (suitable for use in Cloze-test), 8 were selected for use by an algorithm which made sure to provide an even spread of selected words throughout the text. The output of the classifier was converted for use in the online form, as seen in Appendix F.

Finally, the order of the Cloze-tests in the online form was randomized, with the automated Cloze-test ending on position 3.

4.7.2 Test Results and Analysis

From the people reached out to for participation in this user-test, 8 responded (in time). For this particular user-test no personal information/identification was asked, as this was not a factor in this test, and would also provide the participants with anonymity. Which might be important to some of them given that the test involves readability scores/statistics.

The results of each Cloze-test were scored via semantic scoring which allows for both the exact word, synonyms and contextually correct words, to be marked as correct. Before detailing and analysing the results, a reminder of the user-test sequence is added below as this will be integral in the coming discussion and figures[15]

1. Cloze-test #1 Duinen (Dunes) [manual]
2. Cloze-test #2 Treinen (Trains) [manual]
3. Cloze-test #3 Olifanten (Elephants) [automated]

There are multiple ways in this user-test where we could potentially see a distinction between the manually created Cloze-tests and the automated ones. One of which is via the Cloze-test scores, as the automated version could potentially select unsuitable words which are harder/impossible to answer correctly. Looking at the results in Figure 4.9, this does not seem to be the case. On average, at least 6 out of 8 (75%) correct answers were given for both Cloze-tests #1 and #3 (the automated version), and 7 out of 8 (87.5%) for Cloze-test #2.

[13] Docukit: http://www.docukit.nl.

[14] The better scoring combined Custom-NLP (POS) method as described in Section 4.6 was not yet completed at the time of the user-test.

[15] See Appendix F for more details.

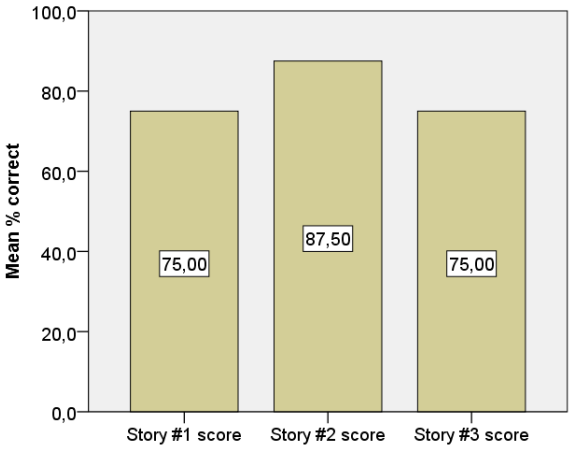


Figure 4.9: Overall Cloze-test scores

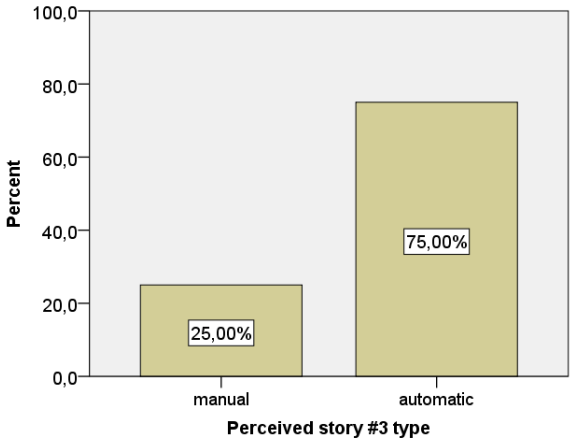


Figure 4.10: Cloze-test #3 type

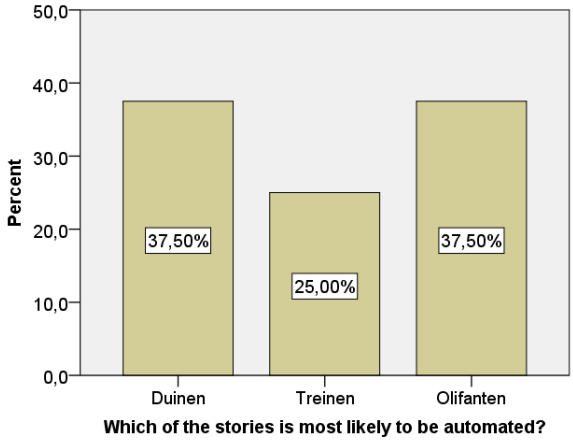


Figure 4.11: Overall automatization choice

A more direct way of assessing the perceived type (manual or automatic) of the Cloze-test was the direct question to the participants after completing the test. In both Cloze-tests #1 and Cloze-test #2, the manually created Cloze-tests, a majority of participants thought after completing the Cloze-test itself that the test was made manually. For Cloze-test #1 this was 62.5% (5 out of 8) of participants and for Cloze-test #2 this was an even larger 87.5% of participants (7 out of 8). For the final Cloze-test, Cloze-test #3 this was the other way around. 75% (6 out of 8) of participants thought after completing the Cloze-test that the test was created automatically, which it was. This seems to indicate that a majority of the participants can still distinguish the automated test from the manually-made ones.

However, this (assumption) is somewhat contradicted when looking at the results shown in Figure 4.11. After completing the 3 individual Cloze-tests, the participants were each asked to select 1 of the 3 completed Cloze-tests which they thought was

most likely to be automated, and also provide a level of confidence in their answer.

What Figure 4.11 shows us is that both Cloze-test #1 and Cloze-test #3 received 3 votes (37.5%), with Cloze-test #2 receiving 2 votes (25%). All relatively similar. When the participants were asked in the confidence of their answer to the previous question, we can see that participants were slightly more confident that Cloze-test #3 was automated (83.3%) compared to Cloze-tests #1 & #2 (70%).

4.8 Discussion

This section contains a brief discussion on the results of the user-test and the future of Cloze-test automation.

Looking at the results from the user-test in the previous section, there are a number of items to discuss. Starting with the average scores, and the differences therein, or lack of. As shown in Figure 4.9 on page 64 there does not appear to be a significant outlier in terms of average scores, be it high or low. Given that all participants were adults with some level of higher education we can assume that the text difficulty itself, or the difficulty of the Cloze-test subject, should not play a part in determining the individual scores. However, this should then also raise the (reasonable) expectation that most, if not all, participants should be able to attain a (near) perfect score, which was not the case. While still above the 'scoring threshold' of 60% which indicates an independent reading level, it appears low. One factor that certainly contributes is the (intentional) small size (8) of the Cloze-test. Any wrong answer has a big (12.5%) effect on the resulting score. Of course it could also indicate that 1 or 2 omitted words in each text were just too difficult, which, certainly for the automated Cloze-test, could be the case. Lastly, while hard to know for sure, it is always possible for a participant to not have had the required time and/or attention span to adequately read, think and then fill-in each answer on the Cloze-test(s).

Another aspect of the similar overall scores is that the automated Cloze-test (#3) did not (heavily) under- or over-perform compared to the other Cloze-tests. Given the BiasedFM- (F_β) score of $\pm$50% which was attained in testing, see Section 4.3, the automated Cloze-test would be expected to contain some misclassified omitted words. However, misclassification does not necessarily result in a difficult or impossible to guess word. It could be the exact opposite by misclassifying a 'obvious' word making it easier to guess.

A more clearer difference in results can be seen in the perceived type (manual vs automated) of each Cloze-test. While Cloze-tests #1 & #2 both have a majority ($>$62.5%) of participants believing the tests were made manually, Cloze-test #3 has a majority (75%) believing it was made automatically. This result is more in line with expectations heading into the user-test. This result might also point us toward the

notion that the performance on the Cloze-test itself is not necessarily an indicator for automation. Given that the average scores for all Cloze-tests are relatively similar as eluded to earlier.

Knowing the individual Cloze-test perception results, it is all the more strange when we look at the results shown in Figure 4.11, where Cloze-test #3 is not (heavily) favoured to be the automated Cloze-test out of the 3. While the participants who thought that Cloze-test #3 was the automated version were slightly more confident (+13%) in their answer, it still seems to directly contradict the individual responses as mentioned in the previous paragraph. A possible explanation might be that there is some primacy/recency effect[16] occurring which makes the participants remember the first and last Cloze-test better but this should then also remind them of their previous choices on the question of automation. This therefore remains a strange 'aberration' in the user-test responses.

Obviously due to the low number of participants no (scientific) significance should yet be attributed to these user-test results. With contradicting results regarding the perceived nature of the automated Cloze-test, it is hard to give a definitive statement on the success of the user-test. However, there are still a number of positives to be taken from this user-test and for the future of Cloze-test automation. The overall scores of the automated Cloze-test were in-line with the results of the other (manual) Cloze-tests, and the automated (classification) method that was used has already been eclipsed in terms of performance scores, as shown in Section 4.6, which shows promise for the future.

[16]The primacy- and recency effect is when you remember items better when they occur in the beginning (primacy) or the end (recency) of a list.

Chapter 5

Conclusions & Future Work

This final chapter of the report contains a discussion on the overall conclusions based on the work done throughout this research project. In the next section we will look at the research goals- and questions posed at the beginning of this document, and examine whether or not those goals have been met and questions were answered.

The final section (Section 5.2) details a discussion on the future work. By looking back at some of the literature and the work done in this project, we can assess where we currently stand, and look forward at ways on how this research can potentially be applied and expanded upon.

5.1 Conclusion

This research project started with a very ambitious goal, developing a large scale online platform where children play and interact with (textual) content, and through the results of their actions (feedback) enable us to improve the assessment of the contents readability. This ultimate goal is the combination of various important topics; readability assessment, gamification and child computer interaction, as researched in Chapter 2 of this report.

This original goal was proven to be unattainable, given the allotted time for this project. The decision was therefore made to focus on several key goals/milestones which provide the theoretical and technical foundation of the original goal. With the milestones being the digital (online) assessment of readability (see Chapter 3) and to research methods for the foundation of an automatic readability assessment method (see Chapter 4). The choice to focus on these particular 'stages' also meant that there was a bigger emphasis on the topic of readability assessment, and less so on the topics of gamification and child computer interaction.

The focus of the first core element/goal in this research was on determining

whether a digital adaptation of a readability assessment method was as effective and scientifically accurate as a similar (regular) offline readability assessment method. Via the use of a user-test this was found to be true, see Chapter 3. A significant correlation was found between the results of the offline- and online Cloze-test which confirmed the chapters hypothesis. Besides that, it also spoke to one of the research goals, as set in Section 1.3, which was the following:

> *Any online or digitized readability result must reflect/correspond to that of (traditional) offline Cloze-test results. I.e. the measurements have to be remain valid.*

This research goal was ultimately attained via the work done in Chapter 3, with one particular caveat. That it only applies to the (digital) adaptation using the Cloze-test type readability assessment method. This research was focussed and build around this method and therefore no conclusions can yet be drawn concerning the digitization of any other readability assessment method.

> *Q1. Is the developed (prototype) system capable of correctly measuring the readability of a text given the users proficiency level?*

Analysis of the user-test results in Chapter 3 also provided us with an answer to the first research question as stated above. Yes, a prototype system using the Cloze-test readability assessment method is capable of correctly measuring the readability of a text.

Given the results of stage A, we were able to continue forward in this research with another step. This step, Stage B, focussed on the 'issue' of automation concerning the Cloze-test readability assessment method. As any large(r) scale future solution using this-, or a similar readability assessment method, would need to be automated in order for it to be viable.

The research in Chapter 4 examined multiple automated methods of emulating the creation of Cloze-tests via human made guidelines. Four different approaches were developed and tested on the same dataset. The scores of which were based on the (in)correct classification of suitable Cloze-test words within a text document. Which resulted in precision- and recall scores that ultimately combined into a single score, the F_β-score, measuring the methods (classification) performance.

Of the four distinct methods that were created to solve the problem of Cloze-test automation, 2 methods reached F_β-scores of around 50%. A custom method that programmatically tries to emulate the (human) guidelines and decision making (51% F_β-score). And a method based on NLP technology, using *Part-of-speech* (POS) to comply with one of the guidelines' rules concerning grammar (48% F_β-score). However, these individual performances were eclipsed when subsequent analysis

showed that a single method comprised of the combination of aforementioned methods, resulted in a 63% F_β-score. With a 100% score, meaning a perfect replication of a manually designed Cloze-test according to the guidelines not being a feasible result, a 63% score is a good result and allows for the prospect of future research- and development using this method as a basis to develop a scalable system. This is even more attainable with potential enhancements/updates to this method resulting in increased performance scores. Concerning the topic of scalability, listed below is one of the research goals which spoke to that topic.

> *The final readability assessment tool must be developed in such a way in that it emphasizes scalability and that it provides a basis for any future gamified adaptations of the tool.*

The automated prototype systems were developed with scalability in mind. Given the performance results of these systems they already provide a solid foundation/blueprint for gamified adaptations and future development. Enhancements to these prototype systems will further improve and solidify this foundation. The work from Chapter 4 also speaks to the second research question.

> *Q2. Assuming we are able to assess the readability correctly, how can the developed system and its results be made scalable?*

We are now able to assess the readability correctly, and a system can be made scalable by automating the creation process of Cloze-tests. A small user-test done in Section 4.7 already showed that the performance of participants on both manually- and automatically created Cloze-tests was very similar. This user-test also resulted in conflicting reports on whether the Cloze-test was automated or not. Given these results, and that since the completion of the user-test a significantly improved automated method has been developed, it shows promise for the future adaptation of automated Cloze-test solutions.

Looking back at the main research goal set in Section 1.3 for this project, we can now examine whether or not this goal was attained. The main research goal was as follows:

> *Delivering the technological and theoretical foundation of a scientifically substantiated readability assessment tool which provides the basis for a future gamified approach of large scale readability assessment.*

This main goal is the culmination of the various sub goals and research questions as previously mentioned in this section. We have established a theoretical and scientifically substantiated foundation to a digital readability assessment method (the

Cloze-test). The automated readability assessment solutions researched and developed in this project can function as a basis for future large scale gamified implementations, but require further development, improvement and testing in terms of performance and reliability before being able to be called a true human replacement method for assessing readability.

5.2 Future Work

Having completed the work for this project, we are now able to look at was has been done, and how to proceed from there.

Starting on the topic of readability assessment, there are a number of items that could do with further examination or exploration. This research focussed specifically on a single readability assessment method, the Cloze-test. However, as examined in Section 2.1, this is not the only readability assessment method. Other options, like the C-task, have not been used in this project as a potential digital assessment method.

So far all readability assessment methods that have been covered in this report are traditional (paper) methods which are currently in use. While this does provide an ideal basis to work off, a method specifically researched, tested and developed to make use of current advances in Artificial Intelligence (AI), machine learning and processing power could potentially result in an incredibly powerful and accurate system without the need for it to be based on (old-school) readability assessment methods like Cloze- or C-tests.

One area which was heavily reviewed and researched before starting this project was the area of gamification. Gamification has only been applied sparingly in places such as the user-test in Chapter 3. This is unfortunate since gamification would most likely be the deciding factor in the eventual success or failure of large scale readability assessment system based on child feedback via a gamified solution. Looking at the research done by people like Louis von Ahn (see Section 2.2), there are numerous options and methods already out there which have successfully been implemented. It would be very interesting to know how these gamification elements could be combined with the underlying principle of readability assessment into a final system which provide a fun and replayable experience to its users, and useful (accurate) readability data to us. This gamification element would eventually need to be the layer on top of the automated readability assessment system(s).

On the topic of child computer interaction (see Section 2.3), there are a few items left to consider when talking about the implementation of a final working system. One of the goals remained to be a readability assessment method/tool which can be used by children at different levels of reading proficiency, from young to old(er).

How do you design a platform in such a way that it correctly assesses readability, all the while being fun and enjoyable for multiple age-groups and levels of (computer) proficiency? And how can modern technology help in this aspect?

Finally, there is one important topic left to discuss that is essential to the success of the final implementation of this system, which is the topic of scalability. For collecting a large amount of feedback through interactions with a system, scalability has to be accounted for. Now, the groundwork for this has been done in Chapter 4 by researching and developing automated solutions for creating Cloze-tests. However, this was only the first of many steps needed to deliver a robust and complete solution that would need to be developed in order to reach the ultimate end-goal of this project, a fully fledged working large scale online gamified readability assessment system.

Bibliography

[1] (2015) Internet users (per 100 people). The World Bank. [Online]. Available: http://data.worldbank.org/indicator/IT.NET.USER.P2/countries/1W?display=graph

[2] R. Pijpers and J. de Haan, *Contact!: kinderen en nieuwe media*. Bohn Stafleu van Loghum, 2010.

[3] R. Appel and A. Vermeer, "Tweede-taalverwerving en tweede-taalonderwijs," 1994.

[4] K. Raaijmakers, ""laat mij het zelf beoordelen". een onderzoek naar de betrouwbaarheid en validiteit van moderne methodieken om tekstbegrip te voorspellen en te meten bij kinderen in de groepen 5 t/m 8 van de basisschool." *University of Utrecht Master Thesis*, 2015.

[5] D. Qian, "Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension," *Canadian modern language review*, vol. 56, no. 2, pp. 282–308, 1999.

[6] L. Verhoeven and A. Vermeer, "Woordenschat van leerlingen in het basis-en mlk-onderwijs," *Pedagogische studiën*, vol. 69, no. 3, pp. 218–234, 1992.

[7] K. E. Stanovich, "Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy," *Reading research quarterly*, pp. 360–407, 1986.

[8] A. Biemiller, "Vocabulary: Needed if more children are to read well," *Reading Psychology*, vol. 24, no. 3-4, pp. 323–335, 2003.

[9] K. Cain and J. Oakhill, "Assessment matters: Issues in the measurement of reading comprehension," *British Journal of Educational Psychology*, vol. 76, no. 4, pp. 697–708, 2006.

[10] W. L. Taylor, "Cloze procedure: a new tool for measuring readability." *Journalism quarterly*, 1953.

[11] R. G. Abraham and C. A. Chapelle, "The meaning of cloze test scores: An item difficulty perspective," *The Modern Language Journal*, vol. 76, no. 4, pp. 468–479, 1992.

[12] M. Hathcock, "Cloze procedure," *esl4teachers.pbworks.com*, 2013.

[13] R. Kraf, L. Lentz, and H. Pander Maat, "Drie nederlandse instrumenten voor het automatisch voorspellen van begrijpelijkheid-een klein consumentenonderzoek," *Tijdschrift voor Taalbeheersing*, vol. 33, no. 3, pp. 249–265, 2011.

[14] P. Gibbons, P. E. T. Association *et al.*, *Learning to learn in a second language*. Heinemann Portsmouth, NH, 1991.

[15] P. Wonghiransombat, "The cloze test and the c-test," *Thammasat University Journal*, vol. 31, no. 2, 2013.

[16] A. S. Gellert and C. Elbro, "Cloze tests may be quick, but are they dirty? development and preliminary validation of a cloze test of reading comprehension," *Journal of Psychoeducational Assessment*, 2012.

[17] J. McGonigal, *Reality is broken: Why games make us better and how they can change the world*. Penguin, 2011.

[18] S. Deterding, D. Dixon, R. Khaled, and L. Nacke, "From game design elements to gamefulness: defining gamification," in *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*. ACM, 2011, pp. 9–15.

[19] E. Halter, *From Sun Tzu to xbox: War and video games*. Thunder's Mouth Press, 2006.

[20] L. Von Ahn and L. Dabbish, "Designing games with a purpose," *Communications of the ACM*, vol. 51, no. 8, pp. 58–67, 2008.

[21] T. W. Malone, "Heuristics for designing enjoyable user interfaces: Lessons from computer games," in *Proceedings of the 1982 conference on Human factors in computing systems*. ACM, 1982, pp. 63–68.

[22] E. A. Locke and G. P. Latham, *A theory of goal setting & task performance.* Prentice-Hall, Inc, 1990.

[23] T. Frattesi, D. Griesbach, J. Leith, T. Shaffer, and J. DeWinter, "Replayability of video games," *IQP, Worcester Polytechnic Institute, Worcester*, 2011.

[24] P. Markopoulos and M. Bekker, "Interaction design and children," *Interacting with computers*, vol. 15, no. 2, pp. 141–149, 2003.

[25] D. S. Acuff and H. Robert, "The psychology of marketing to kids: what kids buy and why," *New York: The Free Press*, 1997.

[26] J. P. Hourcade, "Interaction design and children," *Foundations and Trends in Human-Computer Interaction*, vol. 1, no. 4, pp. 277–392, 2008.

[27] D. J. Cech and S. T. Martin, *Functional movement development across the life span*. Elsevier Health Sciences, 2002.

[28] R. Kail, "Developmental change in speed of processing during childhood and adolescence." *Psychological bulletin*, vol. 109, no. 3, p. 490, 1991.

[29] B. Shneiderman, "Promoting universal usability with multi-layer interface design," in *ACM SIGCAPH Computers and the Physically Handicapped*, no. 73-74. ACM, 2003, pp. 1–8.

[30] B. Shneiderman and C. Plaisant, "Designing the user interface: Strategies for effective human-computer interaction," 1987.

[31] H. E. Jochmann-Mannak, T. W. C. Huibers, and T. J. M. Sanders, "Children's information retrieval: beyond examining search strategies and interfaces," in *The 2nd BCS-IRSG Symposium: Future Directions in Information Access, London*, ser. eWic Series. London: British Computer Society, September 2008, pp. 64–72.

[32] C. L. Borgman, V. A. Walter, and S. G. Hirsh, "The science library catalog: A springboard for information literacy," *School Library Media Quarterly*, vol. 24, no. 2, pp. 105–110, 1996.

[33] A. Druin, "The role of children in the design of new technology," *Behaviour and information technology*, vol. 21, no. 1, pp. 1–25, 2002.

[34] A. van den Bosch, P. Nauts, and N. Eckhardt, "A kids' open mind common sense." in *AAAI Fall Symposium: Commonsense Knowledge*, 2010.

[35] P. Ajideh and S. Mozaffarzadeh, "C-test vs. multiple-choice cloze test as tests of reading comprehension in iranian efl context: Learners' perspective," *English Language Teaching*, vol. 5, no. 11, p. p143, 2012.

[36] J. C. Alderson, "Testing reading comprehension skills (part two). getting students to talk about taking a reading test.(a pilot study)," *Reading in a foreign language*, vol. 7, no. 1, pp. 465–503, 1990.

Appendix A

Comparison of School Year Equivalents (US/UK/NL/FR)

This appendix entry includes a table comparing various different international school systems based on year equivalence.

Comparison of School Year Equivalents

United States, Great Britain, The Netherlands and France

Source*: The American School of The Hague (https://www.ash.nl/)*

Entry age	United States	Grade Level	British System	Grade Level	The Netherlands	Grade Level	French Lycée	Grade level
3	*Elementary School*	Preschool	*Junior School*		*Basisschool*		*Maternelle*	Petite
4		Transition Kindergarten		Reception		Groep 1		Moyenne
5		Kindergarten		Year 1		Groep 2		Grande
6		Grade 1		Year 2		Groep 3		CP
7		Grade 2		Year 3		Groep 4	*Élémentarie*	CE1
8		Grade 3		Year 4		Groep 5		CE 2
9		Grade 4		Year 5		Groep 6		CM 1
10	*Middle School*	Grade 5		Year 6		Groep 7		CM 2
11		Grade 6	*Senior School*	Year 7		Groep 8	*Collège*	6 ème
12		Grade 7		Year 8	*Middelbare School (VWO)*	Brugklas		5 ème
13		Grade 8		Year 9		2e Jaar		4 ème
14	*High School*	Grade 9	*GCSE*	Year 10		3e Jaar		3 ème
15		Grade 10		Year 11		4e Jaar	*Lycée*	2 nde
16		Grade 11	*A Levels*	Year 12		5e Jaar		1 ère
17		Grade 12		Year 13		6e Jaar		Terminale

Appendix B

User-Test Cloze Forms

The included documents in this appendix entry contain all the 10 stories in Cloze-test form which were used in the user-test of stage A, see Chapter 3. These documents were also used for the offline portion of Cloze-test minus a few changes made for inclusion in this report. These changes were as follows:

An identification letter for each story was added next to the tile in brackets and every correct multiple choice answer is highlighted.

[A] Romeinse Geneesmiddelen

Uit een 2.000 jaar oud schipwrak hebben wetenschappers door DNA onderzoek de bron van de Romeinse geneesmiddelen kunnen traceren. De pillen lagen op een Italiaans scheepswrak uit het jaar 120 v. chr. Sporen van wortel, peterselie en wilde **(1)**_____ zijn in de monsters bewaard gebleven. Aangenomen wordt dat de planten gebruikt werden door artsen bij behandeling van de **(2)**_____ op het schip.

Al eerder waren behandelingsmethodes op schrift bekend, maar nog nooit werden de medicijnen zelf gevonden. “Ik vroeg me altijd af of de **(3)**_____ slechts theoretische opmerkingen waren zonder praktische uitwerking of niet”, vertelt Prof. Alain Touwaide, **(4)**_____ van het Instituut voor het Behoud van medische tradities, die 's werelds grootste digitale database van medische **(5)**_____ in het bezit heeft. Prof. Touwaide werkte samen met wetenschappers van het Smithsonian Museum, die de DNA-analyse uitvoerden.

De **(6)**_____ werden in 1974 gevonden aan boord van het gezonken schip Relitto del Pozzino voor de kust van Italië. Waarschijnlijk zonk het schip in een storm. In 2004 werden (**7)**_____ van de pillen overhandigd aan prof. Touwaide.

Voor elke open plek in de tekst staan hieronder 4 mogelijke antwoorden. Omcirkel het woord dat uit de tekst is gehaald. Als je een fout maakt teken dan een pijltje (→) naar het goede antwoord.

1 a. **uien**
b. zwijnen
c. spaghetti
d. boerenkool

2 a. slaven
b. **bemanning**
c. huisdieren
d. aardappels

3 a. tekeningen
b. **teksten**
c. schetsen
d. wetten

4 a. schoonmaker
b. boekhouder
c. **directeur**
d. stagiair

5 a. kleitabletten
b. **manuscripten**
c. bijbels
d. geneesmiddelen

6 a. planten
b. muntstukken
c. **pillen**
d. groenten

7 a. tekeningen
b. **fragmenten**
c. recepten
d. schilderijen

[B] Chinese Muur

Wat doe je als je niet wilt dat vijanden je land binnenvallen? De Chinezen wisten daar wel iets op. Zij **(1)**_____ een muur van duizenden kilometers lang!

De Chinese Muur wordt ook wel de Grote Muur genoemd. En dat is niet voor niks. Het is de langste muur ter wereld. Hij is 6.259 kilometer **(2)**_____ en ligt in het noorden van China. Het gerucht gaat dat je de Chinese Muur helemaal vanaf de **(3)**_____ kunt zien. Maar dat is niet waar. De muur is veel te **(4)**_____ om hem van die afstand te kunnen zien.

Het eerste stuk van de Chinese Muur werd rond 200 voor Christus gebouwd. De muur moest ervoor zorgen dat **(5)**_____ het land niet binnen konden komen. Maar alleen een muur was natuurlijk niet genoeg. De muur moest ook **(6)**_____ worden. En omdat de muur zo lang was, waren daar heel veel **(7)**_____ voor nodig. Meer dan een miljoen! Over de hele lengte staan er meer dan duizend forten en **(8)**_____ op de muur.

Voor elke open plek in de tekst staan hieronder 4 mogelijke antwoorden. Omcirkel het woord dat uit de tekst is gehaald. Als je een fout maakt teken dan een pijltje (→) naar het goede antwoord.

1 a. verzonnen
b. bedachten
c. kochten
d. **bouwden**

2 a. hoog
b. **lang**
c. breed
d. diep

3 a. zon
b. eiffeltoren
c. bergen
d. **maan**

4 a. groot
b. wit
c. **dun**
d. kapot

5 a. chinezen
b. **vijanden**
c. dieren
d. auto's

6 a. **bewaakt**
b. onderhouden
c. gemetseld
d. geschilderd

7 a. nederlanders
b. paarden
c. schilders
d. **mensen**

8 a. snackbars
b. vuurtorens
c. bewakers
d. **uitkijktorens**

[C] Tjernobyl

In 2009 reisden we af naar Oekraïne voor een reportage over het gebied rond de in 1986 geëxplodeerde kernreactor van Tsjernobyl. Het is al een paar jaar mogelijk om de exclusion zone te **(1)**_____ en bijvoorbeeld door de spookstad Pripjat, ten noorden van de geëxplodeerde reactor 4, te **(2)**_____. Maar Anatoliy Pakhlya, het hoofd van de toeristische staatsdienst, wil er veel meer **(3)**_____ naartoe trekken. Een bezoek aan Tsjernobyl moet een vorm van extreem **(4)**_____worden. Volgens Pakhlya is daar een markt voor omdat er overal ter wereld "veel nieuwsgierige mensen op zoek zijn naar **(5)**_____."

De kernramp van Tsjernobyl voltrok zich op 26 april 1986. Na een reeks **(6)**_____, waarbij fout op fout werd gestapeld, explodeerde reactor nummer 4 van de kerncentrale en een wolk radioactief materiaal dreef vervolgens over een groot deel van Noord-Europa. Grote **(7)**_____ in de toenmalige Sovjetrepublieken Oekraïne, Wit-Rusland en Rusland werden **(8)**_____ en zo'n 200.000 mensen werden gedwongen geëvacueerd. Opvallend genoeg kreeg het dorpje waarnaar de **(9)**_____ is genoemd maar een relatief geringe hoeveelheid straling te verduren.

Voor elke open plek in de tekst staan hieronder 4 mogelijke antwoorden. Omcirkel het woord dat uit de tekst is gehaald. Als je een fout maakt teken dan een pijltje (→) naar het goede antwoord.

1 a. vermijden
b. overvliegen
c. **bezoeken**
d. fotograferen

2 a. snowboarden
b. voetballen
c. **wandelen**
d. hockeyen

3 a. **toeristen**
b. kinderen
c. sporters
d. vrouwen

4 a. geweld
b. sporten
c. **toerisme**
d. weer

5 a. vermaak
b. **avontuur**
c. gezelligheid
d. elkaar

6 a. **experimenten**
b. wedstrijden
c. overstromingen
d. spelletjes

7 a. bedrijven
b. boerderijen
c. **gebieden**
d. voetbalclubs

8 a. **besmet**
b. verrast
c. gefeliciteerd
d. genezen

9 a. **centrale**
b. brandweer
c. president
d. boerderij

[D] Vikingen

De Vikingen konden enorm goed schepen bouwen. Deze schepen waren heel geschikt om over zee te varen. De **(1)**_____ hadden een scherpe kiel onder de bodem. Hierdoor waren de schepen goed bestuurbaar. Een groot **(2)**_____ maakte de boten zeer snel. Bij weinig wind moesten de Vikingen zelf **(3)**_____.

Het bekendste Vikingschip was het langschip of drakar. Deze schepen waren ruim 30 meter lang, smal en had een bijna platte **(4)**_____. Omdat zij zo ondiep waren konden de Vikingen er ook eenvoudig mee een **(5)**_____ opvaren. Daardoor waren ook plaatsen die landinwaarts lagen niet **(6)**_____.

Soms was de bevolking gewaarschuwd. Dan vluchtte de mensen naar veilige plaatsen of naar de boerderij van de plaatselijke heer. Hij moest hun **(7)**_____ geven. De Vikingen hielden er niet van om een goed verdedigde en versterkte **(8)**_____ aan te vallen. Belegeringsaanvallen duurden te lang en de Vikingen moesten het vooral hebben van verrassingsaanvallen.

Voor elke open plek in de tekst staan hieronder 4 mogelijke antwoorden. Omcirkel het woord dat uit de tekst is gehaald. Als je een fout maakt teken dan een pijltje (→) naar het goede antwoord.

1 a. wagens
b. huizen
c. **boten**
d. vlotten

2 a. stuurwiel
b. anker
c. **zeil**
d. motor

3 a. zwemmen
b. varen
c. **roeien**
d. lopen

4 a. zitbank
b. tv
c. dak
d. **bodem**

5 a. waterval
b. zee
c. **rivier**
d. oceaan

6 a. bereikbaar
b. **veilig**
c. gevaarlijk
d. vindbaar

7 a. eten
b. geld
c. kleding
d. **bescherming**

8 a. flat
b. **boerderij**
c. bibliotheek
d. restaurant

[E] Duinen

Toeristen verbazen zich er steeds weer over. Nederland ligt lager dan de zee. Is dat niet gevaarlijk, vragen ze zich af. Nee hoor. Ons land wordt tegen de **(1)**_____ beschermd door een lange rij duinen. Samen vormen zij de zeewering. De duinen zijn vanzelf ontstaan.

'Zee, wind, regen en plant, **(2)**_____ samen een duin van zand.' Dat rijmpje klopt precies. Elk duin is begonnen als een klein bergje zand op het **(3)**_____**.** De wind heeft er ander zand tegenaan gewaaid. Daardoor groeiden de bergjes. De regen spoelde ze soms weg, of **(4)**_____ het helmgras in de buurt. Helmgras heeft lange sterke wortels die het duinzand goed vasthouden.

Duinen zijn dus erg **(5)**_____, omdat ze ons land beschermen tegen de zee. Daarom is het belangrijk om er goed voor te **(6)**_____**.** Er is nog een reden om er zuinig op te zijn. Er leven heel veel bijzondere planten en dieren. Van de 1400 soorten **(7)**_____ die er in Nederland zijn, vind je er 850 terug in het duingebied. Van de 190 soorten vogels in Nederland, **(8)**_____ er 140 soorten in de duinen.

Voor elke open plek in de tekst staan hieronder 4 mogelijke antwoorden. Omcirkel het woord dat uit de tekst is gehaald. Als je een fout maakt teken dan een pijltje (→) naar het goede antwoord.

1 a. duitsers
b. regen
c. **zee**
d. ijstijd

2 a. **bouwen**
b. vernielen
c. zijn
d. slopen

3 a. parkeerterrein
b. gras
c. **strand**
d. land

4 a. **voedde**
b. verwoestte
c. verwarmde
d. soms

5 a. gevaarlijk
b. **nuttig**
c. slecht
d. interessant

6 a. betalen
b. plannen
c. **zorgen**
d. debatteren

7 a. **planten**
b. slangen
c. honden
d. palmbomen

8 a. **broeden**
b. leven
c. sterven
d. ontstaan

[F] Schrijven

Kun jij je nog herinneren hoe het was toen je nog niet kon lezen en schrijven? Vast niet. We kunnen het ons niet meer voorstellen, een wereld zonder **(1)**_____ en boeken. Je zou deze woorden niet begrijpen. De televisie en de krant zouden niet bestaan. De **(2)**_____ zou heel anders zijn. Toch bestaat de kunst van het schrijven nog maar vijfduizend jaar. Dat lijkt heel lang. Maar je moet erbij bedenken dat de mensen al honderdduizend jaar **(3)**_____. Er zijn sinds de mensen kunnen schrijven allerlei soorten **(4)**_____ geweest. Maar de letters die jij nu gebruikt, bestaan al meer dan tweeduizend jaar.

Het allereerste schrift bestond uit **(5)**_____. Een tekeningetje van een hoofd betekende ook gewoon 'hoofd'. Het kostte natuurlijk veel tijd om alles te **(6)**_____. Dus werden er andere manieren bedacht om iets op te **(7)**_____. Het werden een soort tekens. Die hadden wel iets weg van letters, maar leken nog helemaal niet op onze manier van schrijven. Het waren een soort krassen. En die krasjes leken wel wat op spijkers. Later noemden de **(8)**_____ het daarom spijkerschrift.

Voor elke open plek in de tekst staan hieronder 4 mogelijke antwoorden. Omcirkel het woord dat uit de tekst is gehaald. Als je een fout maakt teken dan een pijltje (→) naar het goede antwoord.

1 a. werkstukken
b. getallen
c. **letters**
d. rekenen

2 a. school
b. supermarkt
c. **wereld**
d. stad

3 a. eten
b. varen
c. **praten**
d. drinken

4 a. **schrift**
b. tijdschriften
c. boeken
d. honden

5 a. **plaatjes**
b. dieren
c. botten
d. bamboe

6 a. verzinnen
b. **tekenen**
c. verzamelen
d. betalen

7 a. lossen
b. typen
c. **schrijven**
d. nemen

8 a. chinezen
b. kinderen
c. **geleerden**
d. grieken

[G] Treinen

De stoomtrein heeft allang plaatsgemaakt voor elektrische treinen. Die zijn veel sterker en ze kunnen sneller rijden. De elektriciteit komt van kabels boven het spoor. Door de snellere **(1)**_____ moest er betere beveiliging komen. Er werden knipperlichten en spoorbomen gebouwd. Ook kwamen er wissels. Daardoor kon een trein van het ene naar het andere **(2)**_____ gaan. Om het treinverkeer echt veilig te maken hebben alle treinen nu atb, een soort automatische remmen.

Steeds meer **(3)**_____ reizen met de trein. Daarom worden de treinen steeds langer. Sinds 1985 rijden er dubbeldekkers. Die treinen hebben soms wel tien **(4)**_____ met elk honderd zitplaatsen. In totaal zijn dat duizend zitplaatsen. De treinen worden niet alleen langer, ze gaan ook steeds **(5)**_____. In Frankrijk rijden al jaren hogesnelheidstreinen. Op rechte stukken halen die zo'n 350 kilometer per uur. Op onze **(6)**_____ mogen treinen niet harder rijden dan 140 kilometer per uur, omdat er te veel bochten en **(7)**_____ zijn. Daarom is er een aparte lijn aangelegd, de hsl (hogesnelheidslijn). Die loopt van Amsterdam naar België.

In sommige **(8)**_____ gaat het nóg sneller. In Duitsland en Japan rijden magneetzweeftreinen. Deze treinen zweven boven een magnetische baan.

Voor elke open plek in de tekst staan hieronder 4 mogelijke antwoorden. Omcirkel het woord dat uit de tekst is gehaald. Als je een fout maakt teken dan een pijltje (→) naar het goede antwoord.

1 a. auto's
b. reizigers
c. machinisten
d. **treinen**

2 a. land
b. station
c. **spoor**
d. bedrijf

3 a. **mensen**
b. huisdieren
c. vrouwen
d. kinderen

4 a. goederenwagens
b. locomotieven
c. verdiepingen
d. **rijtuigen**

5 a. **sneller**
b. langzamer
c. vaker
d. eerder

6 a. **spoorlijnen**
b. wegen
c. stations
d. bruggen

7 a. stoplichten
b. treinen
c. bergen
d. **wissels**

8 a. provincies
b. **landen**
c. steden
d. werelddelen

[H] Brandweer

In Nederland werken zo'n 27.12 duizend mensen bij de brandweer. Zij zijn samen het brandweer-korps. Een klein gedeelte daarvan is van beroep brandweer-man. Maar de meeste van hen zijn **(1)_____** van de vrijwillige brandweer. Ze hebben ook een andere baan waar ze geld mee verdienen. Als ze aan het **(2)_____** zijn, kan er natuurlijk een brandalarm komen. De vrijwillige brandweer-mannen kunnen dan met een semafoon worden opgepiept. Ze hebben met hun **(3)_____** afgesproken dat ze direct weg mogen als dat gebeurt. In de kazerne heeft elke brandweer-man een eigen **(4)_____** hangen. Ook is er voor iedereen een ademhalings-toestel. In dat toestel zit genoeg lucht voor ongeveer twintig minuten.

Brandweer-mannen doen meer dan alleen branden **(5)_____**. Als er een groot verkeers-ongeluk gebeurt, rukt de brandweer ook uit. Soms moeten ze een slachtoffer uit een auto **(6)_____.** De brandweer komt ook bij overstromingen te hulp. Ze redden dan mensen en huisdieren die door het hoge water hun **(7)_____** niet meer uit komen. Een andere belangrijke taak van de brandweer is het voorkomen van brand. De brandweer adviseert over de **(8)_____** in gebouwen. Ze maken bijvoorbeeld een vlucht-plan.

Voor elke open plek in de tekst staan hieronder 4 mogelijke antwoorden. Omcirkel het woord dat uit de tekst is gehaald. Als je een fout maakt teken dan een pijltje (→) naar het goede antwoord.

1 a. directeur
b. leider
c. **lid**
d. schoonmakers

2 a. blussen
b. koken
c. **werk**
d. tuinieren

3 a. vrouw
b. **baas**
c. kinderen
d. tuinman

4 a. naambordje
b. **pak**
c. gitaar
d. paraplu

5 a. stichten
b. veroorzaken
c. starten
d. **blussen**

6 a. **losknippen**
b. wegvliegen
c. vervoeren
d. verzorgen

7 a. wc
b. tuin
c. pyjama
d. **huis**

8 a. wifi
b. verwarming
c. huurkosten
d. **veiligheid**

[I] Facebook Onderzoek

Volgens een experiment uitgevoerd door het Happiness Research Institute zijn mensen zonder Facebook gelukkiger. De **(1)**_____ verzamelden iets meer dan 1000 proefpersonen en vroeg ze onder meer naar hoe tevreden ze waren met hun leven en hoe **(2)**_____ ze waren. Daarna werd de grote groep proefpersonen in tweeën gedeeld. De helft van de **(3)**_____ mocht een week lang niet Facebooken, de andere helft mocht het sociale medium wel gewoon **(4)**_____ en gebruiken en deed dienst als controlegroep.

Na een week werden dezelfde vragen aan de proefpersonen gesteld. Wat blijkt? De proefpersonen die deel uitmaakten van de **(5)**_____ gaven hun leven een 7,67. Een week later was dat een 7.75. Een lichte **(6)**_____ dus. De groep die een week lang geen Facebook mocht gebruiken, gaf hun **(7)**_____ een 7,56, maar na een week zonder Facebook een 8,12! Ook waren de mensen zonder **(8)**_____ na een week socialer en hadden ze minder concentratieproblemen.

Voor elke open plek in de tekst staan hieronder 4 mogelijke antwoorden. Omcirkel het woord dat uit de tekst is gehaald. Als je een fout maakt teken dan een pijltje (→) naar het goede antwoord.

1 a. **onderzoekers**
b. scholieren
c. werknemers
d. kinderen

2 a. rijk
b. boos
c. **gestrest**
d. slim

3 a. onderzoekers
b. vrouwen
c. mannen
d. **proefpersonen**

4 a. **bezoeken**
b. onderzoeken
c. betalen
d. testen

5 a. **controlegroep**
b. universiteit
c. sportvereniging
d. chatgroep

6 a. daling
b. verlaging
c. uitschieter
d. **stijging**

7 a. werkstuk
b. **leven**
c. cito-toets
d. vrienden

8 a. twitter
b. **facebook**
c. whatsapp
d. snapchat

[J] Olympische Spelen

De oude Grieken waren dol op wedstrijden. Winnen was het belangrijkste, want daarmee toonden ze hun **(1)**_____. We weten dat er in 776 voor Christus Olympische Spelen werden gehouden in Olympia. Die Spelen duurden zes dagen. Op de eerste dag **(2)**_____ de Grieken honderd stieren ter ere van hun belangrijkste god Zeus (zeg: Zuis). In 394 na Christus werden de Spelen **(3)**_____. Verering van meerdere **(4)**_____ zoals op de Spelen gebeurde, paste niet meer bij het **(5)**_____van die tijd.

In de 19e eeuw richtte de Fransman Pierre de Coubertin de Olympische Beweging op. Van **(6)**_____word je een beter mens, vond hij. In 1896 organiseerde hij in Athene de eerste moderne Olympische Spelen.

De moderne Olympische Spelen zijn anders dan die van de oude Grieken. Meedoen is nu belangrijker dan winnen. Ook de **(7)**_____ zijn veranderd. Het gaat er niet meer zo hard aan toe. In tegenstelling tot vroeger mogen nu ook **(8)**_____meedoen.'

Voor elke open plek in de tekst staan hieronder 4 mogelijke antwoorden. Omcirkel het woord dat uit de tekst is gehaald. Als je een fout maakt teken dan een pijltje (→) naar het goede antwoord.

1 a. spieren
b. kennis
c. gezondheid
d. **kracht**

2 a. kochten
b. stalen
c. **slachtten**
d. schilderden

3 a. **verboden**
b. gefilmd
c. verplaatst
d. veranderd

4 a. atleten
b. **goden**
c. landen
d. koningen

5 a. techniek
b. vermaak
c. **geloof**
d. werk

6 a. handelen
b. praten
c. werken
d. **sporten**

7 a. atleten
b. talen
c. **spelregels**
d. records

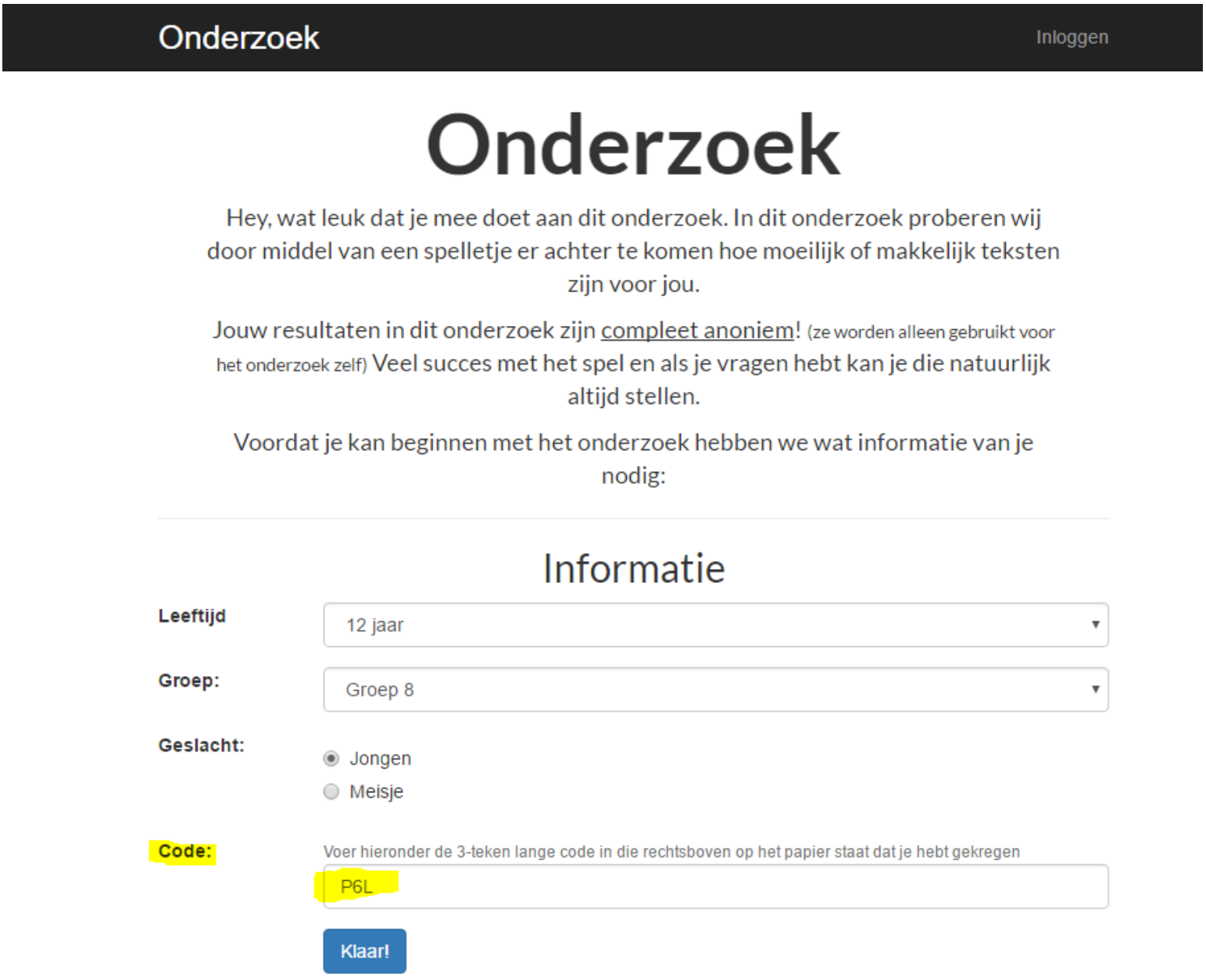
8 a. mannen
b. dieren
c. voetballers
d. **vrouwen**

Appendix C

Online Cloze-test Procedure

The following pages include screenshots of the online Cloze-test as used by the user-test in Chapter 3 of this research. The questionnaire page at the end of online test is not added in this appendix entry, but can be seen in the Appendix D.

Login & Entering Unique ID (highlighted)



Online game explanation page

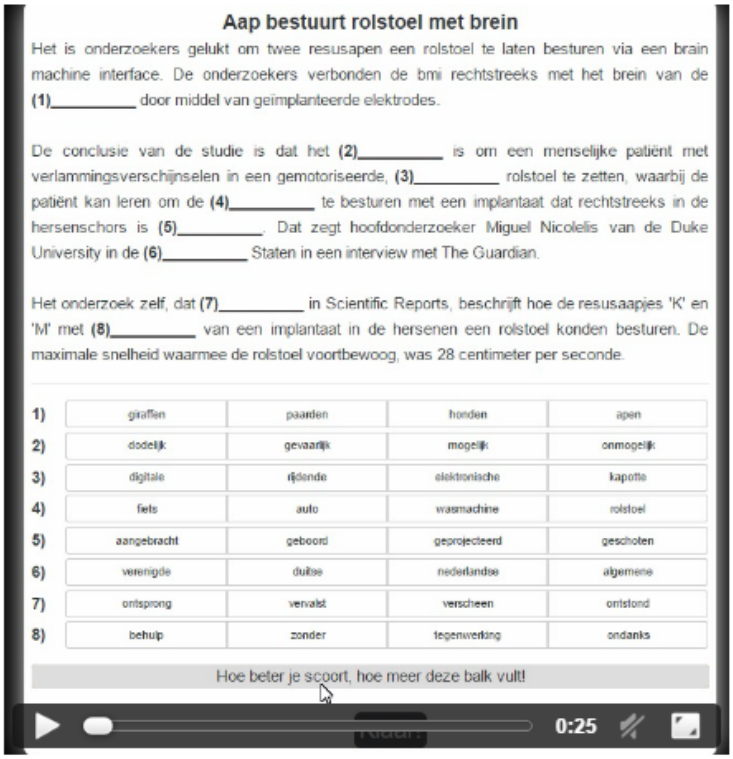
Onderzoek #User-95 Code: P6L

Onderzoek

Op de volgende pagina's krijg je straks 2 testjes met na afloop nog een kort vragenlijstje. Het testje werkt zo:

Je krijgt een tekst te zien met daarin een aantal open plekken (___). Naast elke plek staat ook een nummer die de plek aangeeft van het weggelaten woord. Onder de tekst staan voor elke open plek 4 mogelijke antwoorden. Het is aan jou om het goede antwoord daaruit te kiezen. Als je alles hebt ingevuld klik je op de 'klaar' knop en kan je zien welke van je antwoorden goed of fout zijn. Als je fouten hebt gemaakt kan je die wijzigen. Wanneer je alles goed hebt kan je door naar de volgende pagina!

Hieronder staat een kort filmpje (zonder geluid) die laat zien hoe het werkt. Als je wilt beginnen kan je op de **'Ga Door!'** knop onder het filmpje klikken.



Ga Door!

Game Screen (x2)

Each participant performs the online Cloze-test game twice but using two different stories. For this example only one is shown, the game itself remains identical.

Switch Message

Goed Gedaan!

Je bent klaar met de 2 online verhalen. Heb je de 2 papieren verhalen al gedaan?

Ja Nee

Appendix D

Online Questionnaire

Figure D.1: Online questionnaire as of 09-05-2016

Appendix E

User-Test Questionnaire Results

User-Test Questionnaire Results

After completing both the on- and offline portions of the user-test each participant was asked to fill in a short digital questionnaire form. The results and the explanation of the questions are listed below, when a question has limited number of answers the possible answers are listed in brackets. All questions are translated from Dutch.

User ID	ID of the user in the online database which holds the results for every user for the online portion of the user-test
Q1	*"What did you think of the game?"* [Likert response 1-5: very unenjoyable/unenjoyable/average/fun/very fun]
Q2	*"Was it easy or hard?"* [Likert response 1-5: very hard/hard/normal/easy/very easy]
Q3	*"Was it easier or harder than the paper test?"* [Likert response 1-5: much harder/harder/similar/easier/much easier]
Reading	*"Outside of school, how often do you read?"* [daily/weekly/monthly/yearly/never]
Dyslexia	*"Are you dyslectic (have difficulty reading)?"* [yes/no/unknown]
AVI-level	*"What is your AVI-level?" [AVI- E5/M6/E6/M7/E7/plus/ unknown]*
Comments	The participant was free to type in any thoughts or comments they had about the user-test

User ID	Q1	Q2	Q3	Reading	Dyslexia	AVI-level	Comments (optional)
12	3	4	3	weekly	no	plus	
5	4	4	2	daily	no	plus	
4	3	4	2	weekly	no	plus	
7	5	4	3	daily	no	plus	
10	2	4	4	daily	unknown	plus	dat je iets meer uit kan leggen WAT je met onze resultaten gaat doen want ik weet wel dat je het gemiddelde er uit gaat halen maar wat ga je met het gemiddelde doen?
1	3	3	3	daily	no	plus	waarom heb je niet alles op de computer.
6	3	4	2	monthly	unknown	unknown	
2	5	3	5	weekly	no	plus	
3	3	2	5	yearly	no	plus	
16	4	3	3	daily	yes	unknown	
8	3	4	3	daily	no	plus	het is goed geregld
13	4	3	3	daily	no	plus	
11	3	3	2	daily	no	plus	Sommige vragen wist ik echt niet omdat ze niet zo duidelijk waren. Maar voor de rest was het makkelijk om te verbeteren.
17	5	4	2	daily	no	plus	LEUK!! !! !!!
9	3	3	2	daily	no	plus	Ik vond het best saaie teksten. Misschien moeten er wat leukere teksten komen. Wat ik wel heel cool vond, is dat je zonder Face-Book gelukkiger bent.

18	4	4	3	daily	no	plus	
19	3	4	3	daily	no	plus	nee
15	2	3	4	daily	unknown	m7	doeiiiiiii doeiiiiiii doeiiiiii doeiiiiii doeiiii doeiiii doeiii doeiii doeiii doeiii doeiii
20	3	4	4	daily	no	unknown	ik vondt het wel leuk om jouw zoekmachine te helpen met dit onderzoek
14	4	3	3	weekly	no	plus	het was heel leuk. doeii doeii doeii doeii doeii doeii doeii
21	3	3	3	never	unknown	plus	spanende verhalen een plats dan saaien verhalen voor vroeger of hoe iets is ontstaan.. x'D :)
22	3	3	3	daily	unknown	plus	Nou, bij sommigen moeilijke maar dan daar na ook weer even een makkelijke vraag er tussen. je leert ook meer van da geschiedenis
31	4	3	4	daily	no	plus	nee
28	3	4	4	weekly	no	plus	
25	4	4	3	monthly	no	plus	Het was goed gemaakt vooral ook voor die texten
33	3	3	3	yearly	no	unknown	
27	3	4	3	weekly	no	unknown	het was wel leuk maar het lijkt gewoon een beetje op een les van school
32	3	3	3	weekly	no	plus	
24	3	3	4	weekly	no	plus	ik vond het wel interessant.
26	3	4	3	weekly	no	plus	ik vind het een goed onderzoek.
29	3	3	3	never	yes	unknown	
34	3	3	3	monthly	no	plus	ik vond het een leuk onderzoek.
40	3	3	3	weekly	no	plus	het was wel oke
36	3	4	2	weekly	no	plus	
39	3	5	3	daily	no	plus	
35	3	3	4	weekly	no	plus	
42	3	3	3	weekly	yes	plus	het was wel cool
44	4	5	3	never	no	plus	
45	3	4	3	yearly	no	plus	hoi ik ben cesar ik zit op hockey
46	4	3	4	monthly	no	plus	nee
41	3	4	4	weekly	no	plus	
43	2	4	3	weekly	no	plus	
38	3	4	4	monthly	no	plus	
37	3	4	3	weekly	no	plus	

Appendix F

Cloze Automatization User-Test

This appendix entry contains a copy of the online user-test as talked about in Section 4.7 of this report, formatted for inclusion as appendix.

This user-test contained questions concerning Dutch texts and is therefore written in Dutch as well. The writing style is largely informal, as it was distributed among friends and family.

The type of response required from the participant for a particular (set of) question(s) is shown between brackets ”[ ]” e.g. *[multiple choice response]*. If an answer to a question was required, the following would appear: *required.

Onderzoek Rutger Varkevisser

Hoi, en hartstikke bedankt dat je even een paar minuten de tijd wilt nemen om mij te helpen met mijn onderzoek. De resultaten van dit onderzoekje zullen worden opgenomen in een deel van mijn scriptie voor mijn master diploma bij de Universiteit Twente. Hieronder staat kort beschreven wat en waarom ik dit onderzoek doe. Lees het a.u.b. goed door anders is wellicht later niet alles even duidelijk. (lees anders in ieder geval het stukje over 'de opdracht')

In het kort houdt mijn onderzoek zich bezig met het kijken naar de leesbaarheid van (online) teksten, in het specifiek voor kinderen. Als onderdeel van mijn onderzoek gebruik ik een test, de zogenoemde Cloze-test.

Cloze-test

De Cloze-test is een test waarbij de gebruiker een tekst krijgt met daarin een aantal gaten. Deze gaten beslaan individuele woorden en het is aan de gebruiker om de open plekken in te vullen met het woord dat hij/zij denkt dat verwijderd is, en zodoende in de tekst past. De resultaten van deze test geven een indicatie van het leesniveau van de gebruiker/het niveau van de tekst.

Normaal gesproken worden Cloze-testen handmatig gemaakt aan de hand van bepaalde richtlijnen, zodat de 'goede' woorden worden verwijderd uit een tekst. Dit zijn woorden waarbij de gebruiker aanspraak moet doen op zijn/haar contextuele kennis, zowel binnen als buiten de zin waarin het woord zich bevind. Bijv. geen 'de/het/een', vaak werkwoorden of zelfst. naamwoorden, en het moeten natuurlijk woorden zijn die überhaupt te raden zijn.

Handmatig/automatisch

Wat ik heb geprobeerd als onderdeel van mijn opdracht is om dit proces van woord selectie te automatiseren. Waarbij ik in plaats van handmatig de tekst te doorlopen en volgens richtlijnen woorden te markeren, een systeem een tekst als invoer geef, en als uitvoer een tekst met gaten krijg aan de hand van een aantal algoritmes die 'correcte' woorden voor een Cloze-test zouden moeten selecteren.

De Opdracht

In de komende pagina's staan 3 korte tekstjes (+- 200 woorden) met daarin een 8-tal (genummerde) gaten. Mijn vraag aan jou is om een woord in te vullen dat je denk dat op die plaats in de tekst is weggelaten. Denk hier vooral niet te lang over na, dit is niet het centrale doel van dit onderzoek. Weet je een woord niet, zet dan een streepje (-) of vraagteken (?) of typ iets willekeurigs in. Aan het einde van elke opdracht is mijn vraag of jij, gezien de weggelaten woorden in de tekst/opdracht, denkt of de opdracht handmatig of automatisch (via een systeem) gemaakt is.

De test is anoniem, en het gaat mij ook niet om je individuele score op de opdrachten, maar om de vraag of je (aan de hand van de weggelaten woorden), denkt te weten of elke opdracht automatisch of handmatig gemaakt is.

Sorry voor het lange verhaal, dit was het. Nogmaals bedankt en succes.

p.s. Mocht er iets onduidelijk zijn dan help ik natuurlijk graag, bel/sms/whatsapp me (06########) of stuur me een mailtje (r.a.varkevisser@student.utwente.nl) dan probeer ik zo snel mogelijk te reageren.

1. Duinen

Toeristen verbazen zich er steeds weer over. Nederland ligt lager dan de zee. Is dat niet gevaarlijk, vragen ze zich af. Nee hoor. Ons land wordt tegen de (1)_____ beschermd door een lange rij duinen. Samen vormen zij de zeewering. De duinen zijn vanzelf ontstaan. 'Zee, wind, regen en plant, (2)_____ samen een duin van zand.' Dat rijmpje klopt precies. Elk duin is begonnen als een klein bergje zand op het (3)_____. De wind heeft er ander zand tegenaan gewaaid. Daardoor groeiden de bergjes. De regen spoelde ze soms weg, of (4)_____ het helmgras in de buurt. Helmgras heeft lange sterke wortels die het duinzand goed vasthouden.

Duinen zijn dus erg (5)_____, omdat ze ons land beschermen tegen de zee. Daarom is het belangrijk om er goed voor te (6)_____. Er is nog een reden om er zuinig op te zijn. Er leven heel veel bijzondere planten en dieren. Van de 1400 soorten (7)_____ die er in Nederland zijn, vind je er 850 terug in het duingebied. Van de 190 soorten vogels in Nederland, (8)_____ er 140 soorten in de duinen.

[short-answer text responses]
(1)_____ *required
(2)_____ *required
(3)_____ *required
(4)_____ *required
(5)_____ *required
(6)_____ *required
(7)_____ *required
(8)_____ *required

Kijkend naar de weggelaten woorden in de bovenstaande tekst (in termen van duidelijkheid, moeilijkheid, type woord, zijn ze geschikt voor zo'n soort opdracht), denk je dat deze 'opdracht' handmatig of automatisch gemaakt is? [multiple choice response] *required

- Handmatig
- Automatisch

Als je het zou moeten aangeven op een schaal, hoe zeker ben je dat het manueel/automatisch gemaakt is? [linear scale response 1-5] *required

1. Absoluut zeker dat het manueel gemaakt is
5. Absoluut zeker dat het automatisch gemaakt is

2. Treinen

De stoomtrein heeft allang plaatsgemaakt voor elektrische treinen. Die zijn veel sterker en ze kunnen sneller rijden. De elektriciteit komt van kabels boven het spoor. Door de snellere (1)_____ moest er betere beveiliging komen. Er werden knipperlichten en spoorbomen gebouwd. Ook kwamen er wissels. Daardoor kon een trein van het ene naar het andere (2)_____ gaan. Om het treinverkeer echt veilig te maken hebben alle treinen nu atb, een soort automatische remmen.

Steeds meer (3)_____ reizen met de trein. Daarom worden de treinen steeds langer. Sinds 1985 rijden er dubbeldekkers. Die treinen hebben soms wel tien (4)_____ met elk honderd zitplaatsen. In totaal zijn dat duizend zitplaatsen. De treinen worden niet alleen langer, ze gaan ook steeds (5)_____. In Frankrijk rijden al jaren hogesnelheidstreinen. Op rechte stukken halen die zo'n 350 kilometer per uur. Op onze (6)_____ mogen treinen niet harder rijden dan 140 kilometer per uur, omdat er te veel bochten en (7)_____ zijn. Daarom is er een aparte lijn aangelegd, de hsl (hogesnelheidslijn). Die loopt van Amsterdam naar België.

In sommige (8)_____ gaat het nóg sneller. In Duitsland en Japan rijden magneetzweeftreinen. Deze treinen zweven boven een magnetische baan.

[short-answer text responses]
(1)_____ *required
(2)_____ *required
(3)_____ *required
(4)_____ *required
(5)_____ *required
(6)_____ *required
(7)_____ *required
(8)_____ *required

Kijkend naar de weggelaten woorden in de bovenstaande tekst (in termen van duidelijkheid, moeilijkheid, type woord, zijn ze geschikt voor zo'n soort opdracht), denk je dat deze 'opdracht' handmatig of automatisch gemaakt is? [multiple choice response] *required

- Handmatig
- Automatisch

Als je het zou moeten aangeven op een schaal, hoe zeker ben je dat het manueel/automatisch gemaakt is? [linear scale response 1-5] *required

1. Absoluut zeker dat het manueel gemaakt is
5. Absoluut zeker dat het automatisch gemaakt is

3. Olifanten

Een Afrikaanse mannetjes-olifant weegt ongeveer zesduizend kilo. Dat is net (1)_____ als tachtig mensen. Of zes auto's. Of twaalf grote paarden. Of vijftienhonderd (2)_____. Om dat lijf zo groot en sterk te houden, eet een olifant heel veel. Met zijn slurf zoekt hij de hele dag door naar eten. Hij eet geen vlees. Er staat alleen maar plantenvoedsel op zijn menu. Hij is dol op gras, planten, (3)_____, takken, vruchten en bladeren. Daarvan eet hij elke dag zo'n honderdvijftig kilo. (4)_____ trekken van de ene plek naar de andere. Dat moet ook wel, want al dat groen is op een dag gewoon op door al die grijze veelvraten.

Als een olifant wordt (5)_____, weegt hij honderd kilo. Dat is net zoveel als drie kinderen van tien jaar. Een mensenbaby weegt drie kilo. Omdat olifanten in (6)_____ leven, heeft een olifantenbaby behalve zijn moeder ook heel veel 'tantes'. De vader van het (7)_____ olifantje is na de paring terug gegaan naar de mannetjes-groep. Is de baby-olifant een (8)_____, dan gaat hij naar de mannetjes-groep als hij acht jaar is. Tot die tijd blijft hij bij de vrouwtjes-groep.

[short-answer text responses]
(1)_____ *required
(2)_____ *required
(3)_____ *required
(4)_____ *required
(5)_____ *required
(6)_____ *required
(7)_____ *required
(8)_____ *required

Kijkend naar de weggelaten woorden in de bovenstaande tekst (in termen van duidelijkheid, moeilijkheid, type woord, zijn ze geschikt voor zo'n soort opdracht), denk je dat deze 'opdracht' handmatig of automatisch gemaakt is? [multiple choice response] *required

- Handmatig
- Automatisch

Als je het zou moeten aangeven op een schaal, hoe zeker ben je dat het manueel/automatisch gemaakt is? [linear scale response 1-5] *required

1. Absoluut zeker dat het manueel gemaakt is
5. Absoluut zeker dat het automatisch gemaakt is

Afsluiting

Je bent bijna klaar, nog een paar vragen ter afronding.

Als moet kiezen, van welke van de voorgaande 3 opdrachten ben je het meest zeker dat die automatisch gemaakt is. [multiple choice reponse] *required

- 1) Duinen
- 2) Treinen
- 3) Olifanten

Hoe zeker ben je van deze keuze? [linear scale response 1-10] *required

1. 0% zeker
10. 100% zeker

Heb je verder nog opmerkingen over dit onderzoek? [long answer text response]

Bedankt voor het helpen!

Dat was het. Hartstikke bedankt voor je deelname/hulp aan dit onderzoekje. Ik ben je een toekomstig drankje verschuldigd!

Groet, Rutger