RISE SICS East AB

Using eye tracking and subjective evaluations to determine text difficulty

Cornelia Böhm
2018-07-12

Index

Introduction

This report summarizes a pilot study made on behalf of RISE SICS East AB for the project DigInclude. The project goal is to make online tools available for everyone by creating text simplifications, multi lingual terminology and quick translations, amongst other things[1]. The current study is a part of the text simplification where a text simplifier, using corpuses, is created.

A corpus is a collection linguistic data which can include both transcriptions and written texts. It is in an electronic format and its purpose is to represent a specific part of the language. When using corpuses, it is important that they maintain a certain quality by using a big, representative collection of texts. The texts should be authentic, and thereby not created solely to be included in the corpus ("Korpusar — Språkteknologi.se", 2018).

Purpose and aim

The purpose of this pilot study is to investigate whether the difficulty of texts can be evaluated with the use of eye tracking, together with subjective evaluations, and how this should be done. The study will aim to evaluate the quality of a collection of corpuses to see if they are appropriate to use for text simplification. The collection includes two corpuses of simplified texts and two corpuses of original texts. To consider these corpuses to be of good quality the corpuses of simplified texts will have to be easier to read than the corpuses of original texts.

Eye tracking

Eye tracking is a method of measuring a person's eye movements, often by taking photos of light reflections from the fovea of eyes (Holmqvist & Nyström, 2011). There are three common methods of eye tracking. The first one is the static eye tracker and this includes the remote eye tracker, which is placed in front of the participant, and the tower-mounted eye tracker, which are in close contact with the participant. The second type of eye tracker is the head-mounted eye tracker which is often in the form of glasses on the participant's head. The third eye tracker is the head tracker which is basically a head-mounted eye tracker with an addition of having a head tracker which keeps track of the head's position in space (Holmqvist & Nyström, 2011).

Fixations

A fixation is a state where the eye remains focused on one area over a period of time (Holmqvist & Nyström, 2011). Eye fixations reflect attention and shifts in attention, and fixation duration can indicate the participant's cognitive load as the taxation time increases together with the difficulty of a task (Amadieu, van Gog, Paas, Tricot & Mariné, 2009).

Regressions

Regressions are reading events where the eye movement traces backwards to re-read a part of the text (Holmqvist & Nyström, 2011). Although it is not completely clear what causes all regressions, it is likely that several of them are due to comprehension issues (Rayner, 1998).

Previous Research

Andrzejewska and Stolińska (2016) looked at eye movement as school-children completed tasks of different difficulty. By doing this they wanted to see if eye movements could reveal the difficulty of the tasks. To determine the level of difficulty they used subjective measures (the participants ratings), and behavioural measures (the number of correct answers). The study showed that the number of correct answers correlated with the number of fixations, however, other measures like

[1] https://www.sics.se/projects/digital-inkludering-i-det-uppkopplade-samhallet-for-grupper-med-speciella-behov

blinking did not show any correlation with difficulty level. This indicates that if one wants to look at difficulty level, one should analyse fixations.

Atvars (2017) conducted a study where they used eye movement analyses to find a readability formula. They did this by having children from primary school reading 15 different text with varying level of difficulty. Other than fixations they also found that saccades and regressions could be of use when determining readability of texts. They also thought that fixation duration alone could not determine readability.

Further on, Jarodzka and Brand-Gruwel (2017) also thought that regressions could be of importance when determining the difficulty of a text. Regressions show when participants jumps backwards in the text to re-read, which could indicate the text is difficult. A regression can occur both within a sentence and a single word.

Method

The method used in this study was as follows.

Participants

The study contained 11 participants between the ages of 21 and 31. However, one participant had to be removed from the result since the eye tracking ratio was below 90%, meaning that the eye tracker could not pick up enough data from the participant's eye movements. The participants were recruited by convenience sampling among friends and co-workers and none of them reported having any reading difficulties.

Materials

The eye tracker used in this study was a Remote eye tracker from SMI with 500hz. The program used to launch the eye tracker was SMI iView X, and the program used the run the test was SMI ExperimentCenter. SMI BeGaze, Microsoft Excel and IBM SPSS Statistics 24 was used to analyse the data.

Besides this, the study also contained eight different texts which had been randomized from a collection of corpuses. To make sure that the length of the texts did not affect the participants opinion of the difficulties of the texts all chosen texts were between 5-7 lines long. This means that if the randomized text was too long another randomization was done.

The texts in the corpuses had previously been gathered from different authority- and county websites and the texts were of various length. Some of the original text had simplified versions in the simplified corpuses, but this was not the case for all of them. Two of the four corpuses were based on simplified texts where one of them had texts from authority websites and the other from county websites. The other two corpuses contained original text, also extracted from authority- and county websites. For the experiment, the texts were written in a plain document. The lines where centred and kept their original line breaks. However, double spacing was added to make sure that the eye tracker could differentiate between the lines. The texts can be found in the appendix.

Procedure

Before the test began, the participants read an explanation of how the test would be carried out. They then signed an informed consent and the test could begin. It was a within-group design and for all participants, the order of the texts was randomized for each participant to prevent a learning

effect. Before starting the recording, a calibration was executed until the average deviation was below 1.

The texts were displayed on a computer display and the participant operated the start and finish of every session by themselves. This meant that they pressed a key on the keyboard when they were ready to start reading and the same key when they had finished reading the text. After each text, the screen switched to a pause page, and the participants were asked to retell the content of the text. However, it was stressed that the participants did not have to try to memorize the text but only retell what they could remember. This would make sure that the participants read the texts carefully and not just skimmed them through. After this, the participants were asked to score the text's difficulty between one to seven, where one was very easy and seven was very difficult.

These question sections were carried immediately after the read so that the memory was fresh. The questions were asked and answered carried out verbally, as writing on a paper would mean that the eye tracker would have to be adjusted again. When the two questions were done they could go forth with reading the next text when they felt ready.

When all data had been gathered it was analysed in BeGaze, Excel and SPSS. The subjective evaluations of the difficulty of the texts were in the form of a Likert scale where the participants could answer between one to seven. For the retelling of the texts the participants answered freely, and the test leader wrote down the answers with a specific focus of when the participant remembered the texts incorrectly or not at all.

Results

The result from the study included eye tracking data, subjective evaluations and data about how much the participants could remember from the texts.

Eye tracking data

The eye tracking data was analysed in several ways. One type of analysis which was made was observing scan paths. In figure 1 we can see a scan path of participant 2 reading text three. This is one of the simplified texts from a county website and it is also the text which the participant rated as easiest out of the eight texts. The blue circles show the fixations that the participant has made; the bigger the circle the longer the fixation. To compare with one of the original texts we can look at figure 2. Figure 2 displays a scan path of participant 2 reading the seventh text. This was the text which the participant rated as most difficult out of the eight texts. Compared to image 1, the fixations are visibly longer in the more difficult text and there are also more fixations. When looking at the data of the average number of fixations per word for participant 3 it was 0.70 for text number three and 1.21 for the seventh text. However, the differences in scan paths for the participants were not always as clear as this example.

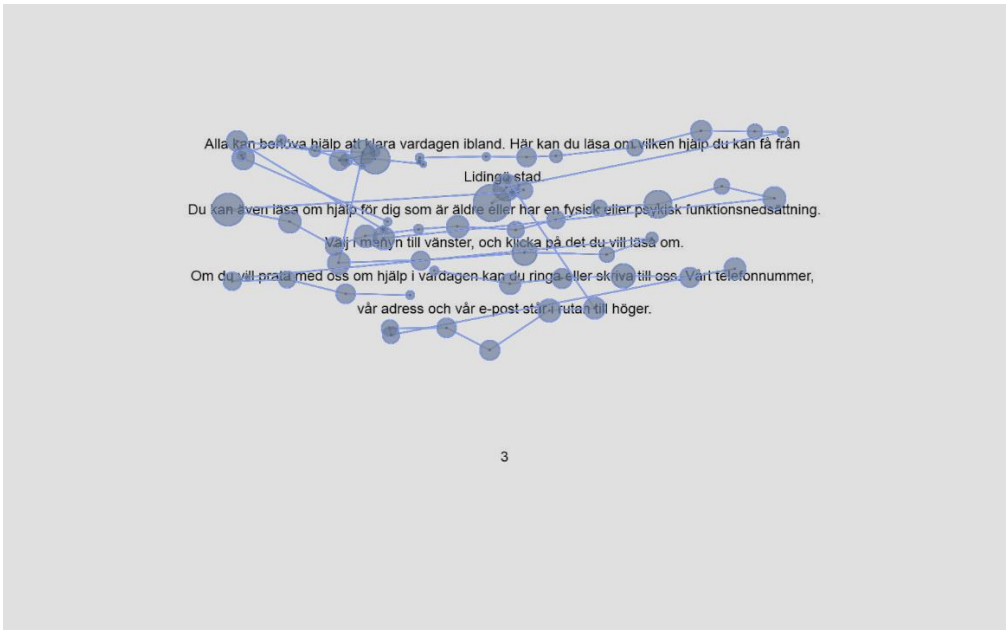

Figure 1. This shows a scan path of participant 2 reading text number three. This was the text which the participant rated as easiest.

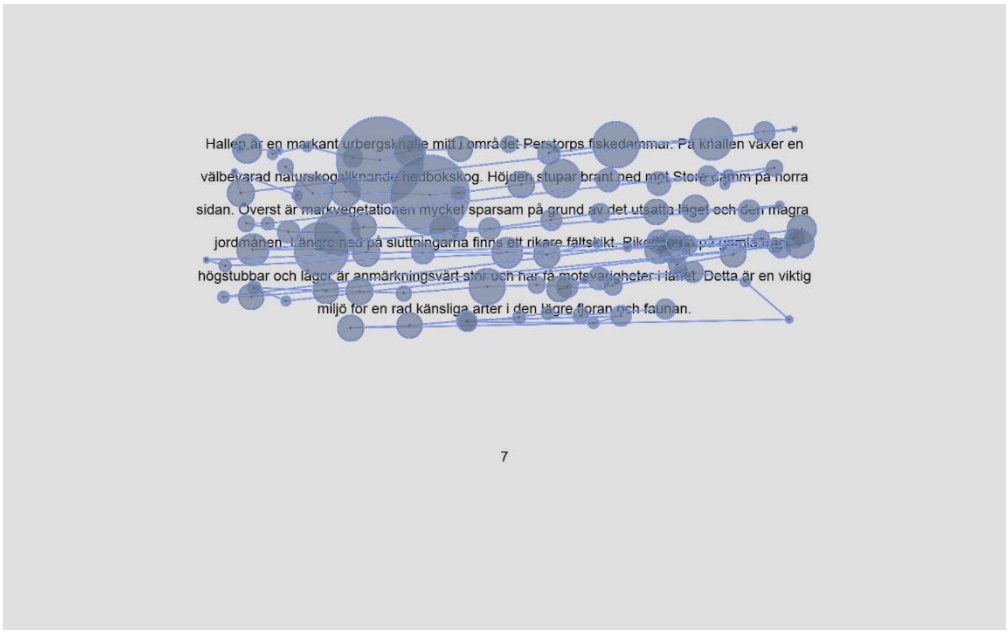

Figure 2. This shows a scan path of participant 2 reading the seventh text. The participant rated this as the most difficult out of the eight texts.

When looking further at the average fixation count per word, a Mann-Whitney U test, revealed a significant difference in the average number of fixations per word between the two conditions (U = 495, p = .003). A comparison between them can be seen in figure 3. Here, we can see that the average number of fixations per word for the simplified texts were 5.28 and for the original texts was 6.80. However, a Mann-Whitney U test on fixation duration showed no significance between the conditions (U = 688, p = .281). There was no significant difference between the simplified or original texts when it came to regressions either (U = 633, p = .108).

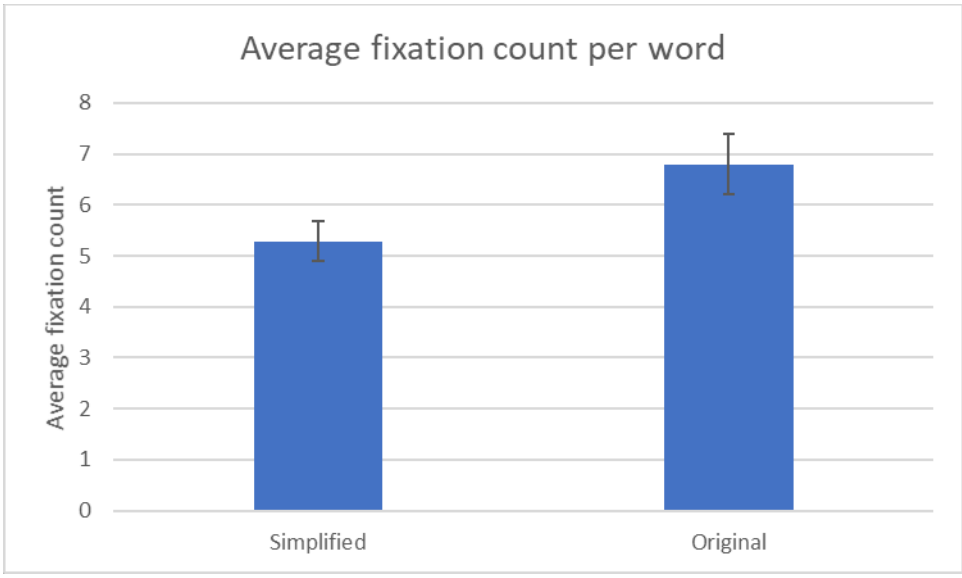

Figure 3. Histogram showing the average fixation count per word for the two conditions. The error bars illustrate the standard deviation.

To see if the average number of fixations correlated with the subjective evaluations, a Pearson correlation coefficient was calculated. It showed a strong positive correlation between subjective evaluations and average number of fixations, but the correlation was not significant (r = .680, n = 8, p = .064).

Subjective evaluations

The average subjective evaluations for each text can be seen in figure 4. It illustrates the participants' evaluations of the difficulties of the texts; one being very easy and seven being very difficult. In the figure we can see that the standard deviations (SD) of the evaluations are rather high in comparison to the means. This shows that many of the participants rated differently on many texts. Text eight was evaluated as the most difficult text with an average of 5.36 and a SD of 0.778. Text number 7 was rated as almost as difficult with an average of 5.18, however, it had a higher SD of 1.11. The text considered as easiest was text number three with an average of 1.45 and a SD of 0.656. Overall it looks like two of the texts had a high difficulty, but the other texts were mainly on the same level.

Further on, in figure 5, the data is presented according to the conditions. Here, we can see that the simplified texts seem to be easier, however, the high SD on both conditions indicate that the participants had varying views on the difficulty. Since the error bars overlap, the difference is not considered significant.

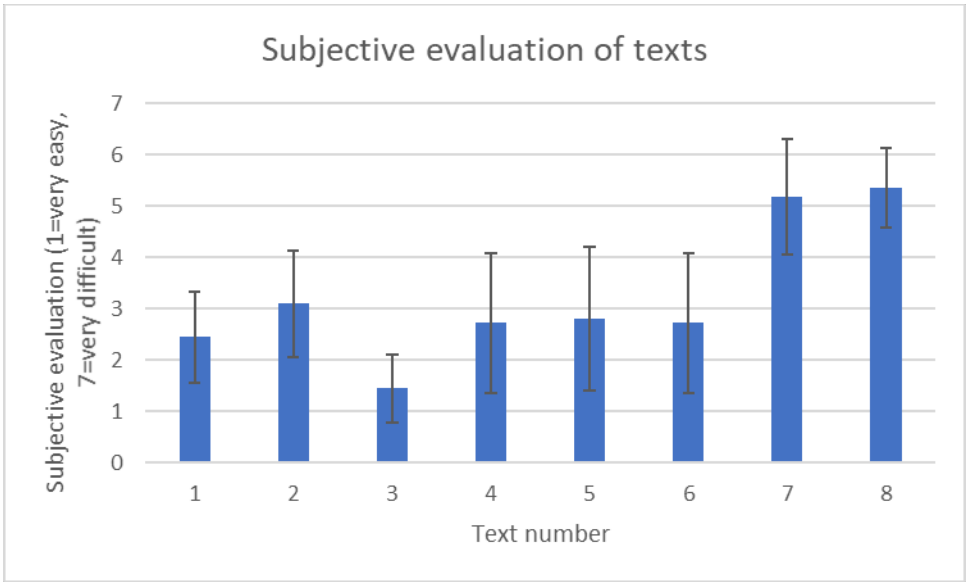

Figure 4. Histograms of the average subjective evaluations of the difficulty of the texts. The error bars illustrate the standard deviation of the evaluations.

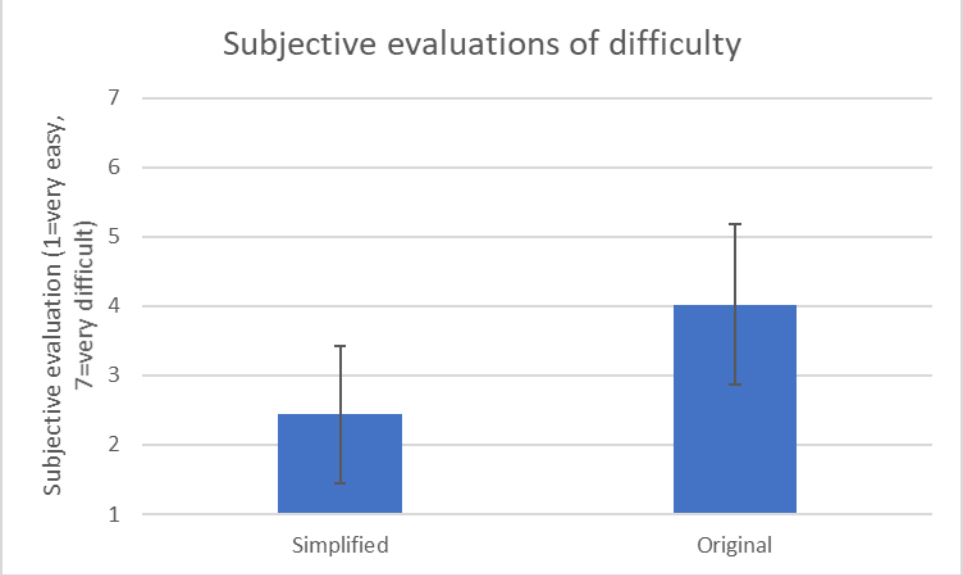

Figure 5. Histogram of the average subjective evaluations of difficulty on the two conditions. The error bars illustrate the standard deviation of the evaluations.

Qualitative data

When it comes to the section where the participants were to recall the texts, most sessions were successes. When it came to the simplified texts, text number two had two participants who did not quite remember what the text was about. They remembered parts of it but were missing some important aspects of the texts. The fifth text had one participant who failed to remember the text correctly, the seventh text had two and text eight had four.

Further on, one participant stated that text number five felt a bit jumpy. Another one said that the sixth text had some unnatural language, and one said that that text was difficult because of the lack of context. Several participants stated that there were many words they did not understand in the seventh text, and that some words were long and heavy. For text eight, one participant said that

there were some unfamiliar words, another one said that it was a lot of jural words, which made it difficult to read. Besides this, a couple of participants stated that it was, overall, difficult to give a number to represent the difficulty of the texts and several mentioned that they felt very tired in the end of the test.

Discussion

This pilot study investigated whether the difficulty of texts extracted from corpuses can be evaluated with the use of eye tracking and subjective evaluations. This was evaluated by measuring participants eye movements as they were reading texts and then asking them to recall what the text was about and rate the difficulty of it. The result revealed a significant difference in average number of fixations per word between the two conditions. There was a strong correlation between the average number of fixations per word and the subjective ratings, however, the correlation was not significant.

The reason why the correlation was not significant may have been the small number of participants. The correlation indicates that there may still be a connection between the ratings and the eye movements and with a bigger number of participants, the correlation may be more evident. If there is a strong correlation between the subjective ratings and the eye tracker, one may conclude that the eye tracker is not necessary, as one can just use subjective ratings instead. However, it may still be very useful. Since some participants worded that it was difficult to put a number on the texts' difficulties, eye tracking may be a good way to get an objective number. Using eye tracker will also overcome the difficulty of people having different scales for their ratings. Some people may often rate high and others may more often rate low even though they find the text equally difficult.

However, this does not mean that subjective ratings and comments are unnecessary. By asking the participants questions you can get at deeper understanding in where the problem lies; is the text complicated because of the words, grammar or lack of context? A recommendation for future studies is to ask specific questions about the texts and what they thought was difficult. This was not done in current study, leading to only having extra information from those participants who chose to talk more about the texts.

Another tip is to have specific limits for when the participant is considered to have remembered the text or not. In the current study, the participants were just asked to retell what the text was about, and when it was evident that he or she did not remember the text this was recorded. However, it was sometimes difficult to estimate how much the participants remembered. From the start, the question was mostly there to make sure that the participants read the text carefully and not just skimmed it through or zooned out as they were reading. After a while I figure out that it was another way to see how difficult the text was; an easy text should also be easier to remember.

To continue, number of regressive fixations and the fixation duration showed no significance even though these measures are supposed to indicate difficulty. It is not certain whether this is because the conditions were similar in difficulty, because the difficulty level can not be seen in these measures or because there were not enough participants in the study. By just looking at the scan paths there do seem to be a difference in fixation duration between the easy texts and those the participants considered to be easy, so further studies are needed to investigate this.

When looking at the result, there seem to be a big difference in the difficulty level among the original texts. Some of them were considered difficult and some of them were considered as easy as the simplified texts. This may be a problem when it comes down to using the corpuses for simplifications.

If there is no difference in difficulty between the corpuses they cannot be used successfully for text simplification. The result showed that there was a significant difference between the original and the simplified text but looking at the numbers it seems like it came down to the seventh and eighth text being difficult enough to raise the average for the whole group. To see whether most text in the corpuses has the same level of difficulty, more texts will have to be tested. It was considered to include more texts in this test, but I did not want the participants to become too tired. Since they had to look straight forward as much as possible, and several reported being tired in the end of the tests, I would have needed to include a paus in the middle if I wanted to include more texts.

Another possible issue with the texts was that only texts consisted of 5-7 lines was included. This, to make sure that the evaluated difficulty was not because the text was long or short. However, many of the original texts were much longer than this, which means that the short texts may not have been representative. It might be the case that the longer texts had more difficult wording than the short ones, which should be evaluated in a future study.

During the tests, one fault in text number seven was found. A comma had, by mistake, been switched to a dot. It led to two participants stopping for a longer time at that specific place. However, it is not considered to have affected the data significantly since they only stopped there for a short while.

To think about in a future study

There are several things to think about in a future study. These are presented below.

Double spacing between the lines

Something that proved to be very important was the spacing between the lines of the text. Despite a good calibration, the equipment was not accurate enough to for single spacing. Without double spacing it was difficult to see which line the participant looked at. It may even be preferred to have more than double spacing or perhaps a bigger font on the text. This would make the analysis easier and more correct. However, having too much space between the lines might make it more difficult to read as it may not be apparent that the lines belong to one coherent text.

Use texts of different length

Make sure to investigate both long and short texts in the corpuses. This may be hard, as there are a bigger number of long texts in the corpuses with original texts than in the simplified ones, but it may be beneficial to look at. There is a big chance that the long texts are more difficult than the shorter ones. An alternative is to extract a few lines from the long texts to make them comparable to the shorter ones.

Have specific questions for each text

Specific questions about each text may make it easier to evaluate whether the participants remember the texts or not. It may also be good to have specific criteria for when the participant is considered to have remember the texts or not. This may be if they remember specific words, just the context or nothing at all.

Have a break

Since several participants felt tired after the tests it may be a good idea to have a break in the middle. This can be something that happens on the screen, such as a game, if one does not want to have to adjust the participant to the eye tracker again.

Measures to look at

Since the fixation time should be an indication of cognitive load this should be investigated. One should also look at the average number of fixations, as this showed a significant difference between the conditions in present study. Other measures to look at is number of regressions, subjective ratings and recall of the text.

Conclusion

To conclude, eye tracking and subjective ratings can be used to evaluate corpuses. By looking at average number of fixations one may see how big cognitive load the participant is under, which indicates the difficulty of the text. This pilot study revealed that there is a big difference between the level of difficulty in the corpuses made up of original texts. This means that several of the texts are probably on the same level of difficulty as the simplified texts, which may be a problem for a future simplification program. However, more tests evaluating more texts is needed to evaluate this further.

Referenser

Amadieu, F., van Gog, T., Paas, F., Tricot, A., & Mariné, C. (2009). Effects of prior knowledge and concept-map structure on disorientation, cognitive load, and learning. *Learning And Instruction*, *19*(5), 376-386. doi: 10.1016/j.learninstruc.2009.02.005

Andrzejewska, M., & Stolińska, A. (2016). Comparing the Difficulty of Tasks Using Eye Tracking Combined with Subjective and Behavioural Criteria. *Journal Of Eye Movement Research*, *9*(3), 1-16.

Atvars, A. (2017). Eye Movement Analyses for Obtaining Readability Formula for Latvian Texts for Primary School. *Procedia Computer Science*, *104*, 477-484. doi: 10.1016/j.procs.2017.01.162

Jarodzka, H., & Brand-Gruwel, S. (2017). Tracking the reading eye: towards a model of real-world reading. *Journal Of Computer Assisted Learning*, *33*(3), 193-201. doi: 10.1111/jcal.12189

Korpusar — Språkteknologi.se. (2018). Retrieved from http://sprakteknologi.se/vad-aer-sprakteknologi/lexikon/korpusar

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372-422. doi: 10.1037//0033-2909.124.3.372

Appendix

Text 1 (enkel myndighetstext)

Kooperativ hyresrätt innebär att en förening äger eller hyr en fastighet. Föreningens medlemmar hyr sina lägenheter av föreningen.

När du flyttar in betalar du en deposition till föreningen, en summa pengar. Depositionen får du tillbaka när du flyttar därifrån. Du kan alltså inte sälja din lägenhet utan du lämnar tillbaka den till föreningen när du flyttar.

Text 2 (enkel myndighetstext)

Länsstyrelsen arbetar med byggnadsvård för att bevara länets kultur-miljö.

Länsstyrelsen har hand om statliga bidrag för att vårda värdefulla kultur-historiska byggnader.

Det kallas byggnads-vårds-bidrag.

För att byggnader ska behålla sitt kultur-historiska värde är det viktigt att sköta om dem på rätt sätt.

Det är också viktigt att använda samma eller liknande material som byggnaderna är gjorda i från början.

Text 3 (enkel kommunal text)

Alla kan behöva hjälp för att klara vardagen ibland. Här kan du läsa om vilken hjälp du kan få från Lidingö stad.

Du kan även läsa om hjälp för dig som är äldre eller har en fysisk eller psykisk funktionsnedsättning. Välj i menyn till vänster, och klicka på det du vill läsa om.

Om du vill prata med oss om hjälp i vardagen kan du ringa eller skriva till oss. Vårt telefonnummer, vår adress och vår e-post står i rutan till höger.

Text 4 (enkel kommunal text)

Pontusbadet , Gammelstads badhus och Råneå badhus . Tempererade utomhuspooler För de som tycker att älv- eller havsvattnet är för kallt så finns det två mycket välbesökta tempererade utomhus-bad i Luleå. Det ena ligger på Örnäset, Aronsbadet , strax intill Örnäsets idrottplats. Arcusbadet , som är lite större och försett med vågmaskin, ligger på Arcus-området utanför Karlsvik. Friluftsbad När sommaren är riktigt varm så lockar de många naturliga badplatserna. På dessa sidor hittar du en kort beskrivning av de flesta av friluftsbaden inom Luleå kommun.

Text 5 (kommunal text original)

Vi har filmer och en del annat material om trafik och trafiksäkerhet till utlåning. Kontakta Erica Moberg (trafiksäkerhetssamordnare) på tel 0660-881 20 om du vill låna materialet eller har frågor kring det.

En dokumentär där vi möter tre unga killar som kört rattfulla men liknande förödande resultat. Filmen är en förkortad variant av en 60 minuters lång dokumentär som sändes i TV4 under vintern 2004/2005. Tid: 16 min.

Text 6 (kommunal text original)

Har du gamla foton, vykort med trelleborgsmotiv och litteratur om Trelleborg med omnejd?

Centralarkivet tar tacksamt emot gamla foton, gamla vykort med motiv från forna tiders Trelleborg. Vi kan förvara dem i vårt kommunarkiv eller kopiera av motiven.

Litteratur om Trelleborg med omnejd tas också gärna emot.

I forumet kan du ställa frågor, lämna förslag och skicka in synpunkter.

Text 7 (myndighetstext original)

Hallen är en markant urbergsknalle mitt i området Perstorps fiskedammar. På knallen växer en välbevarad naturskogsliknande hedbokskog. Höjden stupar brant ned mot Store damm på norra sidan. Överst är markvegetationen mycket sparsam på grund av det utsatta läget och den magra jordmånen. Längre ned på sluttningarna finns ett rikare fältskikt. Rikedomen på gamla träd, högstubbar och lågor är anmärkningsvärt stor och har få motsvarigheter i länet. Detta är en viktig miljö för en rad känsliga arter i den lägre floran och faunan.

Text 8 (myndighetstext original)

Målet gäller identifiering av SMP och åläggande av skyldigheter på grossistmarknaden för mobil terminering.

Länsrätten meddelade dom i målet den 23 mars 2007 vari TeliaSoneras överklagande av PTS grundbeslut avslogs. Kammarrätten har nu beslutat att inte meddela prövningstillstånd vilket innebär att länsrättens dom och PTS beslut står fast.

KR. TeliaSonera ./. PTS. Ej prövningstillstånd vad gäller SMP och skyldigheter för Telia Sonera på grossistmarkanden för mobil terminering. Mål nr 2365-07.