DELFT UNIVERSITY OF TECHNOLOGY

MASTER THESIS

Assisting Experts in Image Description for Visually Impaired People

Author:
Frank VOLLEBREGT

Supervisor:
Dr. Christoph LOFI

A thesis submitted in fulfilment of the requirements
for the degree of MSc. Computer Science

in the

Web Information Systems Group
Software Technology

Student number:	4574818	
Thesis committee:	Dr. Christoph Lofi,	TU Delft
	Dr. Jesper Cockx,	TU Delft

An electronic version of this thesis is available at
https://repository.tudelft.nl/

February 16, 2023

DELFT UNIVERSITY OF TECHNOLOGY

Abstract

Faculty of Electrical Engineering, Mathematics & Computer Science
Software Technology

MSc. Computer Science

Assisting Experts in Image Description for Visually Impaired People

by Frank VOLLEBREGT

There are an estimated 253 million blind and visually impaired people in the world. To grant them access to text publications that contain images, experts are employed to write image descriptions. There is both a societal and a legislative pressure to supply image descriptions to all new and archived publications within a number of years, yet the number of available experts is limited. The image description task for images in textual context is complex, since a fitting description combines salient elements from both the image as well as the context into a description, which also differs depending on the publisher and published medium.

Because of this complex nature, current automated systems are unable to reliably produce desirable results. Instead, this thesis focuses on developing software to assist the experts in their general image description task in order to improve their efficiency. Specifically, we use existing, commercially available automated tools to generate alternative representations of the data. To analyse the system, we develop a user interface to present all of the available data and design an experiment with a small group of experts to investigate the system's applicability and perceived usefulness. We find that such a system has great potential to assist the experts, but that it might be desirable to focus on a solution aimed towards a smaller subset of publications, so that domain-specific information sources can be exploited to improve the information quality.

Acknowledgements

The journey of writing this master thesis has been a long and challenging one for me: From wandering around figuring out which way to go with the work, to the hurdles that arise with working on such a project during a pandemic, I am proud and glad to have finally arrived at the summit of this metaphorical mountain.

A journey that has taught me a lot about the subject of this thesis, but quite possibly even more about the process and challenges of scientific research. I would like to thank my supervisor, Dr. Christoph Lofi, for his guidance and support. His positive mindset and enthusiasm while providing very insightful feedback and guidance has been a great source of motivation and inspiration for me. Additionally, I would like to thank Ioannis Petros Samiotis, who has provided me with frequent feedback and the ability to discuss my ideas and provide guidance during his time as a PhD candidate at the Web Information Systems group.

I would also like to thank everyone at the KB National Library of the Netherlands who participated and helped me during my research. Specifically I would like to thank Ted van der Togt for putting me in contact with the right people and for his guidance in the earlier stages of the project.

Moreover, I would like to thank all of the experts who participated in the experiment to evaluate my system, as well as everyone else at Dedicon that I have had the pleasure to meet and to present or discuss my work for. Specifically, I would like to thank Anne Bottenheft for her feedback, time and genuine interest in my project. It has helped me tremendously to give shape to my work.

A special thank you goes out to my family, for keeping me motivated and occasionally giving me a much needed push to keep going.

Last but not least, I would like to thank my friends for listening to my ramblings about the project and for motivating me to keep going, as well as providing feedback and inspiration in all stages of the project.

Frank VOLLEBREGT
February 2023

Contents

Chapter 1

Introduction

1.1 Context

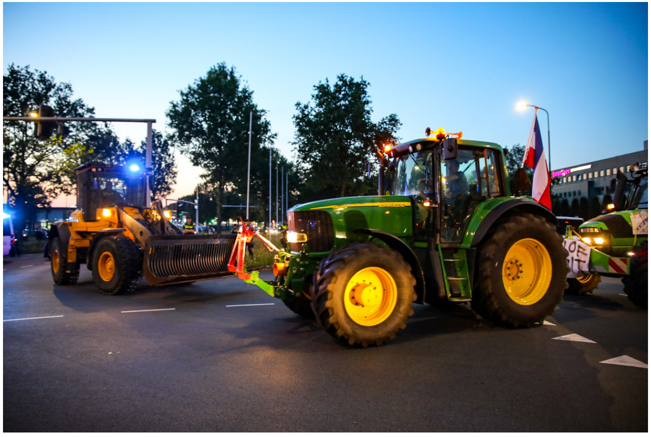
A picture is often said to be worth a thousand words, which is why the text in articles and other publications is frequently accompanied by an image to clarify or illustrate the described topic or simply to capture the readers' attention. Have a look at a partial news article with an accompanying image below, which was published in july of 2022 by Eindhovens Dagblad[1]. It was originally published in Dutch, but for the purpose of reading it here an approximate English translation is provided and part of the article is omitted. The full original text can be found in Appendix A. Some additional context, which the reader may or may not already have, is that excessive nitrogen deposition has been a problem for many years, and that the Dutch government is introducing measures to reduce this deposition[2], which many farmers went to protest against.

[1] www.ed.nl

[2] https://www.government.nl/topics/nature-and-biodiversity/the-nitrogen-strategy-and-the-transformation-of-the-rural-areas

Expert cracks down on farmers' action: "the expiration date of the farmers' protest has passed"

The farmers may want to continue protesting for weeks. But does a country like that? And does it still make sense? Arco Timmermans, professor by special appointment of Public Affairs at Leiden University, does not think so. "The expiration date of this protest has long passed."
Timmermans emphasizes that farmers are naturally sympathetic to a large part of society. Their story is clear. They provide food and want it to last. But now that the actions continue, Timmermans thinks that the shelf life of the actions has expired.
"This applies to politics in The Hague, but also to the population that is affected by the actions, for example in traffic jams." Although according to opinion polls among the latter group, there is still a lot of appreciation for the farmers, according to Timmermans this does not mean that the protest is any longer tenable. "It is very good that there is mediation between farmers and the government, so that the discussion is loosened. Now nothing happens but polarization. Everything that one camp gains is a loss for the other camp."



Farmer Protests in Apeldoorn. © Luciano De Graaf

Reading the article without any prior knowledge tells you that there have been protests, and that the professor does not believe in their longevity. Additionally, these protests have caused traffic jams and other things that affect the general population. In the image, we can see two tractors driving at a protest in Apeldoorn, with blue lights in the background. From the context, we can conclude that this picture likely shows of the unrest that is caused during some of the protests.
Now imagine a visually impaired person wants to learn the information in this same article. Fortunately, techniques like screen readers[3] are readily available to such users. Such a tool exists as a browser extension or installed program on the user's device, and commonly employs a text-to-speech tool that automatically reads the text in the article, as well as the image caption, out loud to the user. However, this leaves the image itself unaddressed.
According to the International Agency for the Prevention of Blindness (IAPB), there are an estimated 253 million blind and visually impaired people in the world, as described in Ackland, Resnikoff, and Bourne, 2017. Using the aforementioned

[3] https://www.afb.org/blindness-and-low-vision/using-technology/assistive-technology-products/screen-readers

screen readers, this group of people can consume the textual information, but the accompanying visual information is missing. As a result, they might miss a part of the article that may provide useful contextual information of this story.

1.2 Common Concepts

During the rest of this work, we will use some terms that are outlined below. While most are common in normal language, some might have a slightly different nuance in this work compared to their familiar meaning.

Visually Impaired People

Visually impaired people, in this work, are people who are unable to consume visual information like images. This group includes blind people, but also people with various visual impairments that make them unable to interpret this information, even though they might see some of it.

Image Context

In published media, an image commonly serves to complement or augment the surrounding story, which we refer to as the image context. For the purposes of this work, this is a body of text, and might occasionally also include an existing image caption from the source.

Image Description and Image Caption

It is important to know that image descriptions, in the context of this work, are different from image captions. A caption usually adds some context: in the example, it mentions where the picture was taken, and at which occasion. On the other hand, a description might, aside from adding context, also mention information about the content of the image. A description for this image could be *two tractors driving on the road. There are met by a yellow shovel truck bearing a blue light. The first tractor is carrying an upside-down Dutch flag.*. There is much more to the contents of an image description, which we investigate during the interview in section 2.1 and summarise in section 2.2.

Image Source

Images exist in different publications. The type of publication and its target audience will influence the length, complexity and process used to describe the image. In this work, we refer to the publication type in which the image is found as the image source. Examples of image sources are news articles, education textbooks or comic books.

1.3 Image Descriptions to the rescue

A solution for visually impaired people, which is already employed in practice, is to provide textual descriptions for images in publications. This way, the existing screen reader can be used to also convey the information in the image, thereby painting a more complete picture for the visually impaired person.

Aside from the ethical motivation of having the information available to as many people as possible, legislative entities are also actively working towards legislation that forces publishers to provide accessible publications: For example, in 2019 the European Union has enacted the European Accessibility Act *Directive (EU) 2019/882 of the European Parliament and of the Council of 17 April 2019 on the accessibility requirements for products and services* 2019, a directive which mandates that products and services offered from 2025 onward meet certain accessibility requirements. The World Intellectual Property Organization (WIPO) describes this in more detail, see Saez, 2020.

Other uses of image descriptions

Aside from the intended benefit of the presence of an image description, that is to help make a publication more accessible for visually impaired people, sighted users as well as computer systems can also benefit from having an image description available. Think, for instance, of certain details or objects in the image that are not immediately clear to a sighted user, which can be clarified by the description. Especially without context about these protests, it may not immediately be clear what is happening in the image. Another example of a system that can benefit from textual descriptions around images is an image retrieval system like Google Images[4] that allows users to enter a text query and retrieves relevant images. Such systems can utilise the textual context around an image, including a description if present, to determine its relevance to the user query.

How are image descriptions created

In practice, image descriptions are usually written by experts working for special libraries and adaptation services. In the Netherlands, the KB National Library[5] is responsible for supplying new and archived publications with image descriptions. They employ Dedicon[6], a foundation specialised in accessible text and images. The experts at Dedicon are the ones writing the image descriptions. The need for image descriptions in publications is specifically high as a result of legislation like the European Accessibility Act mentioned before, making the experts' work very important. Unfortunately, these experts are limited in number, and as a result they can produce only a limited number of image descriptions in a given amount of time. While it may seem obvious to outsource the image description process to an automated system, this proves to be very challenging, owing in part to the complex nature of the image description task. A single image can require a vastly different description depending on numerous different factors such as the source and the target audience. This is explained in some more detail in section 2.2. Still, attempts at defining a system to systematically judge image descriptions have been made, such as the score calculation by Nganji, Brayshaw, and Tompsett, 2013, which uses a set list of heuristics to determine whether an image description is complete. It assumes that if an aspect is present in the image, it should be represented in the description: In practice, this is not necessarily the case depending on the source and audience as mentioned before, but without such assumptions it becomes virtually impossible to reliably rate a given description. Because of the task complexity, automated methods are firstly difficult to create and secondly challenging to use

[4] https://images.google.com/

[5] https://www.kb.nl

[6] https://www.dedicon.nl

directly to perform the image description task. We highlight this some further as part of the related work in section 2.6 and later during the system design in section 4.1. To create image descriptions, we decide that for us, it is not *yet* advisable to step away from using humans, and more specifically, experts. While using non-experts to perform the image description task is possible, experts guarantee a high quality level of the resulting descriptions, a promise that is difficult to make with untrained people. In another thesis, a proposal has been made to instead make use of crowd workers instead, which we also further elaborate on in section 2.6.

Also note that since the source and experts are Dutch, we have an additional challenge when considering existing automated systems, which oftentimes do not take languages other than English into account. Guidelines or similar methods on the other hand should not defer too far from variants in different languages.

1.4 Research Objective

We determined above that given the current state of automated methods we need a system that works in conjunction with the experts to create the image descriptions. The objective of this research is to investigate methods to increase the number of image descriptions produced by the experts while maintaining the same high quality. Broadly speaking, there are two solutions to achieve this goal:

1. Employ more experts
2. Improve the efficiency of each individual expert

The first solution comes with some practical downsides such as the added cost and the lack of available experts. We thus decide to focus on the second solution for this thesis.

In order to improve the efficiency, we aim to select a part of the process to improve using software, which we suspect will benefit the overall efficiency of the image description process. To this end, we interview one such expert in chapter 2 to determine the different steps that comprise the image description process and perceived challenging aspects therein.

We determine that the step of consuming and interpreting all of the information is a good candidate step to investigate, as it causes a high mental load for the experts, where they need to read, watch and augment the information to create a fitting description for the given document in the context of the given publication.

We develop a tool, the process and choices for which are highlighted in chapter 4, that aims to help lower this mental load on the experts during the image description process, by performing some automated preprocessing steps before the document is shown to the experts. During this preprocessing step we retrieve and create alternative representations of the available information. All information, both originally present in the source and added during the preprocessing step, is integrated into a simple user interface for the experts, which we evaluate with a small group of experts working at Dedicon on behalf of the KB National Library.

The experimental design is presented in chapter 5 From the feedback provided by the experts, we learn how they experience working with such an assistive tool and whether we expect further efforts into developing tools to assist the experts are likely to be fruitful.

1.5 Research Questions

Before we can develop a tool that can effectively assist the experts in their image description process, we need to determine what this process currently looks like:

1. What is the general workflow of experts writing image descriptions for visually impaired people in the Dutch language?

   (a) How do the experts determine which images require a description?

   (b) Are there common patterns to determine what information to include in image descriptions?

   (c) What information sources do the experts use and have available to construct an image description?

   (d) What are challenging aspects and steps of the image description task for the experts?

   Informed by the experience of the interviewed expert as well as other experts online, we determine that we are going to lower the perceived mental load by assisting in the process of interpreting the available information. The following is then investigated:

2. How can we use software relying on existing techniques to assist experts in the interpretation step of the image description task for visually impaired people in the Dutch language?

   (a) What different forms (of information) does the output of commercially available automated image and text analysis techniques take? Which of these forms can work as alternative representations of the existing information and how do we account for the non-English language?

   (b) How can all of the available information, comprised of both the original information and the alternative representations, be presented in a software user interface with a high usability?

   (c) How can we, with a limited number of experts available, evaluate the software tool to investigate both its applicability on different image sources as well as how it is perceived by the experts?

   (d) Based on the findings in this thesis, what recommendations can we make about future research on expert written image descriptions for visually impaired people?

1.6 Research Hypothesis

The hypothesis of this work is as follows: When using software to preprocess and display the data and thereby support the information interpretation step of the experts in the image description workflow for visually impaired people in the Dutch language, we can lower the perceived mental load of the experts and a result, improve the efficiency of their image description process.

1.7 Research Contributions

This work contributes an analysis of the expert image description process in the Dutch language for visually impaired people, breaking it up into separate steps. Additionally it introduces a system that can be used by the experts to assist them in this image description process, which is powered by existing automated methods. The code for this system is shared online. Lastly an evaluation of this system is performed, to establish a better understanding of the experts' experiences working with such an assistive tool and gain useful insights for future development of assistive tools for these experts.

Chapter 2

Current Process & Related Work

In this chapter, we aim to answer research question 1. Specifically, we schedule an interview with one of the experts writing image descriptions for Dedicon to discuss the general image description process the experts use in more detail. From this interview, we derive a generalised expert workflow for writing image descriptions as well as some properties of a resulting image description. We then proceed by using the example document from the introduction to illustrate what the image description can look like in practice. Lastly, we identify a challenging step in this workflow, and focus on this step in order to provide an answer to the second research question.

2.1 Interview with an image description expert

2.1.1 Interview Objective and Structure

To gain a better understanding of the image description process of the experts, we want an appropriate candidate to interview. Namely, the interviewee needs to meet the following criteria:

- They need to be experienced writing image descriptions for visually impaired people
- They need to be knowledgeable about the different image sources processed

The KB National Library put us in contact with the experts at Dedicon that write the image descriptions, and we scheduled an interview with one of them to talk about the process and challenges. Ahead of time, we prepared questions that we wanted answered in order to understand the workflow and determine challenges, as per the sub-questions of research question 1 in section 1.5.
Since the interview was conducted in Dutch, subsection 2.1.2 below contains a translated version of the original interview. The Dutch transcript can be found in Appendix B.

2.1.2 Interview Results

The relevant questions and information from the interview has been selected and presented below in English.

What is the first step of the image description process, in general?

“If not yet done, it is selecting images that do or do not need a description, in accordance with the source and publisher. The ratio of time spent describing images and time spent reading the text should reflect the original source. Additionally, the final product should reflect the image of the published medium. In some cases, an image is clearly decorative, and adding a description is not useful."

How do you start when writing the image description itself?

"There are guidelines available in a document, which can and will differ by publication (type). Though usually, they have a lot in common, and we can quite easily extract common aspects for most guidelines: It answers the questions who/what, where and how. The guidelines provide guidance on the order. The length depends greatly on different factors.
To start writing, we use the guidelines and take a look at the image, read the context, which is the text around the image, and determine the answers to the who/what, where and how questions as appropriate, then write them into the description."

Is the available information from the source always sufficient to create a fitting image description?

"Usually it will be, as it is the same that is available to sighted people consuming the image, but there needs to be a balance between objectively describing the image content using only the image, and using external additional information. We do not want to give incorrect information. Imagine a case where there was an image of the Eiffel Tower in Paris, and I want to include the material of the truss structure. Unsure whether it is steel, iron or some other metal, I will then look it up online to be sure: I want to be certain in such cases that the information I give is correct and objective where possible."

What would you say makes the image description process challenging?

"You are processing and interpreting a lot of information at once while writing, which can be overwhelming. Receiving a nudge in the right direction can support the complex process in my head. Usually, you are still processing information as you are in the process of writing the description."

2.2 Properties of an Image Description

From the expert interview, we know that in practice, depending on the image source and publication, there are different guidelines for what an image description should look like. However, there are some common aspects that are present in virtually all image descriptions. If the answers are present in the image itself, the description should answer the following questions:

- **Who/What:** Which people, creatures or objects are depicted in the image? What is happening in the image?
- **Where:** Which location, or which kind of location, is depicted in the image?
- **How:** How is the depicted action performed or how are the contents presented?

Compared to an image caption, an image description goes into more detail, and concerns both the direct contents of the image, as well as the context to help explain

what is going on, especially for more complex and potentially unfamiliar actions and scenes.
Every case is different, but in general we can identify objects and actions in the image itself, and use the context to find information such as the right terms to describe (part of) a specific object, person or action that is depicted, the name of a location, or even some domain-specific jargon that should be used in the publication type at hand.
Adding to this explanation of image descriptions, general guidelines can be found online too, such as those provided by the DIAGRAM Center: *General guidelines* 2016.

2.3 The general image description workflow for experts

From the steps and questions researched during the interview with the expert in section 2.1, we derive the general workflow in Figure 2.1. In this section, we go over and explain each of the depicted actions in more detail.

Select Images to Describe

The first step in the image description process, for a given document, is to determine whether or not to describe any given image if this has not been decided yet. Sometimes publishers have already made a selection of relevant images, but oftentimes this step is also conducted by the experts. Based on the relative space used for images in the original publications, more time and words can be spent describing these elements. Decorative images do not generally get a description, as they do not add more contextual information, but instead serve only to make the article visually appealing. An example of a decorative image in an informational article[1] can be found in Figure 2.2. The article's title translates to *These are the advantages and disadvantages of paying with PayPal*. The included caption says *Image for illustration © Getty Images*. In some rare cases, decorative images may get a description too, such as when the publication is a lifestyle or design magazine, where the image adds to the feeling that is being conveyed.

Interpret the Available Information

To write the description, the expert has to look at the contents of the image and read (part of) the surrounding text, in order to be able to properly describe the image. From this information, the expert can select salient elements to incorporate into the image description. Depending on the publication, there may be a very small or very large amount of context to consider.

Search and Add Information

Depending on the requirements of the description and the complexity of the depicted concepts, it may be required to search for additional information and take this into consideration too. This depends greatly on the publication type and its properties.
To get back to the Eiffel Tower example given by the expert during the interview, an image such as Figure 2.3 may have been shown, where the expert wanted to name the material it was constructed out of.

[1] https://www.ad.nl/tech/dit-zijn-de-voor-en-nadelen-van-betalen-met-paypal abe00f6d/

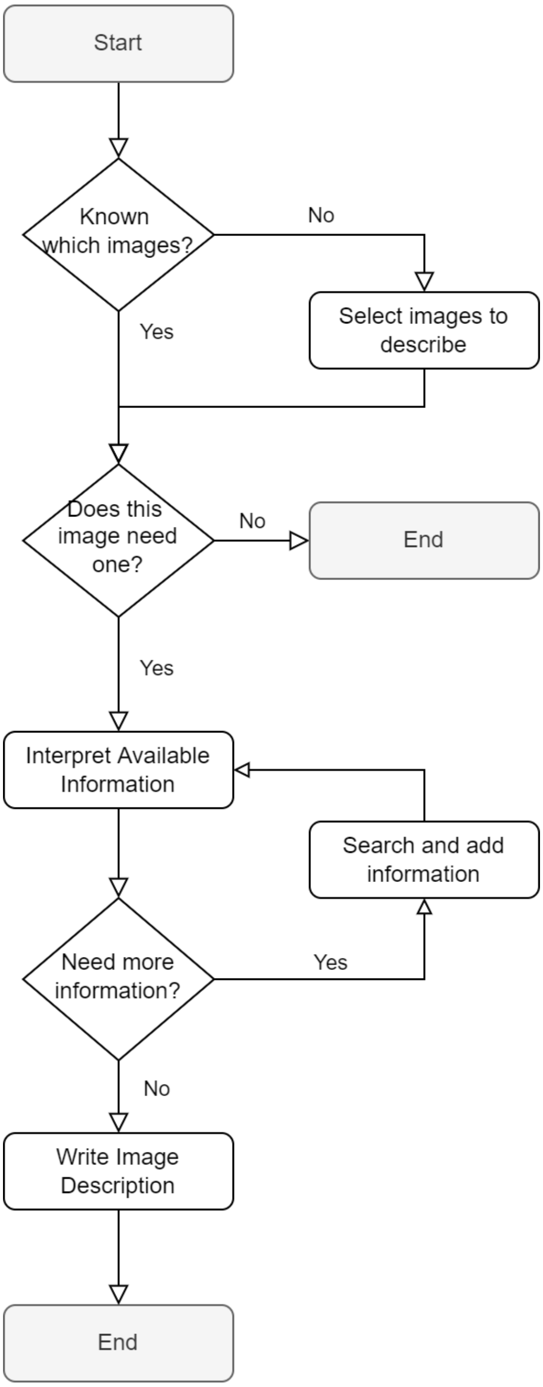
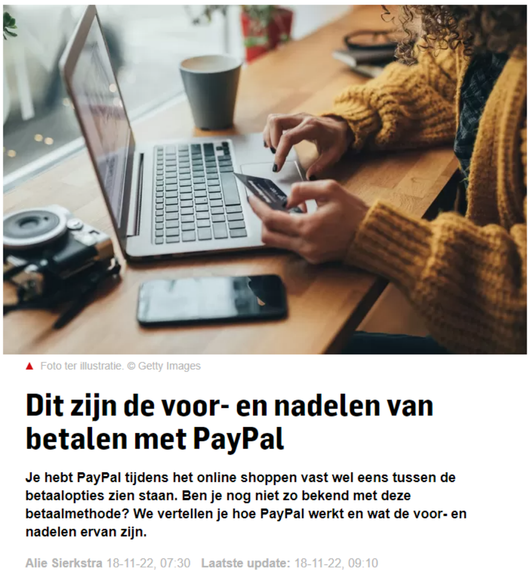

Start
Known which images?
No
Select images to describe
Yes
Does this image need one?
No
End
Yes
Interpret Available Information
Search and add information
Need more information?
Yes
No
Write Image Description
End

FIGURE 2.1: The flow of the expert image description process



▲ Foto ter illustratie. © Getty Images

Dit zijn de voor- en nadelen van betalen met PayPal

Je hebt PayPal tijdens het online shoppen vast wel eens tussen de betaalopties zien staan. Ben je nog niet zo bekend met deze betaalmethode? We vertellen je hoe PayPal werkt en wat de voor- en nadelen ervan zijn.

Alie Sierkstra 18-11-22, 07:30 **Laatste update:** 18-11-22, 09:10

FIGURE 2.2: Example of a decorative image in a Dutch article

Depending on the publication and target audience, this level of detail may be desirable. As it turns out, the Eiffel Tower is constructed out of puddle iron. Such a detail could easily be disregarded in a book for children about the tower, might be interesting to know in a travel guide, and is very important in an education book for civil engineers.

Write Image Description

After having gathered and extracted the required information, the final description can be written. In reality, these last three steps are not as distinct, since information is interpreted and supplemented while writing the description. Since they are separate steps each representing a different task, we distinguish between them in this workflow, so that we may select one of these steps as a candidate for improving the experts' efficiency.

2.4 Applying the expert workflow

Consider again the farmer protest example given in the introduction. We use the workflow derived from the interview to see what goes into creating an image description for this example.

- **Select Images To Describe**: First we determine that this image needs a description. The image provides an example of the execution of one of the farmers' protests, that has led to the reduced acceptance of their cause, according to professor Timmermans. By providing this extra context to the article, we determine that describing the image can add value for visually impaired users. Moreover, since the image constitutes a relatively large portion of the article, simply discarding it is likely not desirable.
- **Search and add information**: The tractors that are driving into the shot from the right side are clearly farmers. However, it is not immediately clear what role the yellow shovel truck on the left plays: Is it supporting the police, as

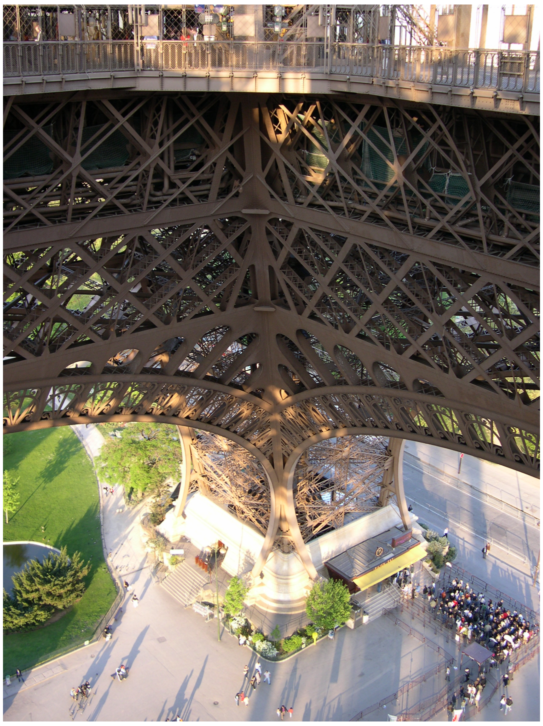


FIGURE 2.3: Example of an Eiffel Tower image as described by the expert

hinted to by the blue light that it bears, or is it part of the farmers' convoy? Looking this up online, it turns out that such a yellow shovel truck was indeed used by the police in Apeldoorn to remain in control of the protests.

- **Interpret information and write Image Description**: We can combine the image contents and the caption, our additional information and the context text to write a full description, ensuring to answer the what/who, where and how questions. It could end up something like *Two green tractors driving on the road during the farmers' protests in Apeldoorn. The first one is bearing an upside-down Dutch flag. There are police vehicles in the background. The tractors are met by a large yellow shovel truck, which is used by the police to prevent the situation from escalating*. This captures both contents of the image such as the tractors and shovel truck, as well as its context such as the fact that it was taken in Apeldoorn during the protests, and that the shovel truck was used by the police.

If the image would be very small on the page, or there would be a lot of text, the description may be a bit shorter. If the image would be presented prominently on the front page of a newspaper, the description would possibly more extensive, perhaps going into more detail on the different objects, the time of day or the location.

2.5 Determining a challenging aspect

Having gained a better understanding of the different steps in the image description workflow and how it can be applied, we can select a challenging step in this process, so we can focus our efforts towards improving this step. The interviewed expert already mentioned that taking in and processing the information while writing the description can be a mentally overwhelming task. A similar statement is provided by Valerie Morrison from the Center for Inclusive Design at The Georgia Institute of Technology during a webinar about image descriptions organised by the DAISY Consortium Orme, Morrison, and Alexander, 2020:

> "My brain is so busy trying to translate that visual information into language, that whatever cognitive process that is trying to edit while I write, is out [of] the window."

It is clear that for the human brain, processing and interpreting the existing textual and visual information to extract salient aspects and forging these together into a fitting image description for the publication can be a mentally daunting task. We thus select this step (interpretation of information) as a fitting candidate, and make our objective to support the experts during this interpretation step in order to improve their efficiency.

2.6 Related Work

Related work in image processing

Existing image processing techniques have gotten increasingly advanced over the past decade. However, most of these techniques are inadequate for directly generating or retrieving image descriptions for visually impaired people, mainly because they lack the ability to take the context into consideration. Bernardi et al., 2016 provide an overview of developments in automated image descriptions, and although the methods addressed in this work are not specifically geared towards visually impaired people, the authors do highlight works that aim to describe the contents of images. The paper distinguishes between two variants: first is the generative approach, whereby visual information, extracted from the image itself, is interpreted and used to build the description. Second is the retrieval approach, in which case there is an existing set of images with descriptions, and descriptions for similar images are considered as candidates for the image at hand. There is also an extra subdivision that is more closely related to this thesis: The authors differentiate between models that use a visual space to retrieve images, where only the image itself is taken into consideration, and a multimodal space, where both the image and text are considered. Of all the papers studied that use retrieval approaches to get image descriptions, more than half use such a multimodal approach. One such approach is work by Hodosh, Young, and Hockenmaier, 2013, in which both images and descriptions are mapped to a common space, and retrieval can occur both ways: A description can be inserted to retrieve a fitting image, and an image can be inserted to retrieve a fitting description. Further developments from the work by Hodosh et al. includes work by Socher et al., 2014 which takes a similar approach but uses Dependency Tree Recursive Neural Networks. According to the authors, this allows their approach to better abstract over word order and syntactic differences that are not relevant for the task. Further derivative works include papers by Karpathy, Joulin, and Fei-Fei, 2014, Lebret, Pinheiro, and Collobert, 2014 and Ushiku et al., 2015, who explore different approaches to map both visual and textual elements to semantic knowledge.

One example of the generative approach is the work by Gupta and Mannem, 2012, who use image tags and annotations to construct a novel image description, and

address the challenges encountered in such an approach. Specifically, they highlight the issues that arise when incorrect labels are used as a source for generating descriptions: Ambiguous or incorrect descriptions may be created.
An example of the simple retrieval approach is outlined by Kuznetsova et al., 2012, which uses image retrieval to get existing descriptions for similar images and composes these to create a novel image description. A later development on this approach by Ordonez et al., 2016 highlights their process of merging the different retrieved phrases, which they evaluate experimentally using crowd workers on Amazon Mechanical Turk[2].
After the paper by Bernardi et al., many advancements have been made in the field of deep learning. For example Kinghorn, Zhang, and Shao, 2019 describe the design and development of a deep learning model which hierarchically builds the description. That is, it considers parts of the image such as objects in the background, persons in the foreground, or a building on one side. It then relates these parts together and compiles them into a final image description. A similar system called Visual Vocabulary, ViVo for short, was developed by Hu et al., 2020. It uses Transformers to generate an image description. The authors claim that unlike earlier efforts, ViVo is also able to handle previously unseen objects. Having been developed by researchers at Microsoft, it is also available online through Microsoft Azure as a commercial on-demand service.
While developing the system in chapter 4, the technique from the aforementioned ViVo paper is also considered as one of the options.
A shared challenge between the majority of these approaches is the uncertainty of the result quality. The descriptions created using the aforementioned work are oftentimes less complex than those that are written for visually impaired people by the process described earlier in this chapter, and even then they oftentimes offer insufficient or incorrect information. In other words, using an automated image description approach does not rule out the human from the image description process. Still, existing approaches offer valuable information that may be utilised by the human image annotator, instead of replacing the entire process.

Related work on writing image descriptions

There are guidelines and tools that can help people writing image descriptions, such as experts, to do it in a more structured and better way. However, even with these guidelines the step of taking in the information and producing a description is left to the annotators themselves.
For example, DIAGRAM Center, short for the Digital Image And Graphic Resources for Accessible Materials Center, provides guidelines for writing image descriptions in accessible media, in *DIAGRAM Center - Making Images Accessible* 2021. As expected, these guidelines align with the information provided by the expert during the interview in section 2.1. Specifically, the following six main guidelines are presented:

1. Context is Key: Descriptions vary by context, avoid repetition and consider how the image fits in the context

2. Consider Your Audience: The same image may get a different description depending on the audience

[2] https://www.mturk.com/

3. Be Concise: As with all writing, do not repeat information and avoid introducing new concepts in the description
4. Be Objective: Allow the reader to form their own opinions and interpretation, do not omit any content
5. General to Specific: With longer descriptions, this allows the reader to get an increasingly better understanding of the image contents if they are interested, or to get the general idea quickly.
6. Tone & Language: Keeping language consistent and picking the right words helps elevate the quality of the resulting description

Such guidelines can help experts structure and form the descriptions more effectively. Another such resource is provided by the Fondazione LIA, a non-profit foundation for accessible Italian books, in their white paper LIA, 2019.
Aside from the resources mentioned above, there is also work that investigates the user requirements for image descriptions by Hollink et al., 2004. They survey people to find out whether they prefer general, specific or abstract image descriptions, and conclude that the former is generally the most sought after descriptions. While this paper does not focus specifically on visually impaired people, it does align with the *general to specific* guideline above, that at first the general contents of the image should be described.
Another work that does focus on the case of visually impaired people is a recent thesis by Chu, 2021, which describes the requirements of a system that can be used to create image descriptions for visually impaired people. In this case, the author focuses on an approach where the users writing the descriptions are not experts, but crowd workers instead.

Related work on image descriptions in context

There are a handful of approaches that try to tie the textual context and image together, which have promising prospects for the future, but the results of which are not necessarily ready to deploy for a use case like the visually impaired people, and oftentimes they have not been designed as such either. Still, for automated tagging to create datasets and train machine learning models, they are already really useful. Examples of this are the works of Feng and Lapata, 2010 and Feng and Lapata, 2012, which use a learning-from-data approach to extract a description from the surrounding text for images in news articles. Similar is the paper by Biten et al., 2019, who advocate the use of an attention system to selectively extract information from news articles, in order to generate the description. Because the attention technique is used, the system can also deal with entities that have not been seen during training, making it more flexible and widely applicable. Once again the authors use humans to judge the results. The main contribution of this work is the GoodNews dataset, which contains 466k samples and is the largest news captioning dataset when it was published. For our work, it is unfortunately not directly applicable, since the dataset is in English, and we specifically consider the Dutch or non-English case.

Chapter 3

Research Approach & Methodology

We have established a better understanding of the approaches taken by experts while writing image descriptions for visually impaired people, and derived their general workflow in section 2.3. Additionally, in section 2.6, we have addressed other work related to this thesis and found that resources aiming to assist experts are for the most part guidelines for writing the descriptions.

From the analysis of the current process in chapter 2, it has become clear that one of the prominent issues with the manual approach used by the experts is that the it can be mentally taxing, specifically to interpret and process the available information while writing the image description itself. To assist the experts in this process, we will develop a software user interface that presents information to the expert in such a way that it helps reduce the mental load and thereby make it easier for the experts to construct an appropriate image description for visually impaired people. Reading a large body of text can easily lead to an information overload, as is already found by Maybury, 1999 in their work on automatic text summarisation. On top of that, thinking of the right words that fit with an image can be challenging, as images may contain context-specific objects and concepts that are less familiar. The approach of our research is to add a preprocessing step to the image description process, addressing these challenging aspects by creating some alternative data representations that can help make the data easier to interpret for the image description experts.

3.1 Requirement Analysis

In order to develop this system, it is important to better understand *what* should be developed. To this end, this section identifies and lists the requirements of the system.

The goal of the software tool is to present different information/representations that may aid in the image description process. For this first iteration the intention is not strictly to create the best possible tool, but rather to try different techniques and representations of the available information and see which ones (if any) best support the experts in their image description process.

One of the main constraints in this process is that the goal is to go from the already available information to the interface and its information representations without any manual steps: Rather, automated methods and techniques should be used, such that scaling up the solution at a later stage will not bring about further limitations. During usage of the software tool, it can collect some simple metadata aside from the descriptions themselves. Namely, the time spent on each task can be stored, to

potentially identify a relation between the time spent on a description and the source of the image description task.
The system must be able to:

- Display the information available from the source. This includes the image, the context text around the image and if present the existing caption of the image.
- Display automatically retrieved alternative representations which are derived from the information available from the source, where appropriate.
- Allow the user to enter an image description for the given image and store it.
- Store useful metadata that can help during the evaluation of the tool's effectiveness.

In chapter 4, we go further into the design process, expanding on each of the requirements listed above.

3.2 Methodology

The goal of this work is thus not simply to develop a tool that can optimally assist experts in the image description process for visually impaired people, but also to establish a better understanding of how we can assist the experts in this process. Put differently, the objective is not to immediately develop some perfect tool that works all the time: this would be ideal but is not realistic. Instead we develop a tool that tries some different variations so we can determine approaches and better guide future efforts of developing tools to assist experts in image description for visually impaired people. As part of this work, we thus need to address each of the following points, as per the sub-questions in research question two:

1. Identify, where appropriate, automated ways to get alternative representations of the available information. This is part of chapter 4.
2. Iteratively design and implement the software tool itself as specified in section 3.1. This is covered throughout chapter 4.
3. Design an experiment where experts' impressions of the software tool are established. This is outlined in chapter 5.
4. Analyse the results of the experiment and draw conclusions about the perceived usefulness of the tool by the experts as well as its applicability on different image sources.
5. This knowledge can then be used to infer recommendations for the design of a software tool to assist experts in image description. These last two points are covered in chapter 6 and chapter 7.

3.3 Testing Data

During the design of the system in chapter 4, we need data to test and validate different techniques and iterations of the solution. For this, we ideally want a dataset that represents different image sources to test with. The KB National

Library kindly provided some real-life data we could investigate in the form of a travel guide for tourists in Amsterdam and part of a lifestyle magazine, both offered along with image descriptions for part of the included images. This data has offered us a great starting point for investigating and deciding between different techniques to utilise. Including most of this data in this publication, however, would require some negotiation with the corresponding publishers, who still own all the rights to the publications. To work around this, we opted to find a dataset with images and surrounding text that is in Dutch. For this purpose, we found the Wikipedia-based Image Text (WIT) dataset published by Google Research in Srinivasan et al., 2021, which is available under the Creative Commons Attribution-ShareAlike 3.0 Unported license.

This dataset contains 37.6 million entries across 108 different Wikipedia languages. While the datasets main purpose is the training of machine learning models, we instead use it to fetch arbitrary articles with images manually. There is no need to use the whole dataset, for the purposes of this work, the 1% training data file is sufficient. It contains 370,373 documents, of which 12,154 are in Dutch. Additionally, we filter to use only those documents that contain a caption for the image, leaving us with 7,579 documents left. It is likely that for training a machine learning system, this would not be a sufficient amount of data, but for retrieving and running the system on some arbitrary rows, it will provide the variation desired. This variation ensures that we take all manner of different images into account, and do not inadvertently get biased by the data provided by the library, which is only composed of a couple dozen examples. These are useful to consider as they represent real data, but are expected to be less representative of all possible cases.

In addition to the Wikipedia dataset and library data, some testing was performed with images I have taken myself. When providing examples while comparing the different image analysis tools in section 4.1, we use some of our own images. In other places, where context text is required, we use the WIT dataset unless mentioned otherwise.

Chapter 4

System Design

As we derived in chapter 3, we aim to design a tool that can assist the experts in the image description task in the Dutch language for visually impaired people, to lower their mental load of extracting salient elements from the available information and therewith increase their efficiency. We determined that to this end, the tool can present the existing information along with alternative representations thereof, the latter retrieved by means of automated methods in a preprocessing step. This chapter describes the considerations and choices made to determine which information is retrieved, and how all of this information is shown to the experts in the user interface.

The tool itself is created by means of an iterative process: We start with a simple version showing only the information already present and supplement this with the additional information from the automated methods as it becomes available, changing it while taking different options into consideration. This chapter highlights the different components and information sources that are integrated in the final iteration. During this development process, we have also kept in touch with the expert we interviewed in section 2.1, to ensure that the resulting system fits in their current workflow. Do note that we do not expose the initially interviewed expert to any version of the tool before the experiment, as this could inadvertently bias the experiment outlined in chapter 5, in which this expert is also participating.

4.1 Representation of the image

It is obvious that the image as-is is shown in the interface. Aside from that, in this section we take a look at some different representations that automated methods can offer based on the image. We investigate several commercially available State-of-the-Art solutions, each of which is used by a number of distinguished companies, and delivers high quality results. Namely, we use Google Cloud Vision AI (used by The New York Times, Box, Texas A&M University), Microsoft Azure Computer Vision API (used by Prism Skylabs, PicCollage) and Imagga Image Recognition API (used by Swisscom, Plex, Fotoware). Each of these three providers offers several different services. An overview of the capabilities of each is provided in Table 4.1. Knowing the capabilities of each provider, we can determine which services can be interesting for us, and use this to decide which provider to use. The goal is to select a few of these services to use in the tool. We go through each and quickly mention what each of them does, after which we can decide whether or not it is a candidate to use in the system.

Service	Google Cloud	Microsoft Azure	Imagga
Object Detection	✓	✓	✗
Labels/Tags	✓	✓	✓
Text Detection	✓	✓	✗
Face Detection	✓	✓	✓
Image Cropping	✓	✓	✓
Colour Scheme	✓	✓	✓
Content Moderation	✓	✓	✓
Image Categorisation	✓	✓	✓
Brand/Logo Detection	✓	✓	✗
Landmark Detection	✓	✓	✗
Image Description	✗	✓	✓*

* Discontinued as of December 2022

TABLE 4.1: Overview of the capabilities of each investigated automated image processing system

Object Detection

Locates and labels objects in the image, like a *cat*, *tree* or *car*. These are an interesting option to consider, as they can the expert by pointing out objects that may be of interest to describe.

Labels/Tags

Are similar to objects in the image, except they usually describe more generally the contents of the image, instead of solely objects located therein. Examples of labels could be *Outdoor*, *City* or *Crowd*. It is also a good candidate for our system, for similar reasons as Object Detection above.

Text Detection

Locates written or printed text within the image. While it may be useful in some cases, the user of the system (the expert) is likely able to read this text in the image without great effort, if so desired. We thus disregard text detection.

Face Detection

Locates human faces in the image, and sometimes also tries to recognise well-known people from them. While it may be useful to know who is depicted, there are two caveats: Firstly, false positives may occur, where the systems incorrectly thinks it recognises a known person. This might confuse and misdirect the expert that is describing the image. Secondly, if there is indeed a well-known person in the image, chances are that this person is already mentioned in the caption or surrounding text. Because of these factors, we also disregard face detection.

Image Cropping

Suggests crops to improve the framing of objects in the image around areas of interest. While useful in its use case, it is not something we can use for our system, since the images are already included in the publication.

Colour Scheme

Detects whether an image is black-and-white, and if it is a colour image, identify dominant colours. For some image sources like lifestyle or design magazines, dominant colours may be useful to denote, and for historic images, it might be beneficial to note whether or not they are black-and-white. Altogether, however, extracting the colour scheme is expected to be not very useful outside these specific cases, and as a result we decide not to use this.

Content Moderation

Detects whether the image is likely to contain adult content of different kinds. These are usually subdivided between adult, racy and gory images.

Image Categorisation

Categorises the contents of the image, usually from a predetermined list of categories. While trying to find similar images to the image at hand it may be useful to use categorisation of this kind. Generally, the categories are not very specific, and for assisting an expert with image description they likely do not add much value.

Brand/Logo Detection

Is similar to face detection, and comes with similar disadvantages: Namely, false positives may occur, and depicted brands, if important, are likely already mentioned in the surround text. It is also not of interest to us.

Landmark Detection

Can be useful in cases where the location depicted in the image is not directly mentioned in the surrounding text. However, similar to the aforementioned services for face and brand detection it is prone to detecting false positives. We do consider landmarks as an interesting option, since *where* is one of the main factors that is usually described in an image description.

Image Description

Aims to output a natural language description for the image. It only takes in the image, so is not able to make use of the text around the image. Still, it is suspected that this description can be a useful starting point or stepping stone for the experts to base their own image description on.

4.1.1 Chosen Services

This leaves us with the following services to consider: *Object Detection, Labels/Tags, Landmark Detection, Image Description*. The first two have a similar output, in the form of key words or short key phrases. What is more, their output can overlap, so we choose one of these two to use in our system. Since the labels also pertain to the more general contents of the image, they offer a broader spectrum of information about the image. We thus decide to use the labels.

Showing detected landmarks to the expert might help in some rare cases, where the location is not yet known, and a well-known location is depicted, which the system can pick up on. However, in the majority of cases, it will either not detect any landmarks, or erroneously think that it has detected some landmarks. Given its limited applicability as a result of the above, we ultimately decide to drop the detected landmarks.

Lastly, we consider the automated image description, which is usually not very extensive or detailed. Still, it offers the experts an opportunity to avoid having to start from scratch for each description. Whether or not this is useful is discussed during the analysis of the experiment results later in this work.

To illustrate the relative difference between the three considered providers, we use five test images and retrieve the labels for each of them, using each of the three providers. The images and outputs can be seen in Figure 4.1, where we show the five highest-confidence labels for each service. These images represent a mix of some nature, animals, and people.

While the tags differ slightly between services, all perform adequately for the experiment. Since we decided to use an automated image description, we unfortunately were unable to use Google Cloud, as they do not offer such a service stand-alone. Note that Google does do image descriptions, but only for their own services (like the screen reader built into Chrome, Google's web browser[1]).

In the end, we decide to use Microsoft Azure to retrieve our image representations for this experimental version, since its descriptions also incorporate landmarks and seem, in our anecdotal experience, to be able to better describe the relation between the different objects in an image. We suspect that this is a result of the visual vocabulary technique that has been applied. The tags are also more general tags about the image contents rather than labels of specific objects therein, which seems to be the case with Imagga. It is important to know however, that these services expand and improve constantly. When developing such a tool in the future it is useful to explore the solutions and capabilities offered at the time. Using a different service that better fits with the needs and use case at hand should be the main consideration.

Translation to Dutch

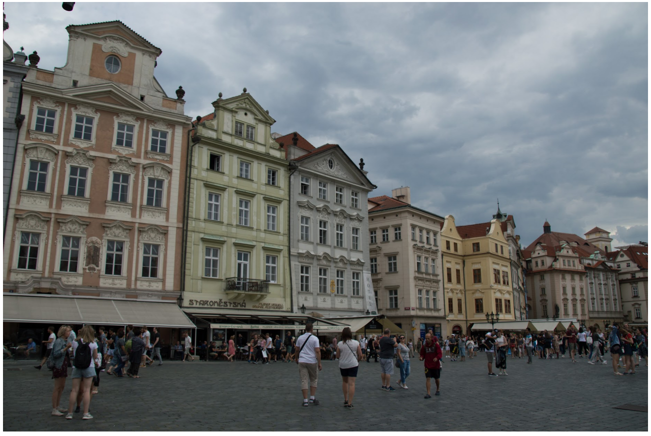
There is one shared challenge with the automated image processing methods used in this section: The output is in the English language, whereas the interface, description and the expert writing this description are all Dutch. To combat this, we need some way to automatically translate this data. For this purpose, any State-of-the-Art machine translation solution should be sufficient. For practical reasons, we chose to use Azure Translator, to keep these tools within the same platform.

4.2 Presentation of the existing image caption

When present, the existing caption can provide some direct context to the image. It differs greatly by source as to how useful this caption is, ranging anywhere between just a copyright notice or a handful of non-descriptive words to an actual description. For instance, the example we have shown in chapter 2 of the decorative image, namely Figure 2.2, uses a caption that does not provide any information about the image or context itself. This is not a big problem for that specific case since the image is decorative. However, in practice publications often

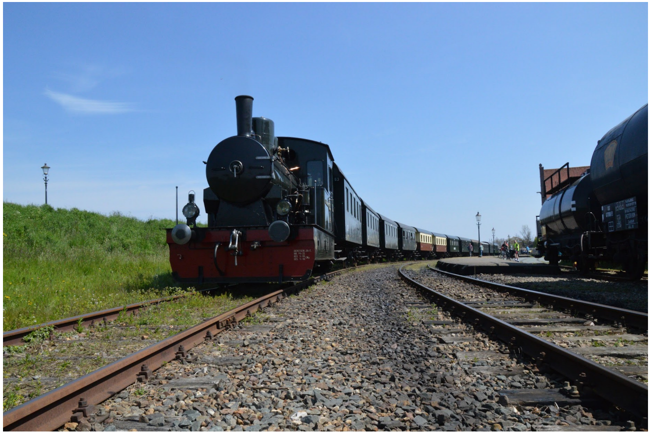
[1] https://support.google.com/chrome/answer/9311597

Google Cloud	cloud, sky, building, daytime, window
Microsoft Azure	building, outdoor, sky, city, people
Imagga	palace, residence, house, building, architecture



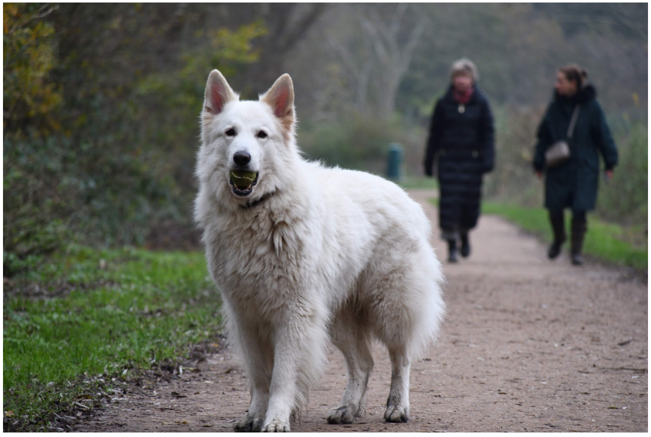
(A) Image #1: Houses

Google Cloud	train, sky, wheel, vehicle, rolling stock
Microsoft Azure	train, sky, track, outdoor, ground
Imagga	steam locomotive, locomotive, track, railroad, railway



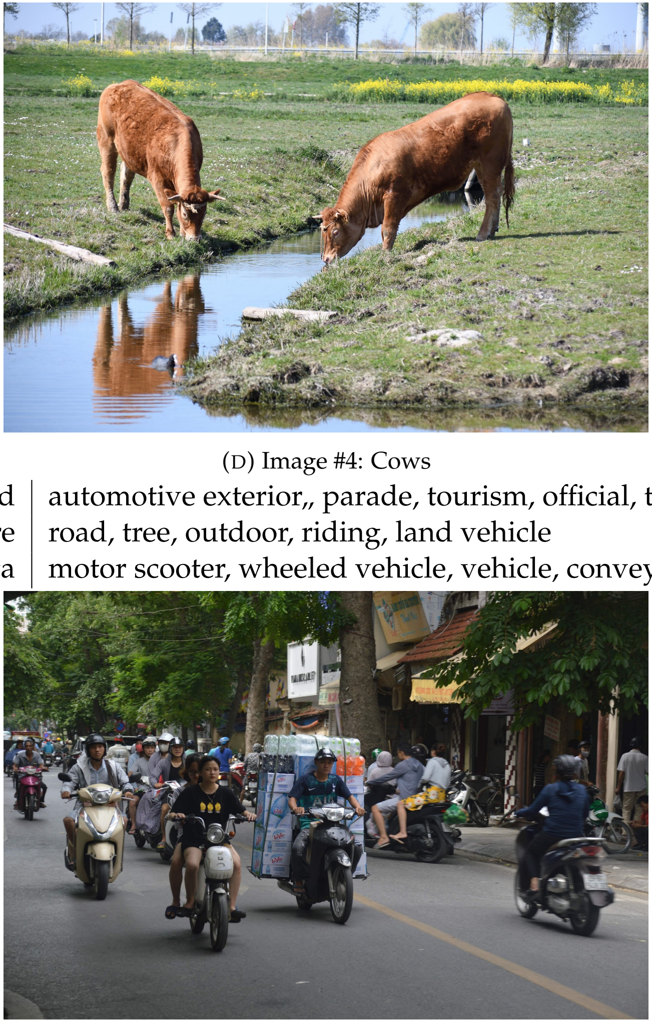
(B) Image #2: Train

Google Cloud	dog, carnivore, dog breed, companion dog, working animal
Microsoft Azure	ground, outdoor, dog, animal, carnivore
Imagga	canine, dog, domestic animal, pet, white wolf



(C) Image #3: Dog

Google Cloud	plant, water, nature, working animal, natural landscape
Microsoft Azure	grass, outdoor, animal, mammal, horse
Imagga	calf, cow, farm, grass, field



(D) Image #4: Cows

Google Cloud	automotive exterior,, parade, tourism, official, transport
Microsoft Azure	road, tree, outdoor, riding, land vehicle
Imagga	motor scooter, wheeled vehicle, vehicle, conveyance, motorcycle

(E) Image #5: Bikes

FIGURE 4.1: Example Images and their labels from the different providers

contain images that are more important and relevant to the surrounding text with either such a caption without information or without a caption altogether. This common absence of image captions was already ascertained by Takagi et al., 2003 when researching usability of the web for visually impaired people. Still, if a caption is present and descriptive it is very likely to be relevant, so it should always be included in the interface. Creating an alternative representation of the caption will likely not yield many useful results, since it is already a relatively short piece of text, making it much less mentally overwhelming to consume.

4.3 Representation of the Context Text

When building an image description, the context text is usually the main source of context. By definition, however, this is a body of text, which can be quite extensive in some cases. This can quickly lead to an information overload, as is already recognised by Maybury, 1999. When going over the challenges faced when there is a need to extract useful information from a very large body of text, a participant interviewed for by the authors mentions the following:

> “There is no time to read everything, and yet we have to make critical decisions based on whatever information is available.”

As a result of experiences like this, text summarisation techniques have been and are being developed. The book by Maybury shows an overview of early advances in this field, but recently many more techniques have been developed and perfected, such as thos by Nenkova and McKeown, 2012 and Vaswani et al., 2017. These techniques can extract sentences, phrases or keywords from a body of text using a variety of approaches.

At this point, we need to remind ourselves that the bodies of context text addressed in this work are not written in the English language, limiting the options that we can choose between. A comparative study of such multilingual methods is provided by Giarelis, Kanakaris, and Karacapilidis, 2021. In the second part of their overview study, they go into a deep learning approach to key phrase extraction, namely KeyBERT by Grootendorst, 2020. It uses a transformer-based machine learning model to extract key words. A great advantage of this approach is that a different language model can be used instead, to use it for different languages and tasks. In our case, since the text is in Dutch, we can make use of BERTje by Vries et al., 2019.

In the end, we can use the keywords extracted using KeyBERT and BERTje as an alternative representation of the context text. We go further into how these are presented to the user in the user interface in section 4.6.

4.4 Information Flow within the System

We now know where all information is gathered from. To summarise this, Figure 4.2 shows the different information sources and their job within the system.

4.5 Practical Considerations

Being able to retrieve all the data to present to the user, the next step is to determine how to store and subsequently show it to the user. The different Azure services are

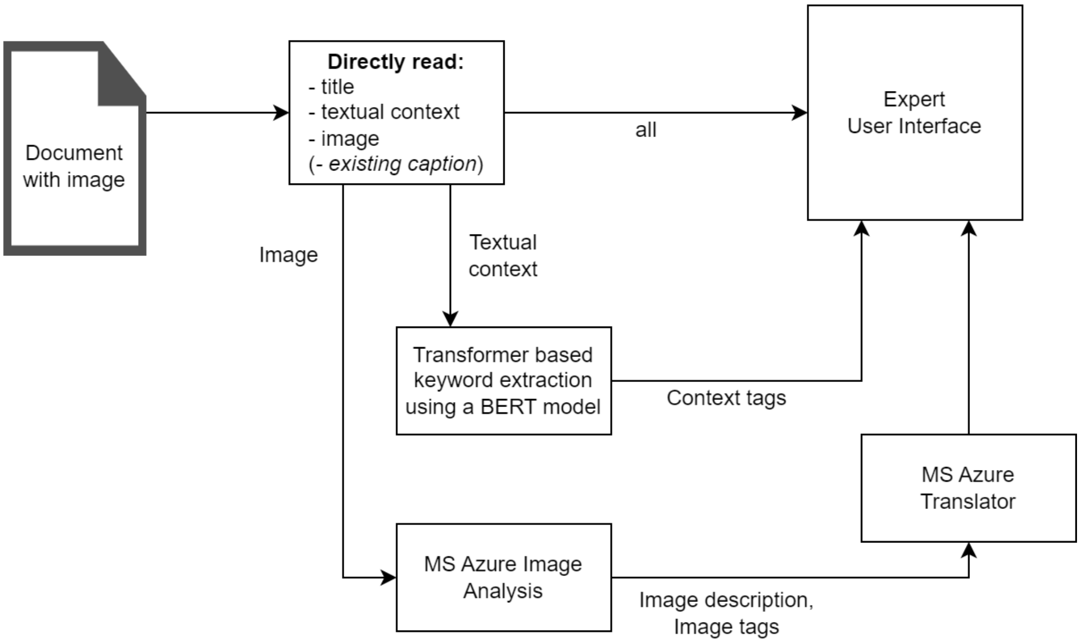


FIGURE 4.2: The flow of data throughout the preprocessing procedure

available via a REST API, and KeyBERT is available as a Python package. We thus created some Python scripts, the source code of which can be found on Github, to combine this data into a python class, which is stored in JavaScript Object Notation (JSON) objects and eventually shown to the user in the interface.

4.5.1 Hosting the Tool

The tool is made available through a web page. This way, there is no need for users to install anything. For the purposes of this work and the experiment therein, GitHub Pages[2] is used, which can host static files. While relatively simple, this approach limits the capabilities somewhat: since this server only accepts simple `GET` requests to retrieve statically hosted files, we cannot `POST` any data to the server, and we cannot host a database to store the different tasks in. However, because it is a prototype this is not much of a problem. The python scripts mentioned above append their data to a JSON file, which is loaded from some JavaScript code. This script then fills the interface that is hosted as a simple HTML file with this data, and keeps track of the submitted results while the user is working. After submitting the final description, the user is shown an end screen, and only one task remains: to collect the data entered by and collected from the user. Since no server is available to `POST` to, a ’download’ button was added to the web page, which downloads the results as a JSON file, which the participants then manually share with the researcher via Microsoft Teams.

4.5.2 Code Repository

The repository with the Python code used throughout this thesis can be found on https://github.com/frankvollebregt/imagedescriptions. This includes code to use the Azure Image Analysis, the KeyBERT keyword extraction, the source code of the

[2] https://pages.github.com/

web page that hosts the tool, and lastly the code that is used during the result analysis in chapter 6.

4.5.3 User Interface Styling

To make the interface somewhat responsive and help graphically style it, we use Bootstrap [3]. This is a simple way to make the interface adapt to the users' browser windows using their Grid system, automatically placing elements further down the page if they do not fit horizontally. For more information about the grid system and different capabilities of Bootstrap, you can read their documentation[4].

4.6 User Interface Design

Up to this point, we have discussed the information sources that are exploited as well as the technologies that are used to create the user interface. The remaining task is to design the interface itself. Aside from validating with users, we can make informed decisions about the design using well established user interface design principles, such as the ones provided by Oppermann, 2002, specifically aimed at learning systems. We highlight the different design principles from this paper which we used in subsection 4.6.1.

4.6.1 User Interface Design Principles

Not all design principles specified in the work by Oppermann, 2002 are implemented, either because they are not relevant to our system or because it was not within the scope for this prototype to include the required functionality for them.

A. **Obvious Start**
The image is placed in the centre of the screen to grab the user's attention. When describing an image, it is clear that the image is the starting point. After this, the attention is drawn by the blue headers above each major section, containing the different information sources that the user may use.

B. **Observe Conventions**
Since Bootstrap is used to back up the interface, conventional components are already used throughout the system. The only action that users can perform is to type and submit a description. The input bar and submit button for this are located in the bottom, with the submit button on the right side, thereby matching mobile chat apps such as *Whatsapp*, *Signal* and *iMessage*. This should help make the process more intuitive and familiar for most users. Additionally, the captions, both the existing one as well as the automatically generated one, are placed underneath the image, which is a familiar location.

C. **Feedback**
Within the system, there are only a handful of instances where the user gets feedback from the system. A submission simply causes the next image with context to be shown, and the only other feedback occurs when the user tries to perform one of the following actions:

[3] https://getbootstrap.com/

[4] https://getbootstrap.com/docs

- Submit without entering a description
- Copy one of the existing captions (see also subsection 4.6.6)

Attempting to perform either of those actions causes a dialog to pop up in the top right corner of the window, notifying the user that the action cannot be performed. Without any feedback, blocking these actions may lead to frustration, so it is important to provide this feedback when appropriate.

D. **Landmarks**
Each of the main information sources (the image tags, context tags and context text) are presented with a prominent blue banner, to make them obvious as areas that are similarly set up to inform the user. This should help to guide the user through the process of consulting these different sources while writing the description.

E. **Proximity**
Similar and related items should be placed together. In practice, this means that both sets of tags (image tags and context tags) are placed on the same side, to the left of the centred image. Additionally, both the existing caption as well as the computer generated caption are placed close to each other under the image. Note that the context tags and context text are conceptually more related, but the choice was made to group all tags together instead, as their information type is similar to consume and utilise.

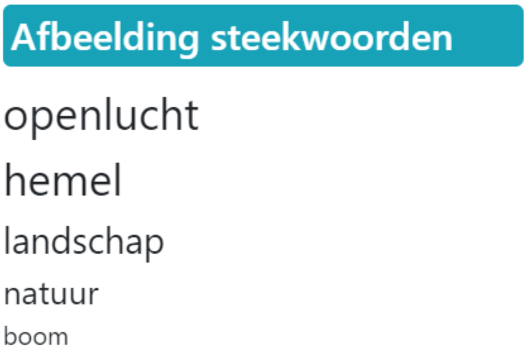
With a general idea of the placement and visual style of each component, we can describe and see an example of each individual component in the system.

4.6.2 Image Tags

Among the alternative representation of the image, we decided to use image tags with confidence scores, as discussed in section 4.1. There are multiple options for displaying these tags in the user interface. The simplest of these is simply listing them in descending order of confidence score. While this does convey the ordering, we can do better. Word clouds are a well established way to convey relative importance of different concepts. Commonly spatial word clouds, sometimes referred to as Wordle-style word clouds, are used in which all words are scattered. An example of such a word cloud, created from the text of the first chapter of this thesis, can be found in Figure 4.3. While this type of word cloud is designed to be visually appealing, it is not the most effective way to convey the information contained in it, as shown in work by Felix, Franconeri, and Bertini, 2017: They compare between this spatial Wordle style layout and a column or row layout in which the words are simply listed in a standard vertical or horizontal list respectively, and find that the column style layout is the best option when the goal is to compare and illustrate the values associated with each word. Another paper by Hearst et al., 2019 do a similar comparison, on a task where there are multiple groups of key words. The authors compare the Wordle style with the column style and find that there is a statistically significant difference in score between them, providing strong evidence that a simple column design that can convey the same information makes it less difficult to extract information.
We thus opt for using as column style representation. We vary the font size with the relative confidence scores: The highest confidence tag gets the maximum font size, the lowest confidence tag gets the minimum font size, and the others scale linearly between them.

FIGURE 4.3: Example of a Wordle-style word cloud

It is important to note that confidence scores, presented as values ranging from 0 (not confident) to 1 (very confident) from automated methods are difficult to compare between different tasks or problems, and one should avoid using their numeric values directly. On the other hand, comparing their values relative to each other within the same result set should not cause us much trouble. An example of the image tag component in the interface is shown in Figure 4.4.



Afbeelding steekwoorden

openlucht

hemel

landschap

natuur

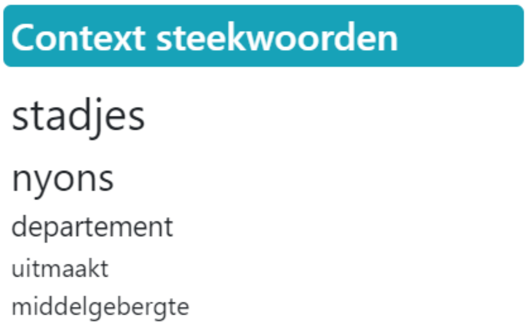
boom

FIGURE 4.4: Example of Image Tags in the interface

4.6.3 Context Tags

As an alternative representation for the context text, the context tags were chosen, as explained in section 4.3. For these, the same word cloud presentation from the image tags mentioned above in subsection 4.6.2 was chosen. By then placing these tags next to each other in accordance with design principle E in subsection 4.6.1, the user of the system can more quickly consume information in both of these components.

In Figure 4.5 an example of the context tag component in the interface is displayed.



Context steekwoorden

stadjes

nyons

departement

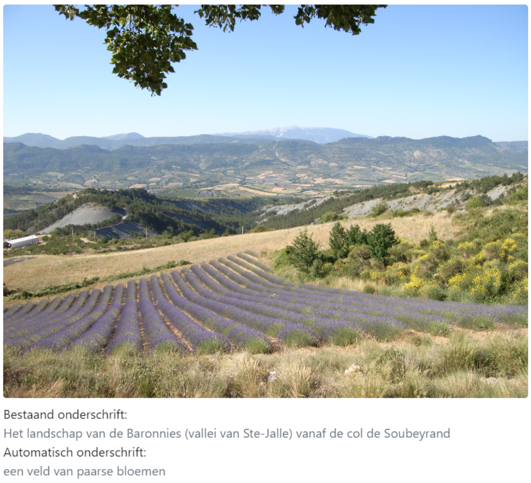
uitmaakt

middelgebergte

FIGURE 4.5: Example of Context Tags in the interface

4.6.4 Image Caption and Context Caption

The image caption that is retrieved from the automated methods as described in section 4.1 is shown directly underneath the image as explained in design principle B in subsection 4.6.1, this is the conventional and logical place for an image caption. The same goes for the context caption, for the very same reason. Additionally, these should be close together as they offer similar information, according to design principle E. An example of the image component and the two captions is shown in Figure 4.6.



Bestaand onderschrift:
Het landschap van de Baronnies (vallei van Ste-Jalle) vanaf de col de Soubeyrand
Automatisch onderschrift:
een veld van paarse bloemen

FIGURE 4.6: Example of the image with its existing and automated caption in the interface

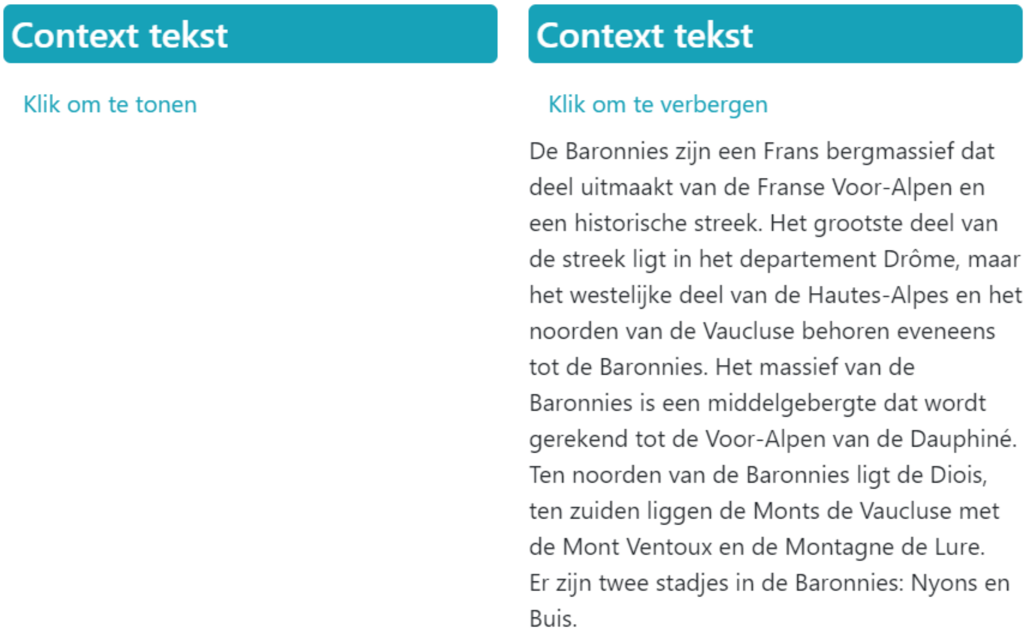
4.6.5 Context Text

One would think that designing and displaying the context text is one of the most straightforward tasks, as it is available with the source. However, instead of simply showing the text to the users immediately, the choice was made to initially hide it. The reasoning behind this is that our aim is specifically to evaluate the helpfulness of the additional information representations found throughout section 4.1 and section 4.3. If the context text is shown immediately, the experts are expected to quickly resort to their existing workflows and ignore these additional representations. On the other hand, we still want the information to be available. We consider different options for this information to show:

- Show after a certain amount of time

- Show after the user clicks a button

As different experts will take different amounts of time to consume a piece of information, it is challenging to determine an amount of time that would work well in the system. Especially so since we want to avoid giving the experts any prior experience before the experiment itself. Additionally, this significantly restricts the process, which is designed to be open-ended, where the expert is free to choose which resources to consult. In the end, using a button press to show and hide the context text is the approach that is chosen. By default, only a button is shown, which, when pressed, shows and hides the context text. An example of this component is shown in Figure 4.7.
When such a system is deployed in practice, components should obviously not be hidden like this, but for the purpose of this experiment we hope it proves useful to gain better insights on the alternative representations of the information.



De Baronnies zijn een Frans bergmassief dat deel uitmaakt van de Franse Voor-Alpen en een historische streek. Het grootste deel van de streek ligt in het departement Drôme, maar het westelijke deel van de Hautes-Alpes en het noorden van de Vaucluse behoren eveneens tot de Baronnies. Het massief van de Baronnies is een middelgebergte dat wordt gerekend tot de Voor-Alpen van de Dauphiné. Ten noorden van de Baronnies ligt de Diois, ten zuiden liggen de Monts de Vaucluse met de Mont Ventoux en de Montagne de Lure. Er zijn twee stadjes in de Baronnies: Nyons en Buis.

(A) Hidden

(B) Shown

FIGURE 4.7: Example of the context text when shown and hidden in the interface

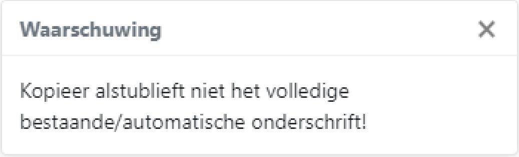
4.6.6 Cheating the System

One risk of having an assistive system is that while it may be well-designed to assist and guide the users, the users may instead end up using it as a replacement. This can occur for one of two reasons: Either the user trusts the system too much or it is simply easier to cheat than to perform the task. The second reason is referred to as the the principle of least effort: An extensively studied theory which states that people will naturally choose the path of least resistance. It was first stated by Ferrero, 1894 and later studied by Zipf, 2016. The theory is stated to apply regardless of the subjects' expertise or proficiency.
Consider for instance Compas, as described by *Predictive justice: When algorithms pervade the law - paris innovation review* 2017. This is a tool that, based on data, calculates the risk of recidivism of a person. In such a high impact case, where

years of a person's life may be influenced by the algorithm's decisions, such concerns are clearer. If a judge puts a lot of trust in the tool, they may unjustly impose a disproportionately large sentence on a suspect, whose case can be different from the cases on which the algorithm bases its judgement. The tool should thus ensure that it is used to *inform* a decision, and not to *make* the decision. In the words of Stéphane Dhonte, a prominent French lawyer:

> "These software, like the common law, will highlight the rule of judicial precedent. But they must never erase the rule written by the legislator. Otherwise, we face the risk of reversing our hierarchy of standards."

The expert tool for image descriptions is susceptible to similar risks, though fortunately with a much smaller impact. Experts may be compelled to simply copy the automatically generated description, instead of using all available sources to build a description themselves. To combat this behaviour, we have implemented simple functionality that prevents the user from directly copying the automatically generated description and shows a small pop-up notification in the top right corner when the user tries to do so. This warning is shown in Figure 4.8. The text translates to *Warning: Please do not copy the whole existing/automatic caption!*. Note that with this system in place, it is still possible to simply type the existing description, but given that this is a bigger effort, we believe that this simple measure will prevent the most obvious way to skip tasks.


Waarschuwing

Kopieer alstublieft niet het volledige bestaande/automatische onderschrift!

FIGURE 4.8: The warning shown to the user when they try to copy a description to the clipboard

Chapter 5

Experimental Design

5.1 Goal of the Experiment

After conducting the experiment and processing the results, we aim to be able to answer research question 2, as introduced in section 1.5. Specifically, we want to answer the following questions:

- Determine whether the experts indeed perceive a reduced mental load, i.e. see added value in having the alternative information representations available?
- Is there a noticeable difference between the usefulness of the different representations (context tags versus image tags versus automated caption), as well as the original data (context text and existing caption).
- Is there a noticeable difference depending on the image source? (education versus news versus informational)

Using the results, we can hopefully make recommendations and guidance for designing such a system for use in practice.

5.2 Experiment Participants

The experiment aims to validate the effectiveness of the tool when used by the experts. To get the participants, we reach out to the expert we interviewed about the current process in section 2.3. She has assembled a total of six (herself included) experts that have varying amounts of experience writing image descriptions, to participate in the experiment. Some of the experts have a specialisation, where they work on images from a certain domain. As an example, one of them specialises in education textbooks. This final specialisation is further highlighted during the discussion in chapter 6.

5.3 Experiment Data

When picking research data to form the tasks during the experiment, some choices have to be made. One of the goals of the experiment is to investigate the effectiveness of the system using some different image sources. On the one hand, this ensures that our results are more likely to be representative for the general case if they indeed turn out to be similar, or on the other hand, we may be able to identify a certain use case that is more challenging in general as an avenue for future research. To properly analyse this, however, there evidently have to be multiple samples for each image source present: A single image will inadequately

represent the source. Thus, there is a trade-off to be made between experimenting with different sources and the representation quality of each source. After all, the experts' time to participate in the experiment is limited.
To get a better understanding of which image sources are commonly processed, we inquired with people at the KB National Library about the publication sources that they process. The response was that their sources reflect the full range of *all* publications, varying from newspapers and magazines to education textbooks, comic books and beyond. However, as already remarked in section 5.2 education books are an image source that has specific experts working on them.
In the end, we opted to select 3 categories: Informational (magazine) article, news article and education book, and for each of these categories, we selected 3 articles with images, to experiment with. Below, we describe where we retrieved the articles for each category and include links to the articles themselves where possible. The full articles may be found in Appendix D, and the full source, in JavaScript Object Notation (JSON) format, is located in Appendix E.

Informational Articles

To represent the informational articles, we selected some of the articles from our initial testing data introduced in section 3.3, originating from the Dutch articles in the WIT Wikipedia dataset. Specifically, they ended up being informational articles linked to locations. The full articles and their source URLs can be found in section D.1.

News Articles

As for news articles, we selected some news articles published by the NOS[1], the Nederlandse Omroep Stichting, which is one of the major news broadcasting organisations in the Netherlands. The full articles and their source URLs can be found in section D.2.

Educational Articles

The articles to represent the educational context were retrieved from teaching modules for high school students, available online on the website of the NEMO Science Museum in Amsterdam[2]. The full documents can be found in section D.3.

5.4 Experiment Variables

It is clear that the process and results are highly complex, and dependent on the individual experts. Still, to find useful results, we outline the variables that are researched in this experiment, as well as those that we control to ensure consistency in the results of the experiment.

5.4.1 Independent Variables

During the experiment, we will study the effects of changing the type of image source. As outlined above in section 5.3, this source is varied to be an informational article, news article and educational article.

[1] https://www.nos.nl
[2] https://www.nemosciencemuseum.nl

5.4.2 Dependent Variables

While varying the image source as mentioned above, we investigate the quality of the interface as a whole, as well as the quality of the different information sources in the interface. Specifically, we gauge whether the participants used each source, and whether they thought its information was useful. We can then see if these differ depending on the type of image source. Resulting from this, we aim to derive whether the system has the ability to help reduce the perceived mental load on the experts during the image description process. Moreover, we observe the time spent on a task depending on the image source, to determine whether certain task types are likely to be more challenging in general. Given the small number of research participants, we are limited in our analysis. This is further expanded on in section 7.3.

5.4.3 Control Variables

The interface and its layout remain constant for all tasks and all experts. The choice to keep this constant was consciously made: Initially, we considered varying the user interface to determine which version, out of multiple, was the best to use, or alternatively to have part of the participants work with the new system, and let a control group use the currently used manual methodology. However, the extra independent variables would only make the analysis process extra complex and less accurate, since the image description process is already highly complex and we are working with a limited number of experts who only have a limited amount of time to describe a limited number of images.

5.5 Experiment Outline

5.5.1 Overview

Figure 5.1 shows the general outline of the experiment. Each of these steps is highlighted further in the rest of this section.

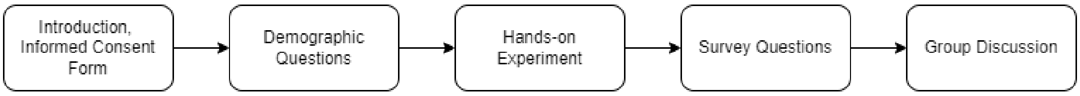


FIGURE 5.1: Outline of the Experiment

5.5.2 Informed Consent Form

To prevent any undue harm, the experiment was submitted for approval by the Delft University of Technology Human Research Ethics Committee. This required a data management plan as well as an informed consent form. As no sensitive personal data like religion or ethnicity or personally identifiable data like a full name or address are recorded, this consent form is relatively straightforward. A blank copy of this consent form can be found in Appendix C. In the results in chapter 6, we mention the experts by a number instead of a name, to continue to prevent using personal data.

5.5.3 Demographic Questions

Before the hands on part of the experiment, we ask three general questions about the experts' demographic. These are outlined below, together with some justification on why these particular questions have been selected.

1. *What is your native language?*
   We ask this question more as a reassurance rather than an extra variable to study. The chances of someone professionally working with accessibility for visually impaired people by means of Dutch spoken or written descriptions to be a non-native Dutch speaker are pretty slim. Still, if it is the case, it is important to take this into account, as it may affect the experience and use of the system for this person.

2. *How many years of experience do you have working with accessible media for visually impaired people?*
   Each of the experts is experienced with accessible media for visually impaired people. The amount of experience varies, however, and this is taken into account as one of the variables to research when looking at the experiment results.

3. *How many years of experience do you have writing and working with image descriptions for visually impaired people?*
   Of the experiment participants, most will have worked with accessible media for a varying amount of time, but not all of them will have professionally written image descriptions at all, or for shorter amounts of time. This experience or lack thereof is also expected to have an effect on the results, and is thus asked from the participants.

5.5.4 Hands-on part

The next part is a hands-on experiment with the experts. During this experiment, the experts are tasked with describing the images highlighted in section 5.3. The experts are instructed that the focus is on the description process rather than the description itself. Because of this, the descriptions should remain relatively short: one or two sentences should suffice. Additionally, the experts are told to keep in mind whether or not they used the different information sources in the interface, such as the provided captions and tags. Like the majority of the experiment, this part is individual, meaning each expert writes their own description for each of the tasks, without the ability to discuss amongst themselves.

5.5.5 Survey

After the hands-on part of the experiment, each expert is asked to individually fill out a survey with different questions. The first part of this survey is an existing, standardised survey called PSSUQ, or Post-Study System Usability Questionnaire. We add this questionnaire because we want to make sure that the interface is effective, and does not limit the ability of the system to perform its task. That is, so we will not have a case where the information is of a high quality but the poor interface still leads to a bad outcome. Additionally, this survey provides a way to compare possible future iterations of the software to this original variant.
We added the second part of the survey to address the specific components of the system, namely the different information sources individually.

Post-Study System Usability Questionnaire

To gain some insight into the general usability of the interface and its stronger and weaker points, we included the Post-Study System Usability Questionnaire, which has gone through multiple iterations: Lewis, 1992, Lewis, 2002 and Sauro and Lewis, 2016. The results of this questionnaire allow us to compare our outcomes, on a relative scale, to other systems, and identify potential points of improvement. Moreover, future iterations can be compared more directly to gain insights in the effectiveness of certain enhancements. We use version 3 of the PSSUQ, which is the most recent version. It contains 16 statements in 3 categories, which can be rated on a seven-point Likert scale ranging from 1 (*Strongly agree*) to 7 (*Strongly disagree*), and a *Not Applicable* option. The PSSUQ produces 4 scores: an overall score, and the System Quality, Information Quality and Interface Quality. Each can be compared between participants, or a specific score over all participants. The results can be found in chapter 6.

Like all quantifiable data, such questionnaires work best with large sample sizes. However, PSSUQ is shown to also work well for smaller experiments with fewer than 15 participants. Specifically, Tullis and Stetson, 2004 found that in 90% of the cases a group size of 12 yielded the same results as a larger group. Still, when analysing the data, we want to be careful not to base strong claims on very scarce amounts of data.

Component Satisfaction Survey

As an extension to the more general PSSUQ questionnaire, we added 14 statements of our own, which are specific to the system itself. To stay consistent, we stick with the same seven-point scale. The statements are listed in Table 5.1. Questions 1-4 and 6-11 are are all pairs of questions related to the usefulness and usage of the different components. The aim of question 5 is to inquire whether the user pressed the button to reveal the full context text. The goal of question 12 is to verify whether the system as a whole provided enough information to write descriptions. Questions 13 and 14 are included to investigate whether there was sufficient information to write a description using respectively the image tags and automated caption, versus the context tags and existing caption.

In the end, we hope to use these results to perform some comparisons between the experts. Note that there are no questions related to each of the specific image sources. Thoughts on these were consciously left for the group discussion outlined below in subsection 5.5.6, because the influence of each specific category for each specific expert on each specific component would yield results with little value for such a small number of experts. This survey thus focuses on the different components in the first place, and how the experts used these across all tasks.

5.5.6 Group Discussion

After all participants have completed the hands-on part and the survey, everyone will reconvene and a group discussion will be started. The goal of this discussion is twofold: In the first place, to discuss the specific software tool used during the experiments, the answers in the survey and thoughts on the different information sources and how they affected the process. The main goal of this first part is to answer research question 2.

After this is clear, we move on to the second part of the group discussion, where we hypothesise about the potential of such a system in the experts' day-to-day lives,

1	The image tags (top left) provided in the interface are useful.
2	I always used the image tags (top left) while writing descriptions.
3	The context tags (bottom left) provided in the interface are useful.
4	I always used the context tags (bottom left) while writing descriptions.
5	I always clicked to reveal the full context text (right).
6	The context text (right) provided in the interface is useful.
7	I always used the context text (right) while writing descriptions.
8	The existing caption from the source (under the image) is useful.
9	I always used the existing caption from the source (under the image) while writing descriptions.
10	The automated image description (bottom one under the image) provided in the interface is useful.
11	I used the automated image description (bottom one under the image) while writing descriptions.
12	The information provided by the system was sufficient to write the description
13	I can write descriptions with only the image tags and automatic caption.
14	I can write descriptions with only the context tags and existing caption

TABLE 5.1: Statements of the component satisfaction survey

how it should work, what would be different about it compared to this prototype. As discussion between the experts is encouraged, no strict interview-like structure is applied, and only these broad topics are planned. The results of this discussion can help us to determine whether the system achieves its goal and to shape the next iteration of this prototype. With the prototype as a reference, the experts are hopefully better able to determine what they like and dislike, and more easily form opinions to fuel this discussion.

Chapter 6

Results & Discussion

The results of the experiment can be found in this chapter. Some are presented for the whole group, whereas others are presented for each individual expert, depending on the analysis that we perform on the results.

6.1 Demographic results

The demographic questions introduced in subsection 5.5.3 yielded the results shown in Table 6.1, where the participants are labelled #1 through #6. Throughout the rest of this chapter, we might be able to relate back to the amount of experience of an expert, if we encounter varying or outlying results. As can be seen in the table, all participants had Dutch as their native language, meaning that this factor does not need to be further taken into account. In terms of experience, they can all be considered experienced working with accessible media. Note that participant #2 is the only one without any experience working directly with visually impaired people.

Question:	**#1**	**#2**	**#3**	**#4**	**#5**	**#6**
Native language	Dutch	Dutch	Dutch	Dutch	Dutch	Dutch
Exp. accessible media	32 yrs	25 yrs	15 yrs	14 yrs	14 yrs	6 yrs
Exp. visually impaired	20 yrs	0 yrs	10 yrs	14 yrs	14 yrs	6 yrs

TABLE 6.1: Answers of the six particpants to the demographic questions

6.2 Hands-on statistics

During the hands-on experiment, some statistics on the time spent on each task was collected. In Figure 6.1, we show for each of the 9 tasks, how much time the experts spent on it. Each bar is split up for each of the individual experts. In this figure, tasks 1-3 are informational articles, tasks 4-6 are educational articles, and lastly tasks 7-9 are news articles.

It is expected that there is a downward trend throughout the experiment, as the experts get more familiar with the interface and how to use it. The very first task took notably more time than all others, which lies within this line of expectation as this is also the very first time the experts are exposed to the tool.

For all the other tasks, there is generally a downward trend. However, there is a spike at task number four. This is the first task that switches the image source to the educational article. Moreover, instead of a photograph, the image included here depicts a schematic drawing, in this case of a hypothetical crime scene, which we

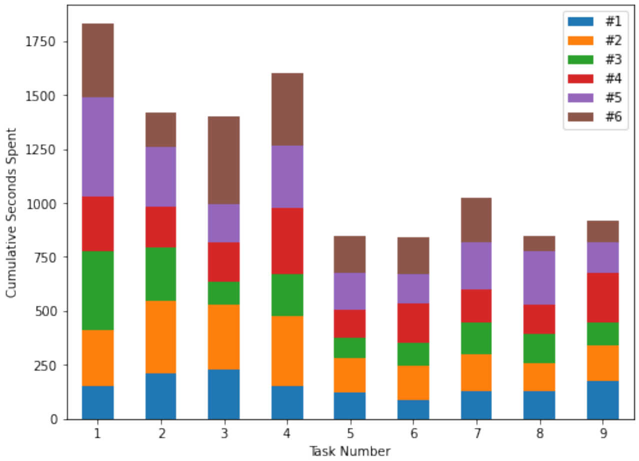


FIGURE 6.1: Plot showing cumulative time spent on each of the hands-on tasks

suspect may be a more difficult task to perform. The same is true for task number 7 to a lesser extent: The cumulative time bumps up slightly, before dropping again at task 8, to a level comparable to tasks 5 and 6 before it.

All in all, there is no significant difference between the three different image sources that is immediately clear from the data: While the first 3 tasks took longer than subsequent ones, it is difficult to tell whether this is caused by these tasks being more complex, or simply because they were the first tasks that the experts encountered. Given more available experts for the experiment, more could be learned about this by varying the order in which the tasks are presented. With only the six participants, however, splitting them up even further would hardly have enabled us to make any strong claims, as the variance of each individual expert would have too much impact.

Each of the experts also had the ability to click on the button to reveal the context text, as mentioned in subsection 4.6.5. The time spent in the task before revealing this has also been recorded, and is shown in Figure 6.2.

There are a number of interesting results to note here, some of which we try to explain. Firstly, it is clear to see that for some tasks, only two out of the six experts actually did reveal the context: This is encouraging, since apparently the context required to write an image description in these cases can be retrieved from the tags and captions alone. However, in the figure, participant #6 has not clicked to reveal the context at all. This happened due to a bug in the user interface page, which would cause it to show *hide context text*, even while it was already hidden, leading this expert to believe that there was no context for the tasks. Regrettably, we only caught this after the hands-on part had already finished. In terms of trends, we can see that most experts did spend some time taking in the information from the sources already presented initially, even if they eventually decided to reveal the complete context. This is encouraging, as it shows that there likely was some useful information in the components that were shown.

6.3 PSSUQ Survey Results

In this section, the survey results of the participants are highlighted, when averaged over all participants. The means of each score are shown in Table 6.2, and

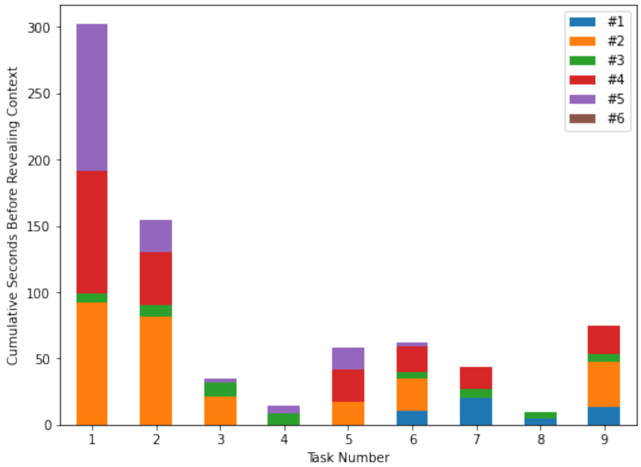


FIGURE 6.2: Plot showing cumulative time spent before revealing the context text

Figure 6.3 shows these scores when mapped as normal distributions. Note that in the whole survey, lower scores are better. While absolute results should not be compared between different studies, as explained by Sauro and Lewis, 2016, relative results may be used for comparing the different image categories. In such PSSUQ scores, a relative difference between each of the three categories is however expected to be present, according to research. In one of the PSSUQ papers, namely by Lewis, 2002, it is noted that they found a mean score of 2.8 for System Usefulness, 3.0 for Information Quality and 2.5 for Interface Quality. To reiterate, these scores should not be compared to ours in an absolute sense, but relatively speaking the expectation is that the Interface Quality has the lowest (best) score, followed by the System Usefulness, and lastly the Information quality has the highest (worst) score. In our case, the Interface Quality is rated relatively poorly. As can be seen in Figure 6.3, though, this category has the highest variance too. Additionally, the mean rating still lies somewhere in the range between *neutral* and *somewhat agree*, so there is no reason to be very concerned about the whole interface acting as a limiting factor to our system's usability.

Quite the contrary, since the System Usefulness was rated exceptionally well compared to the two other categories. All in all, the general rating of the system was positive, and while the Interface Quality seems to be lagging behind the System Usefulness and Information Quality, the System Usefulness received an excellent score, and we can safely say that the overall usability of the system, after using it for about half an hour, was rated positively.

Category	**Score Mean**
System Usefulness	2.1667
Information Quality	3.1190
Interface Quality	3.2778
Overall Score	**2.7917**

TABLE 6.2: Mean PSSUQ (sub-)scores over all participants

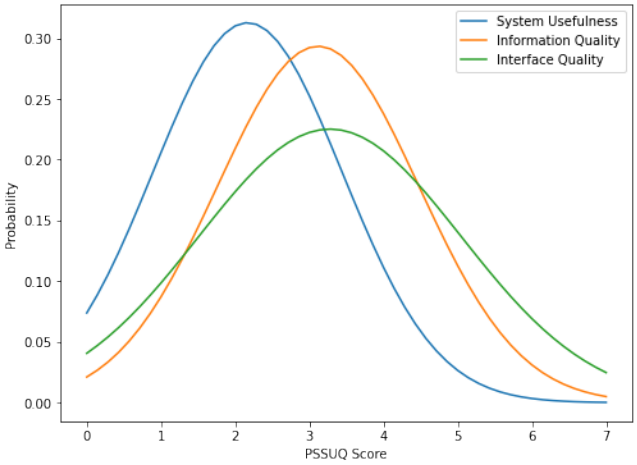


FIGURE 6.3: PSSUQ score distribution over all participants, split by sub-score

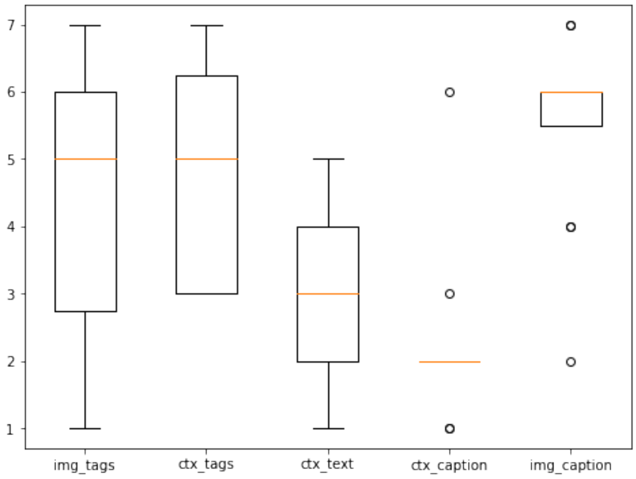


FIGURE 6.4: Box Plots visualising the component satisfaction results

6.4 Component Satisfaction Survey

The second part of the survey was concerned with the usage and satisfaction with the different components showing information in the interface. For each of the five components, namely the image tags, context tags, context text, context caption and automated caption, there were two questions, as already highlighted in Table 5.1. As a result, we only have 12 data points for each component. While this is not a lot, it does provide us with a ground to briefly compare the different components in a relative manner: We show these numbers in box plots in Figure 6.4.
Remember again that in the survey, as with the original PSSUQ, a lower score indicates a positive sentiment. It is immediately clear that the context text and context caption, together comprising the context included with the source, generally receive the most positive rating, indicating that they remain vital when writing a description, as they reliably provide the information.
Both tag lists have a lot of variation in their ratings, indicating that some of the experts considered them to be useful, whereas others did not. While we do not know exactly why this variation exists, we suspect that it a result of these information sources being very useful in some cases, and not useful at all in other

cases. During the group discussion addressed below in section 6.5 we also discussed these different components, and tried to find out more from the experts themselves.
In the end, because of the small number of participants, it is difficult to draw any reliable conclusions from the data of this small survey, but it does provide us with some intuitions on what to focus on when analysing the results of the group discussion.

6.5 Group Discussion

During the discussion the experts where given some points to discuss specifically. Additionally, they were encouraged to respond to each other and discuss amongst themselves.

6.5.1 Remarks about the interface

During the discussion, some practical feedback was provided regarding the program itself, separate from the content presented therein. Overall, the experts shared a positive sentiment towards having such a program available, and found that the system was clear and straightforward to use. This conforms to the results we found from our PSSUQ survey.
There was some practical advice, that would need to be implemented for a potential next iteration, which is listed below in no particular order (in bold), followed by some notes from us about the validity and severity of the presented issue.

- **More flexible scaling**: Two of the experts mentioned that when using the system, they sometimes had issues with the content not fitting on the page. While we took this into account by testing on Full HD (1920x1080px) as well as QHD (2560x1440px) resolutions, we suspect that some of the experts were running the tool in a smaller browser window. A future iteration of such a tool should be adapted to show all of its information regardless of the window size, to a certain point.

- **Zooming in on the image**: One of the expert wanted to take a closer look at the image, and for that they zoomed in the web page. While this works fine, it does cause the other information to automatically get pushed to the bottom of the page. The expert mentioned that they would sometimes forget to scroll down again, and forget to use this information altogether. A future iteration should probably include the ability to click the image to display it in full screen mode, so the user can take a better look at it without affecting the rest of the user interface.

- **Not hiding the context**: All experts agreed that hiding the context text leads to a poor experience for them as users. For obvious reasons, in a real-life scenario where such a system is deployed to assist the experts in their image description process, the context text should be immediately available to the user.

- **Ability to backtrack**: The experts remarked that they would like to have seen the inclusion of the ability to go back to the previous task, or skip ahead, as this is something they commonly do in practice. When developing such a system, it should include the ability to cycle through the questions. As the experiment version of the tool runs as some statically hosted JavaScript code, no complicated navigation had been implemented for this version.

6.5.2 Remarks about the survey

Overall, the survey questions were clear to the experts. One of the remarks, which was already foreseen by us, was that the PSSUQ questionnaire contains some questions about error messages, which, while present in the system, were usually not encountered at all. Ahead of the survey, we already informed the experts that the *Not Applicable (N/A)* option exists in the survey, for questions that do not apply to our system, as well as reiterating it in the header text of the survey, but it still caused some confusion with some of the experts, and could perhaps have been stated more clearly. Mostly, the *Neutral* box was ticked in these cases, which should not greatly affect the overall results, but perhaps it does explain the relatively poor interface quality score. A *Not Applicable* score would have been dropped from the results, whereas a *Neutral* score still counts as a 4 when calculating the scores.

6.5.3 Remarks about the information

The initial reaction was mostly negative: The experts explained that apart from a handful of times, the tags were not useful and sometimes even incorrect. Additionally, the experts shared some tags which they considered to be essential, that were not present in the system. An interesting thing to note for those, is that they also saw value in having some tags that we will refer to as metadata tags. These include some simple information about the image type and colours, such as *black and white image*, or *view from the sky*. With separate techniques, possibly these could be retrieved and linked to a future iteration of the system as well.
One of the experts spoke up, countering the initial negative sentiment: Even though often redundant, she felt strengthened and supported, and got the sense that the automated caption helped steer her in a certain direction. The next question was then whether this was the direction that she would choose herself, where the choice can be made to diverge from the tags and construct the description yourself. She also added that it felt like a luxury to have a collection of different information to pick from, that would normally simply not be present. While you might not use the words themselves, they can initiate the thought process to think about concepts to use. This observation and thought confirms our hypothesis that yes, using an assistive tool like the one developed as part of this work can indeed help lower the perceived mental load on the experts, by jump-starting the thought process.
Between the image tags and the context text, the general thought was that the image tags were more useful, which matches the scores we got before in Figure 6.4. Using a handful of words to capture the meaning of a whole body of text is challenging, and especially for shorter documents the results are not always of a high quality with the technique used during the experiment.
One image that we presented to the experts during the discussion is an image of a hand print on a cave wall, outlined with charcoal powder. For readers that have not seen this image before, this description is likely still very vague, so we included this image, which was in the Wikipedia WIT dataset by Srinivasan et al., 2021 in Figure 6.5.
What makes this image interesting is that it is included in the Wikipedia article of the Cosquer cave in southern France[1]. Interestingly, the article itself does not

[1] https://nl.wikipedia.org/wiki/Grot_van_Cosquer

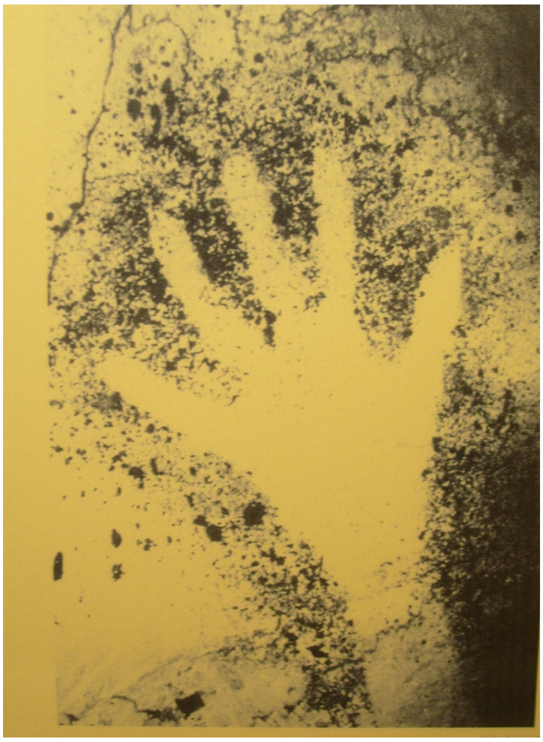


FIGURE 6.5: Photo of the hand print, as included in the WIT dataset

contain any information about the cave painting, but just focuses on the cave in general.
The automated Azure system describes it as a tree, which it clearly is not. The first instinct here would be that this makes the Azure caption useless, but the experts mentioned that this information can still be used for the description. When you would glance at the image, the first thing you might think is *Oh, this looks a bit like a tree*. This first impression can, depending on the context, also be useful, for instance with images of pieces of art like this example. In this sense, even inaccurate results can occasionally be put to use to create a description.

6.5.4 Different image sources

When discussing the different information sources, and the differences between them, it was clear that the most prominent difference of the educational articles, when compared to the other two, is the amount of contextual knowledge that may be required to correctly interpret them.
Where news and magazines are conventionally written to appeal to many people, in education some domain specific knowledge may be required. For example, in a medical handbook, it may or may not be of great importance that there is, for example, a red spot on a person's skin. For this reason, people who write image descriptions in the (higher) education context are usually collaborating with the writers or other people who are knowledgeable in the domain for this very reason. In the majority of cases, automated systems are not trained with these domain-specific cases in mind, so they are very likely to be lacking for those.
Another challenge was identified as education books with exercises: An example as named by the experts: imagine a picture of some different brain scans like MRI and CT scans. Attached to this image, there may be a question asking to label each image with the correct scan type. The image could be the one in Figure 6.6. When

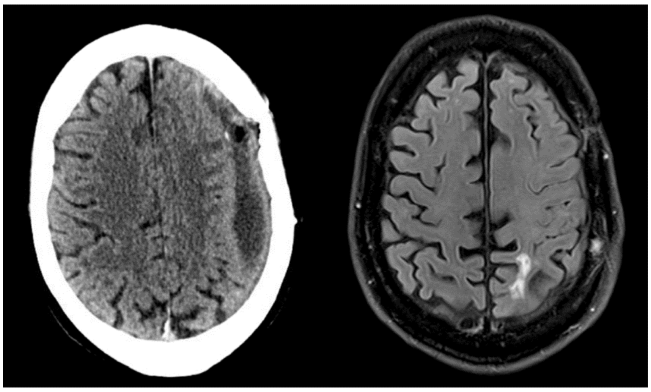


FIGURE 6.6: CT scan and MRI scan of the human brain

writing a description, the expert wants to convey the visual information without giving away the answer. They could mention something like *A grey grainy mass with a sharp white border around it*, but should not accidentally mention that it is, in fact, a CT scan. Aside from this, there may be education books where the required knowledge is in the book, but the exercises are split from the explanations, making it difficult for automated systems to determine what the context of that particular image is.

6.5.5 The future

When discussing how the experts would see such a system implemented for use in practice, there were two clear observations: Firstly, they see great value in having an existing database of image descriptions for visually impaired people to match to, either using tags or using visual information. Such an approach would be more akin to a retrieval system. Still, before such a database is available, it is a compelling idea to use an assistive tool like the one developed here that simultaneously helps the experts perform their task today, while also helping to build this improved system for the future by recording the results.
Secondly, the system here is developed to be as generally applicable as possible. While this offers great value during this experimentation phase, where it is not yet clear what the final product will contain and will be able to do, it proves to be lacking in practice, especially in the quality of the information it provides to the user. The experts hypothesised that using image source and domain specific resources should be added to the system, to improve the overall usefulness. For instance, history books oftentimes contain images of artworks. In the back of the book, a reference to the museum where this artwork is located is usually included. For these cases, the catalogue of the museum in question can be consulted, and may have either more useful information about the artwork in question, or even already have an image description that can be adapted by the expert. Of course there is a balance to be struck between perfectly adapting to each very specific use case to get the information and having a system available that functions on different use cases without further adaptation.

Chapter 7

Summary & Conclusions

7.1 Summary

In this work, we set out to investigate the image description process of experts for visually impaired people and based on this process research and develop the design of a software tool to assist experts in the image description process. The images in question are accompanied by text in publications, in the Dutch language, and there is a body of text and optionally an image caption included in the original document. Automated systems are in practice unable to properly take this context as well as the intended target audience into account, making us dependent on the human experts.

First, we interviewed an expert writing these image descriptions in the Netherlands, where the KB National Library is one of the major organisations responsible for supplying publications with image descriptions, and outlined the general image description workflow used by the experts. Based on this workflow, we selected the step of interpreting the available information to try and improve, as we discovered that this step is considered mentally taxing by the experts.

Next, we try to improve this information interpretation step by providing alternative representations of the available information to the expert, both for the image as well as for the surrounding text. Our hypothesis was that this would make the task more approachable and thereby lower the perceived high mental load. The intuition was that as an alternative to the visual information in the image and the large body of text in the context, the user could instead use alternatives like short sentences and keywords that attempted to convey the same information. To obtain these alternative representations of the existing information, we developed a workflow where the document is preprocessed, and automated systems are used to get the information required. Note that the goal of the system is to develop a better understanding on *if* such a system can lower the mental load, and if so which information is the most useful towards reaching this goal. For this reason, we opted for using state-of-the-art commercially available automated methods, namely Microsoft Azure Analyze Image introduced in Hu et al., 2020 for processing the image, and KeyBERT as described in Grootendorst, 2020 for processing the context text.

After having collected the information, we needed to decide how to present it to the user. To this end, we let user interface guidelines from existing research guide our design process, and determine how different information should be organised in the user interface. In the end, we designed the tool that is stored as statically hosted files, that are served by GitHub Pages so we could let the experts try it out and share their experiences during the experiment.

To evaluate the system, we selected three image sources, namely informational article, educational article and news article. For each of these three categories, we added three different tasks, totalling nine tasks to perform for each of the experiment participants.

We designed the experiment with the main goal of validating our hypothesis: does our software tool help reduce the perceived mental load on the experts during the image description task for visually impaired people? Given the small number of experts that could partake in the experiment, six, the quantitative data that we could collect was limited. Still, to get an idea of the general usability of the system, we used the standardised PSSUQ. We expanded this survey with 14 extra questions to also get some data on the perceived usefulness and usage of the several specific information sources.

To gain more useful insights, we had a group discussion with all six of the participants to discuss in more detail what their thoughts and opinions on this system and such systems in general were, and to discuss future iterations and thoughts. Mixing their different opinions has lead to meaningful insights, reiterated below.

7.2 Insights

While analysing the results, we discovered that not all experts always resorted to using the context text, which means that at least for some tasks, the tool provides sufficient information for the expert to write a description. When looking at the PSSUQ survey results, we discovered that the system usefulness was rated highly, whereas the interface quality received a relatively poor rating, which was likely in part due to the confusion that arose when some questions in this standardised survey did not apply to our system. As for the expanded survey on the different components, the results were within our line of expectations: The existing context was held in the highest regard when used, and the generated tags and automated caption were rated rather poorly.

During the discussion, it became clear that the main problem with these last three resources is that they are useful in some cases, but can be ignored in many cases: It is not so much that they are always useless, but rather that they are not always useful. Still, even when wrong, the information can sometimes be used by the experts to inspire and initiate the thought process. Between the different image sources, the education source was highlighted by the experts to be a challenging one between the three that were selected for the experiment. The main challenge therein is that they commonly require contextual knowledge to understand, making them particularly difficult for a system like ours that relies on very generally applicable techniques. During the discussion we concluded that in order to improve the quality of the information, one useful step would be to develop a system with a particular use case in mind, and employing information sources specific to this use case, making the system more useful.

Additionally, the experts saw great merit in a system that uses a retrieval approach to get previously written descriptions for similar images or articles, as these would provide high quality descriptions that the novel description can be based of off. An actual implementation of a program to assist the experts should be able to send the results to a system that stores them, so that a database of image descriptions is built up over time and can eventually also be used to feed into the system itself.

7.3 Limitations

When working on the experimental design to evaluate the developed system, we ideally wanted to conduct a statistical analysis, resulting in strongly supported conclusions. Unfortunately, the number of experts and their available time to evaluate the system with is limited. To combat this, we have chosen solutions that are proven to work well even with smaller sample sizes, such as the PSSUQ, that was shown to be accurate even for relatively small sample sizes in Tullis and Stetson, 2004. Additionally, our efforts were directed towards fuelling a group discussion amongst the experts, in order to gather qualitative data and feedback. Even so, to make strong claims about how well the system works, an analysis at a large scale would be highly desirable: Ideally, part of the expert pool write the image descriptions in the current manual way, and the others use the system. By comparing their time spent on the different tasks or even determining the mental load while working as described by Kalsbeek and Sykes, 1967.
In the end our group of six experts is too small to make statistically significant claims about the system, but the analysis and development in this thesis opens up opportunities for further study and development of such systems, as the results are promising.
Another limitation of this work is that for the most part, it is focused on photographs and simple illustrations. This decision was made consciously, because our goal was to be able to test the system with a wide variety of image sources. The experts have mentioned that describing infographics for instance are a different challenge altogether. These comprise images such as timelines or process diagrams. While we suspect that a similar system can be used, the automated preprocessing techniques used should be adapted accordingly in order to work for such a use case.

7.4 Conclusions

All in all, we learned that the experts use guidelines to create image descriptions in a structured manner. Within their workflow, the step of interpreting the available information in order to write the description can be mentally tiring. We created a tool to assist the experts during this step in their process. From our experiment, it is clear that for some of the tasks the system helps the experts in their image description process. As a result, we finally conclude that an assistive software tool that helps the experts to write image descriptions for visually impaired people by providing alternative representations along with the available information is able to help improve the experience by the experts during the information interpretation step of the image description process.

7.5 Recommendations

For future efforts to assist experts in the image description process for visually impaired people, we recommend focusing on a particular image source or domain and using information sources and methods that correspond to the image source at hand. This helps improve the quality of the information and leads to a more useful system overall. According to the experts we spoke to during the interview, the education context or a subgroup of it is a particularly good candidate for this. As more improvements are made to automated methods, they can also be integrated within the system with relative ease to further improve its usefulness in the image description task.

Because we limited ourselves to Dutch experts, we ended up with only six participants in our experiment: too few for extensive statistical analysis. We believe that it is valuable to investigate the non-English case, as many techniques are created with the English language in mind. An option for future research is to focus on experts in multiple different languages, to thereby introduce a larger pool of experts and use statistical analysis tools to draw well-supported conclusions, provided that the techniques used are language-independent.

Appendix A

Example News Article

Source: https://www.ed.nl/binnenland/expert-kraakt-boerenactie-houdbaarheidsdatum-boerenprotest-is-verstreken ac75e3a2/

Expert kraakt boerenactie: 'Houdbaarheidsdatum boerenprotest is verstreken'

FIGURE A.1: Boerenprotesten in Apeldoorn. © Luciano De Graaf

Albert Heller 05-07-22, 18:11
De boeren willen misschien nog weken blijven protesteren. Maar trekt een land dat? En heeft het nog zin? Arco Timmermans, bijzonder hoogleraar Public Affairs van de Universiteit Leiden, denkt van niet. „De houdbaarheidsdatum van dit protest is al lang verstreken." Timmermans benadrukt dat de boeren van nature

sympathie hebben bij een groot deel van de maatschappij. Hun verhaal is duidelijk. Ze zorgen voor eten en willen dat dit blijft. Maar nu de acties maar voortduren, denkt Timmermans dat de houdbaarheid van de acties is verlopen.

Wedstrijd om de grootste ballen

„Dat geldt ten opzichte van politiek Den Haag, maar ook ten opzichte van de bevolking die last van de acties heeft, bijvoorbeeld in de file." Hoewel volgens opiniepeilingen bij die laatste groep er nog veel waardering is voor de boeren, betekent dit volgens Timmermans niet dat het protest nog langer houdbaar is. „Het is heel goed dat er bemiddeling komt tussen boeren en overheid, zodat de discussie los getrokken wordt. Nu gebeurt er niets anders dan polarisatie. Alles wat het ene kamp wint, is verlies voor het andere kamp." In Den Haag is heel duidelijk gezegd:

de doelen blijven staan, ook na een eerste ronde van acties door de boeren. „De VVD sprak woorden als: 'We gaan geen bakzeil halen'. Als je dan door blijft gaan met actievoeren, loop je nog meer het risico dat het een wedstrijdje wordt wie de grootse ballen heeft. De boeren of de overheid? Het middel vervangt dan het doel. Als boeren doorgaan, krijgen ze zeker niet wat ze willen."

Eenheid boeren schijn?

Dat er bij de boeren juist eenheid lijkt om lang door te gaan met de acties, kan wel eens schijn zijn, denkt de hoogleraar. „Natuurlijk zegt LTO dat het protesten ondersteunt. Ze moeten wel, omdat ze de hete adem van radicalere groepen als Farmers Defence Force in hun nek voelen. Ze willen niet buitenspel gezet worden in de discussie." Timmermans benadrukt dat hij respect heeft voor de agrarische

sector, de boerenbedrijven waardeert. Het is volgens hem dan óók aan het kabinet om de impasse te beëindigen. „In Den Haag moet empathie zijn voor de boerenbedrijven. Het probleem met de fixatie op de stikstof vanuit de boeren is dat de discussies zich snel gaat vernauwen. Het gaat alleen nog maar om minder dieren, terwijl de problemen rondom stikstof, de natuur en onze voedselvoorziening veel breder spelen."

Appendix B

Expert Interview Transcript

Hoe ziet jullie beeldbeschrijvingsproces er nu uit?

Er zijn richtlijnen die worden gebruikt, per type uitgave. Er zijn bijvoorbeeld richtlijnen voor kranten en tijdschriften, voor educatieve uitgaven, en weer aparte voor jeugdbeeldboeken en graphic novels. Bij de ene uitgave moet nog veel worden geselecteerd, bij andere wordt sowieso bijna alles al meegenomen. Bij de educatieve context is er bijvoorbeeld vaak al een duidelijke rol voor de beelden die erbij zitten. Na de selectie wordt de afbeelding tot je genomen, rekening houdend met de richtlijnen zoals de link met de context. Dingen die al letterlijk in de tekst staan hoeven niet opnieuw genoemd te worden. Het is ook zaak om zo kort mogelijk te zijn. Dingen die niet in de tekst worden genoemd en wel belangrijk zijn. De facetten zijn meestal wie of wat, waar en hoe, dat geldt voor alle producten. De richtlijnen bevatten adviezen over de woordvolgorde. Bijvoorbeeld om te beginnen met hoe iets of iemand erbij staat. Soms, bijvoorbeeld bij sommige tijdschriften, gaat het er vooral om de sfeer te beschrijven, maar dan begin je toch nog steeds met wat of wie. De grootte van de afbeelding, de bladspiegel, en hoe belangrijk de afbeelding is in de context, bepalen ook de lengte van de beschrijving. Ook afhankelijk van de het type product verschilt de lengte, van drie of vier woorden tot enkele regels of een hele alinea. Idealiter zou je als gebruiker kunnen kiezen of overslaan, want het kost veel breincapaciteit om die lange stukken te consumeren.

Hoeveel procent van de beelden in bijvoorbeeld een tijdschrift schat je dat er een beschrijving krijgt?

Er wordt gerekend in minuten. Ongeveer 20 procent van de tijd wordt besteed aan het beeld beschrijven. Deze tijd moet wel aansluiten bij de formule en identiteit van een bepaalde krant. Het moet vooral de juiste verhouding tussen woord en beeld zijn. Als de krant meer ruimte heeft voor beelden zal je er ook meer tijd aan willen besteden. Educatieve artikelen kunnen heel veel tijd in beslag nemen. Het kan zomaar zo zijn dat er een opgave is waarvoor je vrijwel alle afbeeldingen nodig hebt.

Hoe zit het met de benodigdheid van externe bronnen?

Het is een spagaat tussen het objectief weergeven wat je ziet en de juiste informatie geven. Wat je ziet weet je soms zelf ook niet, en je wil niet de verkeerde informatie geven. Dan is er dus checkwerk. Voorbeeld: Je hebt een foto van de Eiffeltoren, en wil de onderdelen beschrijven. Dan kan je overal zeggen "smeedijzer", maar zijn die onderdelen wel van smeedijzer? Je mag en kan ook geen verkeerde informatie geven. Veel tijd van het beschrijven gaat zitten in het selecteren van welke elementen uit de afbeeldingen er wel en niet worden meegegeven. Als je het niet beschrijft is het er gewoon niet. En er gaat dus veel tijd zitten in het uitzoeken. Ik kan soms zomaar 10 minuten dingen opzoeken en daarna in 2 minuten klaar zijn met schrijven. Hier wordt met de tijdsinschatting voor de opdrachtgever ook al rekening mee gehouden.

Denk je dat het een meerwaarde zou kunnen hebben om al een "oppervlakkige" beschrijving beschikbaar te hebben?

Het heeft wel een meerwaarde. De voorwaarde is wel waar het op gebaseerd is. Omdat het soms veel zoekwerk is ernaast. Er zijn een aantal dingen die je toch standaard wel wil meegeven. Alles helpt om het brein van de beschrijver in de goede richting te krijgen, en niet te vergeten. Daarin kan het erg helpen, door het complexe proces in je hoofd te ondersteunen: Je bent de informatie aan het verwerken terwijl je aan het beschrijven bent. Er was ook een onderzoek in Amerika die zoiets al gevonden heeft, het is vrij vermoeiend werk.

Waar denk je dat de winst in zit met een programma dat je helpt tijdens het beschrijven?

Stel je beschrijft sneller, dan zou je nog steeds de verhouding tussen beeld en tekst in het originele werk willen behouden. Je kan dan dus hooguit meer boeken afkrijgen, en zal dan niet per boek meer afbeeldingen behandelen. Het kan ook de kwaliteit ten goede komen. Soms zie ik dingen voorbij komen waarvan ik denk "Hier had gewoon X in moeten staan", een gemiste kans. Het gaat dus om kwaliteit en tijd. Wel hangt het af van de gebruikte bronnen.

Hoe ziet het selectieproces eruit?

Het selectieproces verschilt per uitgave. Bij kranten en tijdschriften maakt een bestaande redactie een selectie van teksten. Het is een vrij nieuw concept, maar ook het beeld zou leidend moeten kunnen zijn in plaats van de tekst, dus dat zie ik in de toekomst ook wel meer gebeuren.

Appendix C

Informed Consent Form

Informed Consent

I have been asked to participate in a study about 'a tool to assist in writing image descriptions'. The goal of the research project is to improve the image description writing process, lowering the mental load required and thereby ideally improving the efficiency and accuracy of the process. The project is a collaboration between the Delft University of Technology and the Royal Library (KB).

If I consent to participate in this study, I have been informed that the study will last approximately two hours and that the study consists of an individual session where the program is used, followed by an individual survey, and finally a group interview with approximately 7 people. Participating in this study entails:

1. Writing image descriptions for the provided images in the program
2. Filling out a survey
3. Voicing my opinions on ideas to assist in the image description writing process

Furthermore I understand:

- That I'm completely free to refuse to answer questions
- That I can decide not to participate at this point and that I can withdraw my participation at any time.
- That the following data will be collected:

    1. Notes that are taken by the researcher
    2. Audio and video recordings. The recordings will be used to analyse the results for purely research purposes and are not to be shared publicly.
    3. The survey you fill out. You will be asked to provide your number of years of experience in the field, and your mother tongue. The data from this survey will be used to analyse outcomes of the study.
- That all individual results will be treated confidentially. Results will only be reported for the group as a whole or in an anonymized way.
- That the anonymised research data will be accessible only to the researcher and responsible researcher via the Microsoft Teams cloud service.

Please tick the appropriate boxes	Yes	No
Taking part in the study		
I have read and understood the study information above, or it has been read to me. I have been able to ask questions about the study and my questions have been answered to my satisfaction.	○	○
I consent voluntarily to be a participant in this study and understand that I can refuse to answer questions and I can withdraw from the study at any time, without having to give a reason.	○	○

Signatures

______________________ ______________________ ________

Name of participant [printed] Signature Date

I have accurately read out the information sheet to the potential participant and, to the best of my ability, ensured that the participant understands to what they are freely consenting.

______________________ ______________________ ________

Researcher name [printed] Signature Date

Study contact details for further information:

Frank Vollebregt

F.C.J.Vollebregt@student.tudelft.nl

Appendix D

Experiment Data

D.1 Informational Articles

D.1.1 Koningsbergen

Source: https://nl.wikipedia.org/wiki/Koningsbergen

FIGURE D.1: Slot van Koningsbergen in 1925

Koningsbergen was een Duitse stad in Oost-Pruisen. Koningsbergen werd gesticht in 1255 en was van 1457 tot 1945 hoofdstad en zowel cultureel als economisch centrum van het oosten van Pruisen. Het was de meest oostelijk en noordelijk gelegen grote stad van het Duitse Rijk. De stad ligt in Samland, dicht bij de Oostzeekust tussen het Wislahaf en het Koerse Haf. De stad is van groot belang geweest in de Duitse geschiedenis en werd, nadat Oost-Pruisen na de Eerste Wereldoorlog van Duitsland gescheiden was, een van de economisch meest vooruitstrevende steden van Europa. Tijdens de Tweede Wereldoorlog werd de stad vrijwel volledig verwoest en daarna geannexeerd door de Sovjet-Unie en herbouwd onder de nieuwe naam Kaliningrad.

D.1.2 Baronnies

Source: https://nl.wikipedia.org/wiki/Baronnies
De Baronnies zijn een Frans bergmassief dat deel uitmaakt van de Franse Voor-Alpen en een historische streek. Het grootste deel van de streek ligt in het departement Drôme, maar het westelijke deel van de Hautes-Alpes en het noorden van de Vaucluse behoren eveneens tot de Baronnies. Het massief van de Baronnies is een middelgebergte dat wordt gerekend tot de Voor-Alpen van de Dauphiné. Ten noorden van de Baronnies ligt de Diois, ten zuiden liggen de Monts de Vaucluse

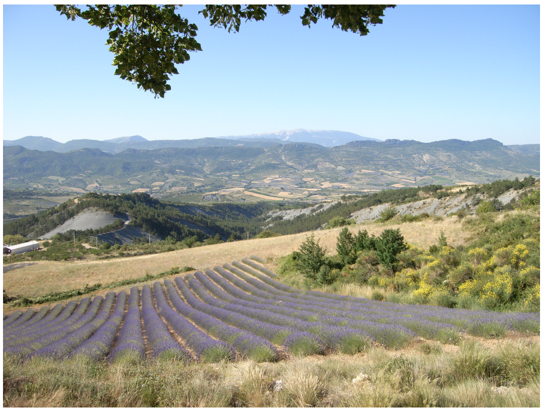


FIGURE D.2: Het landschap van de Baronnies (vallei van Ste-Jalle) vanaf de col de Soubeyrand

met de Mont Ventoux en de Montagne de Lure. Er zijn twee stadjes in de Baronnies: Nyons en Buis.

D.1.3 HH. Laurentius- en Elisabethkathedraal

Source: https://nl.wikipedia.org/wiki/HH._Laurentius-_en_Elisabethkathedraal

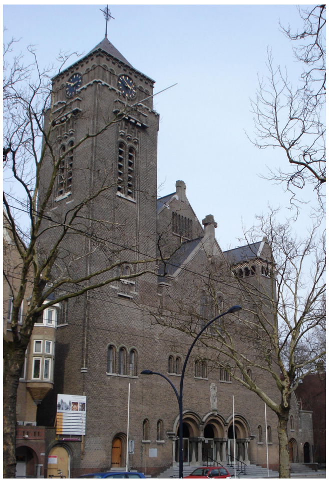


FIGURE D.3: Laurentius- en Elisabethkathedraal

De HH. Laurentius- en Elisabethkathedraal is een rooms-katholieke kerk in Rotterdam, in neoromaanse stijl gebouwd naar een ontwerp van P.G. Buskens. Sinds 1967 fungeert de kerk als kathedraal. De kerk verrees in twee fasen. In de periode 1906-1908 werden koor, transept en schip gebouwd. Het tweetorenfront kwam tussen 1920 en 1922 tot stand. De eerste oorlogsschade liep de kerk op bij een Brits bombardement op 3 oktober 1941, toen de ingang en het dak en de koepel werden beschadigd. Bij een luchtaanval op het gebouw van de Sicherheitsdienst aan de Heemraadssingel op 29 november 1944 raakte een zijbeuk van de kerk zwaar beschadigd. Na de oorlog is deze schade hersteld. Aanvankelijk heette de kerk Sint-Elisabethkerk en diende zij als parochiekerk. In 1956 werd Rotterdam een bisdom. Als kathedraal werd toen de Sint-Ignatiuskerk aan de Westzeedijk gekozen, die werd hernoemd naar de patroonheilige van de stad Rotterdam, Laurentius. Ook de naam van de middeleeuwse Sint-Laurenskerk in Rotterdam verwijst naar deze heilige. Die kerk is echter sinds 1572 protestants. In 1967 werd de Sint-Elisabethskerk de kathedraal van het Bisdom Rotterdam en kreeg zij de huidige naam.

D.2 News Articles

Source: https://nos.nl/artikel/2367020

D.2.1 Dronken man steelt tractor en laat in Duitse plaats spoor van vernieling achter

FIGURE D.4: Kreispolizeibehörde Oberbergischer Kreis

Een dronken 23-jarige man uit het Duitse Radevormwald heeft gisteravond een tractor gestolen en daarna een spoor van vernieling in de gemeente achtergelaten. De tractor was uitgerust met een vier meter brede sneeuwschuiver, schrijven Duitse media.
Hij heeft voor bijna 100.000 euro aan schade aan gebouwen en voertuigen aangericht in de gemeente in de deelstaat Noordrijn-Westfalen. Hij ramde onder meer een gevel van een huis, een schutting, een garage en vijf auto's. Bij de botsing met een auto raakte een 46-jarige bestuurder lichtgewond.
Rond 19.30 uur kreeg de politie oproepen binnen van het tumult. De man liet de tractor achter op een parkeerplaats van een supermarkt. De politie vond de man daar, nog steeds onder invloed, in de buurt van de tractor.

D.2.2 CO2-uitstoot door natuurbranden in 2021 hoogste ooit

Source: https://nos.nl/artikel/2408650

FIGURE D.5: Bosbranden in het zuiden van Frankrijk, afgelopen zomer

Natuurbranden veroorzaakten het afgelopen jaar naar schatting ongeveer 1760 megaton aan CO2-uitstoot. Dat is een record voor wat betreft de uitstoot door

bosbranden, berekende het aarde-observatieprogramma van de Europese Unie, Copernicus.
Door klimaatverandering zullen er steeds vaker bosbranden zijn, denken de wetenschappers van het programma. "Hoge temperaturen, heftige onweersbuien, harde wind en ander extreem weer gaan steeds vaker voorkomen", schrijft het Copernicus.
Ook Sander Veraverbeke, klimaatwetenschapper aan de Vrije Universiteit Amsterdam, ziet een duidelijke trend. "Als ik terugdenk aan vorige zomer met al die grote branden in het Middellandse Zeegebied, in de Verenigde Staten en Canada, en in Oost-Siberië, ben ik niet verwonderd dat dat nu tot records leidt", zegt hij in het NOS Radio 1 Journaal.
Vooral in de zomer op het noordelijk halfrond was er sprake van veel uitstoot, meldde Copernicus eerder. In juli en augustus werden er al records verbroken.
...

D.2.3 Eerste torenvalk geboren in nest op balkon in Terneuzen

Source: https://nos.nl/artikel/2386607

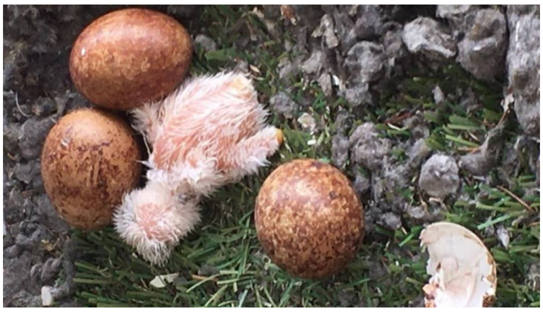


FIGURE D.6: De eieren op het balkon

Het eerste ei van een torenvalk op een balkon in Terneuzen is uitgekomen. De andere eieren in het nest zijn nog heel.
Eind mei bouwde een koppel torenvalken hun nest op het balkon. De torenvalk is een beschermde soort en dus liet de bewoner de vogels in alle rust broeden.
Na ruim dertig dagen wachten, is de eerste torenvalk geboren, net nu de bewoner er niet is. Otto Peltzer, die op het huis past, ontdekte het kuiken vanmiddag. "De moedervalk vliegt altijd weg als de balkondeur opengaat en daardoor zagen we dat er een ei was uitgekomen", zegt hij tegen Omroep Zeeland.
Echt bijzonder Volgens de Werkgroep Roofvogels Zeeland is de situatie op het balkon uniek. "Het is echt bijzonder dat deze valken een balkon in de stad hebben uitgekozen. In Oost-Europa zie je het heel soms wel eens, maar in Nederland niet", zegt een medewerker.
De torenvalk staat sinds 2017 op de lijst van bedreigde soorten. De verwachting is dat de rest van de eieren de komende dagen uitkomt.

D.3 Education Articles

Documents and images for this image source have been retrieved from one of the online learning modules offered by the NEMO Science Museum.

D.3.1 Gegevens lijkvinding

Source: https://www.nemosciencemuseum.nl/nl/onderwijs/voortgezet-onderwijs/lesmateriaal/lesmateriaal/biologie-voor-vo-biowetenschappen-en-maatschappij/

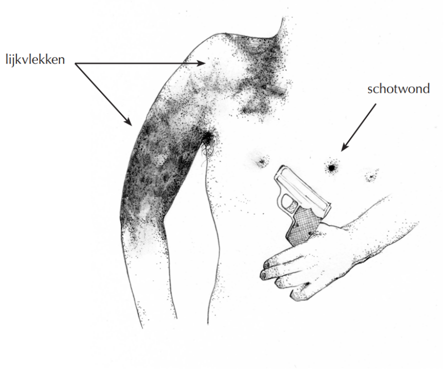


FIGURE D.7: -

Datum en tijd: vrijdag 3 september, 15.40 Ruimte: De kamer waar de overledene is gevonden ligt op de begane grond. Een schuifdeur naar de achtertuin staat op een kier. De temperatuur in de kamer is vastgesteld op 18°C. Ligging: De overledene ligt plat op zijn rug op een driezitsbank. Zijn hoofd ligt op een kussen. De rechterarm van de overledene ligt langs zijn lichaam; de linkerarm ligt op zijn borst. Kleding: De overledene is gekleed in een dunne pantalon, een boxershort en een hemd zonder mouwen. Op de borstkas van de overledene ligt een pistool. Lichaam: Midden op de borst (5 cm links van de rechter tepel) is een schotwond zichtbaar. Op de bovenzijde van de rechterarm, en de rechterschouder zijn lijkvlekken aanwezig (zie afbeelding). Lijkvlekken zijn wegdrukbaar. De lichaamstemperatuur (rectaal) is vastgesteld op 35,0°C.

D.3.2 Getuigenverklaringen

Source: https://www.nemosciencemuseum.nl/nl/onderwijs/voortgezet-onderwijs/lesmateriaal/lesmateriaal/biologie-voor-vo-biowetenschappen-en-maatschappij/
5. Het komt ook voor dat een getuige zich iets herinnert wat helemaal niet is gebeurd. Lees het stukje tekst hieronder.
Stopbord of voorrangsbord? Onderzoekers lieten proefpersonen naar een filmpje kijken waarin een rode auto stopt voor een driehoekig voorrangsbord. Daarna werd aan de proefpersonen o.a. gevraagd of de auto die stopte voor het stopbord nog door een andere auto in werd gehaald. Een groot deel van de proefpersonen bleek zich later 'te herinneren' dat er inderdaad een stopbord was. (Loftus et al. 1978)

D.3.3 Het spel van de Gouden Eeuw

Source: https://www.nemosciencemuseum.nl/nl/onderwijs/voortgezet-onderwijs/lesmateriaal/lesmateriaal/biologie-voor-vo-biowetenschappen-en-maatschappij/
Na een aantal succesvolle reizen van en naar Indië, verging VOC-schip Prins Willem in 1662. De volledige bemanning liet het leven en de hele lading ging verloren.

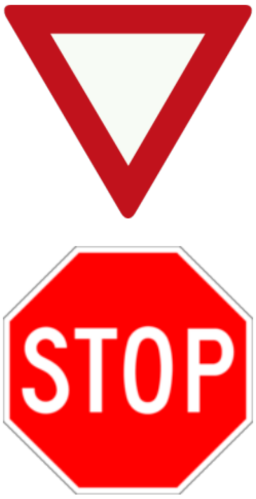


FIGURE D.8: -

FIGURE D.9: Foto: Replica van de Prins Willem / Bron: Dirk van der Made

Appendix E

Experiment JSON Data

```
[
   {
      "ctx_title": "Koningsbergen",
      "ctx_url": "https://nl.wikipedia.org/wiki/Koningsbergen",
      "ctx_caption": "Slot van Koningsbergen in 1925",
      "ctx": "Koningsbergen was een Duitse stad in Oost-Pruisen.
         Koningsbergen werd gesticht in 1255 en was van 1457 tot 1945
         hoofdstad en zowel cultureel als economisch centrum van het
         oosten van Pruisen. Het was de meest oostelijk en noordelijk
         gelegen grote stad van het Duitse Rijk. De stad ligt in Samland,
         dicht bij de Oostzeekust tussen het Wislahaf en het Koerse Haf.
         De stad is van groot belang geweest in de Duitse geschiedenis en
         werd, nadat Oost-Pruisen na de Eerste Wereldoorlog van Duitsland
         gescheiden was, een van de economisch meest vooruitstrevende
         steden van Europa. Tijdens de Tweede Wereldoorlog werd de stad
         vrijwel volledig verwoest en daarna geannexeerd door de Sovjet-
         Unie en herbouwd onder de nieuwe naam Kaliningrad.",
      "ctx_tags": [
         {
            "tag": "oostzeekust",
            "score": 0.5207
         },
         {
            "tag": "gelegen",
            "score": 0.5099
         },
         {
            "tag": "kaliningrad",
            "score": 0.5004
         },
         {
            "tag": "stad",
            "score": 0.4957
         },
         {
            "tag": "herbouwd",
            "score": 0.4775
         }
      ],
      "img": "https://upload.wikimedia.org/wikipedia/commons/0/00/K%C3%
         B6nigsberg_%28Luftaufnahme%29.JPG",
      "img_caption": "een zwart-witfoto van een stad",
      "img_tags": [
         {
            "tag": "tekening",
            "score": 0.9778131246566772
        },
        {
            "tag": "schets",
            "score": 0.9644944667816162
        },
        {
            "tag": "zwart-wit",
            "score": 0.8919709920883179
        },
        {
            "tag": "abstract",
            "score": 0.889147162437439
        },
        {
            "tag": "kunst",
            "score": 0.8362852334976196
        }
    ],
    "type_name": "een informatief boek of tijdschrift"
},
{
    "ctx_title": "Baronnies",
    "ctx_url": "https://nl.wikipedia.org/wiki/Baronnies",
    "ctx_caption": "Het landschap van de Baronnies (vallei van Ste-Jalle)
         vanaf de col de Soubeyrand",
    "ctx": "De Baronnies zijn een Frans bergmassief dat deel uitmaakt van
         de Franse Voor-Alpen en een historische streek. Het grootste
        deel van de streek ligt in het departement Dr\u00f4me, maar het
        westelijke deel van de Hautes-Alpes en het noorden van de
        Vaucluse behoren eveneens tot de Baronnies. Het massief van de
        Baronnies is een middelgebergte dat wordt gerekend tot de Voor-
        Alpen van de Dauphin\u00e9. Ten noorden van de Baronnies ligt de
        Diois, ten zuiden liggen de Monts de Vaucluse met de Mont Ventoux
         en de Montagne de Lure.\nEr zijn twee stadjes in de Baronnies:
        Nyons en Buis.",
    "ctx_tags": [
        {
            "tag": "stadjes",
            "score": 0.5834
        },
        {
            "tag": "nyons",
            "score": 0.5734
        },
        {
            "tag": "departement",
            "score": 0.5523
        },
        {
            "tag": "uitmaakt",
            "score": 0.5463
        },
        {
            "tag": "middelgebergte",
            "score": 0.546
        }
        ],
        "img": "https://upload.wikimedia.org/wikipedia/commons/8/8f/
            Baronnies1.jpg",
        "img_caption": "een veld van paarse bloemen",
        "img_tags": [
            {
                "tag": "openlucht",
                "score": 0.9969696998596191
            },
            {
                "tag": "hemel",
                "score": 0.9944952726364136
            },
            {
                "tag": "landschap",
                "score": 0.8960225582122803
            },
            {
                "tag": "natuur",
                "score": 0.8600541353225708
            },
            {
                "tag": "boom",
                "score": 0.7729629278182983
            }
        ],
        "type_name": "een informatief boek of tijdschrift"
    },
    {
        "ctx_title": "HH. Laurentius- en Elisabethkathedraal",
        "ctx_url": "https://nl.wikipedia.org/wiki/HH._Laurentius-
            _en_Elisabethkathedraal",
        "ctx_caption": "Laurentius- en Elisabethkathedraal",
        "ctx": "De HH. Laurentius- en Elisabethkathedraal is een rooms-
            katholieke kerk in Rotterdam, in neoromaanse stijl gebouwd naar
            een ontwerp van P.G. Buskens. Sinds 1967 fungeert de kerk als
            kathedraal.\nDe kerk verrees in twee fasen. In de periode
            1906-1908 werden koor, transept en schip gebouwd. Het
            tweetorenfront kwam tussen 1920 en 1922 tot stand. De eerste
            oorlogsschade liep de kerk op bij een Brits bombardement op 3
            oktober 1941, toen de ingang en het dak en de koepel werden
            beschadigd. Bij een luchtaanval op het gebouw van de
            Sicherheitsdienst aan de Heemraadssingel op 29 november 1944
            raakte een zijbeuk van de kerk zwaar beschadigd. Na de oorlog is
            deze schade hersteld.\nAanvankelijk heette de kerk Sint-
            Elisabethkerk en diende zij als parochiekerk. In 1956 werd
            Rotterdam een bisdom. Als kathedraal werd toen de Sint-
            Ignatiuskerk aan de Westzeedijk gekozen, die werd hernoemd naar
            de patroonheilige van de stad Rotterdam, Laurentius. Ook de naam
            van de middeleeuwse Sint-Laurenskerk in Rotterdam verwijst naar
            deze heilige. Die kerk is echter sinds 1572 protestants. In 1967
            werd de Sint-Elisabethskerk de kathedraal van het Bisdom
            Rotterdam en kreeg zij de huidige naam.",
        "ctx_tags": [
            {
                "tag": "elisabethkerk",
                "score": 0.5713
            },
            {
                "tag": "bisdom",
                "score": 0.5627
            },
            {
                "tag": "ignatiuskerk",
                "score": 0.561
            },
            {
                "tag": "elisabethskerk",
                "score": 0.5459
            },
            {
                "tag": "tweetorenfront",
                "score": 0.5406
            }
        ],
        "img": "https://upload.wikimedia.org/wikipedia/commons/4/49/
            Rotterdam_mathenesserlaan_kathedraal.jpg",
        "img_caption": "een kerk met een torenspits",
        "img_tags": [
            {
                "tag": "boom",
                "score": 0.9993634223937988
            },
            {
                "tag": "openlucht",
                "score": 0.999226987361908
            },
            {
                "tag": "hemel",
                "score": 0.9988948106765747
            },
            {
                "tag": "gebouw",
                "score": 0.9949307441711426
            },
            {
                "tag": "kerk",
                "score": 0.8539779186248779
            }
        ],
        "type_name": "een informatief boek of tijdschrift"
    },
    {
        "ctx_title": "Gegevens lijkvinding",
        "ctx_url": "https://www.nemosciencemuseum.nl/nl/onderwijs/voortgezet-
            onderwijs/lesmateriaal/lesmateriaal/biologie-voor-vo-
            biowetenschappen-en-maatschappij/",
        "ctx_caption": "-",
    "ctx": "Datum en tijd: vrijdag 3 september, 15.40\nRuimte: De kamer
        waar de overledene is gevonden ligt op de begane grond. Een
        schuifdeur naar de achtertuin staat op een kier. De temperatuur
        in de kamer is vastgesteld op 18 C. \nLigging: De overledene ligt
         plat op zijn rug op een driezitsbank. Zijn hoofd ligt op een
        kussen. De rechterarm van de overledene ligt langs zijn lichaam;
        de linkerarm ligt op zijn borst. \nKleding: De overledene is
        gekleed in een dunne pantalon, een boxershort en een hemd zonder
        mouwen. Op de borstkas van de overledene ligt een pistool. \
        nLichaam: Midden op de borst (5 cm links van de rechter tepel) is
         een schotwond zichtbaar. Op de bovenzijde van de rechterarm, en
        de rechterschouder zijn lijkvlekken aanwezig (zie afbeelding).
        Lijkvlekken zijn wegdrukbaar. De lichaamstemperatuur (rectaal) is
         vastgesteld op 35,0 C.",
    "ctx_tags": [
        {
            "tag": "lichaamstemperatuur",
            "score": 0.6174
        },
        {
            "tag": "linkerarm",
            "score": 0.6024
        },
        {
            "tag": "gekleed",
            "score": 0.5933
        },
        {
            "tag": "lijkvlekken",
            "score": 0.586
        },
        {
            "tag": "schotwond",
            "score": 0.5816
        }
    ],
    "img": "edu1.png",
    "img_caption": "diagram",
    "img_tags": [
        {
            "tag": "schets",
            "score": 0.9955991506576538
        },
        {
            "tag": "tekening",
            "score": 0.9904026985168457
        },
        {
            "tag": "illustratie",
            "score": 0.8790888786315918
        },
        {
            "tag": "tekenfilm",
            "score": 0.798235297203064
        },
        {
            "tag": "inkt",
            "score": 0.7358826398849487
        }
    ],
    "type_name": "een educatief lesprogramma"
},
{
    "ctx_title": "Getuigenverklaringen",
    "ctx_url": "https://www.nemosciencemuseum.nl/nl/onderwijs/voortgezet-
        onderwijs/lesmateriaal/lesmateriaal/biologie-voor-vo-
        biowetenschappen-en-maatschappij/",
    "ctx_caption": "-",
    "ctx": "5. Het komt ook voor dat een getuige zich iets herinnert wat
        helemaal niet is gebeurd. Lees het stukje tekst hieronder.\n\
        nStopbord of voorrangsbord? \nOnderzoekers lieten proefpersonen
        naar een filmpje kijken waarin een rode auto stopt voor een
        driehoekig voorrangsbord. Daarna werd aan de proefpersonen o.a.
        gevraagd of de auto die stopte voor het stopbord nog door een
        andere auto in werd gehaald. Een groot deel van de proefpersonen
        bleek zich later 'te herinneren' dat er inderdaad een stopbord
        was. (Loftus et al. 1978)",
    "ctx_tags": [
        {
            "tag": "filmpje",
            "score": 0.6384
        },
        {
            "tag": "helemaal",
            "score": 0.5796
        },
        {
            "tag": "gehaald",
            "score": 0.5742
        },
        {
            "tag": "proefpersonen",
            "score": 0.5731
        },
        {
            "tag": "waarin",
            "score": 0.5726
        }
    ],
    "img": "edu2.png",
    "img_caption": "pictogram",
    "img_tags": [
        {
            "tag": "stop",
            "score": 0.9955991506576538
        },
        {
            "tag": "bord",
            "score": 0.9904026985168457
        },
        {
            "tag": "abstract",
            "score": 0.8790888786315918
        },
            {
                "tag": "tekst",
                "score": 0.798235297203064
            },
            {
                "tag": "vormgeving",
                "score": 0.7358826398849487
            }
        ],
        "type_name": "een educatief lesprogramma"
    },
    {
        "ctx_title": "Het spel van de Gouden Eeuw",
        "ctx_url": "https://www.nemosciencemuseum.nl/nl/onderwijs/voortgezet-
            onderwijs/lesmateriaal/lesmateriaal/overig-voor-vo/",
        "ctx_caption": "Foto: replica van de Prins Willem / Bron: Dirk van
            der Made",
        "ctx": "Na een aantal succesvolle reizen van en naar Indie, verging
            VOC-schip Prins Willem in 1662. De volledige bemanning liet het
            leven en de hele lading ging verloren.",
        "ctx_tags": [
            {
                "tag": "verging",
                "score": 0.5684
            },
            {
                "tag": "schip",
                "score": 0.5318
            },
            {
                "tag": "liet",
                "score": 0.5314
            },
            {
                "tag": "succesvolle",
                "score": 0.5273
            },
            {
                "tag": "volledige",
                "score": 0.5233
            }
        ],
        "img": "https://upload.wikimedia.org/wikipedia/commons/7/7c/
            Sail_amsterdam_05_stern_prins_willem.jpg",
        "img_caption": "een groot schip aangemeerd",
        "img_tags": [
            {
                "tag": "lucht",
                "score": 0.9934
            },
            {
                "tag": "openlucht",
                "score": 0.9599
            },
            {
                "tag": "boot",
                "score": 0.9402
            },
            {
                "tag": "transport",
                "score": 0.9290
            },
            {
                "tag": "schip",
                "score": 0.8687
            }
        ],
        "type_name": "een educatief lesprogramma"
    },
    {
        "ctx_title": "Dronken man steelt tractor en laat in Duitse plaats
            spoor van vernieling achter",
        "ctx_url": "https://nos.nl/artikel/2367020-dronken-man-steelt-tractor
            -en-laat-in-duitse-plaats-spoor-van-vernieling-achter",
        "ctx_caption": "KREISPOLIZEIBEHORDE OBERBERGISCHER KREIS",
        "ctx": "Een dronken 23-jarige man uit het Duitse Radevormwald heeft
            gisteravond een tractor gestolen en daarna een spoor van
            vernieling in de gemeente achtergelaten. De tractor was uitgerust
             met een vier meter brede sneeuwschuiver, schrijven Duitse media.
             Hij heeft voor bijna 100.000 euro aan schade aan gebouwen en
            voertuigen aangericht in de gemeente in de deelstaat Noordrijn-
            Westfalen. Hij ramde onder meer een gevel van een huis, een
            schutting, een garage en vijf auto’s. Bij de botsing met een auto
             raakte een 46-jarige bestuurder lichtgewond. Rond 19.30 uur
            kreeg de politie oproepen binnen van het tumult. De man liet de
            tractor achter op een parkeerplaats van een supermarkt. De
            politie vond de man daar, nog steeds onder invloed, in de buurt
            van de tractor.",
        "ctx_tags": [
            {
                "tag": "ramde",
                "score": 0.4701
            },
            {
                "tag": "jarige",
                "score": 0.4576
            },
            {
                "tag": "raakte",
                "score": 0.4402
            },
            {
                "tag": "sneeuwschuiver",
                "score": 0.4396
            },
            {
                "tag": "schade",
                "score": 0.4339
            }
        ],
        "img": "https://cdn.nos.nl/image/2021/02/02/712171/2560x1440a.jpg",
        "img_caption": "een tractor met aanhangwagen",
        "img_tags": [
            {
```

```
            "tag": "band",
            "score": 0.9471
        },
        {
            "tag": "landvoertuig",
            "score": 0.9385
        },
        {
            "tag": "wiel",
            "score": 0.9239
        },
        {
            "tag": "voertuig",
            "score": 0.9095
        },
        {
            "tag": "auto-onderdeel",
            "score": 0.8290
        }
    ],
    "type_name": "een krant of nieuwswebsite"
},
{
    "ctx_title": "CO2-uitstoot door natuurbranden in 2021 hoogste ooit",
    "ctx_url": "https://nos.nl/artikel/2408650-co2-uitstoot-door-
        natuurbranden-in-2021-hoogste-ooit",
    "ctx_caption": "Bosbranden in het zuiden van Frankrijk, afgelopen
        zomer",
    "ctx": "Natuurbranden veroorzaakten het afgelopen jaar naar schatting
         ongeveer 1760 megaton aan CO2-uitstoot. Dat is een record voor
        wat betreft de uitstoot door bosbranden, berekende het aarde-
        observatieprogramma van de Europese Unie, Copernicus. Door
        klimaatverandering zullen er steeds vaker bosbranden zijn, denken
         de wetenschappers van het programma. \"Hoge temperaturen,
        heftige onweersbuien, harde wind en ander extreem weer gaan
        steeds vaker voorkomen\", schrijft het Copernicus. Ook Sander
        Veraverbeke, klimaatwetenschapper aan de Vrije Universiteit
        Amsterdam, ziet een duidelijke trend. \"Als ik terugdenk aan
        vorige zomer met al die grote branden in het Middellandse
        Zeegebied, in de Verenigde Staten en Canada, en in Oost-Siberie,
        ben ik niet verwonderd dat dat nu tot records leidt\", zegt hij
        in het NOS Radio 1 Journaal. Vooral in de zomer op het noordelijk
         halfrond was er sprake van veel uitstoot, meldde Copernicus
        eerder. In juli en augustus werden er al records verbroken.",
    "ctx_tags": [
        {
            "tag": "schrijft",
            "score": 0.5801
        },
        {
            "tag": "meldde",
            "score": 0.5627
        },
        {
            "tag": "wetenschappers",
            "score": 0.538
        },
        {
            "tag": "natuurbranden",
            "score": 0.5366
        },
        {
            "tag": "heftige",
            "score": 0.53
        }
    ],
    "img": "https://cdn.nos.nl/image/2021/12/08/810105/2560x1440a.jpg",
    "img_caption": "een groep mensen in een bos",
    "img_tags": [
        {
            "tag": "boom",
            "score": 0.9996
        },
        {
            "tag": "openlucht",
            "score": 0.9992
        },
        {
            "tag": "sneeuw",
            "score": 0.9764
        },
        {
            "tag": "skien",
            "score": 0.9289
        },
        {
            "tag": "brandweerman",
            "score": 0.7981
        }
    ],
    "type_name": "een krant of nieuwswebsite"
},
{
    "ctx_title": "Eerste torenvalk geboren in nest op balkon in Terneuzen
        ",
    "ctx_url": "https://nos.nl/artikel/2408650-co2-uitstoot-door-
        natuurbranden-in-2021-hoogste-ooit",
    "ctx_caption": "De eieren op het balkon",
    "ctx": "Het eerste ei van een torenvalk op een balkon in Terneuzen is
         uitgekomen. De andere eieren in het nest zijn nog heel. Eind mei
         bouwde een koppel torenvalken hun nest op het balkon. De
        torenvalk is een beschermde soort en dus liet de bewoner de
        vogels in alle rust broeden. Na ruim dertig dagen wachten, is de
        eerste torenvalk geboren, net nu de bewoner er niet is. Otto
        Peltzer, die op het huis past, ontdekte het kuiken vanmiddag. \"
        De moedervalk vliegt altijd weg als de balkondeur opengaat en
        daardoor zagen we dat er een ei was uitgekomen\", zegt hij tegen
        Omroep Zeeland. \nEcht bijzonder\n Volgens de Werkgroep
        Roofvogels Zeeland is de situatie op het balkon uniek. \"Het is
        echt bijzonder dat deze valken een balkon in de stad hebben
        uitgekozen. In Oost-Europa zie je het heel soms wel eens, maar in
         Nederland niet\", zegt een medewerker. De torenvalk staat sinds
        2017 op de lijst van bedreigde soorten. De verwachting is dat de
        rest van de eieren de komende dagen uitkomt.",
        "ctx_tags": [
            {
                "tag": "terneuzen",
                "score": 0.5301
            },
            {
                "tag": "uniek",
                "score": 0.4682
            },
            {
                "tag": "past",
                "score": 0.447
            },
            {
                "tag": "uitgekomen",
                "score": 0.4415
            },
            {
                "tag": "geboren",
                "score": 0.4401
            }
        ],
        "img": "https://cdn.nos.nl/image/2021/06/25/756503/2560x1440a.jpg",
        "img_caption": "een groep blauwe en witte ballonnen",
        "img_tags": [
            {
                "tag": "gras",
                "score": 0.9985
            },
            {
                "tag": "openlucht",
                "score": 0.9905
            },
            {
                "tag": "steen",
                "score": 0.9745
            },
            {
                "tag": "bal",
                "score": 0.7693
            },
            {
                "tag": "kerstboom",
                "score": 0.6417
            }
        ],
        "type_name": "een krant of nieuwswebsite"
    }
]
```

Bibliography

Ackland, Peter, Serge Resnikoff, and Rupert Bourne (2017). "World blindness and visual impairment: despite many successes, the problem is growing". In: *Community eye health* 30.100, p. 71.

Bernardi, Raffaella et al. (2016). "Automatic description generation from images: A survey of models, datasets, and evaluation measures". In: *Journal of Artificial Intelligence Research* 55, pp. 409–442.

Biten, Ali Furkan et al. (2019). "Good news, everyone! context driven entity-aware captioning for news images". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12466–12475.

Chu, Yuxin (2021). "Describing Images to Visually Impaired Users: a Requirement Elicitation Approach". In:

DIAGRAM Center - Making Images Accessible (2021). URL: http://diagramcenter.org/making-images-accessible.html.

Directive (EU) 2019/882 of the European Parliament and of the Council of 17 April 2019 on the accessibility requirements for products and services (2019). URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX\%3A32019L0882.

Felix, Cristian, Steven Franconeri, and Enrico Bertini (2017). "Taking word clouds apart: An empirical investigation of the design space for keyword summaries". In: *IEEE transactions on visualization and computer graphics* 24.1, pp. 657–666.

Feng, Yansong and Mirella Lapata (2010). "How many words is a picture worth? automatic caption generation for news images". In: *Proceedings of the 48th annual meeting of the Association for Computational Linguistics*, pp. 1239–1249.

— (2012). "Automatic caption generation for news images". In: *IEEE transactions on pattern analysis and machine intelligence* 35.4, pp. 797–812.

Ferrero, Guillaume (1894). "L'inertie mentale et la loi du moindre effort". In: *Revue Philosophique de la France et de l'Étranger* 37, pp. 169–182.

General guidelines (2016). URL: http://diagramcenter.org/general-guidelines-final-draft.html.

Giarelis, Nikolaos, Nikos Kanakaris, and Nikos Karacapilidis (2021). "A Comparative Assessment of State-Of-The-Art Methods for Multilingual Unsupervised Keyphrase Extraction". In: *IFIP International Conference on Artificial Intelligence Applications and Innovations*. Springer, pp. 635–645.

Grootendorst, Maarten (2020). *KeyBERT: Minimal keyword extraction with BERT.* Version v0.3.0. DOI: 10.5281/zenodo.4461265. URL: https://doi.org/10.5281/zenodo.4461265.

Gupta, Ankush and Prashanth Mannem (2012). "From image annotation to image description". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 7667 LNCS.PART 5, pp. 196–204. ISSN: 03029743. DOI: 10.1007/978-3-642-34500-5_24.

Hearst, Marti A et al. (2019). "An evaluation of semantically grouped word cloud designs". In: *IEEE transactions on visualization and computer graphics* 26.9, pp. 2748–2761.

Hodosh, Micah, Peter Young, and Julia Hockenmaier (2013). “Framing image description as a ranking task: Data, models and evaluation metrics”. In: *Journal of Artificial Intelligence Research* 47, pp. 853–899.

Hollink, Laura et al. (2004). “Classification of user image descriptions”. In: *International Journal of Human-Computer Studies* 61.5, pp. 601–626.

Hu, Xiaowei et al. (2020). “VIVO: Surpassing Human Performance in Novel Object Captioning with Visual Vocabulary Pre-Training”. In: *arXiv preprint arXiv:2009.13682*.

Kalsbeek, JWH and RN Sykes (1967). “Objective measurement of mental load”. In: *Acta Psychologica* 27, pp. 253–261.

Karpathy, Andrej, Armand Joulin, and Li F Fei-Fei (2014). “Deep fragment embeddings for bidirectional image sentence mapping”. In: *Advances in neural information processing systems* 27.

Kinghorn, Philip, Li Zhang, and Ling Shao (2019). “A hierarchical and regional deep learning architecture for image description generation”. In: *Pattern Recognition Letters* 119, pp. 77–85.

Kuznetsova, Polina et al. (2012). “Collective generation of natural image descriptions”. In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 359–368.

Lebret, Rémi, Pedro O Pinheiro, and Ronan Collobert (2014). “Simple image description generator via a linear phrase-based approach”. In: *arXiv preprint arXiv:1412.8419*.

Lewis, James R (1992). “Psychometric evaluation of the post-study system usability questionnaire: The PSSUQ”. In: *Proceedings of the human factors society annual meeting*. Vol. 36. 16. Sage Publications Sage CA: Los Angeles, CA, pp. 1259–1260.

— (2002). “Psychometric evaluation of the PSSUQ using data from five years of usability studies”. In: *International Journal of Human-Computer Interaction* 14.3-4, pp. 463–488.

LIA, Fondazione (2019). *E-books for all. Towards an accessible publishing ecosystem*. Fondazione LIA.

Maybury, Mani (1999). *Advances in automatic text summarization*. MIT press.

Nenkova, Ani and Kathleen McKeown (2012). “A survey of text summarization techniques”. In: *Mining text data*. Springer, pp. 43–76.

Nganji, Julius T, Mike Brayshaw, and Brian Tompsett (2013). “Describing and assessing image descriptions for visually impaired web users with IDAT”. In: *Proceedings of the Third International Conference on Intelligent Human Computer Interaction (IHCI 2011), Prague, Czech Republic, August, 2011*. Springer, pp. 27–37.

Oppermann, Reinhard (2002). “User-interface design”. In: *Handbook on information technologies for education and training*. Springer, pp. 233–248.

Ordonez, Vicente et al. (2016). “Large scale retrieval and generation of image descriptions”. In: *International Journal of Computer Vision* 119.1, pp. 46–59.

Orme, Richard, Valerie Morrison, and Huw Alexander (2020). *The Art and Science of Describing Images*. URL: https://daisy.org/news-events/articles/art-science-describing-images-w/.

Predictive justice: When algorithms pervade the law - paris innovation review (2017). URL: http://parisinnovationreview.com/articles-en/predictive-justice-when-algorithms-pervade-the-law.

Saez, Catherine (2020). *New European directive adds impetus to international efforts to promote accessibility*. URL: https://www.wipo.int/wipo_magazine/en/2020/02/article_0007.html.

Sauro, Jeff and James R Lewis (2016). *Quantifying the user experience: Practical statistics for user research*. Morgan Kaufmann.

Socher, Richard et al. (2014). "Grounded compositional semantics for finding and describing images with sentences". In: *Transactions of the Association for Computational Linguistics* 2, pp. 207–218.

Srinivasan, Krishna et al. (2021). "WIT: Wikipedia-based image text dataset for multimodal multilingual machine learning". In: *arXiv preprint arXiv:2103.01913*.

Takagi, Hironobu et al. (2003). "Accessibility designer: visualizing usability for the blind". In: *ACM SIGACCESS accessibility and computing* 77-78, pp. 177–184.

Tullis, Thomas S and Jacqueline N Stetson (2004). "A comparison of questionnaires for assessing website usability". In: *Usability professional association conference*. Vol. 1. Minneapolis, USA, pp. 1–12.

Ushiku, Yoshitaka et al. (2015). "Common subspace for model and similarity: Phrase learning for caption generation from images". In: *Proceedings of the IEEE international conference on computer vision*, pp. 2668–2676.

Vaswani, Ashish et al. (2017). "Attention is all you need". In: *Advances in neural information processing systems*, pp. 5998–6008.

Vries, Wietse de et al. (2019). "Bertje: A dutch bert model". In: *arXiv preprint arXiv:1912.09582*.

Zipf, George Kingsley (2016). *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio Books.