Improving the automatic generation of Dutch soccer reports using the PASS system

Bachelor Assignment Advanced Technology

Joost Sessink
j.w.m.sessink@student.utwente.nl

June 29, 2020

UNIVERSITY
OF TWENTE.

Human Media Interaction (HMI)
University of Twente

Bachelor assignment committee

Daily supervisor	dr. Lorenzo Gatti
Chairperson	dr. Mariët Theune
External member	dr. Herman Hemmes

Abstract

In this thesis, a process is discussed in which the goal is to improve the PASS system, a data-to-text system that uses match statistics to generate Dutch soccer reports. A study has been done where reports generated by PASS and reports written by reporters were compared. From this study, improvement points were found, some of which have been used to implement new features into PASS; the reports generated by PASS now mention substitutions, which was not present before, the way that players are mentioned has been adapted and many other smaller changes have been made. An evaluation with human participants has been performed to test these implementations, from which it resulted that in general, PASS was successfully improved.

Contents

1 Introduction

1.1 Problem

In this age where technology is ever developing, machines are often taking over tasks that were previously always fulfilled by humans. This automation process of course has the goals to decrease the amount of human labour and to increase the cost- and time-effectiveness of the tasks that need to be fulfilled. However, something that still seems to be done mainly by hand, is sports reporting. Sports are still as relevant as they have ever been [1], so it would only make sense if the process of writing match reports was also automatized in order to save time and resources.

Therefore, a lot of technologies have been developed that are able to automatically generate a match report based on match statistics [2]. These match statistics are for instance the final score of the game, the names of the athletes who scored the goals and at what times these goals were scored.

However, these reports are mostly created with the purpose of giving information as fast as possible. This leads to the problem that people who are interested in the match, especially the fans of one of the teams involved in the match, do not enjoy reading the report as much as they would when an actual reporter would have written it instead of a computer program. This is of course not desirable, because these reports are created for sport enthusiasts after all.

1.2 Current solution

Most of the current text generation systems that already exist are unfortunately closed systems which cannot be accessed by the general public at the moment. For example Wordsmith [1] and Quill [2] are two systems that are currently unavailable for research.

PASS, on the other hand, is an open-source system which is freely available. PASS, which stands for **P**ersonalized **A**utomated **S**occer texts **S**ystem, is a project that has been created by the Tilburg center for Cognition and Communication (TiCC) at Tilburg University [3], who are now collaborating with the Human Media Interaction (HMI) research group at the EEMCS faculty at the University of Twente in the Affective Language Production project (ALP). The PASS system tries to make the generated sport reports more personalised by tailoring them to the fans of the clubs that participated in the games.

It has been shown before that tailoring generated texts towards different audiences is possible. For example, it has been shown that data about the well-being of prematurely born babies in neonatal care can be transformed into different texts, based on who is going to be the reader [4]. For instance, a text that is going to be read by the parents of the child, should contain a lot less technical details and medical language than the text that is meant to be read by the doctor that is giving treatment to the baby.

The PASS system follows this idea as well and uses match data to write 3 different reports, one aimed at the fans of the home team, one at the fans of the away team and one neutral report. The report for the winning team will for instance focus more on the goal(s) scored by the winning team and compliment the team's actions, while the report for the losing team will try to defend the losing team and put things into perspective.

[1] https://automatedinsights.com
[2] https://narrativescience.com

1.3 Research question

During the development of the PASS system, the goal was to make the reports that were generated resemble actual written soccer reports as much as possible. However, the PASS system is still not perfect. This gives rise to the following research question:

Can the automatic generation of Dutch soccer reports by the PASS system be improved?

The hypothesis is that the automatic generation of Dutch soccer reports by the PASS system can be improved, because at first glance, they seem not as complete as would be expected when reading a human-written report.

1.4 Activities

Firstly, many reports were generated using PASS and compared to actual reports written on that match. These results were evaluated and improvement points of the system were found. After these improvement points were found, they were used as a basis for further research on improving the PASS system. Then it was time to get familiarised with the PASS program and implement changes that would improve the text generation. Finally, an evaluation study was performed to check the performance of the new version of the program. At last, all developments in the PASS program and its conclusions were put together in this report.

1.5 Thesis structure

This thesis is structured in the following manner: In chapter 2 some background on the topic is given. In chapter 3, the comparing of the reports generated by PASS and the reports written by human reporters is described. In chapter 4, the implementations that were made to the PASS system are discussed and in chapter 5, the evaluation study is described. All points of discussion and recommendations for further work are mentioned in chapter 6, and at last, in chapter 7, the final conclusions are given.

2 Background

The goal of this assignment is to improve the PASS system that was introduced in chapter 1. In order to do this as efficiently and effectively as possible, knowledge about Natural Language Generation (NLG) is required. Then, it is obviously required to have a good understanding of the PASS system itself. Finally, in order to know how to do an evaluation study and which evaluation study will be most useful to perform, also a better understanding of these kinds of studies is necessary. These topics will be discussed in the following subsections.

2.1 Natural Language Generation

It is difficult to give one specific definition that fully captures the essence of Natural Language Generation [5]. Gatt and Krahmer defined NLG as "the task of generating text or speech from non-linguistic input" [6]. Natural Language Generation can be divided into two different instances:

- Text-to-text generation
- Data-to-text generation

As the names suggest, text-to-text generation requires textual input whilst data-to-text generation uses very variable types of input. This input data can be everything, such as weather data obtained by real-world sensors [7], or, in the case of PASS, data on all events that happened in a soccer match. Because of the fact that PASS falls under the latter category, this subsection will focus on data-to-text generation only.

The Natural Language Generation process is often divided in six different subproblems, which are: content determination, text structuring, sentence aggregation, lexicalisation, referring expression generation and linguistic realisation [8]. Each of these 6 different subproblems is shortly explained below [6].

2.1.1 Content determination

At first, all of the input data needs to be processed and filtered, such that only the vital information or the information that is interesting to be put into the generated text will be used [9]. So, choices need to be made on what data will be used by the NLG system. These choices will be based on what goal this system is trying to accomplish. If for instance the generated texts should be as concise and to the point as possible, then only the most vital data should be selected. However, if the goal of the generated text is to also entertain the reader, then other data should also be selected that is predicted to be enjoyed by the potential readers.

2.1.2 Text structuring

When the content is determined and thus it is known what data is being worked with, the NLG system needs to determine in what order it wants to present its messages. This again depends on the kind of application of the NLG system. For instance, if the order in which certain events in the data happened is important, which would for example be the case when reporting a sports match, then presenting this data in a chronological order would make sense. However, if time is not that

important and the data contains a lot of information, it might be a better idea to group the data on relatedness, i.e. find data that falls in the same category and group this data together.

2.1.3 Sentence aggregation

Not every selected piece of data necessarily needs to map one-to-one to a sentence. To attempt to get rid of redundancies in the phrasing of sentences, multiple data points can be combined into one subject. This process is called sentence aggregation [10]. An example of sentence aggregation in the domain of sports reporting: imagine a player scoring two goals very shortly after each other. Without sentence aggregation, this would be reported in the lines of:

"The player scored a goal after 23 minutes. The player scored a goal after 26 minutes."

As can be seen, this is quite redundant, and could be seen as not very pleasant to read. With sentence aggregation, this same chain of events would be reported similarly to:

"The player scored two goals in the span of just 3 minutes."

In this example, this sentence contains no redundancy, as opposed to the first example. By combining multiple data points into one sentence, the generated sentence could become more pleasant to read [11].

2.1.4 Lexicalisation

Up until this point in the NLG process, exclusively data is considered. Data is selected, structured and potentially some data is aggregated. However, no human language is created yet. That is where lexicalisation will make a start [6]. The processed data can be thought of as being the building blocks of the sentences to be generated. In the lexicalisation process, the words or phrases that can be used to describe the events consisting of these building blocks will be determined. This is not as easy as it might seem, because very often, a single event can be described in lots of different ways in human language. For instance, when reporting a player receiving a yellow card in a soccer match, this action could for example be described as:

"The player received a yellow card"

Or

"The player made a foul and got punished by receiving a yellow card"

Or

"The referee decided to book the player: yellow card"

So, this means that the NLG system has to choose which language to use to describe these events. This choice again has to be made on the basis of the preferred goal of the system. If the main goal is to generate concise reports, it should use short language with as much information as possible. But if the main goal is to generate reports that entertain the readers, it should use language with more variety and perhaps more words or phrases that transition the text from one action to the other.

2.1.5 Referring expression generation

In the lexicalisation process, the sentences that will be the content of the generated final text are constructed. However, these sentences will still contain some gaps that need to be filled in. In the referring expression generation process, these gaps will be filled in by looking at the data and seeing which variables belong to the missing information in the gaps.

The filling of these gaps will depend on the type of output that is desired, similar to the previously mentioned processes. For example, for short and concise reports, the variables will be referred to in a short and efficient way, and with the least amount of words possible that are necessary to successfully distinguish that variable from potential other variables, whilst in the reports that are meant to entertain, the mentioning of the variables will contain more variety or extra information about that specific variable that is not necessarily required to distinguish it, but will increase the enjoyment of those who read the text. To give an example, in a soccer report, 'Lionel Messi' would be enough to successfully identify the player that is being referred to. However, to create more variety, he could also be referred to as something along the lines of: 'The small Argentinian playmaker' or 'FC Barcelona's bearded captain'. With these references, the reader is also able to identify that it is Lionel Messi who is being discussed (assuming that people who read soccer reports about a soccer club have some knowledge about that club and its players), without having to use the same phrase each time he is mentioned. This brings more variety to the generated text.

2.1.6 Linguistic realisation

The last step is the linguistic realisation process. The NLG system is almost finished. The relevant phrases are there, including the relevant words, the only thing that needs to be done is to combine these into a nicely structured sentence. Turning the defined messages into a sentence that resembles an actual human-written sentence is more complex than it might seem on first hand. For instance, the components of a sentence need to be ordered, the verbs have to be modified into their right form depending on the context and possibly punctuation needs to be included [6]. One of the more important problems that the linguistic realisation process has to deal with, is that, in order to make a sentence with correct human language, the output has to contain several components that possibly did not exist in the input.

There are different existing methods, each having their own way at approaching this issue. Some of the most frequently used approaches are statistical approaches, which use statistics on large datasets of human-written content to determine the best generated text [12] [13] [14]. There are differences between the statistical approaches, but the main idea of all these methods is to decrease the human labour involved and increase the overall data coverage, by letting the computer do the work.

Other popular approaches are human-coded grammar-based systems. The idea of these systems is that they make their choices on how to form the sentences based on the grammar of the language that these sentences should be in. Often, this grammar is hand-written by the developers of the NLG software [15] [16]. These kind of approaches contain more human labour than the statistical ones, but this also means that the developer has more direct control on the output, and that it is a bit easier to correct mistakes in the program.

The final popular approach of the linguistic realisation process that will be discussed in this subsec-

tion is the approach that PASS uses, which is the template-based approach. The template-based approach is arguably the most straightforward. This method uses a lot of predefined categories that contain one or multiple sentences that can be used to describe a specific action that falls into that category. These sentences are usually manually defined, but it has also been shown that it is possible to automatically generate template sentences [17]. When an action is described, it is first determined to what kind of category it belongs. Then, one of the sentences is (randomly) selected from that category and used in the output text [18].

This idea can be explained by an example in the soccer report generation context: Suppose that Cristiano Ronaldo scores a goal in the second minute of an arbitrary soccer match. This event could for instance be classified just as 'goal', because the event describes a goal that was scored. A potential template sentence that would describe this action could be:

"<goal scorer> scored a goal in the <minute> minute."

This sentence would for instance be turned into:

"Cristiano Ronaldo scored a goal in the second minute."

But, this action could also be classified as an 'early goal', because the event describes a goal that was scored early in the match. A potential template sentence that would describe this action could be:

"<goal scorer> scored a very early goal, he only needed <minute> minutes",

This sentence would for instance become:

"Cristiano Ronaldo scored a very early goal, he only needed two minutes.".

It is now up to the program to choose which category fits this action best. In this case of a sports reporting system, the latter category would be preferred, because the more details that can be mentioned about an action, the more interesting the text would probably become for the reader. But this is an implementation decision that would again depend on the preferred goal of the output text.

The big advantage that comes when using a template structure, is that you have total control over the language that the system uses, because these template sentences are most often created by hand. If a sentence seems to contain an error, it is very simple to modify that sentence. The biggest disadvantage is however, that if all template sentences are indeed created by hand, that the implementation of these templates will be a lot of work, and perhaps even practically impossible for very big projects.

2.2 The PASS system

As already seen in chapter 1, PASS is an NLG system that automatically generates Dutch soccer reports using match statistics. The interesting thing about PASS is that these reports can be tailored towards the fans of the clubs that participated in the match. PASS is furthermore also capable of writing neutral reports. In this subsection, the PASS system is elaborated upon in more detail [3].

2.2.1 Data

PASS gets the data of the matches from the Gracenote Sports API [3]. Using this API, PASS gets access to the vital data that it needs in order to produce the reports, such as the score of the game and the goal scorers, but it also contains some additional data that can be used to give more variety to the text, such as the nicknames of the players or the name of the stadium where the game is being played in.

In order to make the reports that PASS produces resemble human reports written by reporters, the language used in PASS' templates should be similar to the language used by sport reporters. Therefore, most of the templates used by PASS are inspired by sentences used in reports that were found in the MeMo FC corpus [19]. This corpus is a collection of match reports from the two Dutch professional leagues, 'de Eredivisie' and 'de Eerste divisie' among others. These reports originate from the websites of the clubs participating in the matches. That means that these reports are written for the people who are supporters of these clubs, and will thus contain a preference for these clubs.

2.2.2 Report structure

Using the data mentioned above, PASS generates the soccer reports. Its reports always follow a certain structure. This structure can be divided into four different parts:

Title Mentions whether the match ended in a win, a loss or a draw for the focus team. Also potentially mentions the final score.

Introduction Shortly summarizes the match by for instance mentioning which team won or whether that team won deservedly or not. Also sometimes gives some extra information such as the stadium and the number of attendees.

Game course The body of the report, that also contains the most text. The game course mentions important happenings in chronological order. These are for instance the goals scored, but also a missed penalty can be mentioned.

Debriefing The debriefing mentions all the yellow and/or red cards that were given in the match.

An example of such a report generated by PASS can be seen in Table 1.

[3] https://gracenote.com

Table 1: Example PASS report on the match Roda JC - Ajax of 31/01/16, tailored towards the fans of Ajax

Ajax laat twee punten liggen in Kerkrade
In de wedstrijd van afgelopen zondag heeft Ajax goed stand gehouden tegen Roda JC Kerkrade. Tegen de thuisploeg stond er een 2-2 eindstand op het scorebord.
Milik schoot na 27 minuten op aangeven van middenvelder Gudelj raak. El Ghazi tekende in de 29e minuut voor de 0-2. Na 35 minuten schoot aanvaller Poepon raak, daarna was Maecky Ngombo succesvol voor het doel: 2-2.
Arbiter Serdar Gözübüyük was genoodzaakt 2 gele kaarten te geven, aan Riechedly Bazoer en Davy Klaassen.

2.2.3 System design

Now that the data that PASS uses as input and the structure that it generates as output has been discussed, it is time to discuss how exactly PASS goes from input to output.

The PASS program is all written in the programming language Python. It uses a modular design, meaning that it consists of different modules that all have their own function. The advantage of using a modular design, is that it is nicely organised and that it is therefore easy to make adjustments to the functionality of a module, without negatively interfering with another module. This modular design is represented in Figure 1. The **governing module** is printed in bold. This is done consciously, because this is the central module in which everything is structured. All the other modules are called in the right order from within the governing module. The dashed arrows indicate that there is no direct communication between these modules, but they are there to show the order in which the modules are called from within the governing module.

The text generation process for the four different parts of the report structure is very similar. The only difference is that the content order in the title, the introduction and the debriefing is always fixed, whilst in the game course section, this is not the case. For instance, in the introduction, it is always mentioned first whether the team won, lost or drew, and then always the final score is mentioned. For the game course section, this is not fixed, as the game course section always follows a chronological order. The contents of the game course thus always depend on which events happened in the match, and when they happened.

So, the process corresponding to the text generation for the game course section requires one extra initial step compared to the text generation process for the other sections. This is done in the **topic collection module**. In this module, the match data is scanned and the important events that should be present in the report are extracted. These events are then placed in chronological order, such that there is a list containing events in the order that they should be commented on. The events that should be commented on in the other sections, are already fixed, so for these sections, the Topic collection module is not required.

At this point in time for each report section, PASS knows the events to comment on and in which order to do so. Now, for each of these events, the **lookup module** is used to look at all templates

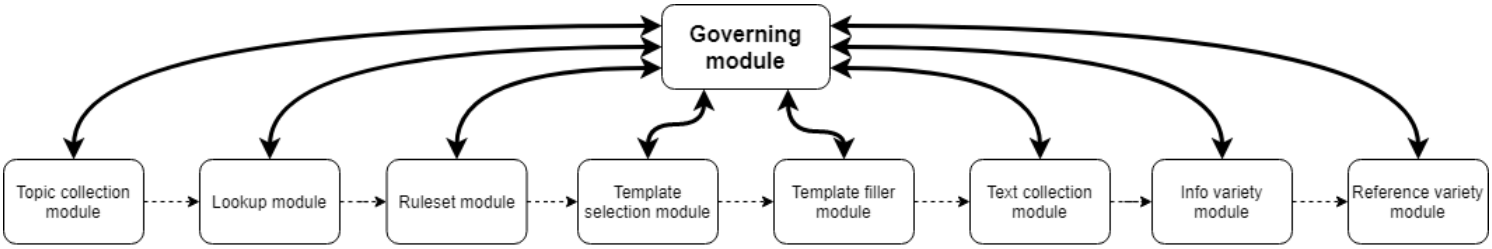


Figure 1: The modular design of PASS.

and collect the possible templates that can be used to describe that event. For instance, if the event is a goal scored for the focus team, then the topic collection module will return all possible templates describing a goal scored by the focus team.

However, perhaps not all templates that correspond to that event are suitable. For instance, a template sentence that describes an equalizing goal being scored and a template sentence that describes a winning goal being scored both correspond to a goal scoring event, but they clearly cannot both be used to describe the same goal scored. Thus, in order to filter the templates collected by the lookup module, PASS uses the **ruleset module**. In this module, for all template categories, it is checked whether they are suitable to describe the current event or not. The governing module will then add all possible templates for that event to a list.

Then, a choice has to be made about which of these possible templates to use in the report. This is done in the **template selection module**. In this module, a choice will be made between the possible templates that were received from using the ruleset module. These templates are assigned a weight corresponding to how detailed they are. A template that describes more information about an event would be preferred, and would therefore be weighted higher. For instance, if the event to be described is a goal scored, then if that goal was for example the winning goal, if it was assisted by another player, or potentially both, a template describing these details will be weighted higher than a template that generally describes a goal scored that could be used in every situation. So, the chance that a more detailed template that is more suited to the current event is chosen is higher than the chance that a less detailed, more general template is chosen, but the more general templates still have some chance to be chosen.

When a template sentence has been selected, it is still not ready to be put in the final text. The sentences still have gaps that need to be filled in. An example of what these incomplete sentences look like could already be seen in subsection 2.1.6. These gaps represent all the data that is dependent on the match for which the report is generated. In the **template filler module**, these gaps are filled in. For each gap, the module goes through the match data to find the information that corresponds to that gap, and then fills it in. To give another example, a template sentence, for the debriefing section, could look like this:

"<referee> had to give <yellow card player> a yellow card."

For the <referee> tag, the process is simple. The template filler module goes through the line-up of the people who participated in the match, until it finds the referee. It chooses a proper way to refer to the referee, and then fills the gap in the sentence. For the <yellow card player> tag, the

template filler module has to look through the match line-up again, until it finds the player that got the yellow card. It then again finds a good way of referring that player, which, in the case of the debriefing is just his full name, and fills in the gap. The returned sentence will then look something like this:

"Referee Antonio Mateu had to give Sergio Ramos a yellow card."

Then the governing module calls the **text collection module**. The only thing that this module does is simply collect the texts for all four report sections and then put them in the right order.

After this step, the report is basically done. All the text has been generated, there are no gaps left and everything is in the right order. However, in Figure 1 there are still two more modules that can be seen after the text collection modules. As the names of these modules suggest, they are used to improve the overall variety of the generated text.

The **info variety module** has the task of filtering out pieces of information that may be redundant. During the template selection process, there may be templates that describe different events, but use the same pieces of information. For instance, if in the introduction, the number of people that were in the crowd is mentioned, and that same information is mentioned again somewhere in the game course section, that would be quite redundant. With the info variety module, this does not happen anymore.

Finally, there is the **reference variety module**. This module is very similar to the info variety module, but instead of information repetition, this module deals with reference repetition. For this version of PASS, the reference variety module goes through the text and tries to find if in two subsequent sentences, two equal references have been found that refer to the same entity. If this is the case, the reference in the second sentence will be changed to another way to refer to that entity. This is a relatively simple implementation that works for this version of PASS, but might become too simple if the PASS system is expanded to create bigger and more complex texts.

2.3 Evaluation

The ultimate goal of this assignment is to improve the PASS system. In order to be able to tell whether this has actually happened, an evaluation study must be done. But how does one evaluate an NLG system? In this subsection, that question is answered. There are actually a lot of ways to evaluate an NLG system [20]. In order to narrow the possibilities down, there are three types of evaluations with different principles that will be discussed in this subsection:

- task-based evaluation
- human ratings evaluation
- metrics evaluation

For task-based evaluation, the principal idea is to test the NLG system by letting users try it, and then measuring afterwards whether a predefined goal has been reached or not. These evaluations can be done in a real-word environment, where the users use the systems according to how and when they are supposed to be used, or these evaluations can be done in an experimental laboratory environment.

For human ratings evaluation, the principal idea is to let users read texts generated by NLG systems and rate them afterwards on properties such as readability or enjoyability. Also these kinds of evaluations can be done in both a real-world environment and an experimental laboratory environment.

Finally, for metric evaluations, the principal idea is to not let humans evaluate the NLG system, but to compare the generated texts to reference texts that were written by humans. These comparisons can be done on a large amount of varying metrics.

In the following subsections, each of these three evaluation types is discussed in further detail.

2.3.1 Task-based evaluation

Task-based evaluations are generally preferred over other kinds of evaluation studies. This is due to the fact that in this case, the system is really being tried out by users, and it can be directly measured if the system has the desired effects.

For these kinds of evaluations, it is important to establish a hypothesis before the evaluation is being conducted [21]. This hypothesis is completely dependent on the goal of the system. For instance, if the system under evaluation is a system that generates text to give car drivers feedback on their safety when driving their car [22], a potential hypothesis for task-based evaluation could be:

"This system will result in safer driving behaviour for the participants."

The variable that should be measured in this example experiment is the driving behaviours of the participants. This could for instance be measured by looking at the number of dangerous situations that these drivers get into whilst they are in their car. This should be measured for two different participant groups, one using the NLG system, and one not using the NLG system, which is the control group. Afterwards, the results for these different groups should be compared, and based on these results, the hypothesis can be accepted or rejected.

The participants should in general be representative of the (desired) userbase of the NLG system that is to be tested, because if people who would usually never use such a system would participate in a task-based evaluation, this could negatively influence the results.

As mentioned before, task-based evaluation can be conducted in a real-world context as well as in an experimental laboratory context. Both of these types have their advantages and disadvantages.

In general, doing task-based evaluations in a real-world environment gives better results compared to task-based evaluations in an experimental laboratory environment, and is therefore generally preferred [20]. This is due to the simple fact that for a real-world experiment, the participant is using the system in the context for which it is meant, while in a laboratory setting, the participant is using the system in an artificial environment, which means that he or she may use the system in a (slightly) different manner than when this system was deployed in a real-world environment.

However, the increased accuracy that comes with the real-world testing also brings downsides with it. In an experimental laboratory context, the researchers have very close control of the participants and their actions. This guarantees that the tasks are fulfilled according to how the researchers would like them to be fulfilled. This prevents a lot of noise, that will almost inevitably be present

when doing real-world experiments, because for these kinds of experiments the participants are not closely supervised. An increase in noise means that, in order to still get meaningful results, more experiments should be done to filter out this noise. This means that real-world evaluation studies take more time and effort to complete, and are in most cases more expensive as well.

2.3.2 Human ratings evaluation

Sometimes, a task-based evaluation study is not feasible, which can have many reasons. The study might cost too much time, effort or money or it requires an unrealistic amount of participants. In these cases, a human ratings evaluation study is a very viable alternative [20].

These evaluations follow a reasonably simple principle: the participants get to either use the NLG system in real-world context or read a couple of texts generated by that system in an artificial context, and are afterwards asked to answer some questions. These questions are usually in such a format, that they can be answered using Likert scales. Likert scales are scales on which the participant can indicate, usually on a scale from 1 to 5 or 1 to 7, to what extent he or she agrees with the statement [23]. Another question variant is to present the participant with two different texts, for instance an NLG text and a human-written control text, and to let the user indicate which one he or she prefers.

The requirements of human ratings evaluations are very similar to those of task-based evaluations. It is very important to establish a proper hypothesis beforehand and the subjects that participate in the evaluation study should be representative of the (desired) userbase of the NLG system.

Human ratings evaluations can also be done in both a real-world environment as well as in an experimental laboratory environment, with both their advantages and disadvantages.

Again, the real-world evaluations are generally preferred over experimental laboratory evaluations. This is because also in this case, evaluating a system by using people that are actually using that system, will give more realistic results compared to using people who are just asked to evaluate texts after reading them [24].

However, the same drawbacks as mentioned for the real-world task-based evaluations apply for the real-world human ratings evaluations. It will cost more time, effort and potentially more money to perform such a study than to perform an experimental laboratory human ratings evaluation study.

The real-world experiments are performed relatively straight-forward. The participant just uses the system and gets asked a couple of questions about it. The laboratory experiments however require a bit more preparation. In this case, the participant is not asked to use the NLG system, but he or she is asked to read a couple of texts and answer questions about those. Thus it needs to be determined which texts the participant is being shown. It can be chosen to show the participant completely random texts, in order to keep the researcher from influencing the results, but it can also be chosen to select specific texts that are selected to cover a range of important phenomena that the researcher wants to be included [25]. Both options are viable approaches, depending on what the goal of the study is. If the main goal is to test the average performance of the system, then it is better to show completely random texts, but if the main goal is to test the performance of a system for a certain context, then it would be better to show texts that correspond to that context.

2.3.3 Metrics evaluation

Finally, there is a third important type of evaluation study, which is the metric evaluation. The big difference with this evaluation study compared to the other studies mentioned in this subsection, is that it requires no human participants. Instead, it compares the texts generated by the NLG system and compares it to a human-written reference text, from which it is known that it is of high quality [26]. Put really bluntly, the more the generated text resembles the reference text, the better.

In reality, the metrics evaluation is a bit more complicated. There exist a couple of different metrics, which use different techniques to determine the similarity between a generated text and a reference text. Some examples of well-known metrics are BLEU [27], ROUGE [28] and METEOR [29]. Doing an evaluation study through metrics is without a doubt the easiest and least expensive way to evaluate an NLG system. However, metric evaluations also have some serious drawbacks.

First of all, it has to be made absolutely sure that the reference text is of a high quality level. Also, even if the metrics evaluation gives a very positive results, it is not absolutely sure that the NLG system has achieved its goal. Suppose that for example the goal of an arbitrary NLG system is to change something in the behaviour of people, for instance an NLG system that tries to help people eat more healthy in order to lose weight. Suppose that this system scores really high on a metrics test, indicating that it produces really high quality texts, but that the majority of the users of the system is not losing any weight. Then the system simply does not work well enough. The texts generated by the system can be of a very high quality, but if the system does not achieve its goals, then it still a bad system.

Also, when an NLG system performs poorly on a metrics test, this would indicate that the produced text is of bad quality. However, this is not necessarily the case, a poor score just means that the produced text is not very similar to the reference text. A text can be very different from the reference text, but still be of high quality, because there are just many ways to express a certain message and many different words can be used to say the same thing. This problem can be tackled by including multiple reference texts which are known to be of good quality. This still does not solve everything, because it is still possible that the output text of the NLG system contains language which is correct and appropriate for the situation, but is not present in the reference texts.

The validity of a number of these metrics has been put to the test in the past [30] [31]. The results of these tests were overall not that positive. It has thus far not been succeeded to officially validate the use of metrics to evaluate an NLG system. That is why in general, metrics should only be used for quick initial feedback to developers of the system. When these developers want a real meaningful evaluation of their system, they should perform an evaluation study using real human participants.

3 Comparing PASS' reports with human-written reports

Before the PASS-system can be improved by making modifications, it is important to have an idea about how real reports are different from PASS'reports. Thus, a lot of reports had been generated by PASS, after which they were compared to reports written on the same match, from the same point of view, by official reporters. These human-written reports were taken from the official websites of the teams that participated in these matches. In this chapter, the study and its results will be discussed.

3.1 Study set-up

Ten different matches were used in this study. Five of these matches were randomly selected from the Eredivisie 2018/2019 season and the other five matches were randomly selected from the Eredivisie 2015/2016 season. These matches have been selected at random to try to cover all the different match scenario's. For instance, some of the selected matches were very exciting for both involved parties, for example a match that ended in a 2-2 draw. In this case, both teams probably have some positive and negative points that can be mentioned in the match report, which makes the report different from a report that is being written about a match that ended in a 5-0 score. Such a report would be very one-sided and also very different depending on whether the report is written for the fans of the winning or the losing team.

For every match, two reports were generated by PASS, one report aimed towards the fans of the home team, and one report aimed towards the fans of the away team. Even though PASS is also capable of generating neutral reports, these were not considered in this study. The reason is that the main goal of the PASS system, and what sets it apart from other NLG systems that automatically generate sports match reports, is to tailor the generated report towards the fans of the clubs that are involved in the matches, and that is why the focus will be on these reports instead of the neutral ones. So in total, 20 reports were generated by PASS. For each of these reports, a corresponding report written by a human reporter was found on the participating clubs' website, meaning that in total, this experiment contains 40 reports.

Then all the reports generated by PASS were paired with the corresponding human-written report. First, the human-written report was read and then the PASS version was read. After reading both reports, general comments on the differences between these two reports were written down. After this had been done for all 20 report pairs, all the individual comments were analyzed, and from these comments a list containing general comments was derived.

3.2 Results

The full general comments on the differences have been summarized into a table with the key points, which can be seen in Table 2. This table contains the most notable differences between the reports written by reporters and the reports generated by PASS, together with the presumable reason of the difference. Each of these differences will be briefly discussed below.

The first thing that was noticed when looking at the reports, is that the real reports were all longer, i.e., contained more words than those generated by PASS. This is probably due to the contents of the reports. PASS' reports mentioned only goals in its game course section (they also mentioned missed penalty kicks, but this is a very rare event), whilst the reporters often mentioned other facts

Table 2: Summary of differences between reports generated by PASS and written by reporters

Difference	**Presumable reason**
The human reports are significantly longer than those of PASS	Human reports use more filler text, PASS almost exclusively comments on goals
PASS only comments on goals or missed penalties, human reports cover a lot more	PASS does not have as much access to data as reporters do, but PASS also does not use all its data yet
In PASS, there is a lot of 'guesswork'	Real reporters observed all match events in detail, PASS only has limited resources
PASS sometimes repeats sentences, human reports do not	If a certain event happens often, PASS does not have enough templates to still be varied
Real reports often mention statistics outside of the match, such as mentioning previous results or upcoming matches, PASS does not	PASS does not have access to this data
PASS' player referring feels random, human reports' player referring looks more structured	Not perfectly implemented in PASS

in their texts, such as substitutions, big scoring chances or general comments on for instance which team has the upper hand at a certain point in the match.

This immediately introduces the next difference, which is indeed that human-written reports contain more information. This has two possible explanations, depending on what kind of information is missing in the PASS reports. In the case of vague information that is difficult to represent in data, such as the atmosphere or whether a team has a high level of teamwork, it is very difficult, or even impossible for PASS to obtain this information. But a lot of these kinds of information are easily observed by a trained reporter. But there is also information that is more easily captured in data that PASS simply does not use yet. For example, there is enough data to mention substitutions in the reports.

Another thing that stood out when comparing the reports was that PASS sometimes uses very specific language to describe certain events. However, the information used in these specific descriptions seem to be based on nothing. For example, PASS once mentioned that a player had scored a goal by shooting with great power. But in the data corresponding to that goal, there is no mention of the speed of the shot or something else from which it could be derived that the shot had great power. The reason why this happens, is that PASS tailors reports towards the fans of the clubs. This means that, when the focus team scores, this will be described more enthusiastically then when the opponent scores. And also, to stick with the previous example, perhaps the fans of the club that scored the goal perceive this as a powerful shot because they are biased towards their favourite team, and genuinely thought of it as a powerful shot, even though that might not have really been the case. So there is some grey area, because a lot of this language is based on the interpretation of the readers, but it could be possible that PASS sometimes gives incorrect information.

Another difference between the reports that were written by reporters and those that were generated by PASS, is that PASS' reports sometimes used very similar sentences or outright identical sentences to describe different events. This happened very rarely, but when it did happen, it decreased the

quality of the report immediately. The presumable reason for this is that when some events happen often in the match, there are not enough templates to describe all these events whilst still resulting in a varied report.

It could also be noticed that human-written reports often contain throwbacks to previous matches played by the focus team or a flashforward to upcoming matches. The reason that this is not implemented in PASS, is that, with the API that PASS uses to collect its data, it is not possible to look at previous matches or upcoming matches.

Finally, the last thing that stood out, is that the way that players are being referred to in the reports generated by PASS feels a bit random. In the reports written by reporters, players are always mentioned by their full name when addressed for the first time. When addressed a second time shortly after the first, often another way of referring to that player is used. For instance, often their nationality, their position in the field or a combination of those is used to indicate which player is being talked about. In PASS, however, this referring to players seemed a lot more random. Sometimes, players were mentioned by their full name, sometimes by their last name only, and sometimes the position of that player was put in front of his name. It did not seem to really follow a structure.

3.3 Conclusions

There are definitely differences between the reports written by reporters and those generated by PASS. Some of these differences are very difficult, if not impossible to prevent. For instance, describing events based on very vague information that cannot be extracted from the data that PASS works with will always be an issue, because this information cannot be directly measured and thus cannot be automatically turned into data that PASS can use. However, some of the points were PASS is lacking according to the differences with the human-written reports, can certainly be improved upon. For instance, the problem of the PASS system commenting only on goals or missed penalty kicks could be fixed by making it also comment on other match events that are present in the data.

4 Implementations

As seen in the previous chapter, the PASS system still has its improvement points. Some of these are very difficult to tackle, but other points are certainly possible to be improved upon. That is what this chapter is about, the actual implementation of features that will hopefully improve the PASS system and its produced reports.

The final modifications that were done on PASS, consist of two larger features and several smaller changes. The two larger implementations that will be discussed in this chapter, are the introduction of substitutions to the reports of PASS and the modifications to the way that PASS refers to the players mentioned in the reports. Furthermore, the smaller changes will also be briefly mentioned.

4.1 Substitutions

The first big implementation that was decided to be introduced to the PASS system was the implementation of substitutions. The goal was to chronologically mention the substitutions in the same order as they happened in the match, similar to the process of describing the goals.

So, following a structure similar to the structure that is used to describe goals scored, the substitutions were successfully implemented. The things that PASS was now successfully able to comment on with respect to substitutions were:

- Mention a regular substitution for the home or away team
- Mention a double substitution for the home or away team
- Mention a triple substitution for the home or away team
- Mention two successive substitutions for the home or away team
- Mention three successive substitutions for the home or away team

The most common form of a substitution is the regular single substitution. The double and triple substitutions are simply two or three substitutions that happen at the same time. This way, if a double or a triple substitution happens, the two or three substitution events can be combined such that only one sentence is required to mention the two or three substitutions at once. The mentioning of two or three successive substitutions is similar to this. Instead of two or three substitutions happening of the same time, they happen at different times, but without interference of other events such as goals scored. So if for instance a substitution is made in the 72nd minute, and then again a substitution is made in the 81st minute, but nothing noteworthy happened in between, these substitutions can be combined such that they are mentioned together in one sentence, similar to a double substitution.

Furthermore, PASS comments on these substitutions depending on the situation. At the moment of describing a substitution, the current score can be determined, as well as the final score of the game. Looking at these two scores, the impact of this substitute can be deduced. For example, if a team is losing 1-0, then a substitution happens, and the final score is a 1-2 win, then PASS can say something along the lines of:

“The home team manager decided to substitute <player 1> for <player 2> in an attempt to turn the tables, something that turned out well in the end.”

Obviously, this is still kind of subjective, because from the data it can't directly be seen whether this new player is the reason that the team suddenly won (except if he scored the two goals in the example, then it would be very plausible that he had a positive impact). But it is a fact that after the substitution the team played better, because they managed to win. So whilst avoiding directly mentioning that this player is the reason everything was turned around, it can be described in a more nuanced way, such as in the given example.

Finally, after looking at the new outputs that PASS produced after the implementation of the substitutions and also looking back at the human-written reports, it was noticed that the substitutions added quite a lot of text to the reports. The proportion of text about the substitutions in PASS' reports was quite a lot higher than compared to human reports. Also, it was noticed that the reports written by reporters, which are aimed at the supporters of one of the one teams, never mentioned substitutions made by the opposing team. So, this feature was adapted again, such that it would only mention substitutions for the focus team. This way, the output would represent the human-written reports more, whilst also reducing the proportion of text dedicated to describing the substitutions.

4.2 Referring expression generation

The other big adaptation of the PASS system that was decided to be made, was changing the referring expression system. Before changes were made, when referring to a player, PASS simply created a couple of possible phrases that could be used to refer to that player and picked one of those phrases in a weighted random order. After filling in all of the gaps, thus after referring to all players present in the report, PASS would check whether the same reference was used twice in short succession. If this was the case, it would find a different way to reference the player. Even though this works, it is not based on how professional reporters or writers work with referencing people.

In an earlier paper, McCoy and Strube [32] study how people are being referred to, by analyzing New York Times news articles. In the same paper, they present an algorithm that tries to automatically generate referring expressions, based on the text from these articles. The way that PASS creates referring expressions is inspired by the algorithm in this paper.

The referring expression process actually starts in the template filler module. Instead of filling in the gaps where a player should be referred, it now simply fills in the name of the player with curly brackets around it, as a placeholder to be changed later [4]. After doing this for all gaps, the PASS process continues as usual until it reaches the reference variety module. At this point, the report is basically finished, except that for all spots where a player is being referenced, it simply says their name with curly brackets around it, instead of a referring expression. This way, it is easy to keep track of how many times a certain player is mentioned and in exactly what sentence they are mentioned.

It then categorizes all players that get mentioned once in the report and all players that get mentioned more than once in separate lists. The expressions for the players that get mentioned only once is very straightforward, they are all simply mentioned by their full name. So, for these players, only the curly brackets need to be removed. For the players that get mentioned more than once,

[4] It also does the same actions for referees and managers that get mentioned, but for the sake of simplicity, only players will be considered in this explanation

their first mention of the report will also be just their full name. This is the case, because for a first or only mention of a player, it is now immediately clear who it is that is being talked about, and because it is his first mention, it can never be repetitive already.

However, the second and potential later mentions of players are a bit more complex. First of all, it is checked for every mention, whether the same player is already mentioned in the same sentence or the sentence before. If this is not the case, then the player's full name or last name is mentioned. Only when the player has been mentioned in that same or previous sentence, then a different referring expression will be used. This is due to the fact that if a player that has not been mentioned for a while is suddenly mentioned without his name being in the expression, there can be confusion about who it should be.

In the case where the player has been mentioned recently, it will be checked whether a pronoun (he or him) can be used. This will be the case when there is no other player being mentioned in the previous or the current sentence. If there is no other player present, a pronoun is sufficient to be able to successfully identify the player, and thus a pronoun will be used to describe this player.

However, in the case that a player has been mentioned recently, but also another player or multiple other players have been mentioned in the same or previous sentence, a pronoun will not be enough, because it could possibly be referring to more than one players. So in this case, a referring expression needs to be generated that can successfully disambiguate the target player from the other players.

Firstly, for all players, some of their characteristics will be compared to see which can be used to differentiate the target player from the other players. These characteristics are:

i. Whether the player is his team's captain or not

ii. The position of the player

iii. The nationality of the player

iv. The shirt number of the player

v. The team the player played for before joining his current team

Some of these characteristics will then be sufficient to successfully tell apart a player. These characteristics will be stored, and a referring expression will then be created using one or a combination of these characteristics. For instance, a player could be referred to as: 'the captain', 'the Brasilian' or 'the Dutch defender'. On their own, these expressions are possibly not sufficient to be able to know which player is being talked about, but within the context of the report, this will be clear. For example, in the following sentence:

"The French striker scored his second goal of the evening."

'The French striker' could be a lot of players without context, but with context, the sentence would look something like this:

"In the 66th minute, Neymar crossed the ball into the box after which Kylian Mbappé tapped it into the goal. The French striker scored his second goal of the evening."

Now it is immediately clear that 'the French striker' is supposed to be Kylian Mbappé (at least for the fans who would read this report).

In this example, using a pronoun is technically possible because from the context of the second sentence, you would be able to figure out who the pronoun is supposed to be referring to. However, to prevent any confusion, it was chosen to never use a pronoun if multiple persons are mentioned in the sentence before or in the current sentence.

4.3 Other work

In the previous subsections, the biggest implementations were covered. However, this was not all that was added to PASS. In this subsection, some of the additional, smaller implementations are discussed.

4.3.1 Extra information, unrelated to the match

As seen in Table 2, PASS is limited to commenting only on things that happened in the match which it is reporting on. However, human reporters regularly mention statistics that relate to the players or the teams, but not directly to the current match. For instance, what is often present in human-written reports, is flashbacks to previous matches or flashforwards to upcoming matches. This is unfortunately something that PASS is not capable of commenting on yet, because with the data available at the moment, it is not possible to determine which match or matches were played by the focus team before the current one, or which match is going to be played next by the focus team.

However, in order to create some extra variety to the match, it was decided to implement some other kind of extra information to the reports. Even though PASS was lacking on information regarding what matches the focus team had been playing or was going to play, it had access to a lot of information about the individual players. For all players, statistics on for instance how many matches they had played or how many goals they had scored were known for all seasons in which they played. So, it was decided to use these statistics, to be able to tell some interesting facts about one of the players of the match, which did not directly relate to the match that was being played.

It was implemented that for all players of the focus team that were mentioned in the report, their previous season and the number of goals that player scored that season were considered. Then, out of these players, the player with the most goals was chosen. After the first mention of this player (or the only mention if this player is mentioned only once), an extra sentence would be added which mentions the number of goals that he scored the previous season. Also the team for which he scored these goals would be mentioned, because this could potentially be a different team than the team he is playing for at the time of the match. Also, these sentences would be different depending on the amount of goals that were scored, because scoring for instance 30 times in one season is a lot more impressive than scoring five times.

4.3.2 Templates

With the adding of substitutions, it meant that there was a whole new event that PASS had to comment on. In order to do this, there should of course be different templates for the different substitutions with template sentences that successfully describe the situation. Templates were made for all five substitution events mentioned in subsection 4.1. However, for each of these substitution events, it was also implemented that PASS would produce different text depending on if the team was winning, tying or losing at the moment of the substitution. That means that these five templates

should contain different sentences depending on the situation. So these five templates were split up into fifteen templates in order to accommodate for these situations. This was however still not enough to include all situations, because it was chosen to also make the produced text depend on the outcome of the match. So, the template sentences also had to be different depending on whether the focus team would win, tie or lose the match. This made it such that, for only including all substitution events, effectively 45 new templates were added. To give an example of one of these templates, a template sentence describing a double substitution when a team is losing in a match that they were eventually going to win is:

"In een poging om de achterstand weg te werken, voerde <focus team manager> een dubbele wissel door: <substitute in 1> en <substitute in 2> kwamen tegelijk het veld in, wissels waardoor <focus team> zelfs nog 3 punten heeft overgehouden aan dit treffen."

This approximately translates to:

"In an attempt to stop lagging behind, <focus team manager> made a double substitution: <substitute in 1> and <substitute in 2> entered the field at the same time, substitutions thanks to which <focus team> even managed to preserve the 3 points."

There were also other templates added, that were affiliated with the substitutes but not directly related to the substitution events. For instance, PASS is now able to comment on when a player scores or assists a goal after he has just been substituted onto the field. Also templates unrelated to substitutes were added. For instance PASS can now also comment on goals that were scored by means of a header or goals that were scored directly from a free kick.

4.3.3 Bug fixes

As with practically all computer programs, also PASS contained some bugs that needed to be fixed. These bugs were for instance found when doing the report comparison study described in chapter 3 or during the testing of the new implementations.

A bug that was fixed, was for instance that PASS added own goals scored by a team as if that team had scored a goal themselves. It also for instance did not recognize goals scored from a free kick at all, and just skipped over them altogether. These bugs, together with more, smaller bugs, were all dealt with accordingly and did not show up again after they were fixed.

5 Evaluation

At this point, the PASS system has gone through quite some changes, which all have had impact on the text that is produced by PASS. These changes were made with the goal of improving the system, but it is not known if they actually did improve the system or not. That is why it was decided to do an evaluation study on the modifications that PASS went through. This study should give insight into whether the PASS system improved or not.

In this chapter, the set-up of the evaluation study will be explained, its results will be discussed and finally, conclusions will be given.

5.1 Type of evaluation study

First of all, the most important question to answer is: What should exactly be measured? The focus of this assignment is on attempting to improve the PASS system. So, it would make sense to compare the reports generated by the new version of PASS with the reports generated by the old version of PASS in some way, instead of measuring the overall quality of texts exclusively produced by the new version of PASS. In this case, the new version of PASS is the most recent version, and the old version of PASS is the most recent version of PASS before the start of this assignment.

The question is then which evaluation study would be the best one to perform. After considering the different possibilities for an evaluation study that can be seen in subsection 2.3, it was immediately clear that it was not going to be in the form of a metrics evaluation, because metrics evaluations simply do not give as meaningful results as evaluations concerning humans do. This leaves a task-based evaluation and a human ratings evaluation left to consider.

For the sake of the PASS system, task-based evaluation is not really an option. This is due to the fact that task-based evaluations measure direct effects that a system has on its users, which is something that PASS does not have nor aspires to have. PASS has the goal of generating soccer reports that people can enjoy, but it does not have a goal to change something in the user's behaviour.

This leaves the human ratings evaluation study. This study is very suited to measure the improvements of the PASS system, because it gives useful information about the user's thoughts on the texts produced by both the old and the new versions of PASS, which can then be compared to determine which version is generally preferred. Another advantage is that it is a lot easier to perform than a task-based evaluation in terms of time and money.

Due to these reasons, it was decided to perform a human ratings evaluation study.

5.2 Evaluation set-up

The next thing that had to be decided, was how to exactly perform this evaluation study. As already seen in subsection 2.3, this can be done in a real-world context as well as in experimental laboratory context, where the former is generally preferred.

5.2.1 Context of the experiment

It has been chosen to do the evaluation study in the form of an online survey that was distributed to peers and relatives of myself. The survey was creating using Qualtrics [5], an online platform which can be used to for instance create surveys. This has been chosen, first of all, because of the COVID-19 restrictions that were active during the time that this evaluation was taking place, which made it nearly impossible to arrange meetings with participants in real life. Luckily, this was not a big issue, because doing the experiment online was also a good option, arguably even a better option than doing it in person. This is the case, because if a soccer fan would read a match report of a match that he or she is interested in, then he or she would most probably be doing this from the comfort of his or her own home. Doing the evaluation in this way, means that the individual experiments could not be controlled closely, which means that perhaps there was some more noise that interfered with the results. However, letting the participants do the experiment at their own homes, enforces them to be in a natural environment, which makes the experiment more realistic.

The texts that the participants were presented, were generated beforehand and thus not entirely random. However, to still ensure realistic results, the matches from which the reports were generated were randomly selected from the Eredivisie 2015/2016 season.

It is difficult to pinpoint whether this evaluation tends more towards a real-world environment or an experimental laboratory environment. The participants get to read the texts at home, which is where the PASS system is primarily meant to be used. This makes the evaluation study tend towards a real-world environment. However, the participant is not really using the system, he or she is just presented some texts that were generated by different versions of this system, which makes the evaluation study tend more towards an experimental laboratory environment. The real nature of the study thus lies somewhere in between, but considering the fact that the participant is merely reading some texts, and not actively using the PASS system, this evaluation will be defined as being conducted in a laboratory context from now on.

5.2.2 Survey structure

At first, the participant is presented a small text, which informs him or her briefly on what PASS is, what he or she is going to do and what is being done with the collected data. After reading this, the participant gives consent for their answers to be used in the study by proceeding to the first question.

At first, the participant is asked whether he or she is a soccer fan, and if so, which Dutch club they support if they support any. These questions were added to be able to for instance filter out report entries if participants were shown a text involving the club that they support, which could potentially influence the ratings that he or she would give to the text.

Then, the participant was presented with a text generated by PASS on a randomly selected match from the database of twelve matches from the Eredivisie 2015/2016 season that were included in the survey. This text could be either from the old version of PASS or from the new version of PASS, which was randomly decided by Qualtrics. The participant was not aware of which version of PASS the report came from. After reading this report, the participant was asked to rate the system on four different criteria on a 7 point Likert scale. These criteria consisted of the readability,

[5] https://www.qualtrics.com/

the correctness, the variety and the enjoyability of the text. After rating this text, the participant was then presented with a second report on the exact same match, aimed at the same team as in the first report, but this time, the report was generated by the other version of PASS[6]. The participant was then asked to also rate this text on the four same criteria as he or she did for the other text. Then, thus after reading and rating both versions of the reports corresponding to the same match, the participant was asked to select which of the two reports he or she generally preferred. This process was then repeated once more, so the participant read and rated in total four texts, divided between two matches. An example of this survey could look like can be seen in Appendix A.1

Due to limitations of the free version of Qualtrics that was used to conduct this experiment, the survey had to be split into three parts. So instead of one survey having a database of texts corresponding to twelve different matches, there were three surveys which each had a database containing texts corresponding to four different matches. Ultimately, this was not a big problem, because the three different surveys were distributed relatively evenly and the results of the three surveys were afterwards put together again so that it seemed as if just one large survey had been conducted.

5.3 Results

In total, the surveys were filled in 41 times. The first survey got 11 responses, the second survey got 12 responses and the third survey got 18 responses. As mentioned in the previous subsection, each survey contained four different matches, from which the participant would read reports on two of these four matches. The selection of these two matches per survey was evenly spread out by Qualtrics, such that all matches would be equally represented in the surveys. However, because the third survey got 18 responses, which is a bit more than the other two, this means that each match in that survey was covered nine times[7], while the the matches in the first survey were covered five or six times and the matches in the second survey were covered six times. So, the four matches that were used in the third survey are slightly more represented in the final results than the other eight matches. This should not be that big of a problem though, because after all, these matches were also randomly selected.

For each individual survey entry, the participant had to rate four separate texts (two matches with each two texts). All of these individual ratings were saved as entries in one big table. The IP address of the users were used as a way to separate the different entries. For each entry, it was stored whether the participant was a soccer fan or not, from which match the report was generated, from which version of PASS the report was generated and of course the different scores that the participant rated the text with. A small part of this table can be seen in Figure 2 (with blanked out IP addresses for privacy reasons).

[6] So, if the participant was first presented a report from the old version of PASS, he or she would now be presented a report on the same match from the new version of PASS, and vice versa.

[7] Each match would be covered 1 out of 4 times. Each participant would be presented reports from 2 matches. $\frac{1}{4} \cdot 2 \cdot 18 = 9$

IP ADDRESS	FAN	MATCH	PASS VERSION	Leesbaar	Correct	Afwisselend	Aantrekkelijk
	Nee	RJC - Aja	Old	4	3	4	6
	Nee	RJC - Aja	New	4	5	6	7
	Nee	Vit-Fey	Old	4	6	7	2
	Nee	Vit-Fey	New	2	3	5	7
	Ja	RJC - Utr	Old	4	3	3	2
	Ja	RJC - Utr	New	6	5	5	4
	Ja	Wil - Fey	Old	6	6	6	4
	Ja	Wil - Fey	New	5	4	5	6
	Ja	RJC - Aja	Old	4	6	3	3
	Ja	RJC - Aja	New	4	5	5	5
	Ja	Vit-Fey	Old	5	6	6	5
	Ja	Vit-Fey	New	5	5	6	6
	Ja	Exc - Twe	Old	5	2	3	3
	Ja	Exc - Twe	New	4	4	1	3
	Ja	Gro - Utr	Old	3	4	4	4
	Ja	Gro - Utr	New	5	5	6	4
	Ja	Her - Hee	Old	4	4	3	3
	Ja	Her - Hee	New	5	5	5	5
	Ja	Wil - Fey	Old	3	5	6	2
	Ja	Wil - Fey	New	3	2	3	5

Figure 2: Part of the table containing ratings results with blurred IP addresses

First of all, the means were taken for each rating category for both the old version of PASS and the new version of PASS. These results can be seen in Table 3. Also, on these rating results, a three-way ANOVA statistical test has been done using the statistical software SPSS[8]. Such a test can be used to see whether one of three independent variables or a combination of two or three of these independent variables has an effect on a dependent variable. In the case of this evaluation, the independent variables are:

- The IP Address of the participant
- The match which is being reported on
- The version of PASS that generated the text (Old or New)

The dependent variable would be the score that was given for one of the four rating categories. So, in total, four different three-way ANOVA tests were done on the four different rating categories. The results of these tests can be seen in Appendix A.2, and a summarized version containing only the significance values can be seen in Table 4.

[8]https://www.ibm.com/products/spss-statistics

Table 3: Means of the four different rating categories for the old version of PASS and the new version of PASS

	Old PASS	**New PASS**
Readability	5.18	5.21
Correctness	5.23	5.39
Variety	4.66	5.10
Enjoyability	4.41	5.06

Table 4: Significance values for the four different ANOVA-tests. All entries below 0.05 are printed in italics

	Readability	**Correctness**	**Variety**	**Enjoyability**
IP address	*0.004*	*0.000*	0.306	*0.045*
Match	0.397	*0.045*	0.895	0.261
PASS version	0.857	0.110	0.124	*0.004*

In Table 4, in the most left column, all independent variables can be seen. It was chosen to leave out the combinations of the variables in this summary, because for the context of this evaluation study, they are not important. In the top row, the four different dependent variables, (which are the four different rating categories) can be seen.

Several significant effects (p¡0.05) have been found; the IP address has a significant effect on the readability, correctness and enjoyability of the reports, the match has a significant effect on the correctness of the reports and the version of PASS has a significant effect on the enjoyability of the reports.

Every time the participants read the two reports on the same match, originating from both versions of PASS, they were asked to indicate which of the two reports they preferred. The results of this question can be seen in Figure 3.

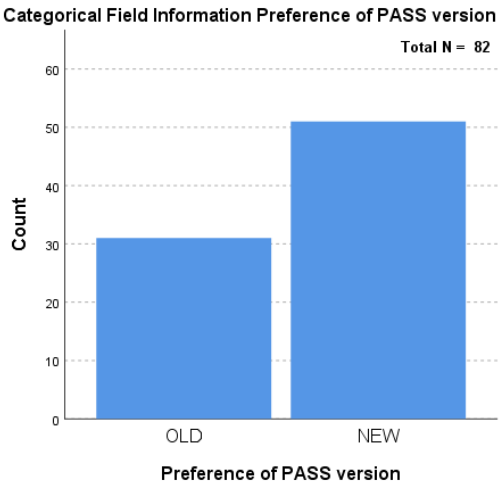


Figure 3: Preferences of the participants towards the old and the new version of PASS

As can be seen, 31 participants preferred texts that came from the old version of PASS and 51 participants preferred texts that came from the new version of PASS.

In order to see if these results were significant, a binomial test was done in SPSS as well, in order to determine if there is a preference for one of the two values. The results of this test can be seen in Appendix A.3. The null hypothesis was:

"The preferences of the PASS version for either the old or the new version occur with probabilities .500 and .500 respectively."

The outcome of the test was a p-value of 0.036, indicating that the null hypothesis could be rejected.

5.4 Conclusions

Now that all results are gathered, it is time to look at what they mean. In this subsection, conclusion will be drawn from the results that were shown in the previous subsection.

5.4.1 Ratings' means comparison

The first results that will be discussed in this subsection can be found in Table 3. On all four different categories, the new version of PASS seems to score better than the old version of PASS. But just looking at the means is not enough evidence that therefore the new version of PASS is actually better than the old one, which is due to various reasons.

The first reason is that there is statistical noise that influences the results due to the fact that the experiments could not be observed. For instance, people could be focused whilst reading the first text, but then be distracted by something whilst reading the second text, which could result in

them rating the second text in a different way than they would normally do, influencing the quality of the results. So if the values of the two to be compared means would be very close to each other, then there would be no significant difference.

Another reason is the random selection of the matches. Since there is a limited selection of matches that were used in the report, each match will have a relatively high impact on the result. This means that if for one match, the generated text by one of the versions of PASS is qualitatively a lot better than the generated text by the other version, which is always possible due to the fact that some templates might contain qualitatively better or more enjoyable text and the fact that they are randomly selected, this will be quite noticeably present in the results. There are some more statistical reasons why you cannot just compare means, but, in this case, these are the more notable ones. Thus, even though Table 3 gives an initial idea on which version might be better, this is not enough to statistically prove this.

5.4.2 Three-way ANOVA

Therefore the three-way ANOVA test was also conducted. First consider the results for readability in Table 4. The only significant value is found at the IP address. This means that the value that was filled in for readability, depended very much on the participant. The readability does not seem to be dependent on the version of PASS, which would mean that the readability has not significantly increased with the new version of PASS. This sounds logical, because while a lot of template sentences were added, hardly any existing template sentences were modified. This means that both versions of PASS select template sentences from a database of sentences that is, for the most part, the same.

For the correctness category, the results were once again very dependent on the participants. This means that each participant had a separate way of rating the correctness. However, this time, the results also depended on the version of the match. A possible reason for this phenomenon, is that for each match, both versions of PASS comment on the same things except for the substitutions, which are only present in reports of the new version of PASS. So, because a lot of the same match events are being commented on, there are a lot of similar templates that are selected, which all contain similar language. This means that if a report from one version of PASS is rated high on correctness, then the other version, corresponding to the same match, will probably be rated high on correctness as well, and vice versa. Thus, the selected match has an influence on the correctness.

The results of the three-way ANOVA test on the variety are straightforward, there seems to be no significant influence of any independent variable on the variety of the reports. The independent variable that comes closest to influencing the variety of the reports, is the version of PASS. But, as was discussed in subsection 5.3, the p-value is not low enough in order to conclude that the version of PASS had a significant influence on the variety.

This leaves the results on enjoyability. It seems that IP address and the version of PASS are the two independent variables that have a significant influence on the enjoyability that the participants experienced whilst reading the texts. The IP Address again indicates that the way of rating the enjoyability of the texts was different for each participant. However, the version of PASS has not been seen as a variable with an influence on one of the dependent variables before. This means that it can confidently be said that the extent to which the participants enjoyed the texts depended on the version of PASS from which they were reading a text. When taking into account the results in

Table 3, it can be concluded that the reports generated by the new version of PASS were significantly more enjoyed by the participants than the reports generated by the old version of PASS.

5.4.3 Binomial test

Finally, a binomial test was conducted on the answers to the question where the participants were asked to indicate which version of the text they preferred. The null hypothesis of the test assumed that the chance of either preferring the old or the new version of PASS is 50%. In other words, it assumed that the two different versions are equally likely to be preferred. This makes sense, because if this null hypothesis were to be true, then that would mean that the system has not improved by much, because then both versions would be preferred equally likely.

The significance value was found to be 0.036. $0.036 < 0.050$, which means that the null hypothesis can be rejected, as is also concluded in the results. This means that apparently the two versions of PASS were not equally likely to be preferred, but rather one of the two was chosen significantly more often. When this conclusion is set against the results as seen in Figure 3, the final conclusion can be drawn that the participants of the survey overall significantly preferred the texts that were generated by the new version of PASS.

6 Discussion & Recommendations

A lot of work has been done during the course of this assignment. The procedures and results of this work could be read in all the previous chapters. In this chapter, some of these procedures or results are being called into question, by discussing them more in depth. In addition, recommendations for potential further research or potential new implementations to the PASS system are given.

6.1 Assumptions of evaluation study

A big part of this thesis was the evaluation study described in chapter 5, because there it was determined how to answer the research question. In order to get a conclusion out of the results, statistical tests were performed. However, for the three-way ANOVA test, an assumption had to be made. Normally, for a three-way ANOVA, the dependent variable, which in the case of this evaluation were the ratings given for the four categories, should be measured at the continuous level. The most important assumption that is made for continuous variables, is that the degree of difference between values is always equal [33]. An example of a continuous variable is the temperature in degrees Celsius. For instance, the difference between 10 °C and 20 °C can be assumed to be the same as the difference between 30 °C and 40 °C. So for this evaluation, it was assumed that the values obtained through measuring with a 7 point Likert scale are continuous. However, this might not always be the case, because it cannot safely be said that the difference between a 6 and a 7 on a Likert scale is perceived by the participants as being equally large as the difference between a 3 and a 4 on the same scale.

6.2 Reports used in evaluation study

For the evaluation study, a database of 24 reports of 12 matches from the Eredivisie 2015/2016 season were used. This is quite a small database, considering that PASS can generate reports about every soccer match of which match statistics are available. However, because it was not possible to ask participants to use the PASS system by themselves, the reports had to be generated in advance. Also, because of the limitations in the Qualtrics software, there was only space for reports on 12 different matches.

If a bigger database could have been used, then the results would be more realistic, because then more examples of PASS would have been tested, but considering the situation, this form of evaluation was the best alternative.

6.3 Participant selection of evaluation study

The target audience of the PASS system consists of course of soccer fans, so, if one was to evaluate this system using human participants, it would make sense to do this evaluation with soccer fans. For the evaluation described in this thesis, the surveys were distributed amongst peers and relatives of mine, a lot of which are soccer fans. Although to be sure, as an extra check, at the beginning of the survey it was asked whether the participant was a soccer fan or not, in order to potentially filter these responses out if necessary.

From the 41 respondents, 30 respondents had indicated that they were a soccer fan and the other 11 respondents had thus indicated that they were not a soccer fan. So, the majority of the participants

was a soccer fan, but also a sizeable portion was not. However, to keep the sample size as large as possible, it was decided to take all of the survey entries into account when conducting the statistical test, including the non-soccer fans. It can be argued that these entries should have been filtered out, because they are not filled in by people who belong to the target audience, but it can be argued as well that people who are not necessarily in the target group are also capable of rating soccer texts, because you do not have to be a soccer fan to be able to understand what is being written, or in this case, automatically generated.

6.4 Recommendations for future research/work

Even though the implementations described in this thesis have already helped to improve the PASS system, there is always room for more improvement. This subsection discusses recommendations for further research or further work on the PASS program.

6.4.1 More information in the reports

As noted in chapter 3, there are still a lot of statistics that are being mentioned in human-written reports that are never present in the reports of PASS. Perhaps the most common in the reports written by reporters that is missing in PASS is the mentioning of previous matches and especially the mentioning of upcoming matches. As of now, PASS does not have access to this data. So, in future work, it should be attempted to get hold of this data.

On a similar note, in chapter 5 it could be seen that the variety of PASS did not significantly improve between the two versions. So, another recommendation is to do some more research to see what other information could potentially be added to the reports in order to increase the variety.

6.4.2 Incorrect information

Sports reports can be either objective or subjective. In the case of PASS, it is obvious that the reporting style is more subjective than objective, because the main purpose of PASS is to tailor the generated reports towards the fans of one of the two clubs who played the match. However, as also touched upon in subsection 3.2, sometimes PASS perhaps goes a bit too far with regard to subjectiveness. It is perfectly fine to be biased towards one of the involved parties of the reports, but only as long as the reports are still based on facts, which is now not always the case. So, in further work on the PASS system, it should be checked for all template sentences whether the facts they mention are actually based on match statistics or just randomly reported.

7 Conclusion

The goal of this chapter is to summarize all conclusions that were made throughout this thesis and to give an answer to the research question: "Can the automatic generation of Dutch soccer reports by the PASS system be improved?" .

The report comparison study of chapter 3 showed that there was room for improvement concerning the PASS system, especially when it came to increasing the amount of content that was present in the reports generated by PASS.

From the results of the evaluation study conducted in chapter 5, it was concluded that it could not be shown that the PASS system had improved significantly when it came to the readability, correctness, or variety of the generated reports. For the readability and the correctness of the reports, this was as expected, because on those categories the PASS system was not changed much. However, the results for the variety were a bit disappointing. Concerning the work on the referring expression generation, this can be explained. The reports that PASS generates are still relatively short and it is rare that a player gets mentioned more than once, let alone more than once in two consecutive sentences or the same sentence. Therefore, the direct results of this implementation were not immediately noticeable. However, the substitutions that were added to the match reports were definitely immediately noticeable in the reports. They made it such that PASS did not only comment on the goals by adding an extra event to the reports, which would indicate that the variety of the reports would also improve, but this was not the case.

From the results of three-way ANOVA test concerning the enjoyablity of the generated texts, it was concluded with high confidence that the enjoyablity of the texts generated by the new version of PASS was higher than the enjoyability of the texts generated by the old version of PASS.

This leaves the results of which version of PASS was generally preferred in the evaluation study. These results also point in the favour of the new version of PASS. It was concluded that the reports generated by the new version of PASS were generally confidently preferred over the reports generated by the old version of PASS.

To conclude this thesis, an answer will be given to the research question: Yes, the automatic generation of Dutch soccer reports by the PASS system can be improved, and it has been improved.

References

[1] R. W. McChesney, *Media made sport: A history of sports coverage in the United States*, pp. 49–69. Sage Newbury Park, CA, 1989.

[2] J. Gong, W. Ren, and P. Zhang, "An automatic generation method of sports news based on knowledge rules," in *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*, pp. 499–502, IEEE, 2017.

[3] C. van der Lee, E. Krahmer, and S. Wubben, "PASS: A Dutch data-to-text system for soccer, targeted towards specific audiences," in *Proceedings of the 10th International Conference on Natural Language Generation*, pp. 95–104, 2017.

[4] S. Mahamood and E. Reiter, "Generating affective natural language for parents of neonatal infants," in *Proceedings of the 13th European Workshop on Natural Language Generation*, pp. 12–21, 2011.

[5] R. Evans, P. Piwek, and L. Cahill, "What is NLG?," in *Proceedings of the second international conference on natural language generation*, pp. 144–151, 2002.

[6] A. Gatt and E. Krahmer, "Survey of the state of the art in natural language generation: Core tasks, applications and evaluation," *Journal of Artificial Intelligence Research*, vol. 61, pp. 65–170, 2018.

[7] E. Reiter, S. Sripada, J. Hunter, J. Yu, and I. Davy, "Choosing words in computer-generated weather forecasts," *Artificial Intelligence*, vol. 167, no. 1-2, pp. 137–169, 2005.

[8] E. Reiter and R. Dale, "Building applied natural language generation systems," *Natural Language Engineering*, vol. 3, no. 1, pp. 57–87, 1997.

[9] S. G. Sripada, E. Reiter, J. Hunter, and J. Yu, "Generating English summaries of time series data using the Gricean maxims," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 187–196, 2003.

[10] H. Dalianis, "Aggregation in natural language generation," *Computational Intelligence*, vol. 15, no. 4, pp. 384–414, 1999.

[11] H. Cheng and C. Mellish, "Capturing the interaction between aggregation and text planning in two generation systems," in *Proceedings of the first international conference on Natural language generation-Volume 14*, pp. 186–193, Association for Computational Linguistics, 2000.

[12] A. Cahill, M. Forst, and C. Rohrer, "Stochastic realisation ranking for a free word order language," in *Proceedings of the Eleventh European Workshop on Natural Language Generation*, pp. 17–24, Association for Computational Linguistics, 2007.

[13] A. Ratnaparkhi, "Trainable methods for surface natural language generation," in *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pp. 194–201, Association for Computational Linguistics, 2000.

[14] I. Langkilde, "Forest-based statistical sentence generation," in *1st Meeting of the North American Chapter of the Association for Computational Linguistics*, 2000.

[15] J. A. Bateman, "Enabling technology for multilingual natural language generation: the KPML development environment," *Natural Language Engineering*, vol. 3, no. 1, pp. 15–55, 1997.

[16] B. Lavoie and O. Rainbow, "A fast and portable realizer for text generation systems," in *Fifth Conference on Applied Natural Language Processing*, pp. 265–268, 1997.

[17] G. Angeli, C. D. Manning, and D. Jurafsky, "Parsing time: Learning to interpret time expressions," in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 446–455, Association for Computational Linguistics, 2012.

[18] T. Becker, "Natural language generation with fully specified templates," in *SmartKom: Foundations of Multimodal Dialogue Systems*, pp. 401–410, Springer, 2006.

[19] N. Braun, M. Goudbeek, and E. Krahmer, "The multilingual affective soccer corpus (MASC): Compiling a biased parallel corpus on soccer reportage in English, German and Dutch," in *Proceedings of the 9th International Natural Language Generation Conference*, pp. 74–78, 2016.

[20] E. Reiter, "Types of NLG evaluation: Which is right for me?." https://ehudreiter.com/2017/01/19/types-of-nlg-evaluation/, Jan 2017.

[21] E. Reiter, "How to do an NLG evaluation: Task-based (extrinsic) performance in real-world context." https://ehudreiter.com/2017/04/27/task-based-real-world-nlg-eval/, Apr 2017.

[22] D. Braun, E. Reiter, and A. Siddharthan, "Creating textual driver feedback from telemetric data," in *Proceedings of the 15th European Workshop on Natural Language Generation (ENLG)*, pp. 156–165, 2015.

[23] A. Joshi, S. Kale, S. Chandel, and D. K. Pal, "Likert scale: Explored and explained," *British Journal of Applied Science & Technology*, vol. 7, no. 4, p. 396, 2015.

[24] E. Reiter, "How to do an NLG evaluation: Human ratings in real-world contexts." https://ehudreiter.com/2017/02/23/real-world-human-ratings-evaluation/, Feb 2017.

[25] E. Reiter, "How to do an NLG evaluation: Human ratings in artificial context." https://ehudreiter.com/2017/01/09/human-ratings-nlg-evaluation/, Jan 2017.

[26] E. Reiter, "How to do an NLG evaluation: Metrics." https://ehudreiter.com/2017/05/03/metrics-nlg-evaluation/, May 2017.

[27] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 311–318, Association for Computational Linguistics, 2002.

[28] C.-Y. Lin, "ROUGE: a package for automatic evaluation of summaries," in *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*, 2004.

[29] S. Banerjee and A. Lavie, "METEOR: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

[30] A. Belz and A. Gatt, "Intrinsic vs. extrinsic evaluation measures for referring expression generation," in *Proceedings of ACL-08: HLT, Short Papers*, pp. 197–200, 2008.

[31] E. Reiter and A. Belz, "An investigation into the validity of some metrics for automatically evaluating natural language generation systems," *Computational Linguistics*, vol. 35, no. 4, pp. 529–558, 2009.

[32] K. F. McCoy and M. Strube, "Generating anaphoric expressions: pronoun or definite description?," in *The Relation of Discourse/Dialogue Structure and Reference*, 1999.

[33] A. Field, *Discovering statistics using SPSS*, ch. 1, p. 8. Sage, Third ed., 2009.

A Appendix

A.1 Example survey

A.1.1 Introduction text

Deze enquête gaat over voetbalverslagen gegenereerd door het PASS-systeem, een computer-systeem dat automatisch voetbalverslagen schrijft aan de hand van statistieken van voetbalwedstrijden. PASS schrijft verschillende versies, gepersonaliseerd naar fans van de clubs die de wedstrijd speelden.

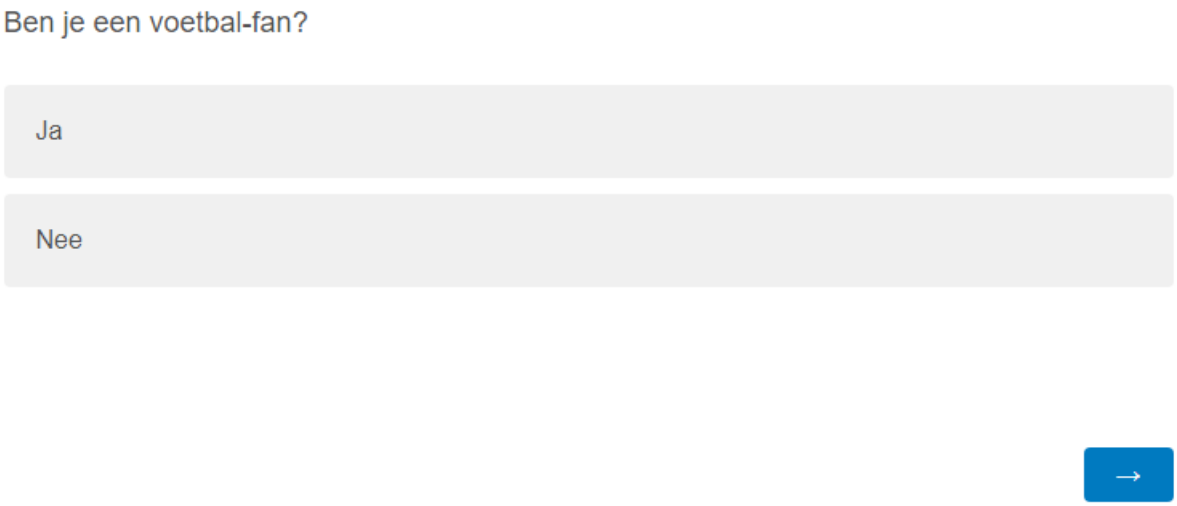
Je krijgt eerst 1 of 2 korte introductievragen te zien. Daarna ga je 2 keer 2 voetbalverslagen lezen van een willekeurige eredivisie-wedstrijd van het seizoen 2015/2016. Deze verslagen zijn allemaal automatisch gegenereerd door een versie van PASS. Na het lezen van elk verslag beantwoord je een aantal vragen.

Deze enquête zal ongeveer 10 minuten duren.

Door verder te gaan, geef je toestemming dat je antwoorden gebruikt kunnen worden voor een data-analyse. Er worden geen persoonlijke gegevens gevraagd en alle verzamelde data blijft anoniem. Deze data wordt veilig opgeslagen en zal niet publiekelijk gemaakt worden, slechts gemiddeldes van alle ingevulde enquêtes zullen openbaar worden. Mocht je nog verdere vragen hebben, stuur dan een e-mail naar: j.w.m.sessink@student.utwente.nl

→

A.1.2 Introduction question 1



A.1.3 Introduction question 2

A.1.4 Text 1, PASS version 1

Excelsior - FC Twente 05/12/2015 (gericht aan de fans van Excelsior)

Excelsior moet genoegen nemen met punt

Meer dan een puntendeling zat er niet in tegen FC Twente: na negentig minuten voetbal in een moordend tempo kon vriend en vijand met deze uitslag leven. Ondanks dat Excelsior grote gedeeltes van het duel beter was, stond er een 1-1 eindstand op het scorebord in Stadion Woudestein.

In de 38e minuut mikte Tom van Weert op aangeven van Bas Kuipers zijn inzet in het doel van Marsman: 1-0. In de 75e minuut schoot middenvelder Ziyech de gelijkmaker binnen: 1-1.

Kevin Vermeulen van Excelsior en Jeroen van der Lely van FC Twente pakten een gele kaart.

	1	2	3	4	5	6	7
Hoe leesbaar is de tekst? 1 = Absoluut onleesbaar, 7 = perfect, totaal duidelijk.	○	○	○	○	○	○	○
Hoe correct is de tekst? 1 = Erg veel taal/grammatica fouten, 7 = foutloos qua taal/grammatica.	○	○	○	○	○	○	○
Hoe afwisselend is de tekst? 1 = Erg veel herhalend taalgebruik, 7 = veel afwisseling in de tekst.	○	○	○	○	○	○	○
Hoe aantrekkelijk is de tekst om te lezen? 1 = Er is niks aan, 7 = zeer aantrekkelijk.	○	○	○	○	○	○	○

A.1.5 Text 1, PASS version 2

Excelsior - FC Twente 05/12/2015 (gericht aan de fans van Excelsior)

Excelsior deelt punten met FC Twente

Excelsior moest na een leuk voetbalgevecht afgelopen zaterdag genoegen nemen met een punt. Voor ruim 3400 toeschouwers speelde de club uit Rotterdam op zaterdagavond met 1-1 gelijk tegen FC Twente.
In de 38e minuut mikte Tom van Weert op aangeven van Bas Kuipers zijn inzet in het doel van Nick Marsman: 1-0. De aanvaller wist vorig seizoen al 13 doelpunten te maken voor Excelsior. In de 75e minuut kwam FC Twente op gelijke hoogte, toen Hakim Ziyech de bal tegen de touwen schoot: 1-1. Een wissel aan de kant van Excelsior bij een gelijke stand na 82 minuten: Stanley Elbers kwam het veld in voor Daryl van Mieghem. Helaas wist deze wissel niet genoeg teweeg te brengen en zouden de punten uiteindelijk gedeeld worden. Terwijl het na 85 minuten nog gelijk stond in Stadion Woudestein, wisselde Groenendijk Jeff Stans voor Kevin Vermeulen. Dit had echter niet veel invloed op de einduitslag.

Kevin Vermeulen van Excelsior en Jeroen van der Lely van FC Twente pakten een gele kaart.

	1	2	3	4	5	6	7
Hoe leesbaar is de tekst? 1 = Absoluut onleesbaar, 7 = perfect, totaal duidelijk.	○	○	○	○	○	○	○
Hoe correct is de tekst? 1 = Erg veel taal/grammatica fouten, 7 = foutloos qua taal/grammatica.	○	○	○	○	○	○	○
Hoe afwisselend is de tekst? 1 = Erg veel herhalend taalgebruik, 7 = veel afwisseling in de tekst.	○	○	○	○	○	○	○
Hoe aantrekkelijk is de tekst om te lezen? 1 = Er is niks aan, 7 = zeer aantrekkelijk.	○	○	○	○	○	○	○

A.1.6 Text 1, preference question

Welke van de 2 teksten vond u in het algemeen beter?

Excelsior - FC Twente 05/12/2015 (gericht aan de fans van Excelsior)

Excelsior moet genoegen nemen met punt
Meer dan een puntendeling zat er niet in tegen FC Twente: na negentig minuten voetbal in een moordend tempo kon vriend en vijand met deze uitslag leven. Ondanks dat Excelsior grote gedeeltes van het duel beter was, stond er een 1-1 eindstand op het scorebord in Stadion Woudestein.
In de 38e minuut mikte Tom van Weert op aangeven van Bas Kuipers zijn inzet in het doel van Marsman: 1-0. In de 75e minuut schoot middenvelder Ziyech de gelijkmaker binnen: 1-1.
Kevin Vermeulen van Excelsior en Jeroen van der Lely van FC Twente pakten een gele kaart.

Excelsior - FC Twente 05/12/2015 (gericht aan de fans van Excelsior)

Excelsior deelt punten met FC Twente
Excelsior moest na een leuk voetbalgevecht afgelopen zaterdag genoegen nemen met een punt. Voor ruim 3400 toeschouwers speelde de club uit Rotterdam op zaterdagavond met 1-1 gelijk tegen FC Twente.
In de 38e minuut mikte Tom van Weert op aangeven van Bas Kuipers zijn inzet in het doel van Nick Marsman: 1-0. De aanvaller wist vorig seizoen al 13 doelpunten te maken voor Excelsior. In de 75e minuut kwam FC Twente op gelijke hoogte, toen Hakim Ziyech de bal tegen de touwen schoot: 1-1. Een wissel aan de kant van Excelsior bij een gelijke stand na 82 minuten: Stanley Elbers kwam het veld in voor Daryl van Mieghem. Helaas wist deze wissel niet genoeg teweeg te brengen en zouden de punten uiteindelijk gedeeld worden. Terwijl het na 85 minuten nog gelijk stond in Stadion Woudestein, wisselde Groenendijk Jeff Stans voor Kevin Vermeulen. Dit had echter niet veel invloed op de einduitslag.

Kevin Vermeulen van Excelsior en Jeroen van der Lely van FC Twente pakten een gele kaart.

A.1.7 Text 2, PASS version 1

FC Groningen - FC Utrecht 17/01/2016 (gericht aan de fans van FC Groningen)

FC Groningen verliest in eigen huis van 6tal FC Utrecht

Ondanks het feit dat FC Groningen een tijd met een man meer speelde, heeft het zondag de thuiswedstrijd tegen FC Utrecht niet in winst om kunnen zetten. Op bezoek bij de ploeg van manager Erik ten Hag stapte de club uit Groningen na een slechte start en enkele discutabele beslissingen van scheidsrechter Björn Kuipers met een 1-4 nederlaag van het veld.

Na 16 minuten kwam FC Utrecht op voorsprong toen Ramselaar scoorde: 0-1. Ruiter had na 27 minuten geen antwoord op een doelpoging van Sørloth na een mooie aanval: 1-1. FC Utrecht kwam door twee gelukkige treffers van Kum en middenvelder Andreas Ludwig op een 1-3 voorsprong.

Er werden 2 gele kaarten uitgedeeld aan de zijde van FC Groningen voor Michael de Leeuw en aan de zijde van de uitploeg voor Andreas Ludwig. Na 76 minuten spelen moest hij met zijn tweede gele kaart richting de douche.

	1	2	3	4	5	6	7
Hoe leesbaar is de tekst? 1 = Absoluut onleesbaar, 7 = perfect, totaal duidelijk.	○	○	○	○	○	○	○
Hoe correct is de tekst? 1 = Erg veel taal/grammatica fouten, 7 = foutloos qua taal/grammatica.	○	○	○	○	○	○	○
Hoe afwisselend is de tekst? 1 = Erg veel herhalend taalgebruik, 7 = veel afwisseling in de tekst.	○	○	○	○	○	○	○
Hoe aantrekkelijk is de tekst om te lezen? 1 = Er is niks aan, 7 = zeer aantrekkelijk.	○	○	○	○	○	○	○

A.1.8 Text 2, PASS version 2

FC Groningen - FC Utrecht 17/01/2016 (gericht aan de fans van FC Groningen)

FC Utrecht zegeviert in Groningen

Ondanks dat FC Groningen vrijwel de gehele wedstrijd de bovenliggende partij was, is het zondagmiddag toch tegen een onnodige nederlaag aangelopen. De thuisploeg verloor na een hoopvol begin met 1-4 van FC Utrecht.

Na 16 minuten belandde de bal via het hoofd van Bart Ramselaar achter Robbin Ruiter. Alexander Sørloth schoot FC Groningen in de 27e minuut weer naast FC Utrecht. Vorig seizoen nog legde de spits 13 keer de bal in het net voor FK Bodø/Glimt. In de 48e minuut schoot Timo Letschert raak uit een vrije trap: 1-2. Manager Van de Looi probeerde tevergeefs nog het tij te keren door in de 56e minuut een frisse kracht in te brengen: Jesper Drost kwam het veld in voor Desevio Payne. De vrijstaande Chris Kum schoot op aangeven van Yassin Ayoub raak. Toen Andreas Ludwig koel de 1-4 aantekende was de wedstrijd natuurlijk beslist. In een poging om de achterstand weg te werken, voerde manager Erwin van de Looi 2 wissels achter elkaar door: Rasmus Lindgren en Danny Hoesen kwamen het veld in. Zij wisten echter uiteindelijk het tij ook niet meer te keren.

De wedstrijd eindigde met 2 gele kaarten, voor Andreas Ludwig en Michael de Leeuw. Andreas Ludwig werd in de 76e minuut na zijn tweede gele kaart uit het veld gestuurd.

	1	2	3	4	5	6	7
Hoe leesbaar is de tekst? 1 = Absoluut onleesbaar, 7 = perfect, totaal duidelijk.	○	○	○	○	○	○	○
Hoe correct is de tekst? 1 = Erg veel taal/grammatica fouten, 7 = foutloos qua taal/grammatica.	○	○	○	○	○	○	○
Hoe afwisselend is de tekst? 1 = Erg veel herhalend taalgebruik, 7 = veel afwisseling in de tekst.	○	○	○	○	○	○	○
Hoe aantrekkelijk is de tekst om te lezen? 1 = Er is niks aan, 7 = zeer aantrekkelijk.	○	○	○	○	○	○	○

A.1.9 Text 2, preference question

Welke van de 2 teksten vond u in het algemeen beter?

FC Groningen - FC Utrecht 17/01/2016 (gericht aan de fans van FC Groningen)

FC Groningen verliest in eigen huis van 6tal FC Utrecht
Ondanks het feit dat FC Groningen een tijd met een man meer speelde, heeft het zondag de thuiswedstrijd tegen FC Utrecht niet in winst om kunnen zetten. Op bezoek bij de ploeg van manager Erik ten Hag stapte de club uit Groningen na een slechte start en enkele discutabele beslissingen van scheidsrechter Björn Kuipers met een 1-4 nederlaag van het veld.
Na 16 minuten kwam FC Utrecht op voorsprong toen Ramselaar scoorde: 0-1. Ruiter had na 27 minuten geen antwoord op een doelpoging van Sørloth na een mooie aanval: 1-1. FC Utrecht kwam door twee gelukkige treffers van Kum en middenvelder Andreas Ludwig op een 1-3 voorsprong.
Er werden 2 gele kaarten uitgedeeld aan de zijde van FC Groningen voor Michael de Leeuw en aan de zijde van de uitploeg voor Andreas Ludwig. Na 76 minuten spelen moest hij met zijn tweede gele kaart richting de douche.

FC Groningen - FC Utrecht 17/01/2016 (gericht aan de fans van FC Groningen)

FC Utrecht zegeviert in Groningen
Ondanks dat FC Groningen vrijwel de gehele wedstrijd de bovenliggende partij was, is het zondagmiddag toch tegen een onnodige nederlaag aangelopen. De thuisploeg verloor na een hoopvol begin met 1-4 van FC Utrecht.
Na 16 minuten belandde de bal via het hoofd van Bart Ramselaar achter Robbin Ruiter. Alexander Sørloth schoot FC Groningen in de 27e minuut weer naast FC Utrecht. Vorig seizoen nog legde de spits 13 keer de bal in het net voor FK Bodø/Glimt. In de 48e minuut schoot Timo Letschert raak uit een vrije trap: 1-2. Manager Van de Looi probeerde tevergeefs nog het tij te keren door in de 56e minuut een frisse kracht in te brengen: Jesper Drost kwam het veld in voor Desevio Payne. De vrijstaande Chris Kum schoot op aangeven van Yassin Ayoub raak. Toen Andreas Ludwig koel de 1-4 aantekende was de wedstrijd natuurlijk beslist. In een poging om de achterstand weg te werken, voerde manager Erwin van de Looi 2 wissels achter elkaar door: Rasmus Lindgren en Danny Hoesen kwamen het veld in. Zij wisten echter uiteindelijk het tij ook niet meer te keren.

De wedstrijd eindigde met 2 gele kaarten, voor Andreas Ludwig en Michael de Leeuw. Andreas Ludwig werd in de 76e minuut na zijn tweede gele kaart uit het veld gestuurd.

A.2 Three-way ANOVA results

Tests of Between-Subjects Effects

Dependent Variable: Leesbaar

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	275.256[a]	155	1.776	3.157	.041
Intercept	3813.096	1	3813.096	6778.838	.000
IPADDRESS	131.070	34	3.855	6.853	.004
MATCH	7.567	11	.688	1.223	.397
PASSVERSION	.020	1	.020	.035	.857
IPADDRESS * MATCH	24.958	32	.780	1.387	.329
IPADDRESS * PASSVERSION	32.625	34	.960	1.706	.218
MATCH * PASSVERSION	14.242	11	1.295	2.302	.123
IPADDRESS * MATCH * PASSVERSION	29.642	32	.926	1.647	.235
Error	4.500	8	.563		
Total	4706.000	164			
Corrected Total	279.756	163			

Figure 4: Results three-way ANOVA on 'Leesbaarheid' (Readability)

Tests of Between-Subjects Effects

Dependent Variable: Correct

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	286.640[a]	155	1.849	5.918	.005
Intercept	4016.370	1	4016.370	12852.385	.000
IPADDRESS	137.485	34	4.044	12.940	.000
MATCH	11.888	11	1.081	3.458	.045
PASSVERSION	1.010	1	1.010	3.231	.110
IPADDRESS * MATCH	25.367	32	.793	2.537	.084
IPADDRESS * PASSVERSION	27.984	34	.823	2.634	.076
MATCH * PASSVERSION	15.537	11	1.412	4.520	.021
IPADDRESS * MATCH * PASSVERSION	32.777	32	1.024	3.278	.041
Error	2.500	8	.313		
Total	4915.000	164			
Corrected Total	289.140	163			

Figure 5: Results three-way ANOVA on 'Correctheid' (Correctness)

Tests of Between-Subjects Effects

Dependent Variable: Afwisselend

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	305.561[a]	155	1.971	.789	.736
Intercept	3498.930	1	3498.930	1399.572	.000
IPADDRESS	122.548	34	3.604	1.442	.306
MATCH	12.153	11	1.105	.442	.895
PASSVERSION	7.380	1	7.380	2.952	.124
IPADDRESS * MATCH	47.728	32	1.492	.597	.857
IPADDRESS * PASSVERSION	38.158	34	1.122	.449	.950
MATCH * PASSVERSION	13.560	11	1.233	.493	.863
IPADDRESS * MATCH * PASSVERSION	62.829	32	1.963	.785	.708
Error	20.000	8	2.500		
Total	4228.000	164			
Corrected Total	325.561	163			

Figure 6: Results three-way ANOVA on 'Afwisselendheid' (variety)

Tests of Between-Subjects Effects

Dependent Variable: Aantrekkelijk

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	356.226[a]	155	2.298	2.451	.085
Intercept	3239.862	1	3239.862	3455.853	.000
IPADDRESS	101.014	34	2.971	3.169	.045
MATCH	16.382	11	1.489	1.589	.261
PASSVERSION	15.303	1	15.303	16.324	.004
IPADDRESS * MATCH	35.148	32	1.098	1.172	.436
IPADDRESS * PASSVERSION	68.415	34	2.012	2.146	.129
MATCH * PASSVERSION	17.966	11	1.633	1.742	.220
IPADDRESS * MATCH * PASSVERSION	75.329	32	2.354	2.511	.087
Error	7.500	8	.938		
Total	4045.000	164			
Corrected Total	363.726	163			

Figure 7: Results three-way ANOVA on 'Aantrekkelijkheid' (Enjoyability)

A.3 Binomial test results

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The categories defined by Preference of PASS version = OLD and NEW occur with probabilities .500 and .500.	One-Sample Binomial Test	.036	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .050.

Figure 8: Results of binomial test on preference of the version of PASS