Facial expressions of story tellers

Generating embodied agent scripts from text only input

Adri Wiekens
a.wiekens-1@student.utwente.nl

ABSTRACT
In this paper, we describe results of the experiments with automated emotion detection in pure text input.

Keywords
Virtual storyteller, emotion detection, EmotionML, embodied agents, script generation.

1. INTRODUCTION

A few years ago the concept of the virtual storyteller was thought up by a student [15]. This concept was to build a multi agent system that could generate a story in natural language. But since the concept was a "virtual storyteller" and not a "story generator" the program is somewhat incomplete without an agent telling the story like a real human being. Thus there rose the need for an embodied agent that could actually present the story in the same way as that human being does with the same story telling techniques. A human story teller can use tone and inflection in his voice to influence his audience, can make hand gestures express actions in the story, use simple gestures with his face like the movement of an eyebrow or a wink with an eye to add subtleties to a story and even just the basic expression on his face can express the feelings of a story character or invoke feelings of suspense or surprise with the audience. This is but a short list of techniques that are either already under research or have been researched, yet there may be more. If a virtual story teller is to ever be an equal to the real human story teller, it needs to master all of these techniques.

In this research the center of focus are two techniques that are used in storytelling. The first one: facial gestures i.e. winking, proper use of eyebrows, nodding and shaking of the entire face. The second one are facial expressions i.e. the facial representation of the emotional state of the storyteller. The main problem that needs to be tackled in this research is to automatically generate a standard form input based on only (Dutch) text. So our starting point is pure text in the form of a dutch story (in particulair a fairy tale but why this specific domain is chosen is elaborated later) and the required output is a set of instructions for an automated story teller, this in the form of an embodied agent. This output is preferably generated in the form of an internationally accepted standard tag such as an EmotionML tag [12] or a standardized form of BML. Then the next step would be to convert this standard form of emotional tags to a specific form of input that can serve as a form of control for an embodied agent in order to actually express the found and generated emotional content.

The goal of this research is to determine whether it is possible to create a script for an embodied agent with just natural language as input. In order to put limits on the area of research we will restrict ourselves in this goal to the field of fairytales as natural input. Fairytales have a certain structure that a computer program should be able to recognize with the proper program setup. In order to be a more complete storyteller it is necessary for the system to make the right expressions at the right time, expressions here being emotional facial expressions and facial gestures.

2. OVERVIEW

In the next paragraph we will lay out the preexisting research and related work. This is followed by the analysis of emotion evocation, on what techniques should be used by an embodied agent to achieve certain effects with the audience. After discussing these preparation techniques there follows the analysis of the software that was designed for this experiment. This so called "taggertool" will serve as the central program for the execution of the experiment. The internal workings of this software are discussed in the sections after the initial introduction of the software. The last section of the taggertool discusses the several pieces of third party software that is used for some inner workings. The final sections of the paper treat the experiment set up, procedures. participants, execution of experiment and results. The conclusiveness of the experiment by itself has been used to determine the successfulness of this research; after which there is a section on points of discussion such as the possibilities for improvement. The last section then gives a conclusion.

3. RELATED WORK

This research deals with two major elements which have been addressed seperatly in other works, namely the detection of emotion in text, and the evocation of emotion during

story telling. This paragraph will deal with related work and previous research done for those two elements. The evocation of emotions will be described on the level of generating facial expressions and facial gestures.

3.1 Emotion detection in text

In one research experiment performed by Tao Jianhua [14] the suggested system was to perform a complete context analysis. This meant that they took one element from the sentence and compared how another element of that same sentence was influenced by the first. The method they used to analyse the two was by finding context information from an external source of information such as online encyclopedia references. For example if there were two nouns in one sentence and after context searching it is revealed that one is a predator and the other there is the prey they conclude that the prey should be fearful and the predator should be content that he is near prey.

There are several other papers that also suggest context information as crucial in finding emotional content [7]. So from this point on a context dictionary will be considered as necessary. The different papers have different approaches to get information from text, but they all agree on the point that the main words that carry information for emotion and sentiment. The crucial words for this are for context free checking are nouns, while nouns can be informative for sentiment and emotion when checked from context. The third type of word that was of important is the adverb. This could serve as an effective modifier of existing emotional content, as is shown in another paper where they determined that adverbs in combination with adjectives is far more effective than checkingfor adjectives alone [2].

Next to the analysis of pure emotions there is also the analysis of sentiment that can be relevant to the telling of the story. This can be used to analyze who the protagonist and who the antagonist is, which further on can be used to invoke the proper emotion at the right time. For example, in the story of little Red Riding hood the story ends with the death of a clear cut antagonist, the big bad wolf. If we look purely neutral without favoring the outcome of either, this will be considered a sad event because one of the story characters dies. However since this is supposed to be the happy ending of the story it should be brought to the audience in a manner that invokes happiness. So the presentation of the outcome of an event should be brought in a manner that depends on whether it is favorable by the protagonist.

A good algorithm for sentiment analysis was demonstrated in a paper by Peter Turney [16] and later refined in a second one [2]. Where it here describes a "thumbs up" or "thumbs down", similarly an algorithm can be constructed as describing "favorable to antagonist" and "favorable to protagonist". This would require extra information in the suggested dictionary of the previous paragraph.

3.2 Facial expression

As mentioned before in paragraph 2 of the previous section, there are several papers involved in defining a proper model for generating the facial expressions. So once the emotional content has been determined, it should be possible to feed just information to an embodied agent in order to present it.

3.3 Facial gesture generation

When it comes to eyebrows, a paper focused on several western European languages including Dutch found correlation between language use and the movement of eyebrows. More accurately when there was a need to apply extra emphasis or a change in pitch the eyebrows also made a short gesture[4]. During a part of story that implies denial or confirmation storytellers often use the basic head gestures that in Dutch culture enforce this i.e. the shaking and nodding of the head. An important note is that the meaning of these gestures is different in Russia for example, since the language of the storytelling is Dutch we also use Dutch gesture language.

4. EMOTION EVOCATION

In order to do a complete analysis of the text for emotional content we need a deterministic model to describe the emotional state of the story teller. In the section I shall describe which model will be used for the research about text analysis and I shall explain for each element why they are used in this way.

4.1 Addition to the first person perspective

The technique of storytelling when telling from a first person perspective is to make use of character impersonation. I.e. the storyteller pretends to actually be the character whose line he or she is currently reading. So in order to model the emotional state that the storyteller is supposed to express we only need to use a standard character emotional model in six standard emotions [10][11], under ideal circumstances the stsoryteller would make use of alternative text to speech, but unfortunately this is not an option within this particular research, since the realisation of BML relies on making use of one form of text to speech. It is possible to tamper slightly this would be outside of the researched domain and would go beyond the idea of creating the script fully automatic.

In a work by Elizabeth A. Winston [18], the writer pointed out a way to enhance the experience towards the listeners even more. She took the story about the turtle and the hare. The main characters of this wellknown story are, as the title tells us, a turtle and a hare. To improve the quality of the experience of this story the author suggests that, next to normal techniques, it can also be effective to change the default or neutral expression a bit to express unique traits of every character. For example, when we look up what the characteristics are of turtles, we see that turtles are slow and have a very long life. This is often used in stories to express that they are never in a hurry. A story teller can demonstrate this by, whenever speaking as the turtle, using a somewhat droopy face as a normal disposition and perhaps even lowering the rate of speech. While when appearing as the hare we can use the same character analysis to indicate that when speaking as the hare, the story teller should breathe and speaking more quickly and seem a bit restless [13].

4.2 Other perspectives

With characters in the story the standard approach is to apply the first person perspective. With the narrator this

is more complicated. T.C. Nijmeijer wrote a paper [9] on the varying perspectives in narrative story telling. In his paper on story generation he concluded which perspectives in story telling we can distinguish. If in creation of story ONLY these perspectives occur then in telling the stories we can accordingly assume that we have to be only in one of those perspectives. After all we can only read a story to the audience after it has been created. This method will from here be refered to as the empathic method.

4.3 Frey versus the writer

For this research we take stories with a relatively straightforward narrator: the all-seeing and all knowing one. An important statement that Nijmeijer made in his paper on perspectives is that the writer and the narrator are the same person. In the book "How to write a damn good story" by Frey [5] [6], there is a different opinion on the matter on who the writer should be compared to the narrator. He states that they are two very distinct entities. One being a real person with real emotions, while the other is simply a special state character in the book designed to relay stories and evoke proper emotions. As such the narrator can have his own opinions and emotional states to accomplish this.

When we take such a position we can take a different method of evoking emotions. While an standard response is to invoke fear by mimicking fear and hoping for an empathic result. The approach suggested by Frey opens up the idea of instead of using fear to create fear, intimidate the listener at moments where fear is appropriate. It also means that evoking these emotions will result in them being aimed at the story teller in case of anger and fear, during the experiment we will see whether this has a negative, positive or neutral impact on the performance of the story teller.

In order to test the effectiveness of this strategy two distinct algorithms are devised in the final stages of selecting emotional expressions. One from now on we will call the "Frey-method" and a second we will call the "Empathic-method".

5. TAGGERTOOL

The piece of software that automatically performs the task of creating a script for the agent that tells the story has been named the tagger tool. The task it needs to perform is contains three stages. First it needs to analyse the text in order to find the protagonist and the antagonist in the story, and it needs to check which part is told from third and which part from first person perspective. The second part is to determine which emotion needs to be invoked at what part of the story. The last stage is to weave the proper emotions and facial gestures into the story to make it into a script that can be performed by an embodied agent.

5.1 Program design

The developed program can handle two basic forms of input. First of all there is the Unicode plain text input. This is sufficient as input in concordance with the overall objective of the experiment: to generate an embodied agent script from just this input.

However during the experiment it showed that just using this form of input is impractical. Since every individual step that is taken cannot be stored as a conveniently loading object. So each experiment step needs to be repeated for each experiment. These experiments included the automatic segmentation of the input text into emotional parts i.e. parts that in itself contain no shifts in emotion for the story teller, more on that later , and the insertion of something that could start a gesture. Since this was a very time costly procedure (for even a small fairytale of several pages like little red riding hood it takes several minutes), this proved to be rather unpractical.

Instead of just taking input and processing that into output. We take the pure text and load this into an object that can be observed with the interface and could be modified with the controllers who call upon the tagging algorithms. The created object can be serialized and be written to a disc.

6. ALGORITHM IMPLEMENTATION

The crucial components of the program are the text processing algorithms. Each of the programs will tag the input data with emotionML [12] tags, what in turn can be converted to a full script to control an embodied agent.

6.1 Text segmenter

The first part of each of the algorithms including the random one is the text segmentation element. This segmentation process functions by searching for interpunction, after each occurence of a '.' or ',' or ';' or ':' or ' " ' a new segment will start, if 2 forms of interpunction are next to eachother for instance in the case of ' :" ' they will be taken as 1 element of interpunction. Each occurence of ' " ' will also serve as a marker for a change of character perspective, from first to third and vice versa.

6.2 Random algorithm

To serve as a baseline comparison and a test environment the first algorithm is based on a randomizer. It functions as followed:

1. The text is segmented.
2. Each segment is fed to the random script one by one
3. Each segment is returned with attached a tag containing: one random emotion, with a random intensity, the emotion is linked to the speech fragment and placed at the end of the script (this is needed for the scheduler of the BML realisation).
4. To achieve a gesture rate of about 1:10 words, the text is processed word by word and given a 90 % chance to not have a gesture following, and a 10 % chance to have a gesture following. The 1:10 chance is a baseline that after experimenting seemed not to crowded. It can be adjusted by the interface. These gestures are added next to the emotion where they are being used.
5. If a word has a following gesture the randomizer picks a random gesture with random intensity.

The result is a text that can easily be turned into a BML script with scripted text accompanied by random gestures and emotions at certain time intervals appropriate for emotional changes.

6.3 Getting the proper parameters

In the previously mentioned random tagger algorithm, the basic functionality is working. However the return information is useless since it is not based on the input (other than the segmentation algorithm). To obtain the proper emotion and gesture tags the next part of the algorithm is needed where we obtain the proper story context on which the awarding of emotions, intensities of emotions and facial gestures is based.

6.4 General information algorithm

After getting to the conclusion that in order for the program to function properly we need to find general information first before we can start processing the text for tags, I created an extra framework step. This is the general information searching algorithm. It searches for tags that describe the story title and author, and it tries to determine who the protagonist and antagonist within the story is. Because this program is restricted to the telling of fairytales, we have several good guidelines to finding these. First of all, finding the protagonist is relatively straightforward:

1. Collect all noun phrases within the text.
2. Rank them by how often they are mentioned.
3. From the most popular noun phrases. From the most popular noun phrases(the threshold for this can be set in the program but the default threshold is that it needs to be mentioned at least 30 % of the times of the most popular phrase,) we take the one t hat is mentioned first in the story. The default threshold of thirty percent was established by testing with several input stories.
4. Without knowledge of the context and definitions we now set this noun-phrase as the protagonist. This by turned out to be rather accurate though exactly how accurate this is has not been determined.

In order to optimize the story there is an option to use a dictionary. In this case it will take the same set of noun phrases as found in the algorithms and then compares it to a dictionary of common protagonists and antagonists combined with their nemesis. If it finds a set in the dictionary that matches the popular noun phrase set it will use this instead of the most common phrase. Since all test inputs were also a subject of the protagonist antagonist index the accuracy here is 100 Determining the antagonist is a little bit more complex than the protagonist. As mentioned before if the protagonist/antagonist dictionary is used it will just compare it to this and find a good match. However if it cannot use this(because it is turned off), it will try to find the noun phrase that is last mentioned in a story sentence where there is also a word that involves dying mentioned. That noun phrase will be placed as antagonist. This works as long as the vocabulary of demise words is large enough, and the antagonist actually dies in the story, which is very common in fairytales but not always the case. The downside is that since in some cases another not uncommon noun phrase is involved in killing (so the fragments contains a death related word) the antagonist. In this case if the one of the options is the already identified protagonist it will be the other noun phrase. However if there is a second common noun phrase which is not the protagonist, This can cause some confusion in determining the right noun phrase. The solution I chose for this is to simply pick the last occurring one, which seemed to work for the test inputs[4], however this can be a case of (over)fitting.

6.5 Emotional content determination

After the general story information has been collected we have enough information to start applying character emotions and emotional intent to the story. This is done by in two distinct different ways depending on the perspective of the story fragment (i.e. first person or third person). In case of the first person way it depends on whether there is a colon present. If there is an adjective and a verb that means pronunciation followed by a colon then that can be of high influence. If this found adjective is known in one of the dictionaries that determine emotional load for first person perspective it will determine the entire load of the following fragment. For example:

Fearful he said:"oh my, I think I am lost."

The first half of this fragment would be neutral while the first person fragment, the second half, would be characterized by the word fearful thereby placing âĂIJfearâĂİ emotional content tags in front of this fragment. Otherwise the emotion determination algorithm will take the sum of all emotionally loaded words (take individual words and compare them to the emotional dictionary) and normalize this (contradicting emotion like happiness and sadness will cancel each other out if their sum intensities is equal). The third person perspective is a three stage algorithm. The first step is to take the emotional load of the sentence in very much a similar way as step two in the first person perspective. It will determine for each part of speech what emotional load it default invokes, and will sum and normalize this. The second step is determining how this relates to the protagonist and antagonist, and depending on the tagging method pick a proper invoking emotion. The final step is determining what expression to use to accomplish this.

6.6 Gesture determination

The last part for generating a complete tagged text is the adding in of the gesture tags. There are several parts to the determination of where gestures are appropriate. First of all there is the indication of a questioning sentence. To find this the gesture scanning algorithm simply scans for the basic words indicating a question(who where what, etc.). Next to this there is the attempt to find the emphasis of the sentence. To do this an external program was used that is able to determine stress in a sentence. Unfortunately the base text of this experiment was in Dutch while this program functioned in English. So in order to make this work the fragment is sent to a translation server. Afterwards it is translated to English, and after that it is sent to the server to determine the stress of this fragment. After that it will be send back to translate the word with most stressed in an attempt to add a emphasis gesture here. The final gestures are done at the place of interpunction. By scanning for exclamation mark we can decide to make the stress more expressive, or by finding question marks we can add an extra quizzical remark instead of a stress remark in this segment.

7. FILE PROCESSING

After processing some stories and doing some test runs the inefficiency of the program itself dawned. The problem being that every time a story needs to be processed, first it needs to be analyzed, the general information found, the tagging algorithms both need to do their jobs and then the story is finished and it can be exported or broadcast to a network port. Since the algorithm is complex and intense this will take up a lot of time which can be avoided by simply creating an object writer around the already existing story model object.

7.1 Export/import versus save/ load

At this point in the development process an extra option was build into the system to not only import and export stories, but also to store fully analyzed and tagged stories on the local system so they can be adjusted more easily. This functionality also helps saving time with the testing of the several expression algorithms. As mentioned in the Emotional content paragraph2 there are several different methods that be used to create the right emotional response from the audience. Since the underlying information, i.e. "try to invoke emotion x" is independent from the technique used for it, it is useful if we can just load the underlying information several times and just directly generate the correct expressions from that instead of doing the first steps several times over again.

7.2 Third party software

Firstly to do part of speech tagging the Alpino[3] system was used. In this case the windows version was used. By making use of the console interface the to-be-analyzed text was part of speech tagged, written to a file which file is after that loaded again in the tagger tool to gain all the necessary part of speech information such as the recognition of adjectives and noun phrases. This is very valuable as a filter for important words however it still tells us very little about the context by itself. For this at first an attempt was made with WordNet, more on this in a later paragraph, this development stage of the software failed. After the WordNet attempt another approach was chosen. By making use of a translation online by Google translation, and using the retrieved information for context analyses by an existing Affective Dictionary. This is not extremely accurate but should be enough to serve as a proof of concept.

8. USABLE RUNTIME DICTIONARIES

A special feature to make the system more adaptable is the implementation of the runtime dictionaries. To make sure that the system has the proper vocabulary to perform its script building tasks we have created several dictionaries that can be adjusted while the system is running. In order to make everything fast enough the dictionaries are ordered alphabetically automatically with a quick sort[7] during the loading process, if they are not in correct order, after checking (and if needed sorting) a binary search can be used to find items making worst case search times Log(n). For future work and more versitility it might be desireable to create these dictionaries in a seperate (SQL based) database, this would enable multi user support and would make it easier to adjust the system.

8.1 Character dictionary

The first database is the one that is optionally usable, the story character database. This dictionary contains an alphabetical name of known common story characters, like wolves, witches, little Red Riding hood, Snowwhite, sleeping beauty, etcetera. Each common character is connected to zero or more nemesis. Aside from this, each character is marked as most likely being a protagonist, most likely being an antagonist or both or none, being selected as none however will set the character to inactive. The last thing they have is a list of used synonyms, which solves a problem if there are several well known characters in the same story that only the one first named will be the protagonist.

8.2 Filter dictionary

Next to this there is the "standard useless filling"-words dictionary. These are words that are rather common in fairytales but serve little to no purpose for determining emotional content. Words like the forest as in the place where it takes place is a common early named noun which with the current used algorithm can make the forest as the protagonist and therefore needs to be filtered. This happened to be the case in the dutch version of the snowwhite story. Next to standard personal pronouns and similar referring words need to be filtered as much as possible.

8.3 Emotional implication dictionaries

The third dictionary within the system is the basic emotions dictionary. This is a set of words that invoke certain emotions by some people[9]. This list can be modified for a long time and made larger easily, but due to the size of this part of the research I choose to keep this of limited size and use external resources, more on. The list consists out of three different types of words. The first one and most important one is the adjectives. After experimenting with the tagger tool and comparing my results to that of other people in this field of research I came to the conclusion that this is the most heavy weight component when it comes to invoking emotions. Secondly certain nouns can cause specific emotions in fairytales. In this domain there are standard objects that serve a very specific purpose, for example poison apple is always supposed to invoke fear, as with the snowwhite fairytale, likewise for the "Spinnewiel" of Sleepingbeauty. Because nouns are generally far less expressive, the probability that they determine the overall emotion of the segment is lower.

8.4 Adverb dictionaries

Lastly there is the adverb that by itself does not determine the emotion but rather modifies the strength of the adjective or noun directly followed by it. Example: "The extremely scared girl ran into the forest" is more intense than just "the not so scared girl ran into the forest". So unlike the adjective and noun they do not influence what emotions are present in the segment, they just amplify, reduce or invert the intensity with the proper amount. The choice to use this method is based on the research[2] that emotional content determination is more accurate on adverbs in combination with adjectives than adjectives alone.

8.5 Character elimination dictionary

The last modifiable dictionary is the one required for the finding of the non listed antagonist. If the user of the system wants the system to determine who the antagonist is by itself without the aid of an antagonist index, then the system will do this by finding an often used noun in a text segment that involves the termination of a story character. In general fairytale storytelling as mentioned before in the paragraph about the finding of the antagonist, usually the antagonist meets its demise near the end of the story or is gotten rid of in some other way. The termination in any form of this character can be then seen as the victory of good over evil. In some stories (like in several versions of the sleeping beauty story for example) the antagonist does not die, so there this system does not work and the listed antagonist still needs to be used. The concept for this algorithm is based on existing fairytale structures and experiments performed during this research. Another interesting problem case is when a slightly different kind of fairytale is picked. For example if we take the little mermaid in its original form by Hans Christian Anderson, the protagonist is actually put as antagonist because she is the one who meets her demise at the end of the story.

9. EMBODIED CONVERSATIONAL AGENT

At first Deira[8] was the chosen embodied agent that would execute the created scripts. The choice of this agent was based on the fact that this agent had been specifically written for the virtual story teller. The embodied agent could provide basic emotional expressions as well as well as other facial movements that could serve as gestures. This was also the agent for which the original socket connection was written. However during the research there appeared to be several licensing problems with this agent causing a shift to another embodied agent, ELCKERLYC.

9.1 The ELCKERLYC agent

The ECA that performs the role of story teller in this experimental story teller system is the Elckerlyck agent [17] as developed by the university of Twente. This embodied agent is able to combine facial gestures with several different forms of text to speech engines, in this specific system we will be making use of Loquendo text to speech engine for Dutch language. It is able to express emotions, speak at various rates of pronunciation and is capable of showing several gestures based on BML input. By providing a proper script to the embodied agent we can create a video performance that suits the needs of the experiment. It has been demonstrated to be capable of real time realisation of the scripts and it is also capable of being expanded with extra emotions, gestures and expressions if required.

10. EXPERIMENT

In order to determine the answer to the research questions put forward, i.e. "Can we determine the emotional content of pure text by algorithm?" and "What approach should we use to evoke the proper emotions with the listeners and viewers of the story teller?" an experiment was set up. First of all we need to put forward a baseline for the null hypothesis, this will be a randomly tagged text. In order to see whether we can convincingly tag the pure text input with the algorithm the non-random should at least perform better than the randomly tagged text. Secondly to answer the question of what evocation approach we need to use we need to compare which of the evocation tagging methods is better. The last explicit text is one that mirrors the empathic version, this should perform the worse since the emotions should make no sense at all.

10.1 Procedure

In order to perform the actual experiment we needed to get a test input that suit the domain. The three fairytales that were picked that were chosen, red riding hood, sleeping beauty and snowwhite, were choosen because they are well known and one has been used before in the virtual story teller environment. Because they are well known it made it easy to just take a fragment from the story instead of the entire story, without having to worry about the lack of knowledge about the story from the viewer and listener.

The three story were used as input in dutch text and processed according to four algorithms, one random baseline, secondly the empathic version where we mimic the emotion that the reader is supposed to experience and the story teller represents the writer, thirdly the Frey method where the story teller represents a character that evokes specific emotions and a mirrored version of the empathic, to see whether an inverted representation of a more common method is experienced as less.

After obtaining the output scripts we can continue to process them with the BML realizer platform of the Elckerlyc agent. Important here is that this is still the full story, in order for the algorithm to function it needs to full text from beginning to end, after realizing it with the agent it can be cut into fragments that can be presented to the test subjects. In previous attempt to measure presentation quality of an embodied agent and test runs we had already determined that full text duration would come at a cost of interest and would already cause a lower rating than when we measure just a small part of a story. Since we want to measure the difference between all the methods it is important that other external factors, such as a loss of interest, do not disrupt the results of the experiment.

10.2 Measurements - what constitutes "better"

When comparing the several presentations that the virtual storyteller gives with eachother we need to define what it actually means for method A is better than method B. The use of five or seven points Likert scales on the factors of anthropomorphism, animacy, likeability, perceived intelligence and perceived safety have been validated to be successful in determining the performance of automated interaction systems that make use of embodied agents[1]. Cronbach's alpha values found in previous tests using these questionares are all above 0,70 showing their reliability.

10.3 Participants

The participants in the experiment were all part of a convinience group, all were either fellow students, former fellow students and colleagues. Due to the fact that all story telling methods are produced in the same way and presented in the same way there should be no influence of social bias. The average age was 31 years, with a standard deviance of 3 years. The male to female ratio was 2,3:1.

10.3.1 Subject tasks

The general set up of the experiment was done in a digital presentation format, with use of PowerPoint. The subjects were presented with six slides. The first slide contained their subject number and an opening logo. This was done in order to make sure that the subject number matches the number on the polling paper. The second page shows a short introduction. Telling them what to expect and what they should try to focus on. This is done in order to avoid the person judging the performance of the embodied agent within the parameters of this research and not on the known shortcommings the system has.

Following these first two introduction sheets are the sheets of the stories themselves. In order to prevent that the users have to deal with the possible imperfections of the tagger tool and the duration of generation of script and movie, prerendered movies have been used. They watch three short movies. Each of the movies is one single fragment of one fairytale. So we have three fragments from three fairytales. The choice to use three different stories is to further give a proof of concept that the system is independent of the story input, and it should also demonstrate that the performance is consistent per method of tagging independent of the type of story. Each story was tagged with each method, this enables to make cross examine the performance for each tagging method.

The test subjects were asked to perform this test in a controlled environment in such a way that they could only open the presentation and were able to ask questions if anything remained unclear. This also enabled the observer to make notes during the experiment.

The participants were asked to fill in the questionare that can be found in the appendix. Next to that they were personally interviewed on reasons why they rated each element the way they did, this in order to see if there were any side effects. For instance it is possible that even the performance was bad, it can still be rated as pleasant since the awful performance was unintentionally hilarious due to all the incorrect display of emotions, such as a clown making funny faces during a supposedly sad moment.

Next to the information that allows us to draw a conclusion they were also asked to fill in information about age and experience with embodied agents. This allows us to search for possible patterns on why certain people judged the performance the way they did. And aside from that they had the possibility to give an open response to the asked questions to elaborate their answer further.

10.4 Used tagging methods

The three other methods (beside the random ones) are the Frey methods[5][6], the empathic method[9] and the Frey method where the emotional opposites are flipped (happiness becomes sadness, anger becomes fear and disgust becomes surprise, henceforth to be described as the mirrored Frey methods). The choice for the first two methods (Frey method and empathic) specific methods is because they are according to the research done in preparation of this experiment the most likely candidate for an optimal result. The last method with mirrored emotions should have a performance that is worse than the suggested methods and perhaps even worse than random.

10.5 Drawing conclusions

After getting the results from all stories we can group them by evocation method and derive mean values from each of the catagories we can compare the different evocation methods with eachother. If we see a significant difference in a single tailed T-test between the two groups where method A is better (as defined by the subsection on measurements) than method B we can accept conclude that on that specific catagory it performed better. The desired outcome is that at least both non-random methods perform significantly better than the random one.

11. EXPERIMENT RESULTS

The results from the experiments were rather disappointing. In all cases the participants graded the performance as very low. There were some interesting differences here, for example the part of anthromorphism where people did rate the realism of the ECA rather high, while at the same time they were considered as machinelike. A very common complaint was that the staring was in some cases uncomfortable very machinelike. The pauses between sentences were also considered too long for the participants, although they did match the pauses used between story parts of a human story teller (0.3 seconds between story parts).

When we look at all results we can see the following summary

Random	mean	deviation
anthropomorphism	2.33	0.577
animacy	2.14	0.412
likeability	2.1	0.750
perceived intelligence	2.33	0.225
perceived safety	6.33	1.100

empathic	mean	deviation
anthropomorphism	2.65	0.612
animacy	2.315	0.412
likeability	2.250	0.500
perceived intelligence	2.6	0.540
perceived safety	5.91	0.900

Frey	mean	deviation
anthropomorphism	2.50	0.215
animacy	2.50	0.577
likeability	2.45	0.628
perceived intelligence	2.8	0.444
perceived safety	5.50	1.200

Mirror	mean	deviation
anthropomorphism	2.45	0.500
animacy	1.81	0.178
likeability	2.00	0.450
perceived intelligence	2.33	0.788
perceived safety	6.00	1.409

When comparing all of the algorithems in pairs by T test, there are no real significant differences. Though the differences that are there, slightly favor the empathic behaviour for realism, and the Frey method for intelligence and likability. None of the methods used resulted in any form of intimidation, which worked as a strong disadvantage for the Frey method which used this tactic to evoke fear with the participants.

Another possible source of the relatively low rating is the fact that nine out of the twenty participants had already seen the story teller in a more personal environment (namely as a behavior change support coach). The suddenly less personal method of talking and the lacking of adressing the participant by name was found to be rather inpersonal since this had always been present in their previous encounters.

12. DISCUSSION

In this research there were several goals set to achieve. First goal was to create a instruction script for an ECA from pure text fully automatic. The second goal was to make this script a good one that performed the roll of a storyteller well. In the following paragraph some successes are being discussed, some failures and some places where some assumptions could be further analyzed, such as the further development of the emotion dictionary.

12.1 Development successes

The major success is basically that there is a system that functions on just pure text input that can generate a script for an embodied agent without any need for external input beside this. The interface for doing this contains only a limited amount of adjustable settings, enough to "tweak" the algorithms yet not too much to make it confusing and difficult to use.

12.2 Concessions

The initial development design stated that it should be an individually program with perhaps some external sources directly connected to it. Unfortunately the use of the third party software prevented this. Next to this due to the inability to insert WordNet, the context analysis of the story has be seriously restricted, making the system less accurate than desired. So the first concession that was made was that the system can only work if there is an Alpino[3] version installed on the same machine and with a proper path which need to be individually configured. The second concession is that the program can only work as long as it is connected to the internet since it needs Google and Affective Dictionary Services in order to function properly. The third concession that had to be made is the accuracy of the system is lower than desired due to the lack of a professional dictionary environment. This last concession could however be fixed with simply future development by adding in more words and more precisely calibrating the existing internal dictionary. This is rather time intensive and has to remain a future work element.

12.3 Future development

At this part the development of the tagger tool for this particular research has been sufficient to do initial testing. However if there was a wish for further use of this tool there are several points of future development that can be addressed.

1. Better third party software integration. With some extra time for development there could be better use of context analysis and the part of speech tagging could be better integrated making the system more easily installable and perhaps less platform dependent. The current construction only allows windows use due to the console construction.
2. Multiple-document loading. With the program in its current form, only one story can be loaded at a time. A small change in the interface and internal storage of the documents should enable multiple document storage.
3. Larger set up user tests. The current timeframe does not allow a large set of user test to determine the effectiveness of the individual approaches of Agent script generation.
4. Creating larger dictionaries with more accurate emotional references.
5. Making use of a multi role Embodied agent, that can directly interpret emotions and can talk using multiple voices. Also a more comical or cartoonish style agent could be more effective since it would allow for demonstrating more extreme emotions without looking "weird".

13. CONCLUSION

The overall research can at least be called partly successful. With the information that can be attained from pure text input we can automatically generate a script that can be used by an embodied agent. The information gathering that can be done on the domain of fairytale storytelling proved to be fairly accurate, as in the algorithm in nearly all cases correctly identified the protagonist and antagonist, and when we look at the models, the empathic and Frey method of story telling both put the right emotions at the right places. If the information that was already present within the virtual storyteller be parsed to the tagger tool, there would actually be less need for this process. The experiment failed in execution however, the chosen medium in combination with the chosen test participants, where the experiment yielded no significant results. The previous experience of the participants with the embodied agent seemed to prevent them from objectively looking at the different character the same agent was portraying, while the chosen agent was not really expressive enough with the generated scripts to act naturally.

14. REFERENCES

[1] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived

Intelligence, and Perceived Safety of Robots. *International Journal of Social Robotics*, 1(1):71–81, January 2009.

[2] Farah Benamara, Carmine Cesarano, Antonio Picariello, Diego Reforgiato, and V. S. Subrahmanian. Sentiment analysis: Adjectives and adverbs are better than adjectives alone. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, 2007. Short paper.

[3] Gosse Bouma, Gertjan Van Noord, and Robert Malouf. Alpino: Wide-coverage computational analysis of dutch. *Language and Computers*, 37(1):45–59, 2001.

[4] P. Ekman. About brows: Emotional and conversational signalsek. *Human ethology*, pages 169–202, 1979.

[5] James N. Frey. *how to write a damn good Novel,*. St. Martin's Press, December 15, 1987. ISBN - 13: 978-0312010447.

[6] James N. Frey. *how to write a damn good Novel volume 2*. St. Martin's Press, March 15, 1994. ISBN-13: 978-0312104788.

[7] Bonnie Neff Gregory Neff and Paul Crandon. Assessing the affective aspect of languaging: The development of software for public relations. In *the 52nd Annual Conference of the International Communication Association in Seoul, Korea,*, 2002.

[8] F.L.A. Knoppel, A.S. Tigelaar, D. Plass-Oude Bos, T. Alofs, and Z.M. Ruttkay. Trackside deira: A dynamic engaging intelligent reporter agent (demo paper). In L. Padgham, D. Parkes, J. Mueller, and S. Parsons, editors, *Proceedings of the Seventh International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS08)*, volume 1, pages 1681–1682, Estoril, 2008. The International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS). ISBN=978-0981738116.

[9] Thomas Christiaan Nijmeijer. Narrative perspectives in the virtual storyteller. Capita Selecta paper HMI.

[10] Aldo Paradiso and Marcello L'Abbate. A model for the generation and combination of emotional expressions. In *Workshop on Representing, Annotating, and Evaluating Non-Verbal and Verbal Communicative Acts to Achieve Contextual Embodied Agents, Fifth International Conference on Autonomous Agents, May 29, 2001*. McGraw-Hill, 2001.

[11] Isabella Poggi and Catherine Pelachaud. Performative facial expressions in animated faces. In *Embodied Conversational Agents*, pages 155–188. MIT Press, 2000.

[12] Marc Schröder, Paolo Baggia, Felix Burkhardt, Alessandro Oltramari, Catherine Pelachaud, Christian Peter, and Enrico Zovato. Emotion Markup Language (EmotionML) 1.0. W3C Working Draft, July 2010.

[13] Marie L. Shedlock. *The Art of the Story-Teller*. Echo Library, January 24, 2007. ISBN-13: 978-1406815221.

[14] Jianhua Tao. Context based emotion detection from text input. *INTERSPEECH2004*, pages 1337–1340, 2004.

[15] M. Theune, S. Faas, D.K.J. Heylen, and A. Nijholt. The virtual storyteller: Story creation by intelligent agents. In S. Göbel, N. Braun, U. Spierling, J. Dechau, and H. Diener, editors, *TIDSE 2003: Technologies for Interactive Digital Storytelling and Entertainment*, pages 204–215, Darmstadt, 2003. Fraunhofer IRB Verlag. ISBN=3-8167-6276-X.

[16] Peter Turney. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of ACL-02, 40th Annual Meeting of the Association for Computational Linguistics*, pages 417–424, Philadelphia, US, 2002. Association for Computational Linguistics.

[17] H. van Welbergen, D. Reidsma, Z.M. Ruttkay, and J. Zwiers. Elckerlyc - a bml realizer for continuous, multimodal interaction with a virtual human. *Journal on Multimodal User Interfaces*, 3(4):271–284, 2010. ISSN=1783-7677.

[18] Elizabeth A. Winston. *Storytelling and conversation: discourse in deaf communities*. Gallaudet University Press, 1999. ISBN-13 978-1563680816.

Figure 1: Page 1 questionaire

Deelnemer nummer:.....
Leeftijd: ..... jaar
Geslacht: Man/Vrouw (doorstrepen wat niet van toepassing is)
Ervaring met agent: Ja/nee (doorstrepen wat niet van toepassing is)

U gaat zo deelnemen aan een experiment, de verdere uitleg hiervan zult u zien op de eerste sheet van de presentatie, die nu voor u te zien is op het scherm. Veel luister- en kijkplezier.

Figure 2: Page 2 questionaire

Verhaal 1

Ik vind de verhaal verteller:	1	2	3	4	5	6	7	
Onvriendelijk	O	O	O	O	O	O	O	Vriendelijk
Incompetent	O	O	O	O	O	O	O	Competent
Verantwoordelijk	O	O	O	O	O	O	O	Onverantwoordelijk
verstandig	O	O	O	O	O	O	O	Dwaas

Door de verhaalverteller voel ik mij:	1	2	3	4	5	6	7	
Zenuwachtig	O	O	O	O	O	O	O	Ontspannen
Geïrriteerd	O	O	O	O	O	O	O	Kalm
rustig	O	O	O	O	O	O	O	Verrast

De verteller is in mijn ogen:	1	2	3	4	5	6	7	
Levendig	O	O	O	O	O	O	O	Dood
Onbewust	O	O	O	O	O	O	O	Zelf bewust
Kunstmatig	O	O	O	O	O	O	O	Natuurlijk
Machine achtig	O	O	O	O	O	O	O	Mensachtig

De manier van vertellen en Emotie vertoon ervaar ik als:	1	2	3	4	5	6	7	
Realistisch	O	O	O	O	O	O	O	Nep
Inert	O	O	O	O	O	O	O	Interactief
Aangenaam	O	O	O	O	O	O	O	Onaangenaam
Verschrikkelijk	O	O	O	O	O	O	O	Leuk

Het emotie vertoon, en de bewegingen zijn:	1	2	3	4	5	6	7	
Elegant	O	O	O	O	O	O	O	Stijf
Mechanisch	O	O	O	O	O	O	O	Organisch
apathisch	O	O	O	O	O	O	O	Reactief
Stilstaand	O	O	O	O	O	O	O	Bewegelijk

Ik vind de verhaal verteller:	1	2	3	4	5	6	7	
Aardig	O	O	O	O	O	O	O	Onaardig
Onwetend	O	O	O	O	O	O	O	Wijs
Onvriendelijk	O	O	O	O	O	O	O	Vriendelijk
Dom	O	O	O	O	O	O	O	Intelligent

Figure 3: Page 3 questionaire

Verhaal 2

Ik vind de verhaal verteller:	1	2	3	4	5	6	7	
Onvriendelijk	O	O	O	O	O	O	O	Vriendelijk
Incompetent	O	O	O	O	O	O	O	Competent
Verantwoordelijk	O	O	O	O	O	O	O	Onverantwoordelijk
verstandig	O	O	O	O	O	O	O	Dwaas

Door de verhaalverteller voel ik mij:	1	2	3	4	5	6	7	
Zenuwachtig	O	O	O	O	O	O	O	Ontspannen
Geïrriteerd	O	O	O	O	O	O	O	Kalm
rustig	O	O	O	O	O	O	O	Verrast

De verteller is in mijn ogen:	1	2	3	4	5	6	7	
Levendig	O	O	O	O	O	O	O	Dood
Onbewust	O	O	O	O	O	O	O	Zelf bewust
Kunstmatig	O	O	O	O	O	O	O	Natuurlijk
Machine achtig	O	O	O	O	O	O	O	Mensachtig

De manier van vertellen en Emotie vertoon ervaar ik als:	1	2	3	4	5	6	7	
Realistisch	O	O	O	O	O	O	O	Nep
Inert	O	O	O	O	O	O	O	Interactief
Aangenaam	O	O	O	O	O	O	O	Onaangenaam
Verschrikkelijk	O	O	O	O	O	O	O	Leuk

Het emotie vertoon, en de bewegingen zijn:	1	2	3	4	5	6	7	
Elegant	O	O	O	O	O	O	O	Stijf
Mechanisch	O	O	O	O	O	O	O	Organisch
apathisch	O	O	O	O	O	O	O	Reactief
Stilstaand	O	O	O	O	O	O	O	Bewegelijk

Ik vind de verhaal verteller:	1	2	3	4	5	6	7	
Aardig	O	O	O	O	O	O	O	Onaardig
Onwetend	O	O	O	O	O	O	O	Wijs
Onvriendelijk	O	O	O	O	O	O	O	Vriendelijk
Dom	O	O	O	O	O	O	O	Intelligent

Figure 4: Page 4 questionaire

Verhaal 3

Ik vind de verhaal verteller:	1	2	3	4	5	6	7	
Onvriendelijk	O	O	O	O	O	O	O	Vriendelijk
Incompetent	O	O	O	O	O	O	O	Competent
Verantwoordelijk	O	O	O	O	O	O	O	Onverantwoordelijk
verstandig	O	O	O	O	O	O	O	Dwaas

Door de verhaalverteller voel ik mij:	1	2	3	4	5	6	7	
Zenuwachtig	O	O	O	O	O	O	O	Ontspannen
Geïrriteerd	O	O	O	O	O	O	O	Kalm
rustig	O	O	O	O	O	O	O	Verrast

De verteller is in mijn ogen:	1	2	3	4	5	6	7	
Levendig	O	O	O	O	O	O	O	Dood
Onbewust	O	O	O	O	O	O	O	Zelf bewust
Kunstmatig	O	O	O	O	O	O	O	Natuurlijk
Machine achtig	O	O	O	O	O	O	O	Mensachtig

De manier van vertellen en Emotie vertoon ervaar ik als:	1	2	3	4	5	6	7	
Realistisch	O	O	O	O	O	O	O	Nep
Inert	O	O	O	O	O	O	O	Interactief
Aangenaam	O	O	O	O	O	O	O	Onaangenaam
Verschrikkelijk	O	O	O	O	O	O	O	Leuk

Het emotie vertoon, en de bewegingen zijn:	1	2	3	4	5	6	7	
Elegant	O	O	O	O	O	O	O	Stijf
Mechanisch	O	O	O	O	O	O	O	Organisch
apathisch	O	O	O	O	O	O	O	Reactief
Stilstaand	O	O	O	O	O	O	O	Bewegelijk

Ik vind de verhaal verteller:	1	2	3	4	5	6	7	
Aardig	O	O	O	O	O	O	O	Onaardig
Onwetend	O	O	O	O	O	O	O	Wijs
Onvriendelijk	O	O	O	O	O	O	O	Vriendelijk
Dom	O	O	O	O	O	O	O	Intelligent