ABSTRACT
In this paper, we present MARVIN, a futuristic newsroom simulation. MARVIN was developed with the aim to find out if insights into future methods of interaction with 'Robot Journalism' can be obtained by researchers through a newsroom simulation. The simulation that was developed consists of three scenarios that involve typical journalistic tasks: Data Analysis, Writing and Interviewing. A total of nine test participants with journalistic backgrounds tested either one or two of these scenarios at a specially prepared room at the University of Amsterdam, which represented the 'futuristic newsroom'. We found that MARVIN allowed journalists to experiment with various interactions with tools that could one day play an important role in their day-to-day work. MARVIN allowed them to think about these tools, while sparking conversation and debate about various possibilities and preferences. MARVIN therefore succeeded as an approach for researching journalists’ preferences regarding forms of interaction with Robot Journalism in the future.

Keywords
Robot Journalism; Simulation; Research Methodology; Wizard of Oz; Storytelling;

1. INTRODUCTION
In 2009, the L.A. Times created a program called Mapping L.A. that aimed to automatically present and update maps regarding demographic data, crime rates and school statistics for each of the 272 neighborhoods that make up the country. In 2010, Larry Birnbaum, Stuart Frankel and Kristian Hammond founded Narrative Science, a software Company that provides newsrooms with automatically written news (Levy, 2012). In 2013, The Washington Post announced they would be creating Truthsller, a live fact-checking platform for political speeches (Truth Teller, 2013).

These examples are part of the development of the field Nicholas Diakopoulos (2011) calls computational journalism, or “the application of computing and computational thinking to the activities of journalism including information gathering, organizing and sensemaking, communication and presentation, and dissemination and public response to news information, all while upholding core values of journalism such as accuracy and verifiability”. More specifically, they belong to an area within computational journalism called ‘Robot Journalism’. We define Robot Journalism as a form of automation in which algorithms are used within at least one phase of the journalistic process, wherein the algorithm makes at least one autonomous decision somewhere within this process.

In this context, the goal of the research presented in this paper, is to find out how insights into future methods of interaction with ‘Robot Journalism’ can be obtained by researchers. This objective arose as part of Fontys FutureMediaLab’s pursuit to research and develop tools and algorithms that respond to the needs of journalists in the future. As a starting point for this research, project leader Eefje Op Den Buysch developed a concept for an immersive news environment (Op Den Buysch, 2016). Our research takes her ideas a step further by proposing an approach for implementing such an environment in the form of an interactive ‘Robot Journalism’ simulation, which received the name of MARVIN. Through MARVIN, as we will show in this paper, researchers can observe how journalists would interact with prototypes (interface and interaction designs) of several central technological tools of the future of journalism. Journalists will then engage with these tools as part of the simulation, and reflect on their experiences afterwards.

In the next chapter, related works are discussed and some key concepts are defined. Then, the rationale behind the chosen simulation strategy is explained and the overall system components are described. In the following chapter, both the design and the content of the simulation are explained. Furthermore, the tests and evaluation methods that were used will be discussed. Finally, we will discuss our findings and draw a conclusion in relation to the goal that we set out with.

2. RELATED WORK
The design and creation of a robot journalism simulation entailed, first of all, an understanding of what robot journalism is, as well as an understanding of how journalists would react when they first encountered the system. In the next section, we will present a discussion of these issues, as well as an overview of the current state of academic research regarding robot journalism.

2.1 Definition and Scope of Robot Journalism
Robot journalism (or automated journalism) is a concept that emerged around 2009, when the New York Times published a story about Stats Monkey, a project conducted by students at the Intelligent Information Laboratory at Northwestern University. Stats Monkey was the first system that allowed for systematic compiling of sports statistics and computer-written news stories based on those statistics (Carr, 2009). The technique adopted by the students is also known in the Computer Science field as Natural Language Generation (NLG) (Van Der Kaa & Krahmer,
The introduction of Stats Monkey greatly shaped the concept of Robot Journalism, as the term has mainly been used to refer to automatic content generation within online journalism ever since, in subjects with a significant amount of structured data, such as finance, polling results and sports (Graefe, 2016). However, to scope the definition of Robot Journalism to only include NLG techniques entails excluding processes of automation that are being pursued in areas without as much structured information, as well as in other types of media, such as print or broadcast, which have space and time limitations, and therefore no use for NLG (Carlson, 2014).

In this regard, robot journalism can better be considered as a subfield of computational journalism (Cohen, Hamilton & Turner, 2009; Young & Hermida, 2015), which refers to “the application of computing and computational thinking to the activities of journalism” (Diakopoulous, 2011). Specifically, Graefe (2016) defines robot journalism as “a relatively new phenomenon in the area of computational journalism”, that mainly refers to the use of algorithms to “automatically generate news stories without human intervention” but that also “allows for automation in each step of the news production process, from the collection and analysis of data, to the actual creation and publication of news”.

For the purpose of this research, the second part of the definition provided by Graefe (2016) will be prioritized, and reformulated in order to fit the purpose of this research. Therefore, from this point forward, Robot Journalism will be considered as a form of automation in which algorithms are used within at least one phase of the journalistic process, wherein the algorithm makes at least one autonomous decision somewhere within this process.

2.2 Journalism and Technology Development

The design of a system that should allow researchers to get insights into what tools they should develop for future journalists, mandatorily entails dealing with the relationship journalists have with technological advancements in their field.

First of all, it is necessary to consider how technology has influenced journalism historically. In this regard, Pavlik (2000) clearly states that technology has had a deep influence over the way news production has been conducted. According to him, technology has not only affected the way journalists work, their daily practices, but also what they write about and how the news stories circulate once published. Cottle and Ashton (1999) and Brill (1995), however, point out that the introduction of new technology in newsrooms doesn’t necessarily entail improvements: the development of new skill sets for journalists as well as adapting to new practices can also prove prejudicial to journalists’ activities.

One of the reasons this happens is that the introduction of technology does not impact its acceptance. Quite to the contrary, Zhou (2008) argues that there are four potential reactions journalists can have to the introduction of new technology in their newsrooms: voluntary adoption, forced adoption, resistant non-adoption, and dormant non-adoption.

In the case of Robot Journalism, journalists have reacted with scepticism regarding the benefits it might bring to their field (Carlson, 2014). When reporting about automation in newsrooms, the postures oscillate between a fear of replacement to a promise of augmentation of their labour: repetitive tasks would be carried out by computers, either replacing low level workers or freeing up much of the journalists’ time for higher level content production (Carlson, 2014; Van Dalen, 2012). Gynnild (2013) also argues that “the impact of computational exploration in journalism is less dependent on technological creation than on news professionals’ values, goals, and interactional skills development”. Therefore, journalists’ backgrounds and perceptions regarding the incorporation of new technologies into their work environments have a profound relationship with their effect. In this regard, when it comes to the adoption of Robot Journalism, Carlson (2014) argues that “automated journalism complicates [technology adoption] dynamic[s] through intervention of non-journalistic actors”, which in his research referred to the firm Narrative Science—an NLG provider—, and within the context of this paper can refer to both its authors and Fontys FutureMediaLab.

2.3 Research Regarding Robot Journalism

Finally, because of robot journalism’s novelty, there has not been much research conducted in the field. The research has been particularly scarce with regard to its future implementations.

Most of the focus of academic research has been around the reader’s acceptance and perception of automatically generated content (Clerwall, 2014; Van Der Kaa & Krahmer, 2014; Van Dalen, 2012). Similarly, certain authors analyze the reactions of journalists to automatic content generation, and how traditional news writing compares to computer generated content (Graefe, 2016). Ethical and legal aspects of the rise of robot journalism have also been dealt with by academia: Anderson, Bell and Shirky (2012) discuss algorithmic transparency in news creation, while Weeks (2014) analyzes how automated journalism is covered by current jurisprudence in the United States.

The current paper, aims to expand knowledge regarding the future design and development of Robot Journalism, in relation to journalists’ needs and preferences.

3. APPROACH

3.1 Basic Principles

The development of a research method that could provide a means for obtaining insights into future methods of interaction with ‘Robot Journalism’ entailed defining a set of non-negotiable principles, that were primarily based on the conclusions drawn from the literature review.

First of all, most of the research into Robot journalism deals with the automation of news writing processes, which journalists often perceive as a threat to their job. In this regard, considering that journalists’ participation was indispensable for a successful simulation, it was mandatory that our design was one that did not alienate journalists. Therefore, we decided to focus on constructing scenarios composed only by tools that would aid journalists in their work—and never replace them.

Second, the collective “journalists” covers various backgrounds and a wide variety of tasks within newsrooms and media outlets. Considering that Fontys FutureMediaLab does not explicitly prioritize one type of journalist over another, the simulation had to be adaptable for any potential participant. This meant that any simulation would have to be relatable to a number of different participants. However, in order to reduce the scope of the simulation into a feasible task, it was decided that the simulation would focus on what is known as ‘hard news’: “politically relevant [news] item [that is reported in a thematic way, and] focuses on the societal consequences of events, is impersonal and unemotional in its style” (Reinemann, Stanyer, Scherr & Legnate, 2012).
Third, in order to evaluate different means of interaction with futuristic tools, it was necessary to design a simulation that allowed for variation, and wasn’t limited by the current state of technological development.

These three principles were central through the whole research and development process, as they provided the baseline for all the stages, from interview questions and research observations to the design requirements that led to the construction of MARVIN as the final simulation.

3.2 Methods

The development of MARVIN consisted of three stages carried out over two periods: an eight week ideation stage, and four weeks of iterative development and testing stages. The ideation of MARVIN was based on two inputs: personas constructed based on the researchers’ previous knowledge of journalistic activities, and field research and interviews in Dutch newsrooms.

The personas were built based on recent news events, such as the U.S. Presidential election, a conflict in a press conference given by an Argentinian football team, and a local event in Amsterdam. These personas aimed to illustrate the roles journalists could potentially take on, as well as provide grounds imagining potential Robot Journalism interactions and tools that could prove useful to them.

In parallel, a qualitative research approach was adopted to further explore journalistic practices and confirm (or refute) our pre-formulated ideas regarding journalism and newsroom environments. Newsrooms and journalists were contacted and visits were scheduled with two ‘hard news’ outlets: RTL Nieuws, a Dutch national news broadcaster, and AT5, a regional news agency based in Amsterdam. In both instances an observation approach was taken, where it was possible to observe a normal work day in the newsroom. Additionally, unstructured interviews were conducted after both observations by two members of the research team. The purpose of the interview was to obtain insights into the newsroom’s dynamics, as well as the interviewee’s usual work day. This allowed us to formulate a clearer idea of what the work of a journalist involves, and how it can vary from one journalist to another. Finally, this was a key moment to gather information concerning their perceptions (and preconceptions) with regard to Robot Journalism. From them, we could gather that journalists are sceptical but intrigued by Robot Journalism, confirming what the literature review had already portrayed.

The outcomes from the newsroom visits led us to decide that, while the personas that had been constructed were reasonably accurate (if a bit idealized), differentiating potential interactions based on subjects—for example politics, sports and regional news—would prove difficult to generalize to a broad spectrum of journalists, and therefore would conflict with our second principle (see section 3.1). Consequently, the focus of the personas was redirected towards the construction of scenarios. These, instead of being based on a particular event, were based on general tasks that all journalists deal with at some point of their career: analysing data, interviewing and writing a news story.

Basing MARVIN on task-centred scenarios instead of specific events freed the simulation from future constrictions based on newsworthiness: for example, while the U.S. Presidential election was a current topic while creating MARVIN in 2016, it would not be so five years later. Therefore, any simulation based on the said subject would be outdated. Task-centred scenarios, however, allowed for updates in the storylines that provide the basis for the simulation, without altering MARVIN’s functionality.

After the scenarios had been constructed, the development (and subsequent testing) stages began. In this period, it was found that existing technologies provided a limitation to the types of interaction that would be possible. It was then decided that a “Wizard of Oz” approach (Kelley, 1983) was the best strategy for the interaction design: by ‘faking’ the technologies, the journalists would have absolute freedom to try any type of interaction they deemed comfortable and adequate without any limitations.

A priority scheme was determined for the development and testing of the scenarios, based on the variety of interactions that were plausible within them. Data Analysis and Writing provided the opportunity to journalists to interact with a visual as well as audio based tools. Because of the complexity of the potential interactions, Data Analysis was given priority over Writing. Finally, Interviewing, was composed by a single audio based tool only. Therefore, it was decided that it would receive lowest priority in the construction of MARVIN.

The quality of MARVIN was to be tested, in order to verify its general usability, as well as its ability to provide insights into potential interactions for the researches at Fontys FutureMediaLab. A first prototype testing phase was conducted with media and technology students in order to verify the usability of the product. The results of these tests can be found in the evaluation section. Once all necessary improvements had been incorporated into the simulation, a second round of testing was conducted in order to analyse MARVIN’s research potential.

For maximum quality results, the testing of the simulation had to be done with journalists that had experience in either data analysis, writing, or interviewing. Also, they preferably wrote on more serious, ‘hard’, news stories. In the weeks before the actual testing was conducted, the team members used their personal networks to find participants that were willing to cooperate. Also, several news agencies were contacted by email. This resulted in a total amount of seven journalists, from the following news agencies: BNR Nieuwsradio, De Groene Amsterdammer, Het Parool, AT5, Veronica, and DutchNews.nl. Also, two students with journalistic backgrounds participated, as they had experience with the tasks in the simulation. The results from the tests and MARVIN’s final design can be found in the following sections.

4. MARVIN

MARVIN is a futuristic robot newsroom simulation, and consists of three scenarios: Data Analysis, Writing and Interviewing. Each scenario had a specific objective—pitching a story angle, writing a brief story, and obtaining a quote from an interviewee, respectively—based on a fictional news story. In order to reach each simulation’s goal, the simulation participants could utilise the tools as needed. The ‘Robot Journalism’ tools that were made available to them were a Summarizer Dashboard, a Chatbot, an automatic Fact-Checker and an Auto-Tagger.

In order to reach the final design that is hereafter presented, a series of pilot tests were conducted with media and technology students, who provided feedback on their experiences with the system. From this, a we constructed a list of potential improvements, which will be explained further in the ‘Evaluation’ section.

Every scenario consisted of three main phases: an introduction to the tools and the scenario, the simulation itself and a post interview. First, the test participant was welcomed and given an explanation of the simulation and the fictional news story they would have to work on. Then, the journalist was taken to the ‘newsroom’ where MARVIN had previously been set up. Two of
the researchers—hidden inside the ‘newsroom’—then carried out the ‘Wizard of Oz’ interactions, while the remaining three observed and listened to them through a live video feed. Finally, after the scenario was completed, a post-test interview was conducted in which the journalists are asked for feedback about their experience in the simulation.

In the following sections, all three scenarios and their respective phases will be described in detail.

### 4.1 Step 1: Introduction

Upon arrival, the participants were greeted by three of the five researchers in a room separate from the simulation. Here they would get an introduction to the project and a short explanation on their role in it. Then, upon entrance in the simulation, the journalist received instructions from one of the researchers. First, the goal was explained and an additional short introduction to the available tools was given.

### 4.2 Step 2: Simulation

#### 4.2.1 Scenario 1: Data Analysis

The Data Analysis scenario revolves around the work of a desk research journalist. The simulation for this scenario took place in a room that represented a newsroom, which functioned as a quiet workplace for the journalist to work on their story. The journalist was given the objective to write a pitch for a story, concerning a (fictional) statement given by the mayor of Amsterdam at a press conference: “Tourists will no longer be allowed in Amsterdam”.

Once all the researchers who did not fulfil the role of MARVIN left the room, the actual simulation would start. If the journalist started a dialogue right away, MARVIN would respond to them. If not, MARVIN would take the initiative and greet them. Once the dialogue started, MARVIN would offer to retrieve data from the internet about the subject that they would have to write about. Figure 1 shows how MARVIN would build up the data on the Dashboard. Behind the scenes, this was done by showing a PowerPoint on the screen.

When the data was complete, the journalist could start working on his/her pitch. This was the phase where the journalists could explore the environment and work with the tools in whatever way they preferred.

#### 4.2.1.1 Environment

Within the Data Analysis scenario, two main tools were available for the journalists, a Summarizer Dashboard and a Chatbot. Figure 2 illustrates the flow of information between MARVIN (portrayed by the researcher) and the journalist.

### 4.2.2 Summarizer Dashboard

As previously mentioned, the setting for this scenario was a desk setting in a newsroom. Therefore, the Summarizer Dashboard was designed on a large screen and contained a large amount of data about the subject of the scenario. The journalists could search through this by using many rich interactions, such as gestures, touch and voice. Because the researchers are in control of the Dashboard, they could respond to any type of interaction, allowing the journalist to experiment freely.
In terms of interface, many options were considered over the course of development. The hexagonal design that can be seen in Figure 3 was chosen due to its simplicity and modularity. However, it should be noted that the exact visual interface was not the focus of this research, and was merely a means to an end. The Dashboard was created and controlled with Sketch.

**Chatbot**

In addition to the Summarizer Dashboard, MARVIN also offered Chatbot interaction to the journalist. Through a microphone and speakers, the researcher could converse with the journalist in a robotized voice during the simulation. This robotic effect was created with a very short and quickly repeating delay in the Digital Audio Workstation software 'Ableton Live'.

The journalist could ask MARVIN for more information, for which MARVIN could then provide answers through the Chatbot interface, or by showing the already existing information on the Summarizer Dashboard. Most of the responses that MARVIN gave through the Chatbot interface, were improvised by the researcher.

### 4.2.2 Scenario 2: Writing

The Writing scenario revolves around the work of journalists who write stories about various topics. The simulation for this scenario took place in a room that represented an office, where the journalist could work on their laptop to write a story about the fictional news event. The goal of this simulation was for the journalist to write a story, based on the statement of the mayor and two related interview transcripts.

Before starting the simulation, the journalist had the time to read through the statement of the mayor and the interviews, to get acquainted with the 'news events'. Once all the researchers who did not fulfill the role of MARVIN left the room, the actual simulation would start. If the journalist started a dialog right away, MARVIN would respond to them. If not, MARVIN would wait for the journalist to start typing on the story. Functioning as a Fact-Checker and an Auto-Tagger, MARVIN would look at the content while the journalist was typing, and would suggest tags based on what they wrote. Additionally, whenever the journalists asked for feedback on either a fact or something else within the story, MARVIN could respond to this via voice and help them with their question.

#### 4.2.2.1 Environment

The 'Writing' environment had three main tools available for the journalist, through two different interfaces. The first two were a Fact-Checker and an Auto-Tagger, which were available within the writing interface of the journalist. Figure 4 shows the flow of information between MARVIN and the journalist through these tools.

**Fact-Checker and Auto-Tagger**

The largest part of actually writing a story happens within a computer environment. Therefore, we placed the Fact-Checker and the Auto-Tagger tools within the writing environment of the journalists. Figure 4 shows that the researcher had a computer that was connected to the computer of the journalist, through which the researcher could monitor the journalist's progress, and fulfill the role of the Fact-Checker and the Auto-Tagger. Figure 5 shows the writing interface on the researcher's side (visually, it looks almost identical to the journalist's one).
connected through a Node.js websocket server. The local network ensured speed and reliability, which were crucial for letting the connection with the Fact-Checker and the Chatbot look seamless.

**Chatbot**

As in the Data Analysis scenario, MARVIN also offered Chatbot interaction to the journalist. The journalist could again ask MARVIN for more information, for which MARVIN could then provide answers through the Chatbot interface, or by typing a new comment through the Fact-Checker interface. In this scenario, all responses that MARVIN gave through the Chatbot interface were improvised by the researcher.

**4.2.3 Scenario 3: Interviewing**

Finally, the Interviewing scenario revolves around the work of journalists who regularly interview people about all kinds of subjects. The goal of this simulation was for the journalist to conduct an interview and make notes of it for suggested later use. The participant of the simulation was presented with a scenario in which he or she had to gather opinions and/or facts about the news story at hand, by conducting an interview.

First, the journalist had time to read through an interviewee profile and a series of suggested questions for the interview. Once all the researchers who did not fulfill the role of MARVIN or the role of the interviewee had left the room, the actual simulation would start. The journalist would interview the researcher who played the interviewee, according to their supposed expertise on the subject of the fictional news story. The interviewee would occasionally answer questions of the journalist with false information, which MARVIN could then point out through the headphones that the journalist was wearing. The journalist could then either address this lie, or merely make note of it.

**4.2.3.1 Environment**

![Figure 6: Interviewing Scenario Environment](image)

The interaction in this task mainly facilitates the flow of information from the ‘Chatbot’ to the participant, as one tool only was used: a voice controlled Fact-Checker.

**Fact-Checker**

As can be seen in Figure 6, the Fact-Checker communicated with the journalist from a microphone through headphones. When the interviewee answered one of the questions of the participant with a lie, MARVIN informed the participant (through a headphone, so it would not intervene with the interview) about this directly. As with the other scenarios, the voice was again slightly robotized.

**4.3 Step 3: Post Interview**

The ending of the simulation was indicated by either the journalists themselves (if they told MARVIN that they finished their pitch or their story), or MARVIN (by telling the journalists that the things they had so far were fine and could be sent to the editor). Then, the other researchers would come back to collect the journalist from the ‘newsroom, and lead them to a separate room, where they could interview the journalists about their experience in the simulation.

**5. EVALUATION**

In order to determine the effectiveness of the simulation for allowing researchers to observe and learn from the interactions journalists have with MARVIN, two test rounds—pilot tests and the actual tests with journalists—were conducted. For the official tests with journalists, we followed the principle proposed by Nielsen (2000), and aimed to select five users for each of the tasks. This number of users is sufficient for testing the simulation as it enables us to detect a lot of potential usability problems.

**5.1 Iteration 1: Pilot Tests**

During the development phase of the simulation, pilot tests were done with five information and media students and one teacher in the information studies domain. These tests allowed us to determine whether the simulation was sufficiently effective in its purpose, and provided insights into aspects of it that needed improvement. Additionally, because of the ‘Wizard of Oz’ strategy (Kelley, 1983) for the interaction methods, there would be a lot of involvement from the researchers in the actual implementation of the scenario. These tests, therefore, also worked as practice grounds for developing the necessary acting skills for the roles that were to be filled by the researchers.

Five pilot tests were conducted, following the priority scheme detailed in the Methods section. Four test participants tested the Data Analysis scenario, while two of them tested the Writing scenario. Because of time constraints, it was not possible to develop all three scenarios for the pilot testing day and, therefore, Interviewing could not be tested.

Based on these tests, a range of caveats surfaced that needed consideration. First, in Data Analysis, a proper introduction to MARVIN and the tools that the test participants could use was intentionally left out. The idea behind this strategy was to try and minimize external influences over potential interaction strategies that the participants might try. However, it also caused confusion and unease for the test participants as they were unsure how to start working with the system. All the participants emphasized their discomfort when questioned about their experience, so it was decided that a different ‘starting’ strategy would have to be implemented in the second round of testing.

Second, in the Data Analysis scenario, multiple test participants pointed to the fact that they wanted to save information, look into additional sources or access the original articles. This was not possible during pilot testing because the Summarizer Dashboard only consisted of one screen that would not allow for multiple windows. Consequently, it was decided to include an extra screen for the journalists’ tests that could be used by the participants as an addition to the Summarizer Dashboard.

Also, it was apparent in the Writing pilot tests that the users did not know where to direct their attention to when the Chatbot talked to them. This was solved by giving the Chatbot a physical presence: a strip of light was placed on a table that lit up when the Chatbot talked.

While these were the most prominent caveats in the simulation that could be addressed before the testing round with journalists, some other issues arose that should be noted here. Some of the pilot test participants noted that they would like to have the
possibility to compare sources within the Summarizer Dashboard, they wanted to see the timeframe of events, and they wanted to make graphs themselves from the information that was presented to them. Due to time limitations, these issues were not solvable within the scope of this research. Finally, the pilot tests provided the grounds for creating the observation sheets (see Appendix 1) that were later used in the journalists’ test to take structured notes, analyse the journalists’ behaviour and draw conclusions regarding MARVIN’s effectiveness as a research method. In order to produce said notes, it was necessary to identify and structure the different stages the test participants would go through while in the simulation. In this regard, five stages were identified within the Data Analysis scenario, and three in the writing Scenario. For the Data Analysis scenario, the stages were (0) understanding the task, (1) understanding the tool, (2) exploring the content, (3) creating material, and (4) wrap up, and for the Writing scenario, the stages were (1) understanding the tool, (2) creating material, and (3) wrap up. The stages differ from one scenario to another mostly because of the complexity of the task and the participants’ unfamiliarity with the Data Analysis tools. Lastly, on the observation sheet, for each of the stages, questions were proposed that served as a guidance for the researchers during and after the observations.

5.2 Iteration 2: Journalist Tests

A total amount of nine participants tested the improved version of MARVIN. Most of them tested just one of the tasks, because of time constraints regarding their work schedule. However, two of the participants had the time and willingness to test multiple tasks. All in all, this resulted in five tests for the Data Analysis scenario, five tests for the Writing scenario, and one pilot test for the Interview scenario. Each journalist was assigned to a task based on the type of work they were most familiar with, which was determined by consulting the journalists prior to the test days. So, for example, journalists that were responsible for writing (news) stories in their daily work would be assigned to the writing scenario. In Table 1, an overview is presented on the test participants.

In order to assess the effectiveness of the simulation, benchmarks were created for each of the scenarios. The benchmarks were based on the goals that the research set out with, i.e. to find out whether a simulation can provide researchers with insights into future methods of interactions with ‘Robot Journalism’. In the following section, ‘Results’, these benchmarks are discussed further.

5.3 Results

In order to evaluate MARVIN’s effectiveness, structured notes were taken for each test from ‘live’ feed that was transmitted from the ‘newsroom’ to the observing researchers. Through this strategy, the researchers that were not in charge of MARVIN could analyze the behavior of the journalists. Finally, a post simulation interview was conducted with each test participant allowing the journalists to verbalize their impressions and provide explicit feedback on the experience.

These observations and interviews were used to evaluate each scenario, through the use of an evaluation rubric composed of six benchmarks (see tables 2 and 3). These benchmarks were, in this project, the requirements to determine the success of MARVIN as a method for doing research on robot journalism. Each benchmark was graded as insufficient if the requirements it states were not met, sufficient if there was a minimal fulfilment, and good if it was abundantly present. The benchmarks marked with an asterisk (*) are the mandatory evaluation metrics. A test was considered successful if at least four of the benchmarks—which must include the mandatory ones—were characterized as, at least, sufficient.

5.3.1 Scenario 1: Data Analysis

A total of five test participants engaged in the Data Analysis scenario with the Summarizer Dashboard and the Chatbot. Table 2 shows the specific benchmarks of the Data Analysis scenario. Out of all five only three were deemed successful. In this section, those results will be explained in detail.

<table>
<thead>
<tr>
<th>Type of Journalist</th>
<th>Institution</th>
<th>Tested Scenario</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Participant 1</td>
<td>Reporting Journalist</td>
<td>BNR Nieuwsradio</td>
</tr>
<tr>
<td>Test Participant 2</td>
<td>Research Journalist</td>
<td>De Groene Amsterdammer</td>
</tr>
<tr>
<td>Test Participant 3</td>
<td>Writing Journalist</td>
<td>Het Parool</td>
</tr>
<tr>
<td>Test Participant 4</td>
<td>Communication Student</td>
<td>University of Amsterdam</td>
</tr>
<tr>
<td>Test Participant 5</td>
<td>Research Journalist</td>
<td>AT5</td>
</tr>
<tr>
<td>Test Participant 6</td>
<td>Journalism Student</td>
<td>Utrecht University</td>
</tr>
<tr>
<td>Test Participant 7</td>
<td>Allround Journalist</td>
<td>Dutchnews.nl</td>
</tr>
<tr>
<td>Test Participant 8</td>
<td>Writing Journalist</td>
<td>AT5</td>
</tr>
<tr>
<td>Test Participant 9</td>
<td>Writing Journalist</td>
<td>Veronica Superguide</td>
</tr>
</tbody>
</table>

Table 1: Test Participant Details

---

1 ‘Understanding the task’ was numbered as stage 0 (zero) because in essence it was external to the actual simulation: it was presented outside the newsroom and did not serve as actual input for the construction of the news pitch, but merely functioned as background information to the Data Analysis scenario. If the journalist did not comprehend the task, they would have been unable to achieve the scenario’s goal. Writing, differently, did not rely as strongly on an introductory instance, and could more easily be comprehended as a task within the actual simulation.
First, during the tests we saw only two of the test participants attempt multiple interactions with the dashboard. Most journalists stuck to the first method of interaction they found: voice. This was very different from the pilot tests, where every participant attempted multiple options. However, in hindsight this was not surprising, as one important change was made after feedback from the pilot tests showed that the participants felt really awkward at the start of the simulation. The change in question was that MARVIN would introduce himself without prompting from the participant, which we therefore believe lead to the journalists responding and sticking only to voice as the primary method of interaction. The fact that the voice interface was the sole conscious way of interacting with the system for some of the journalists however did not mean that this would be the preferred way of interaction in a future newsroom, as is evidenced from the feedback from the interviews. For example, test participant #5 stated in the post-interview: “I kept saying to zoom in. I think it would’ve been useful [to make gestures for this]”. This indicates that the test participants were, during the simulation, not aware of all the possibilities, but they could communicate their preferences to us afterwards. Also, the observations of the data analysis scenario shows that the journalists also had unconscious interactions with MARVIN, such as pointing to the screen and shifting their bodies in the direction they wanted the dashboard to move.

Second, we saw that all journalists attempted to converse with the Chatbot. This is evidenced by the fact that they continuously asked MARVIN for help and even seemed to consider MARVIN as their colleague. To illustrate this point we present some quotes from the test participants: “Okay, so let’s get to work on this story” (test participant 2); “Why did you show me this?” (test participant 2); “Hey, and another thing…” (test participant 4); “I am going to write this down by myself, you are fine with that, right Marvin?” (test participant 1) and “Bring it on!” (test participant 2).

Third, during the post-interview (which can be found in the Appendix) all of the journalists discussed various potential interactions at length. For example, test participant 1 expressed that, if he had been aware of the possibility, he would have preferred to gesture at or touch the Summarizer Dashboard as opposed to instructing MARVIN by voice. Furthermore, test participant 2 expressed that he didn't really like the voice commands and would prefer a desktop environment with mouse and keyboard for example. Other remarks can be found in Appendix 1.

Fourth, some of the test weren't observable live, due to technical difficulties. As it's crucial for a research method to have the possibility of observation, these tests, accordingly, could not be deemed successful. However, there were recordings available, which allowed us to make observations nonetheless. Naturally, for the fifth benchmark this was the same. As the tests couldn't be observed live, the researchers were naturally also not able to take notes. Again, this was possible after the test through the recordings.

Finally, to determine the successfulness of the scenario, it was important for us to be able to understand the thinking of the journalist. Because of this, it was necessary for the observing researchers to discover patterns in the journalist's interactions with MARVIN. For all of the tests it was visible in what phase they were and in what manner they were discovering the tools.

Other interesting observations we noted were that different type of journalists interacted in different ways during the data analysis scenario. For example, test participant 2, a research journalist from De Groene Amsterdammer, continuously asked the Chatbot to pull up the original articles on the second screen, so that it could be read. In comparison, test participant 1, a reporter from BNR Nieuwsradio, did not use the second screen at all and just read the summaries from the Summarizer Dashboard and engaged in a conversation with the Chatbot about the content. These different interactions make sense, because of the difference in the type of work that the journalists usually engage in (respectively investigative journalism and short term journalism). While these results don't directly relate to the validity of MARVIN as a research method, they confirm the second principle upon which MARVIN was constructed - that the simulation would have to be relatable to test participants with different journalistic backgrounds. Also, these results provide interesting insights for future implementations of MARVIN.

5.3.2 Scenario 2: Writing

Five journalists tested the Writing scenario. Here they engaged with the Fact-Checker, the Auto-Tagger and the Chatbot. Table 3 shows the specific benchmarks of the Writing scenario. As can be seen, all of the tests were deemed successful.

<table>
<thead>
<tr>
<th>Requirements Data Analysis</th>
<th>Test 1</th>
<th>Test 2</th>
<th>Test 3</th>
<th>Test 4</th>
<th>Test 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>The journalist attempts multiple interactions with the Dashboard.</td>
<td>sufficient</td>
<td>insufficient</td>
<td>insufficient</td>
<td>good</td>
<td>insufficient</td>
</tr>
<tr>
<td>The journalist attempts to converse with the Chatbot</td>
<td>sufficient</td>
<td>good</td>
<td>good</td>
<td>good</td>
<td>good</td>
</tr>
<tr>
<td>The journalist formulates and expresses opinions about the interactions in the post interview*</td>
<td>good</td>
<td>good</td>
<td>good</td>
<td>good</td>
<td>good</td>
</tr>
<tr>
<td>The researchers can easily observe the journalist during the test based on movement/interaction and thoughts*</td>
<td>insufficient (mic off)</td>
<td>good</td>
<td>good</td>
<td>insufficient (no live images)</td>
<td>good</td>
</tr>
<tr>
<td>The researchers can take structured notes on the provided form</td>
<td>insufficient (mic off)</td>
<td>good</td>
<td>good</td>
<td>insufficient (no live images)</td>
<td>good</td>
</tr>
<tr>
<td>The established phases are apparent during the test</td>
<td>good</td>
<td>sufficient</td>
<td>good</td>
<td>good</td>
<td>sufficient</td>
</tr>
</tbody>
</table>

Table 2: Data Analysis Scenario Evaluation Matrix
First, while writing the story, all of the journalists were confronted with the Fact-Checker when they introduced false data from one of the interviews into the story. In all of the tests it was clear that the journalist actually read the suggestions and took them for facts. They all utilised the Fact-Checker's suggestions and tips.

Second, two of the journalists asked the Chatbot for help during the test. The fact that three of the journalists did not do this was later explained in the interviews as based on the fact that they did not have any doubts. For most of the journalists, the content they had to work with was deemed sufficient and any necessary questions were already answered by the Fact-Checker.

Third, as with the data analysis scenario, it was very important that the journalists would discuss multiple interactions and preferences about those during the post interview. All five of the journalists discussed potential options extensively. For example, test participant 3 said that she would not be interested in talking to the Chatbot extensively, because the Auto-Tagger and the Fact-Checker on screen were sufficient. On the other hand, test participant 8 noted how the voice interaction made him feel as if he was collaborating on the work with a colleague.

Fourth, there were no difficulties in observing any of the journalists while they were writing the story. Observations could be easily made due to the monitoring that was going on while they were writing. Additionally, afterwards all of the recordings were available.

This ease of observation also allowed the researchers to take structured notes for further analysis. Therefore, the tests also sufficed for the fifth benchmark.

Finally, as with the Data Analysis scenario, it was important for us to be able to understand the thinking of the journalist. Through the observations, the researchers were able to discover patterns in the journalists interactions with MARVIN that were easily classified into the phases presented in the observation sheets.

5.3.3 Scenario 3: Interviewing

Finally, following the priority scheme that was described in the section 'Methods', only one test was done of the interview scenario. However, because we hadn't tested this before during the pilot test phase, this was essentially a pilot test, even though the test participant was in fact a journalist. The test was only done because one of the journalists actually did have the time and a fitting expertise.

During the test we saw that the journalist was able to converse with the interviewee while also listening to a voice version of the Fact-Checker through headphones. We were able to extensively observe the journalist while they were doing this, as well as discuss its successfulness and various other possibilities for interaction during the post-interview. For example, one journalist (participant 7) suggested that if the conversation is in real life, the Chatbot on an earpiece would be the preferred way to go. However, for a Skype interview for example, it would be interesting to have the Fact-Checker on screen. The journalist was even interested in getting more information from the Chatbot than only facts. For example: interesting follow-up questions.

Since there were no benchmarks for this test, we can't say if it was successful or not. However, since the test could be observed, interactions were discussed afterwards, and there was a clear structure in the test (with phases), the interviewing scenario is expected to be another interesting simulation.

6. DISCUSSION

The results, as they are described in the previous section, remain up for discussion.

First of all, because of time constraints we only conducted one full round of testing. The results then, do not present an infallible research method, but do prove that it is an effective one. However, future testing would be necessary to improve issues that arose from both the pilot tests and the journalists’ tests, and that could not be covered within the scope of this research.

Secondly, although the interfaces were not the focus of our investigation, it remains a possibility that the hexagonal design of the Summarizer Dashboard and the Chatbot's voice and technologies (speakers and headphones) could have significant influence on journalists’ interactions. Therefore, the interactions in our simulation do not necessarily point to what should be developed in terms of (types of) interfaces.

Thirdly, because of the way MARVIN was constructed, the journalists only engaged in interactions they already knew were possible. This excluded the possibility that there might be better ways to do things that did not come up in the current implementation of MARVIN. However, this was done intentionally, as the approach that was taken actually let the journalist decide upon their preferred type of interaction. Therefore, the possible unimaginative choices of the journalist remain a natural tradeoff.

<table>
<thead>
<tr>
<th>Requirements Writing</th>
<th>Test 1</th>
<th>Test 2</th>
<th>Test 3</th>
<th>Test 4</th>
<th>Test 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>The journalist accepts the Fact-Checker's help.</td>
<td>sufficient</td>
<td>good</td>
<td>sufficient</td>
<td>good</td>
<td>sufficient</td>
</tr>
<tr>
<td>The journalist attempts to ask the Chatbot for help</td>
<td>insufficient</td>
<td>good</td>
<td>insufficient</td>
<td>sufficient</td>
<td>insufficient</td>
</tr>
<tr>
<td>The researcher formulates and expresses opinions about the interactions in the post interview*</td>
<td>good</td>
<td>good</td>
<td>good</td>
<td>good</td>
<td>good</td>
</tr>
<tr>
<td>The researchers can observe the journalist during the test based on movement/interaction and thoughts*</td>
<td>good</td>
<td>good</td>
<td>good</td>
<td>good</td>
<td>good</td>
</tr>
<tr>
<td>The researchers can take structured notes on the provided form</td>
<td>good</td>
<td>good</td>
<td>sufficient</td>
<td>good</td>
<td>good</td>
</tr>
<tr>
<td>The established phases are apparent during the test</td>
<td>good</td>
<td>sufficient</td>
<td>good</td>
<td>good</td>
<td>good</td>
</tr>
</tbody>
</table>

Table 3: Writing Scenario Evaluation Matrix
Then, a drawback of a simulation that is set up internally at the University of Amsterdam, is that it is an artificial setting— the journalists have to imagine that they are in a ‘futuristic newsroom’. Reactivity, i.e. the possibility that the test participants alter their behavior because of the presence of researchers, is also a potential issue. Especially because they were aware that everything was being recorded. These factors might have had an effect on the behavior of the journalists. Naturally, we aimed to minify this effect by ensuring that no researchers were visibly present during the test rounds in which the journalists interacted with the tools, but the issue still persists nonetheless. However, since we are not interested in the actual behavior of the journalists, but want to find out whether the simulation is usable as a research tool to learn about the future of journalistic work, reactivity of the test participants during their interactions with MARVIN is not a big issue.

Finally, as already mentioned in the results section, the results of the Interview scenario can not be deemed trustworthy. It should therefore only be seen as an interesting new topic of research. It will therefore be further discussed in the section Future Work.

7. CONCLUSION
When this research started, we set out with the goal to find out how insights into future methods of interaction with 'Robot Journalism' can be obtained by researchers. Our simulation, MARVIN is an attempted method at achieving this, and was tested in two phases within the research. We found that MARVIN allows journalists to experiment with various interactions with tools that could one day play an important role in their day-to-day work. Although journalists were aware of the ‘theatre’ in the simulation, MARVIN allowed them to think about these tools, while sparking conversation and debate about various possibilities and preferences. MARVIN therefore succeeded as an approach for researching journalists’ preferences regarding forms of interaction with Robot Journalism in the future. However, many possibilities have not been explored and become therefore interesting fields for future work and inquiry.

8. FUTURE WORK
Looking at the results of the simulation tests, various topics have still not been discussed or researched enough and therefore remain interesting future topics and focuses for research.

First of all, feedback that we received from the test participants suggested that there were a range of functionalities missing in MARVIN that would have been useful to the journalists. These functionalities were mainly related to the visualization of information and the timeframe of the articles and data sources. Some of the journalists really specifically asked for features, such as combining data/sources or plotting graphs. When we look at these occurrences, we can see that journalists also have a strong and useful opinion about features. So this could be an interesting direction for future inquiry, as developers do not only need to know about interaction, they also need to know about features. Future researchers who will continue to build on the idea of a newsroom simulation are thus encouraged to consider its value for assessing not only the interactions that journalists have with it, but also the potential it has for assessing what features are important for journalists.

Second, only one test was conducted for the Interview scenario. We thus cannot conclude anything definitive about the usefulness of this scenario in the simulation, but our initial results were promising. As interviewing is one of the main tasks for journalists to obtain information and opinions, it is fruitful to find out how journalists would like to do this when they have access to futuristic tools. Therefore, other researchers are encouraged to do multiple tests for the interview scenario, so that its usability as a research tool for robot journalism can be assessed.

Finally, a lot of new interesting points for the post interview were discovered. For example, no questions were asked so far about the unconscious interactions of the journalists, which was actually discovered as an interesting topic. A future study on Robot Journalism should therefore ensure that the post interviews are updated in a structured manner based on the findings of previous researches and on their own. This would contribute to the value of the insights obtained from these post interviews.

9. REFERENCES


10. APPENDIX

10.1 Appendix 1: Observations during test and post interviews

Test Participant number 1  
BNR Nieuwsradio - Data Analysis Scenario

For this test, the observation was done from the video footage after the actual test.

<table>
<thead>
<tr>
<th>Simulation</th>
<th>Observations</th>
<th>Post sim feedback</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase 0: understanding the task</td>
<td>Do they understand the fictional news story? Check. Reads the statement again.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Do they understand what the tools can do? yes.</td>
<td></td>
</tr>
<tr>
<td>Phase 1: understanding the tool</td>
<td>When the journalist tries to get the dashboard working... What do they do? The bot starts by itself, and he engages in a conversation. Checks if it’s a person or a computer. Answer: it’s a computer. “Let’s get started... weird remark by the mayor (NOT exact quote)”. Didn’t ask for the info explicitly, MARVIN fed it into the dashboard. Tells the bot about the story.</td>
<td></td>
</tr>
<tr>
<td>Are they comfortable?</td>
<td>Yes.</td>
<td>Yes.</td>
</tr>
</tbody>
</table>
| What method of interaction do they try? And prefer? (touchscreen, gestures, talking, ...) | Talking. | natural talking
<p>|                             | Asked for info, and chat bot went to the info without being prompted to move by gestures or commands. | Didn’t try to do gestures or touch because the talking was enough. HE would have used it if he knew he had the possibility. He would’ve preferred the touch screen. |
| (PHASE 2/3) He instinctively points at the dashboard, while asking out loud. (7:00) Looks like a conversation with an actual person, gestures accompany talking, but there are no commands. |                   |                   |
| Phase 2: Exploring the content | What information do they ask for and in what format? Missed info about how the mayor wanted to do this. SEE PHASE 3 |                   |
|                             | What is their thinking strategy: talking out loud, writing in a notepad, using second |                   |
|                             | Reads the statement every once in awhile.                                    |                   |
|                             | Bot suggested article on second screen                                       |                   |</p>
<table>
<thead>
<tr>
<th>Phase 3: creating material</th>
<th>How do they use the content?: change how it’s organized, delve deeper, ask for original sources...</th>
<th>Query based. Asked about info he wanted. Took about 10.30 minutes to realize that he could ask for additional information, a minute later he started asking for more info.</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Thinking strategy: does it differ from the previous stage?</td>
<td>asks for permission to write in a paper. Thinking out loud and trying to understand the data. Bouncing ideas with the bot, by talking. 13:30 pitches a direction to marvin. uses his own notes to pitch. Once he establishes it, he starts going deeper into it. *sort of tell the bot what to aim for Shows the TV the paper with his notes.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phase 4: wrap up</td>
<td>How do they signal that they’re done?</td>
<td>I’m going to tell you what I’m going to write about. ‘Dank je marvin” Doesn’t dictate anything.</td>
<td></td>
</tr>
</tbody>
</table>

**Post interview**

<table>
<thead>
<tr>
<th>questions asked</th>
<th>Would you use this system in your own work environment</th>
<th>Knowing that they were being observed affected his behaviour. Would act a bit different in his own environment. “I felt like I was in a test”. of course I would use it” in my day to day job.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Did the directions/angles that we gave you helpful?</td>
<td>It helps, but it makes it less free.</td>
<td></td>
</tr>
<tr>
<td>Extra notions</td>
<td>He felt a little bit rushed to choose an angle for the story. similar to his day to day job... ‘ If there is an</td>
<td></td>
</tr>
</tbody>
</table>
**Test Participant number 2**  
De Groene Amsterdamer - Data Analysis Scenario

<table>
<thead>
<tr>
<th>Simulation</th>
<th>Observations</th>
<th>Post sim feedback</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Phase 0: understanding the task</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Do they understand the fictional news story?</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Do they understand what the tools can do?</td>
<td>Yes.</td>
<td>Took a bit of time.</td>
</tr>
<tr>
<td><strong>Phase 1: understanding the tool</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>When the journalist tries to get the dashboard working...</td>
<td>“Hello?” “let's get to work on this story”</td>
<td></td>
</tr>
<tr>
<td>What do they do?</td>
<td>Tells the bot what the story is about. Asks a specific question.</td>
<td></td>
</tr>
<tr>
<td>Are they comfortable?</td>
<td>Bit strange at first, but once he figured out how it worked he could do it quickly.</td>
<td></td>
</tr>
<tr>
<td>What method of interaction do they try? And prefer? (touchscreen, gestures, talking, ...)</td>
<td>Talks to chat bot. chat bot gives info directly, without being prompted to move the camera. Instead of zooming, walks towards the tv. 7:08 - [screen flickered] ‘Are you ok, Marvin? Your screen was shaking’ Once he heard an answer, he kept going with the pitch.</td>
<td>Didn't think about touching or gesturing because it responded well to commands.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Not a fan of voice commands. Can’t imagine it working, though. So, comment should be taken lightly, problem is mostly with him not imagining it. Problems with editorial floor and voice commands. Click system or touch screen would be better. Can’t imagine that voice screen would ever work.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Clear and useful interface.</td>
</tr>
<tr>
<td><strong>Phase 2: Exploring the content</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>What information do they ask for and in what format?</td>
<td>Specific info. Follows up on info in dashboard.</td>
<td></td>
</tr>
<tr>
<td>What is their thinking strategy: talking out loud, writing in a notepad, using second screen, etc.</td>
<td>Uses second screen, and reads articles in-depth. Gives specific orders.</td>
<td></td>
</tr>
<tr>
<td>Phase</td>
<td>How do they use the content?: change how it’s organized, delve deeper, ask for original sources...</td>
<td>asks extra info. asks about sources on additional screen. INTERACTION: Second screen. Commands: scroll up, down. Main dashboard: asks questions, doesn’t really give commands for movements, says ‘could we...’. Nyn pulled up info on the second screen without being prompted and threw off the journalist: he asked ‘why are you showing me this’ with sort of a tone in his voice. Wants related articles and timeframe of articles</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Thinking strategy: does it differ from the previous stage?</td>
<td>talks out loud but summarizing data, and asking questions. Doesn’t show his train of thought.</td>
<td></td>
</tr>
<tr>
<td>Phase 4: wrap up</td>
<td>How do they signal that they’re done? ‘I think I’m ready now’</td>
<td></td>
</tr>
</tbody>
</table>

**Post interview**

**questions asked**

<table>
<thead>
<tr>
<th>Can you imagine this in your day to day work</th>
<th>Lacking some functions: first phase of work, collecting info, grouping it and organizing it. But then, it would be useful to order the articles by date (historical overview). saves a lot of time. I Want to make a timeline, and compare different data.</th>
</tr>
</thead>
<tbody>
<tr>
<td>If you could use the system in different way, how would you use the system, ideally.</td>
<td>i'm not a huge fan of voice, complicated to operate it. but did not see a additional value of the voice control, would be chaotic in a newsroom.</td>
</tr>
</tbody>
</table>
**Test Participant number 3**

Het Parool - Writing Scenario

<table>
<thead>
<tr>
<th>Simulation</th>
<th>Observations</th>
<th>Post sim feedback</th>
<th>timeline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase 1: understanding the tool</td>
<td>How do they interact with it initially?</td>
<td>Just writing</td>
<td>Starts with writing right away</td>
</tr>
<tr>
<td>Are they comfortable?</td>
<td>Jumped a bit when chatbot spoke</td>
<td>But she knew that there were people on the other side of the screen.</td>
<td></td>
</tr>
<tr>
<td>Do they look for/react to a physical presence?</td>
<td>the light is really cool, it feels like a real robot, pretty cool.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Phase 2: creating material | Do they believe the chatbot? | | Not using the chatbot |
| Do they use the suggestions? | | | |
| Do they ask the bot for extra information? | | Almost no response to the chatbot |
| Do they ask for other functions? (spell check, format, etc) | Would like to have a synonym function | And dictating at the end |

| Phase 3: wrap up | How do they show the task is complete? | Robot told her to 'check' the article. | |

**Post interview questions asked**

| Did you feel pressure because of the robot? | no pressure, calm voice |
| what did you think about the tools? | fact checker is nice, somethings are really nice to know. |
| Why did you not start a conversation | didn’t feel like it, but maybe in another scenario? |
It would be easy to dictate when you are finished. Especially for a pitch.

**Test Participant number 4**
Communicatie student - Writing Scenario

<table>
<thead>
<tr>
<th>Simulation</th>
<th>Observations</th>
<th>Post sim feedback</th>
<th>timeline</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Phase 1:</strong> understanding the tool</td>
<td>Chatbot starts. Gives normal responses</td>
<td>prefer to use voice most of the time.</td>
<td>starts with reading, then writing</td>
</tr>
<tr>
<td>How do they interact with it initially?</td>
<td>Looks like he is feeling comfortable and relaxed.</td>
<td>after 5 minutes</td>
<td>chatbot asks,’ can I help you’. responses normally but surprised</td>
</tr>
<tr>
<td>Are they comfortable?</td>
<td>noticed it</td>
<td>but didn’t make a difference</td>
<td></td>
</tr>
<tr>
<td>Do they look for /react to a physical presence?</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| **Phase 2:** creating material | Do they believe the chatbot? | seems so | He did |
| Do they use the suggestions? | yes. | he did | chatbot added a tag, he used it. confirmed. |
| Do they ask the bot for extra information? | yes, there was a conversation. Asks Marvin whether a specific number is correct, and ask for additional information. | yes, a lot | asked some question about van der Laan, and used the information. “ okay” uses the fact checker trough voice |
| Do they ask for other functions? (spell check, format, etc) | comments showing longer and not disappearing. | | |

| **Phase 3:** wrap up | How do they show the task is complete? | chatbot told him | |

**Post interview**

**questions asked**
<p>| what did you think about the suggestions from MARVIN? | they were very useful, but after a while they disappeared and didn’t came back. |
| I needed to ask for it again, and then MARVIN gave a short | |</p>
<table>
<thead>
<tr>
<th>Simulation</th>
<th>Observations</th>
<th>Post sim feedback</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Phase 0: understanding the task</strong></td>
<td>Do they understand the fictional news story?</td>
<td>yes, because of the previous test</td>
</tr>
<tr>
<td></td>
<td>Do they understand what the tools can do?</td>
<td>yes, especially the chatbot, because of the previous test</td>
</tr>
<tr>
<td><strong>Phase 1: understanding the tool</strong></td>
<td>When the journalist tries to get the dashboard working... What do they do?</td>
<td>Chatbot starts with: ‘Hoi’</td>
</tr>
<tr>
<td></td>
<td>Are they comfortable?</td>
<td>yes</td>
</tr>
<tr>
<td></td>
<td>What method of interaction do they try? And prefer? (touchscreen, gestures, talking, ...)</td>
<td>He prefers to talk. But seamlessly using gestures, probably without noticing.</td>
</tr>
<tr>
<td><strong>Phase 2: Exploring the content</strong></td>
<td>What information do they ask for and in what format?</td>
<td>‘can you show me the article on the second screen’ didn’t work so he apologized to marvin.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>‘can you put the tweet from… on the second screen.’</td>
</tr>
<tr>
<td></td>
<td></td>
<td>‘can you open a word document so i can write the speech.’</td>
</tr>
<tr>
<td></td>
<td>What is their thinking strategy: talking out loud, writing in a notepad, using second screen, etc.</td>
<td>thinks out loud and uses the second screen a lot.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>has a conversation with Marvin.</td>
</tr>
<tr>
<td><strong>Phase 3: creating material</strong></td>
<td><strong>Do they trust the information or question it?</strong></td>
<td><strong>How do they use the content?: change how it’s organized, delve deeper, ask for original sources...</strong></td>
</tr>
<tr>
<td>--------------------------------</td>
<td>-------------------------------------------------</td>
<td>---------------------------------------------------------------</td>
</tr>
<tr>
<td></td>
<td>he trust it.</td>
<td>he’s using content from an article and wants a piece copied and paste in his word doc.</td>
</tr>
<tr>
<td><strong>Phase 4: wrap up</strong></td>
<td><strong>How do they signal that they’re done?</strong></td>
<td>(speaking to the Marvin, in the word doc.) ‘we’re going to close it with: eberhard gaat voor eigen gewin’. Marvin asks is this your pitch? ‘yes, this is all the information that’s needed for the pitch’.</td>
</tr>
</tbody>
</table>

**Post interview**

<table>
<thead>
<tr>
<th><strong>questions asked</strong></th>
<th><strong>which interaction did you prefer?</strong></th>
<th><strong>initially commands, ‘but I knew that from the previous scenario. ‘ Also swipe intention.</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>was it easy to use?</strong></td>
<td><strong>it reacted a little slow but it was not annoying. Marvin is like google but easier.</strong></td>
</tr>
<tr>
<td></td>
<td><strong>What do you think of a male voice?</strong></td>
<td><strong>it’s fine, it’s like a buddy that is here to help you. I would not like a female voice.</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Did you missed the light presence?</strong></td>
<td><strong>no, I knew it was there.</strong></td>
</tr>
<tr>
<td><strong>Extra</strong></td>
<td>It would be nice to be able to write during the analysis. write things by yourself</td>
<td></td>
</tr>
<tr>
<td>Phase 0: understanding the task</td>
<td>Do they understand the fictional news story?</td>
<td>Yes</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Do they understand what the tools can do?</td>
<td>No, not immediately. so she asked.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Phase 1: understanding the tool</th>
<th>When the journalist tries to get the dashboard working...</th>
<th>Using commands</th>
<th>At first I had no idea what to do. Solved: “just start asking. That’s how it always works, right?”.</th>
</tr>
</thead>
<tbody>
<tr>
<td>What do they do?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Are they comfortable?</td>
<td>it seems so</td>
<td>bit weird. I like the way it worked.</td>
<td></td>
</tr>
<tr>
<td>What method of interaction do they try? And prefer? (touchscreen, gestures, talking, ...)</td>
<td>talking, commanding</td>
<td>Didn’t know that she could’ve touched it. “1:41. I kept saying zoom in...” I think it would’ve been useful. You’re not used to it.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Phase 2: Exploring the content</th>
<th>What information do they ask for and in what format?</th>
<th>initially only using the dashboard.</th>
<th>Second screen: when you’re researching you don’t want to read the full article. So I can see in the summary that chatbot pulls out article</th>
</tr>
</thead>
<tbody>
<tr>
<td>What is their thinking strategy: talking out loud, writing in a notepad, using second screen, etc.</td>
<td>writing in a notepad. using the second screen.</td>
<td></td>
<td>writing on notepad</td>
</tr>
<tr>
<td>Do they trust the information or question it?</td>
<td>it seems she trust it</td>
<td>Yes, but she wants to know the source.</td>
<td></td>
</tr>
</tbody>
</table>

| Phase 3: creating material | How do they use the content?: change how it’s organized, delve deeper, ask for original sources... | the chatbot pulls it out, she didn’t ask for it | Have a lot of questions that haven’t been answered. Used to look for information myself. wants to be able to pull up info and numbers from the original source by tapping them. sometimes you want to compare things |  |
Thinking strategy: does it differ from the previous stage?  
She is more to the point, probably thinks before she asks/says something.

Phase 4: wrap up  
How do they signal that they’re done?  
She asks for pitch length.  
She is telling her pitch to the chatbot.

Post interview

**Questions asked**

<table>
<thead>
<tr>
<th>Did you thought about using the screen/system in a different way, like gestures or touching the screen.</th>
<th>No, was that possible? If I knew before hand maybe I would have done that.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Was the second screen useful?</td>
<td>During writing or research you don’t want to read the whole article. Mostly summaries.</td>
</tr>
<tr>
<td>Who would you trust more, a computer or a person?</td>
<td>Depends on person and information. I don’t want to write something that’s also somewhere else. I don’t know if a computer searches everywhere.</td>
</tr>
</tbody>
</table>

Extra

**Test Participant number 6**  
Journalism student - Data Analysis Scenario

<table>
<thead>
<tr>
<th>Simulation</th>
<th>Observations</th>
<th>Post sim feedback</th>
<th>Timeline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase 0: understanding the task</td>
<td>Do they understand the fictional news story?</td>
<td>Yes</td>
<td>Chatbot started, is showing data</td>
</tr>
<tr>
<td></td>
<td>Do they understand what the tools can do?</td>
<td>Has some trouble, looks like she feels a bit uncomfortable</td>
<td>Uses voice: “zoom in tweets” “the other one” “economical perspective”</td>
</tr>
<tr>
<td>Phase 1: understanding the tool</td>
<td>When the journalist tries to get the dashboard working... What do they do?</td>
<td>Talking insecure questions about subjects. Not giving commands</td>
<td>“How does it work?” “Can I select something”</td>
</tr>
<tr>
<td></td>
<td>Are they comfortable?</td>
<td>Not much</td>
<td>It was weird in the beginning, talking to a computer feels strange. I had the feeling I was talking to a wall.</td>
</tr>
<tr>
<td>Phase 2: Exploring the content</td>
<td>What method of interaction do they try? And prefer? (touchscreen, gestures, talking, ...)</td>
<td>only talking. At some point used gestures! swiping on main screen. also tried scrolling on the second screen.</td>
<td></td>
</tr>
<tr>
<td>Phase 2: Exploring the content</td>
<td>What information do they ask for and in what format?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phase 2: Exploring the content</td>
<td>What is their thinking strategy: talking out loud, writing in a notepad, using second screen, etc.</td>
<td>talking to herself</td>
<td></td>
</tr>
<tr>
<td>Phase 2: Exploring the content</td>
<td>Do they trust the information or question it?</td>
<td>yes, chatbot: ‘this is a summary of an extensive article’ ‘oh okay, thanks’</td>
<td></td>
</tr>
<tr>
<td>Phase 3: creating material</td>
<td>How do they use the content?: change how it’s organized, delve deeper, ask for original sources...</td>
<td>want comparisons</td>
<td></td>
</tr>
<tr>
<td>Phase 3: creating material</td>
<td>Thinking strategy: does it differ from the previous stage?</td>
<td>talks to herself and writes on notebook.</td>
<td></td>
</tr>
<tr>
<td>Phase 4: wrap up</td>
<td>How do they signal that they’re done?</td>
<td>‘I think I got it!’</td>
<td></td>
</tr>
<tr>
<td>Phase 4: wrap up</td>
<td></td>
<td>she asked if chatbot is going to write or she is. she is dictating, but loses threat.</td>
<td></td>
</tr>
</tbody>
</table>

Post interview

| questions asked | why would you prefer a physical presence of the chatbot. like a light. | yeah guess so, like on a website with a chat for help, they have an avatar. It gives you the feeling you talking to someone. |
| questions asked | did you noticed that you were using gestures? | not didn’t noticed it. you’re talking on a screen and at some point when I noticed when I use my hands the computer will move with me. |
| questions asked | what would you prefer, interaction wise? | I think it would be better for this, to just use you hands and scroll by yourself. and that the computer gives more background information and commentary. |
| questions asked | Why did you gestured scrolling to the second screen and not on the big screen? | It felt like the big screen was the computer/robot and the second screen was mine. and it was a article so that’s why i can scroll. |
could you see yourself working with this system? | this system has all the information organized, you could google it, but this has it all in one overview. it was like a mind map that you can alter by yourself.

can you think of something for this system, a function or so? | maybe some more pictures or videos, so you can watch it with the text.

Extra

dictating a pitch and then it would pop-up in an extra box in the screen sort of list for notes in bullet points. so it summaries you as well during the session. at the end you can see what kind of thoughts you had

Participant number 7

Dutchnews.nl - Writing Scenario

<table>
<thead>
<tr>
<th>Simulation</th>
<th>Observations</th>
<th>Post sim feedback</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase 1: understanding the tool</td>
<td>How do they interact with it initially?</td>
<td>goes straight for writing. When Marvin talks she jumps a bit, and laughs. keeps going normally after.</td>
</tr>
<tr>
<td></td>
<td>Are they comfortable?</td>
<td>seems like it. a bit frazzled when Marvin talks for the first time.</td>
</tr>
<tr>
<td></td>
<td>Do they look for/react to a physical presence?</td>
<td>yes. looks for the origin of the voice. doesn’t look towards the lights, but upwards.</td>
</tr>
<tr>
<td>Phase 2: creating material</td>
<td>Do they believe the chatbot?</td>
<td>Yes: accepts tags and accepts/uses suggestions from Marvin</td>
</tr>
<tr>
<td></td>
<td>Do they use the suggestions?</td>
<td>Yes. uses comment (Reaction: 17:10) (remark pops up: 18:20) (Immediate correction). Immediate changes after both signals.</td>
</tr>
<tr>
<td></td>
<td>Do they ask the bot for extra information?</td>
<td>No.</td>
</tr>
<tr>
<td></td>
<td>Do they ask for other functions? (spell check, format, etc)</td>
<td>No.</td>
</tr>
<tr>
<td>Phase 3: wrap up</td>
<td>How do they show the task is complete?</td>
<td>Chatbot ask if draft can be sent. Julia says that she wants to proofread, doesn’t trust the chatbot for this. a team member arrives early and Julia is a bit hesitant to leave her writing.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>one part did not know to respond. on the other hand it feld it was not worth it. “ clarifies things for you” that’s the purpose of the technology.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>synonyms</td>
</tr>
</tbody>
</table>
### Post interview

<table>
<thead>
<tr>
<th>questions asked</th>
<th>What did you think of the tools</th>
<th>the fact checker is really good. sometimes it’s not clear what the source is, so a fact checker would be a good solution. especially when you need to write something quick, so it needs to be instant.</th>
</tr>
</thead>
<tbody>
<tr>
<td>what did you think about the chatbot?</td>
<td>I wasn’t sure to speak to it. I thought maybe it’s pre-programmed.</td>
<td></td>
</tr>
<tr>
<td>what did you think about the light presence ?</td>
<td>yeah I noticed that, it was nice.</td>
<td></td>
</tr>
<tr>
<td>did you understood the assignment we gave you?</td>
<td>yes, only wasn’t sure in the beginning what kind of article it needed to be.</td>
<td></td>
</tr>
<tr>
<td>did you understand the autotagger?</td>
<td>first I was not sure what to do with Do them, is it a google tag? need more of an explanation. It would be nice if you can click on them and it would expose more tags or information or synonyms.</td>
<td></td>
</tr>
<tr>
<td>are the tags maybe biased?</td>
<td>No they only influence you in the way that you use the word/tag multiple times in the article then you would’ve done. but not influences the direction of your writing, the tags are neutral, and it’s a fact.</td>
<td></td>
</tr>
<tr>
<td>Extra</td>
<td>I liked the part that the chatbot gives you a reminder. sometimes you get drifted away during writing, so somebody who reminds you is helpful. it’s better that it is a voice, because a email or some sort, you could ignore that.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Talks back and answers questions.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Doesn’t let the chatbot determine the time she spends on a story.</td>
<td></td>
</tr>
</tbody>
</table>

### Interview Pilot

Observing:

Don’t uses information from chatbot immediately, but writes the information down for later use.

Only uses the first suggested question from Marvin, after that it is only follow-up questions. But during the interview, the real time feedback from Marvin is used.

Uses phrasings from chatbot: “It’s even unconstitutional”

Julia thought that the feedback from Marvin was spoken by an actual computer (after someone wrote it)

Q: How was it?
A: bizarre, but really good!

Q: was the voice of the chatbot intervening?
A: Not at all, it was good. I was thinking about the previous quotes so I used that knowledge. And added the things from the chatbot, by saying something contradicting. so you don’t accuse somebody.

“ I didn’t look at the questions at all, because it’s not polite to look at your phone during a live interview. When it is for a skype interview or telephone interview you can use a second screen with texts or information on it.
Q: Did you trust the information the chatbot was giving you?
A: yes, if it is a solid system with solid sources, I would trust it.

Extra: Journalists are using shorthands. when writing they use sort of symbols.

Test participant 8
AT5 - Writing Scenario

<table>
<thead>
<tr>
<th>Simulation</th>
<th>Observations</th>
<th>Post sim feedback</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase 1: understanding the tool</td>
<td>chatbot started with talking, response normal, eventually asks for more information.</td>
<td>weird, works pretty well.</td>
</tr>
<tr>
<td>How do they interact with it initially?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Are they comfortable?</td>
<td>it seems so</td>
<td>first not that much, didn’t know what to do.</td>
</tr>
<tr>
<td>Do they look for /react to a physical presence?</td>
<td></td>
<td>for me, it has no added value. I can imagine in a busy newsroom you need headphones, so you don’t need a physical presence.</td>
</tr>
<tr>
<td>Phase 2: creating material</td>
<td>he accepted the tags. “yes, it’s fine”</td>
<td></td>
</tr>
<tr>
<td>Do they believe the chatbot?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Do they use the suggestions?</td>
<td>seems so</td>
<td></td>
</tr>
<tr>
<td>Do they ask the bot for extra information?</td>
<td>not yet, after awhile he asked for for confirmation about a fact/ or given information.</td>
<td></td>
</tr>
<tr>
<td>Do they ask for other functions? (spellcheck, format, etc)</td>
<td>also asks for results from a poll, in to the screen next to the writing part.</td>
<td></td>
</tr>
<tr>
<td>Phase 3: wrap up</td>
<td>chatbot asked if it was okay to send it in.</td>
<td></td>
</tr>
<tr>
<td>How do they show the task is complete?</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Post interview

| questions       | did you liked working with it?       | yeah, the whole collaboration works really nice and is useful. |
do you trust the system

what did you think of the chatbot, was it to intervening?

why didn’t you ask?

_extra

I could think of: synonyms, spell check, visual/image suggestions, related content, relevance (if the sentence or writing is relevance

**Test participant number 9**

Veronica - Writing Scenario

<table>
<thead>
<tr>
<th>Simulation</th>
<th>Observations</th>
<th>Post sim feedback</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase 1: understanding the tool</td>
<td>How do they interact with it initially?</td>
<td>starts reading shortly after writing, chatbot starts talking, reaction: ‘hmmm yeah’. eventually didn’t responded anymore. no response at all.</td>
</tr>
<tr>
<td></td>
<td>Are they comfortable?</td>
<td>not that much</td>
</tr>
<tr>
<td></td>
<td>Do they look for /react to a physical presence?</td>
<td></td>
</tr>
<tr>
<td>Phase 2: creating material</td>
<td>Do they believe the chatbot?</td>
<td>seems so, but not responding.</td>
</tr>
<tr>
<td></td>
<td>Do they use the suggestions?</td>
<td>yes, she does immediately</td>
</tr>
<tr>
<td></td>
<td>Do they ask the bot for extra information?</td>
<td>I didn’t think that asking for more information was possible</td>
</tr>
<tr>
<td>Phase 3: wrap up</td>
<td>Do they ask for other functions? (spell check, format, etc)</td>
<td>normally I would search for images, videos, in social media, or tweets to put in the article or quotes to embellish the text.</td>
</tr>
<tr>
<td>-----------------</td>
<td>------------------------------------------------</td>
<td>-------------------------------------------------</td>
</tr>
</tbody>
</table>

Post interview

<table>
<thead>
<tr>
<th>questions asked</th>
<th>what did you think of the robot talking?</th>
<th>it was fun, first I didn’t know if I needed to respond to it.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>was the voice okay or do you prefer messages</td>
<td>voice and messages are not distracting or at least less.</td>
</tr>
<tr>
<td></td>
<td>did you had questions during the writing?</td>
<td>didn’t know it was possible to ask info.</td>
</tr>
<tr>
<td></td>
<td>if you can go in again, would you do something different? ideal in writing</td>
<td>I would like to say to the system, “ I want this part and this part from several articles in my doc’. the system does the heavy lifting, and then you can edit the text maybe.</td>
</tr>
<tr>
<td></td>
<td>useful for daily job?</td>
<td>especially the fact checker with the pop ups, “ aah thank you’.</td>
</tr>
<tr>
<td></td>
<td>was the talking good for you or can you think of other interactions.</td>
<td>talking when you’re alone is great. but not for in a newsroom, or with earplugs.</td>
</tr>
</tbody>
</table>

Extra
