How Many Utterances Are Needed to Support Time-Offset Interaction?

Ron Artstein[1] and **Anton Leuski**[1] and **Heather Maio**[2]
Tomer Mor-Barak[1] and **Carla Gordon**[1,*] and **David Traum**[1]
[1]USC Institute for Creative Technologies, 12015 Waterfront Drive, Playa Vista CA 90094-2536, USA
[2]Conscience Display, 1023 Fifth Street, Coronado CA 92118, USA
[*]Now at California State University Long Beach, 1250 Bellflower Blvd, Long Beach CA 90040, USA

Abstract

A set of several hundred recorded statements by a single speaker is sufficient to address unrestricted questions and sustain short conversations on a circumscribed topic. Statements were recorded by Pinchas Gutter, a Holocaust survivor, talking about his personal experiences before, during and after the Holocaust. These statements were delivered to participants in conversation, using a "Wizard of Oz" system, where live operators select an appropriate reaction to each user utterance in real time. Even though participants were completely unconstrained in the questions they could ask, the recorded statements were able to directly address at least 58% of user questions. The unanswered questions were then analyzed to identify gaps, and additional statements were recorded to fill the gaps. The statements will be put in an automated system using existing language understanding technology, to create the first full working system of time-offset interaction, allowing a live conversation with a real human who is not present for the conversation in real time.

Introduction

For the past 150 years, people have been able to engage in direct conversation when separated by vast distances, using technologies such as the 19th-century telegraph and telephone (Bell 1876), 20th-century analog and digital video conference systems, and 21st-century applications (such as Skype and Hangouts) which enable high-fidelity multimodal voice, video and text interactions on consumer-grade electronic devices. However, all of these technologies require that conversation participants be available for conversation at the same time. Recently, Artstein et al. (2014) presented a concept of *time-offset interaction*, which removes this contemporaneity requirement while preserving the synchronous nature of conversation. The basic premise of time-offset interaction is that when the topic of conversation is circumscribed, the utterances of the participants are predictable to a large extent (Gandhe and Traum 2010). Knowing in advance what an interlocutor is likely to say, a speaker can record a large set of statements in advance; during the actual conversation, a computer program selects recorded statements that are appropriate reactions to the interlocutor's utterances. The selection of statements can be done in a similar fashion to existing interactive systems with synthetic characters (Leuski and Traum 2011).

Artstein et al. (2014) implemented a dialogue system that illustrates their concept of time-offset interaction. However, their system is highly restricted in content, comprising a total of 19 recorded statements; it can be used to demonstrate a conversation, but the user needs to know exactly what recorded responses are available. The present paper takes the first step to substantiate the conjecture that a larger set of recorded statements enables sustained conversation with participants who are not familiar with the recorded material. Our domain of interaction is questions and answers with a Holocaust survivor, such as arise in a typical museum setting after the survivor tells his or her story to a group of people. We recorded over 1400 utterances by Pinchas Gutter, a Holocaust survivor, elicited specifically to provide broad coverage within a narrowly circumscribed domain – his personal life experiences before, during and after the Holocaust. We then used these recordings in conversation with approximately 120 uninitiated participants in a Wizard-of-Oz setting, and found that the recorded statements were able to adequately address at least 58% of the users' utterances. The questions collected from the participants were used to inform a second round of recording, intended to fill the major gaps in the content of Mr. Gutter's statements; the new statements, combined with dialogue management techniques to address unknown questions, will enable unmediated conversation with an automated system (presently under development), allowing future generations to experience a face-to-face conversation with a Holocaust survivor.

The principal contribution of this paper is to outline a process for capturing the statements required to make time-offset interaction work. Real-time selection of statements in reaction to user utterances will use algorithms that have been deployed successfully in many dialogue systems in multiple conversational domains (Leuski and Traum 2011), and is not the focus of this paper. For systems that are to be deployed in real-world situations, the data used to develop and train the system are just as important as the reasoning algorithms. We therefore concentrate on the data collection method and analysis, which enables an interactive, automated dialogue system that simulates a conversation with a real person. The paper describes the procedure for eliciting appropriate speaker statements and participant data, presents an analysis of the coverage, and outlines further steps needed to make time-offset interaction a reality.

Method

Builders of dialogue systems are faced with a chicken and egg problem: for optimal performance, the system needs to have clear expectations of what users might say, but what users say is affected by the behavior of the system. The way around this problem is through iterative development (Rapp and Strube 2002): user interaction data are collected in stages as the system is developed, so that each successive stage has user data corresponding to the most recent system version. Iterative development can continue until the system is stable enough so that further changes in the system don't cause substantial changes in the users' reactions to it.

The same principle of iterative development can be used for time-offset interaction: each iteration would involve refinement of the speaker's statements and the collection of conversational data from new users, until the set of statements reaches a sufficiently broad coverage. Development of the system described in this paper was limited to two iterations by logistical and budgetary considerations: the end system requires very high quality recordings of the speaker statements, involving hiring external specialists and bringing the speaker from Toronto to a studio in Los Angeles. Given that only two iterations were possible, we designed a process to maximize the utility gained from these two iterations (Figure 1). The first iteration involved collecting potential user questions from a variety of sources, crafting an elicitation script based on the collected materials and other key points we wanted our speaker to address, and recording speaker statements elicited through the script. The recorded statements were then processed, categorized, and put into an interface that allowed a human operator ("wizard") to rapidly access and play each statement. The second iteration used the wizard system to collect user questions in conversation, with one or more operators playing the speaker's statements in real time in response to user utterances; the collected user utterances were then annotated, and a new elicitation script was created to address the important gaps identified through the process, followed by a second recording. The statements from the second recording are presently being assembled into a fully automated interactive dialogue system, to enable independent time-offset conversations with the speaker. The remainder of this section describes the content development process in detail.

Initial question collection

The time-offset interaction we are developing is intended to replicate a very specific type of conversation – a question-answer session that typically occurs in museums after a Holocaust survivor tells his or her story. Our starting point for collecting statements for time-offset interaction is our speaker, Pinchas Gutter, who has been telling his story in a variety of forums for well over a decade, and thus has a lot of knowledge about which statements work well with various audiences, which facts and stories he wishes to share, and how to deliver these. For a dialogue system to be successful, however, it needs to not only deliver the desired content, but to do so in a way that addresses the concerns of the conversation participants; we therefore need to know what the most common user concerns are. In the absence of a working dialogue system to elicit user questions, our first iteration begins with collecting such questions from a variety of available sources.

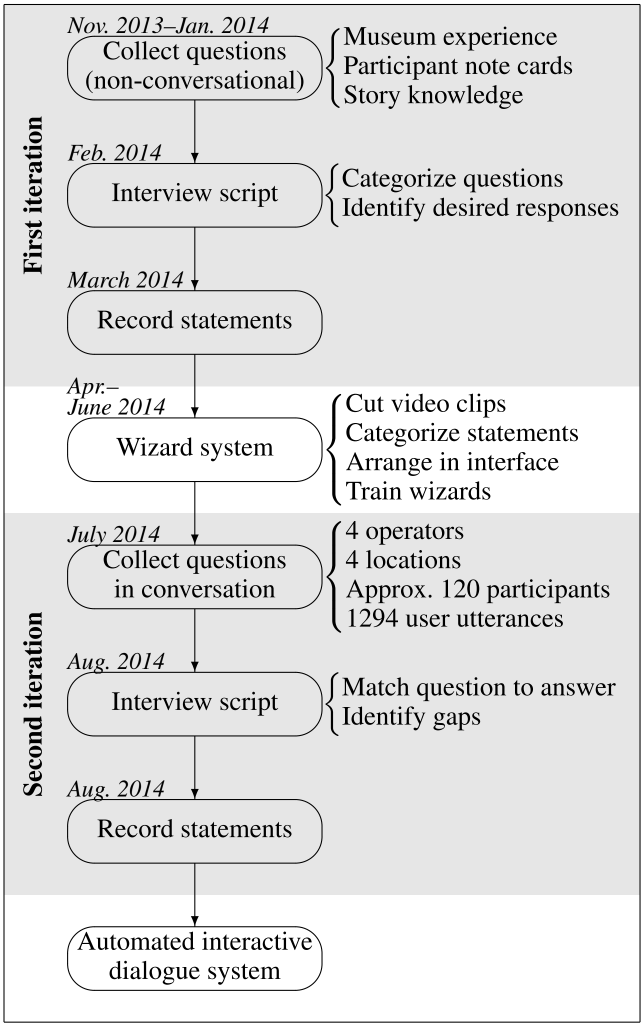


Figure 1: Content development process and timeline

Question-answer interactions happen on a daily basis at many Holocaust museums. Based on 20 years of field experience, the third author drafted an initial list of commonly asked questions; that list was then sent out to various experts in the fields of Holocaust testimony preservation, history, genocide studies, trauma specialists, Holocaust museum education staff, and representative target audiences, who were asked to add to that list. Those additions were reviewed and consolidated by the third author to come up with a list of the Top 99 questions, reflecting the most salient issues that our speaker should address in his recorded statements.

A second source of questions presented itself when our speaker participated at a Holocaust commemoration event at the College of Saint Elizabeth in Morris County, New Jersey, in front of an audience of local students (ages 12–19, accompanied by a small number of adults teachers and chaperones).[1] The event featured a film about Mr. Gutter's return voyage to Poland nearly 60 years after the Holocaust, followed by a live question-answer session with Mr. Gutter and Stephen Smith, executive director of the USC Shoah Foundation. The audience were provided with note cards and asked to write down 3 questions they had for Mr. Gutter. A total of 334 note cards were collected and transcribed, providing 746 questions relating to Mr. Gutter's story from a wide variety of users with different backgrounds. This set of questions will be referred to as the NJ questions.

In addition to the collected questions, we crafted a set of questions designed to elicit specific stories and other bits of information, based on our prior familiarity with our speaker. This was done in recognition that any collected set of questions will have some gaps, and that a good story can often serve as a response to a question that did not ask for it specifically. Both the collected and the devised questions were categorized according to themes and arranged into a set of interview scripts for recording.

First recording

Time-offset interaction employs the *selection approach* to dialogue management, where at each conversational turn, the system picks an appropriate utterance from a corpus of available utterances (Gandhe and Traum 2010). Our speaker, Pinchas Gutter, is not an actor, and is not able to deliver prepared statements in a natural, conversational tone. Therefore the questions were arranged in an elicitation script and given to an interviewer, and our speaker responded naturally to the interviewer's questions.

The recording took place over five days in March 2014. Most of the interviewing was done by Stephen Smith, due to his long-established rapport with the speaker. However, there are also disadvantages to this acquaintance, because the speaker's responses may presuppose a lot of shared knowledge. We therefore had some of the interview sessions conducted by a person not familiar to the speaker. An important target audience is children and young adults; in order to capture statements addressed specifically to a younger crowd, we brought in some younger interviewers – a college student, two high school students, and several children of staff associated with the project (ages 8–12).

A specific type of statement that proved particularly difficult to elicit was off-topic reactions such as *could you please repeat that* or *I don't know*. It is important for a dialogue system to have a wide variety of off-topic reactions to use when the user asks a question for which the system does not have a direct answer, or when the system does not understand what the user had said (Artstein et al. 2009). Such reactions did occur naturally during the interview, but not with enough frequency and variety to satisfy the needs of the eventual dialogue system. We therefore had to resort to other methods of eliciting off-topic reactions, for example asking the speaker explicitly to repeat certain statements, or asking questions while instructing the speaker to not answer them directly.

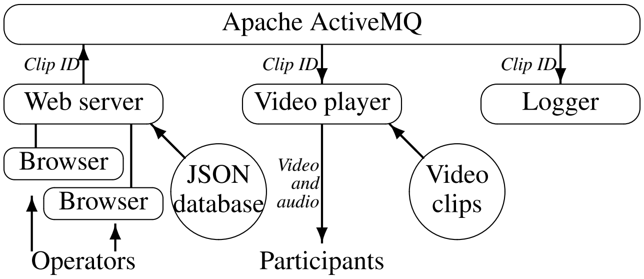


Figure 2: Wizard system architecture

Following the recording, the full interviews were transcribed and segmented into individual statements. The footage from the interviews is intended for high-resolution, high-quality video display in the final dialogue system, but processing of the footage would not begin until the second recording was in place. We therefore performed a rough cut of the raw video, resulting in a total of 1420 clips of individual statements, ranging in length from under one second (statements like *yes*, *no*, and *go ahead*) to several minutes. The rough-cut video clips were then used for a set of "Wizard of Oz" sessions as a mock-up of the eventual dialogue system, in order to collect user questions in conversation.

Wizard interface

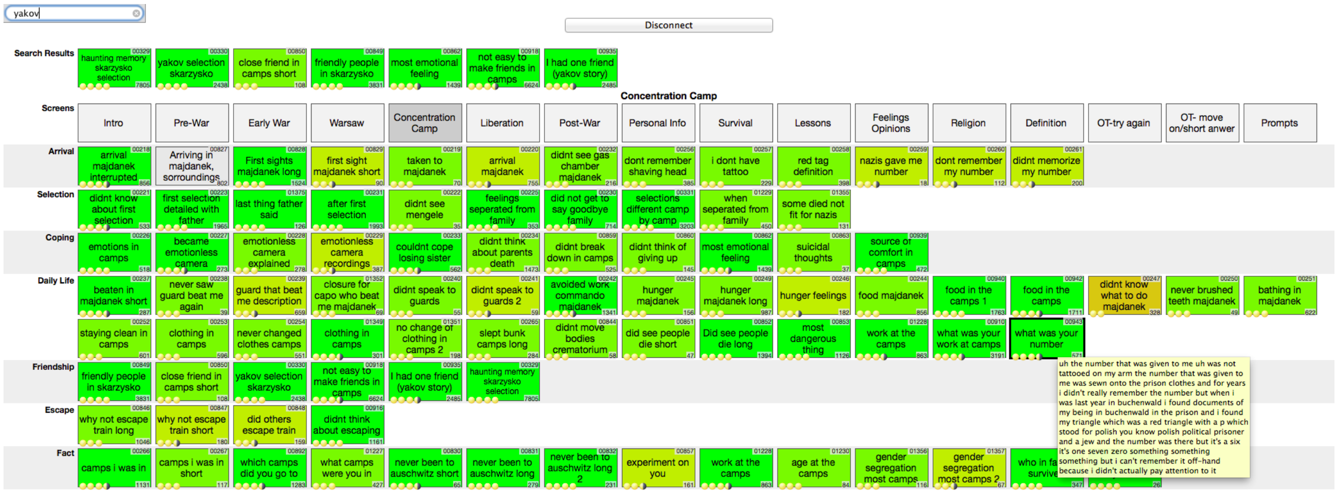
To collect user utterances in conversation we designed a system that allows operators ("wizards") to quickly access a large number of video clips for rapid playback. The system includes two main components – a wizard interface and a video player, which communicate using the VHMsg messaging protocol (Hartholt et al. 2013) built on top of the ActiveMQ message broker.[2] The wizard interface is implemented as a web application that can reside on a server or a local machine. It presents a collection of clickable buttons in a web browser window. Operators use the interface buttons to trigger playback of individual utterances. Multiple clients can connect to the web application at the same time, allowing several wizards to control the interface simultaneously. The wizard interface sends VHMsg messages to the video player with instructions to play specific clips (Figure 2).

The large number of video clips makes it impossible to access all of them from a single screen. The interface therefore contains alternative screens, and specific buttons (tabs) to switch between screens. An operator can get to a desired clip with just two mouse clicks – one to reach the correct screen, and a second to select the video clip. The screen switching buttons are in the second row of each screen.

The screens are organized according to theme, and within each screen, the individual buttons are arranged in rows according to sub-themes; the same button may appear in more than one place, if the associated clip fits multiple themes. Each button displays several characteristics to aid in quick identification: a short label text on the button, a color code, badges on three corners, and a tooltip that appears when the mouse hovers over the button. A search facility provides a

[1] http://www.nj.com/independentpress/index.ssf/2013/11/holocaust_survivor_pinchas_gut.html

[2] http://activemq.apache.org


yakov
Disconnect
Search Results
haunting memory skarzysko selection
yakov selection skarzysko
close friend in camps short
friendly people in skarzysko
most emotional feeling
not easy to make friends in camps
I had one friend (yakov story)
Concentration Camp
Screens
Intro
Pre-War
Early War
Warsaw
Concentration Camp
Liberation
Post-War
Personal Info
Survival
Lessons
Feelings Opinions
Religion
Definition
OT-try again
OT- move on/short anwer
Prompts
Arrival
arrival majdanek interrupted
Arriving in majdanek, sorroundings
First sights majdanek long
first sight majdanek short
taken to majdanek
arrival majdanek
didnt see gas chamber majdanek
dont remember shaving head
i dont have tattoo
red tag definition
nazis gave me number
dont remember my number
didnt memorize my number
Selection
didnt know about first selection
first selection detailed with father
last thing father said
after first selection
didnt see mengele
feelings seperated from family
did not get to say goodbye family
selections different camp by camp
when seperated from family
some died not fit for nazis
Coping
emotions in camps
became emotionless camera
emotionless camera explained
emotionless camera recordings
couldnt cope losing sister
didnt think about parents death
didnt break down in camps
didnt think of giving up
most emotional feeling
suicidal thoughts
source of comfort in camps
Daily Life
beaten in majdanek short
never saw guard beat me again
guard that beat me description
closure for capo who beat me majdanek
didnt speak to guards
didnt speak to guards 2
avoided work commando majdanek
hunger majdanek
hunger majdanek long
hunger feelings
food majdanek
food in the camps 1
food in the camps
didnt know what to do majdanek
never brushed teeth majdanek
bathing in majdanek
staying clean in camps
clothing in camps
never changed clothes camps
clothing in camps
no change of clothing in camps 2
slept bunk camps long
didnt move bodies crematorium
did see people die short
Did see people die long
most dangerous thing
work at the camps
what was your work at camps
what was your number
Friendship
friendly people in skarzysko
close friend in camps short
yakov selection skarzysko
not easy to make friends in camps
I had one friend (yakov story)
haunting memory skarzysko selection
Escape
why not escape train long
why not escape train short
did others escape train
didnt think about escaping
Fact
camps i was in
camps i was in short
which camps did you go to
what camps were you in
never been to auschwitz short
never been to auschwitz long 2
never been to auschwitz long 2
experiment on you
work at the camps
age at the camps
gender segregation most camps
gender segregation most camps 2
who in fa survive

uh the number that was given to me uh was not tattooed on my arm the number that was given to me was sewn onto the prison clothes and for years i didn't really remember the number but when i was last year in buchenwald i found documents of my being in buchenwald in the prison and i found my triangle which was a red triangle with a p which stood for polish you know polish political prisoner and a jew and the number was there but it's a six it's one seven zero something something something but i can't remember it off-hand because i didn't actually pay attention to it

Figure 3: Wizard control interface. A 1920-pixel-wide screen allows the display of up to 16 columns per row. The buttons in the second row are used to change screens; the remaining buttons play individual video clips. The top row of buttons is the output of the text search in the top left corner. Information on each button includes a label text, color coding, and three badges at the corners. The tooltip appears when hovering over the button, displaying the full text of the associated video clip.

partial character match with the label and tooltip texts. We used the color coding to indicate the approximate quality of the video clip; the original idea was to help operators make a choice when several clips were appropriate, though as operators became familiar with the layout of the individual screens, the colors mostly served as familiar landmarks. Our badges indicated the clip ID on the top right (a number between 1 and 1422), an approximate length of the video on the bottom right, and a graphic indication of quality on the bottom left. The tooltip contained the full transcribed text of the video associated with each button (Figure 3).

The content and organization of buttons on the wizard interface is edited and stored in an Excel file. We also created a utility application to convert the Excel file into the JSON format that is read by the web application. Following the processing and categorization of the video clips we put them into the wizard interface and started training the wizards on operating the system. Throughout the training (which took several weeks) we refined the categorization and organization. In its final state, the wizard interface contains buttons for 1251 clips, distributed among 16 screens (this number was chosen to allow the screen tabs to fit in a single row on a 17-inch laptop). Most of the buttons (1149) appear just once in the interface; 99 buttons appear twice, and 3 buttons appear three times.

Wizard data collection

The data collection effort took place in 11 days spread over a period of 31 days in June and July 2014. Four people were trained as operators ("wizards") for the collection of conversational data; we aimed to have at least two operators working simultaneously in order to achieve a normal conversational pace by reducing the time searching for responses. Participants talked to the speaker individually or in small groups; they were given a short introduction about the purpose of the data collection, and instructed to converse normally with the speaker on screen. We did not collect any personal information about the participants, but attempted to ask for basic demographic information (gender, age, religion, ethnicity); this proved impractical with larger groups or in settings where participants were flowing freely into and out of the room, and therefore our demographic information is incomplete. We did not collect structured feedback about the wizard sessions. Most participants asked to start with the 5-minute introduction clip where the speaker tells his basic life story. We did not time the individual sessions (and in free-flowing settings, the concept of "session" is not well defined), but we estimate that typical participants engaged with the speaker for about 20–30 minutes. Approximately 120 people participated in the conversational sessions.

Collection sites included a museum, a middle school, a nearby university, as well as bringing participants to our location (some of the participants were recruited through online ads and were paid $15; the rest, including all off-site participants, were volunteers). Our mobile set-up included one laptop to run the wizard system components (activeMQ, web server, video player, and logger), a flat-screen 22-inch monitor with built-in speakers to display the video clips, two 17-inch laptops for the operators, a black fabric screen to hide the operators from the participants' view, and a Marantz PMD-660 recorder to record the user utterances. If the room had a fixed television or projector, we used that instead of the portable monitor. If the room was arranged like a classroom, the operators sat at the back of the room, behind the participants, and no occlusion was necessary.

A total of 1350 user utterances were recorded and transcribed; we excluded 25 demographic summaries directed at the experimenter and 31 backchannels (*hmm*, *um* etc.), leaving 1294 utterances which were directed at the speaker. Speaker statements were recovered from the logs and aligned with the transcribed user utterances. A total of 1711 speaker statements were selected by the operators, representing 551 unique statement types (these numbers include some statements used for testing when no participant was present and some statements that were immediately canceled by the operators, but are missing statements from one day of testing when the logger was not working properly). The distribution of statement types is far from even – the 6 most frequent speaker statements ($N \geq 22$) account for 9.8% of the data and the 25 most frequent statements ($N \geq 10$) account for 24%, while at the other end, 109 statements were only used twice by the operators, and 238 statements were only used once.

The transcribed user utterances, along with the Top 99 and NJ questions from the first iteration, were annotated to identify gaps in the recorded statements. The material was split between two annotators (who had also served as wizards); each utterance was matched with an appropriate response if such a statement was available, or with a response "can't answer" if no direct response was found in the recorded clips. In this effort the annotators used a total of 420 recorded statements, including 5 versions of "can't answer". For each user utterance not addressed by existing recorded statements, a decision was made as to whether it required an elicited response or if it could be left unanswered and handled by an off-topic response. The decisions were made based on perceived importance: in general, if a question was asked by more than one user it was deemed important, and a subjective judgment call was made for the singleton questions. The important questions were arranged into an interview script for the second recording (further analysis of this annotation effort appears in the results section).

Second recording

The second recording took place over two days in August 2014. Again, most of the interviewing was done by Stephen Smith. However, an important set of questions identified in the gap analysis were naïve questions, which show little understanding of the speaker and his circumstances; to elicit sincere responses to such questions we had them asked by a child interviewer who had participated in the first round and demonstrated the ability to elicit useful statements.

The second recording was transcribed and segmented, and rough cut video clips from both recordings are now being assembled into an automated dialogue system. While no new speaker statements will be recorded, development will continue iteratively on the user input side, along the lines outlined by Rapp and Strube (2002): user interactions with the system will be used to refine the system's language understanding and dialogue management components, and each successive refinement will be used to collect new user data.

Results

An initial evaluation was conducted in the midst of the wizard collection effort, to assess the frequency in which the known questions were occurring in conversation. The annotation materials included all the utterances collected during the first 6 days of wizard testing (a total of 738 user utterances). Two annotators were tasked with matching the collected utterances to the Top 99 list: an utterance was considered to be essentially the same as one or more of the Top 99 if a speaker's answer to one was likely to also serve as an answer to the other. If an utterance was not essentially the same as a Top 99 question then it was marked as "none" if it was a question, or "naq" if it was not a question at all (examples of utterances marked as "naq" include *goodbye*, *thank you*, and *yeah music is very important in life*). Inter-annotator reliability was $\alpha = 0.82$ (Krippendorff 1980), calculated on 727 utterances (one of the annotators had failed to mark 11 utterances), and ignoring disagreements on which of the Top 99 questions an utterance was mapped to. Overall, about 29% of the user utterances were essentially the same as the Top 99 questions, 42% were other questions, and 29% were not questions. This annotation confirmed that top questions identified by experts do occur in high frequency in conversation, but not so high that addressing these questions alone would be enough to sustain a conversation.

Question source	Top-99		NJ		Wizard	
	N	%	N	%	N	%
Answer available	73	74	366	49	756	58
No answer	18	18	348	47	376	29
Both	1	1	5	1	21	2
Unannotated	7	7	27	4	141	11
Total	99	100	746	100	1294	100

Table 1: Questions answered by the first recording

A separate analysis was conducted based on the question-response annotation described in the section on wizard data collection. Table 1 shows the results of the annotation effort, broken down by question source; for each source the table shows the number of questions or utterances for which a direct response was available and those which did not have a direct response. A small number of utterances were annotated with both, usually because the utterance had multiple parts, only some of which were addressed by existing statements. The table shows a wide disparity between the different sources. Not surprisingly, the source with the highest coverage is the Top 99 list, where at least 73 questions (74%) are directly addressed by speaker statements. It is interesting to note that the recorded statements address a higher proportion of utterances from the wizard collection than from the NJ data set, even though the NJ data were used to guide the recording. The analysis shows that at least 58% of utterances made in conversation are directly addressed by recorded statements.

Some of the gaps identified were specific bits of information that were missed in the interview script from the first iteration. For example, four independent participants asked the same question with the exact same wording: *Where do you live now?* While our speaker did mention his current abode in several of the statements from the first iteration, none was a direct answer, so a direct answer was recorded in the second round (he lives in Toronto). Several participants asked questions with the (false) presupposition that the speaker lived in or had immigrated to the United States, so a direct reaction was recorded to correct this misconception.

There were also some general themes that emerged from the wizard data. While the purpose of the interaction is primarily to educate people about the Holocaust and the events surrounding it, it turned out that many of the participant questions were about the speaker's life today, his profession, and his family and children. Since this appears to be a genuine interest in this type of conversation, we elicited many additional statements on these themes. Of course, these statements give information that is current as of the time of recording; in this sense, the interaction is archival, and will need to be interpreted by future participants as grounded in a specific time.

Other participant questions are very specific follow-ups to individual statements. For example, in talking about the Nazi labor camps, the speaker tells how the guards made the prisoners carry heavy stones; one participant asked a follow-up question: *What was the purpose of um carrying these stones back and forth?* This question is likely to only ever appear as a follow-up, but as such, it may well appear again. However, it is impractical to load a dialogue system with all the likely follow-ups to all system utterances, and this is the kind of question which is best handled by an off-topic response (e.g. *I have nothing to say about that topic*).

Finally, we note that there is a long tail of unseen questions, some with non-negligible frequency. For example, our speaker says he was born in Łódź, Poland. About two weeks after the wizard data collection we conducted a large demo, where two users independently asked about the location of Łódź within Poland. According to our criteria, a frequency of 2 implies that a question is important – but this particular question had not been encountered in the formal testing. The existence of important but unseen questions underlines the need for a robust off-topic mechanism, to adequately handle unknown questions that cannot be addressed directly.

Discussion

The coverage analysis has demonstrated that already after the first round of recording, at least 58% of user utterances in conversation can be directly addressed by the recorded statements. Since the second recording addresses the most frequently encountered gaps in coverage, we expect the proportion of covered material to rise. This would leave about 20%–30% of the utterances without a direct answer among the recorded material, and those will be addressed through dialogue management, indirect answers and off-topic responses, to ensure that the conversation proceeds smoothly. We thus have support for the conjecture that a large but fixed set of statements is sufficient to enable time-offset interaction. The precise size is to be determined with further testing, but given the numbers we have seen in our wizard testing (551 unique statements used by wizards) and annotation (420 statements), we estimate that around 600–700 recorded statements should provide the required coverage.

The annotation results underscore the importance of collecting questions from real users engaged in conversation: both the Top 99 and the NJ collections missed important questions and themes, which were revealed through the wizard data collection, enabling the closure of important gaps in the coverage.

The next step will be the creation of an automated dialogue system. This involves training a classifier to identify appropriate responses based on matched questions (Leuski and Traum 2011). New user questions will be collected with the automated system to refine the classifier and dialogue management.

Time-offset interaction has a large potential impact on preservation and education – people in the future will be able to not only see and listen to historical figures, but also to interact with them in conversation. The content development process described in this paper has only been performed for one person, and for one intended domain of conversation. Future research into time-offset interaction will need to generalize this process, in order to identify which of the common user questions are specific to the person, which are specific to the dialogue context or conversation topic, and which are of more general application.

References

Artstein, R.; Gandhe, S.; Gerten, J.; Leuski, A.; and Traum, D. 2009. Semi-formal evaluation of conversational characters. In Grumberg, O.; Kaminski, M.; Katz, S.; and Wintner, S., eds., *Languages: From Formal to Natural. Essays Dedicated to Nissim Francez on the Occasion of His 65th Birthday*, volume 5533 of *Lecture Notes in Computer Science*. Heidelberg: Springer. 22–35.

Artstein, R.; Traum, D.; Alexander, O.; Leuski, A.; Jones, A.; Georgila, K.; Debevec, P.; Swartout, W.; Maio, H.; and Smith, S. 2014. Time-offset interaction with a Holocaust survivor. In *IUI '14: Proceedings of the 19th International Conference on Intelligent User Interfaces*, 163–168.

Bell, A. G. 1876. Improvement in telegraphy. U.S. Patent 174,465.

Gandhe, S., and Traum, D. 2010. I've said it before, and I'll say it again: An empirical investigation of the upper bound of the selection approach to dialogue. In *Proceedings of SIGDIAL 2010: the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 245–248.

Hartholt, A.; Traum, D.; Marsella, S. C.; Shapiro, A.; Stratou, G.; Leuski, A.; Morency, L.-P.; and Gratch, J. 2013. All together now: Introducing the virtual human toolkit. In *International Conference on Intelligent Virtual Humans*.

Krippendorff, K. 1980. *Content Analysis: An Introduction to Its Methodology*. Beverly Hills, CA: Sage. chapter 12, 129–154.

Leuski, A., and Traum, D. 2011. NPCEditor: creating virtual human dialogue using information retrieval techniques. *AI Magazine* 32(2):42–56.

Rapp, S., and Strube, M. 2002. An iterative data collection approach for multimodal dialogue systems. In *Proceedings of LREC 2002: The Third International Conference on Language Resources and Evaluation*, 661–665.