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

Time-offset interaction is a new technology that enables conversational interaction with a person who is not present, using pre-recorded video statements. Statements were recorded by Pinchas Gutter, a Holocaust survivor, talking about his personal experiences before, during and after the Holocaust. Participants interacted with the statements through a "Wizard of Oz" system, where live operators select an appropriate reaction to each utterance in real time; unanswered questions were analyzed to identify gaps, and additional statements were recorded to fill the gaps. Even though participant questions were completely unconstrained, the recorded statements from the first round directly addressed at least 58% of the questions; this number rises to 95% with the second round of recording, when tested on newly elicited utterances. This demonstrates the feasibility for a system to address unseen questions and sustain short conversations when the topic is well defined. The statements have been put into an automated system using existing language understanding technology, to create a preliminary 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[1] and Hangouts[2]) which enable high-fidelity multimodal voice, video and text interactions on consumer-grade electronic devices. However, all of these technologies require that the parties be available for conversation at the same time. In Artstein et al. (2014) we presented the 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 known, the participants' utterances are predictable to a large extent (Gandhe and Traum 2010). Knowing what an interlocutor

is likely to say, a speaker can record statements in advance; during 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).

The dialogue system in Artstein et al. (2014) illustrates the concept of time-offset interaction but is highly restricted in content (19 recorded statements in total); to demonstrate a conversation, a user needs to know exactly what recorded responses are available. The present paper takes a first step to support the conjecture that a larger set of recorded statements enables sustained conversation with participants who are not familiar with the recorded material. We focus on 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 survivor Pinchas Gutter, elicited to provide broad coverage on a fixed topic – his personal experiences before, during and after the Holocaust. We used these recordings in conversation with approximately 120 uninitiated participants, and found that they adequately addressed at least 58% of the users' questions. The user utterances informed a second round of recording, intended to fill the major gaps in the content of Mr. Gutter's statements; the new statements increase coverage to 95% on newly elicited utterances. Combined with dialogue management techniques to address unknown questions, this 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 a two-phase process for capturing statements for time-offset interaction: initial recording, interactive elicitation, and pick-up recording. Automated selection of statements in conversation uses algorithms that have been deployed successfully in many dialogue systems (Leuski and Traum 2011), and is not the focus of this paper. For real-world applications, 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. This paper describes the procedure for eliciting appropriate speaker statements, presents an analysis of the coverage, and outlines further steps needed to make time-offset interaction a reality.

[1]http://www.skype.com
[2]https://plus.google.com/hangouts

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.

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 sufficiently broad coverage. Our development was limited to two iterations by logistical and budgetary considerations: the end system requires very high quality recordings, involving hiring external specialists and bringing the speaker from Toronto to a studio in Los Angeles. We therefore 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 wizards 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

Our time-offset interaction is designed to replicate a question-answer session that typically occurs in museums after a Holocaust survivor tells his or her story. Our speaker, Pinchas Gutter, 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. Our first iteration begins with collecting such questions from a variety of available sources.

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
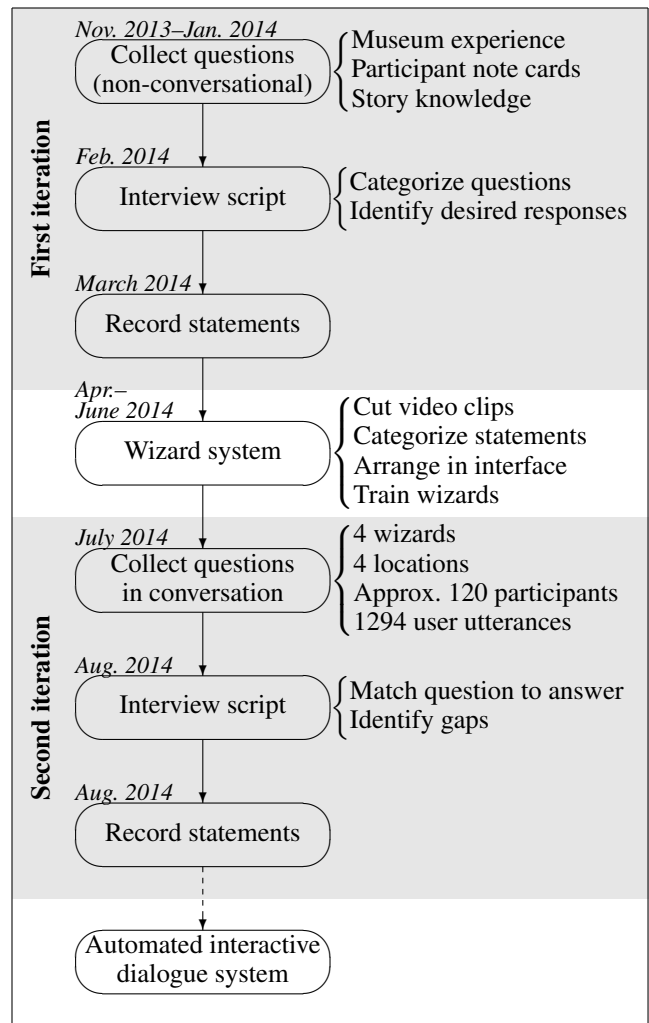

Figure 1: Content development process and timeline

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. Additions to the list were consolidated by the third author to make 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). 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 the 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

To record the speaker utterances in a natural, conversational tone, we arranged the questions in an elicitation script that was given to interviewers; our speaker responded naturally to the interviewers' 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 people 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).

We also elicited 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 also elicited off-topic reactions by asking the speaker explicitly to repeat certain statements, or asking questions while instructing the speaker to not answer them directly.

Following the recording, the full interviews were transcribed and segmented into individual statements. Processing of high-resolution footage would not begin until the second recording was in place, so we 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 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
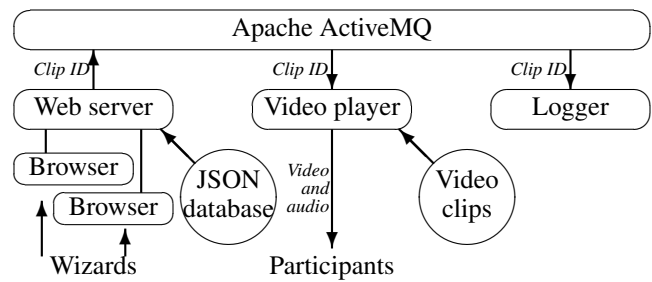

Figure 2: Wizard system architecture

broker.[3] 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. Wizards 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. A wizard 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 by theme, and individual buttons are arranged in rows according to sub-themes; a 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 partial character match with the label and tooltip texts. Color coding indicates the approximate quality of the video clip; the original idea was to help wizards make a choice when several clips were appropriate, though eventually the colors mostly served as familiar landmarks. Badges show 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 contains 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. Wizard training took several weeks, and throughout this period we refined the categorization and organization of the interface. In its final state, the wizard interface contains buttons for 1251 clips, distributed among 16 screens (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.
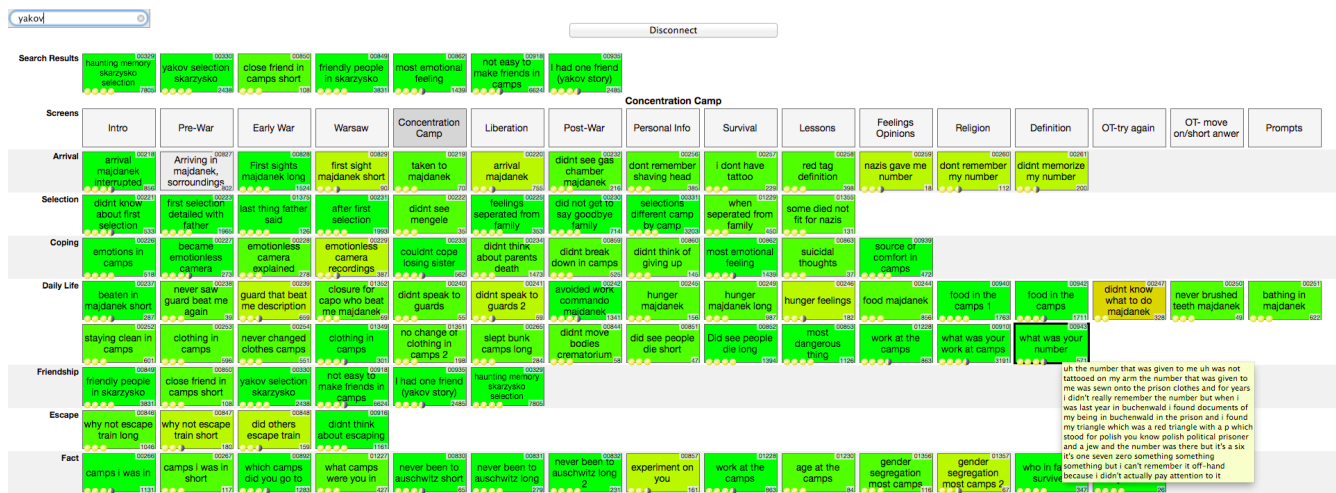
---

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

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.

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 wizards for the collection of conversational data; we aimed to have at least two wizards 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 experience. 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 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 university, and our own institute. Some participants were recruited through online ads and paid $15; the rest, including all off-site participants, were volunteers. We used one laptop to run the wizard system components, a flat-screen 22-inch monitor with built-in speakers to display the video clips, two 17-inch laptops for the wizards, a black fabric screen to hide the wizards 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 wizards 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 system logs and aligned with the transcribed user utterances. A total of 1711 speaker statements were played by the wizards, using 551 distinct video clips (including some statements from system testing and some statements that were immediately canceled, but missing statements from one day when the logger was not working properly). The distribution of statements is far from even – the 6 most frequently used video clips ($N \geq 22$) account for 9.8% of the data and the 25 most frequent clips ($N \geq 10$) account for 24%, while at the other end, 109 clips were only used twice by the wizards, and 238 clips 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 one 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.

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, resulting in a total of 597 additional video clips. Clips from both recordings were then assembled into a preliminary automated dialogue system. While no new speaker statements will be recorded, development continues iteratively on the user input side, following Rapp and Strube (2002): user interactions are used to refine the system's language understanding and dialogue management, and each successive refinement is 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 matched 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.

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 propor-

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[a]	7	7	27	4	141	11
Total	99	100	746	100	1294	100

[a]Utterances left unannotated do to a technical error

Table 1: Questions answered by the first recording

tion 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 statements recorded in the first round.

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.

Some general themes also 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, 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. It is impractical to load a dialogue system with all the likely follow-ups to all system utterances, so such questions are best handled by an off-topic response (e.g. *I have nothing to say about that topic*).

There remains 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.

Following the second recording, we implemented a preliminary automated system and collected additional data

Best Response	Dec. 10		Jan. 15		Jan. 18	
	N	%	N	%	N	%
1st rec.	452	69	427	65	203	58
2nd rec.	187	29	214	32	129	37
Off-topic	15	2	21	3	19	5
Total	654	100	662	100	351	100

Table 2: Source for the best response in three days of data collection, winter 2014–2015

from people's interactions with the system; annotators subsequently identified the best response for each user utterance, for the purpose of retraining the language understanding components. Table 2 shows the results of this annotation for three days of data collection at a local university and a local museum. While the majority of the utterances are addressed by questions from the first recording, a substantial minority (29–37%) are best addressed by questions from the second recording. The proportion of utterances that cannot be addressed directly with material from either recording is 5% or less. The annotated responses represent 426 distinct statements, of which 42 were used on all three days, 109 were used on two days, and 275 statements were used on one day. In the full annotation of interactions with the automated system, 701 distinct statements represent the best responses to a total of 4081 participant utterances.

Discussion

The coverage analysis has demonstrated that after the first round of recording, at least 58% of user utterances in conversation can be directly addressed by the recorded statements. Coverage increases to 95% with the inclusion of statements from the second recording; this large increase is probably due to the fact that the second recording was informed by 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. We thus have support for the conjecture that a large but fixed set of statements is sufficient to enable time-offset interaction. The total number of clips recorded in this effort – slightly more than 2000, including all duplicate statements and retakes – should serve as an upper bound.

While the analysis shows that a direct answer *exists* for 95% of the user utterances, an automated system is likely to *find* the right answer for a lower proportion, due to deficiencies in speech recognition and gaps in language understanding. Remaining utterances will be addressed through dialogue management, indirect answers and off-topic responses, to ensure that the conversation proceeds smoothly. A preliminary automated dialogue system has been built, and we are now collecting new user questions to train and refine the matching between user questions and speaker responses (Leuski and Traum 2011).

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*. 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 *Proc. IUI '14*, 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 *Proc. SIGDIAL 2010*, 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.

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 *Proc. LREC 2002*, 661–665.