This is a preprint of the manuscript entitled "Explicit feedback from users attenuates memory biases in human-system dialogue", to be published in the International Journal of Human-Computer Studies. Minor differences between this version and the final version of the article might be found. For the final version, please refer to the actual publication:

Knutsen, D., Le Bigot, L., & Ros, C. (in press). Eliciting explicit feedback from users to attenuate memory biases in human-system dialogue. ***International Journal of Human-Computer Studies.*** **doi:10.1016/j.ijhcs.2016.09.004**

Running head: Memory biases in human-system dialogue

Explicit feedback from users attenuates memory biases in human-system dialogue

Dominique KNUTSEN

Department of Psychology, University of Essex, UK

Ludovic LE BIGOT & Christine ROS

Université de Poitiers, France; CNRS (CeRCA, UMR 7295), France

Corresponding author

Dominique Knutsen

University of Essex

Wivenhoe Park

Colchester CO4 3SQ

United Kingdom

Email address: dknutsen@essex.ac.uk

Abstract

In human-human dialogue, the way in which a piece of information is added to the partners' common ground (i.e., presented and accepted) constitutes an important determinant of subsequent dialogue memory. The aim of this study was to determine whether this is also the case in human-system dialogue. An experiment was conducted in which naïve participants and a simulated dialogue system took turns to present references to various landmarks featured on a list. The kind of feedback used to accept these references (verbatim repetition vs. implicit acceptance) was manipulated. The participants then performed a recognition test during which they attempted to identify the references mentioned previously. Self-presented references were recognised better than references presented by the system; however, such presentation bias was attenuated when the initial presentation of these references was followed by verbatim repetition. Implications for the design of automated dialogue systems are discussed.

Keywords

Human-system dialogue, dialogue memory, common ground, memory biases, feedback, Wizard-of-Oz studies

Highlights

- Human dialogue memory is subject to presentation and acceptance biases.
- The present study examined whether this is also the case in human-system dialogue.
- Participants interacted over the phone with a simulated dialogue system.
- Participants were subject to a presentation bias which was attenuated in some cases.
- Implications for the design of automated dialogue systems are discussed.

Explicit feedback from users attenuates memory biases in human-system dialogue

1. Introduction

Human-system dialogue is a goal-oriented activity during which a human being (usually referred to as a user) uses language to interact with an automated dialogue system. Such interactions are increasingly frequent, as it is not uncommon nowadays to interact with a system using natural speech or keywords in order to buy a train ticket or to book a flight (see Barrett & Jiang, 2012; Grudin, 2005; Pieraccini & Huerta, 2008; Zhou, 2007).

The psychological processes at play in human-system and human-human dialogue are supposed similar, as users' expectations and beliefs regarding dialogue system are analogous to those held by human partners engaged in dialogue (e.g., Bergmann, Branigan, & Kopp, 2015; Branigan, Pickering, Pearson, & McLean, 2010; Branigan, Pickering, Pearson, McLean, & Brown, 2011; Branigan, Pickering, Pearson, McLean, & Nass, 2003; Brennan, 1991, 1996; Cavedon et al., 2015; El Asri, Lemmonier, Laroche, Pietquin, & Khouzaimi, 2014; Iio et al., 2015; Johnstone, Berry, Nguyen, & Asper, 1995; Kiesler, 2005; Koulouri, Lauria, & Macredie, 2015; Le Bigot, Caroux, Ros, Lacroix, & Botherel, 2013; Powers et al., 2005; Suzuki & Katagiri, 2007; van Lierop, Goudbeek, & Krahmer, 2012; Zoltan-Ford, 1991). In this sense, dialogue psychology provides important insight for the development of automated dialogue systems. In particular, one major finding is that human dialogue partners attempt to produce partner-adapted utterances as they interact (Brennan & Clark, 1996; Clark & Wilkes-Gibbs, 1986; Yoon & Brown-Schmidt, 2012). To do so, they rely on their memory for past interactions, or dialogue memory (e.g., Gibbs, 1986; Keenan, MacWhinney, & Mayhew, 1977; Le Bigot et al., 2013; Pasupathi & Hoyt, 2010). However, such memory is subject to a number of biases which cause some of the encoded pieces of information to

become less readily accessible than others (Knutsen & Le Bigot, 2015; Knutsen, Ros, & Le Bigot, in press). The first goal of the current study is to determine whether these biases are also observed when a human user interacts with a dialogue system. Verifying this assumption would imply that users are likely to systematically have difficulty remembering part of the information produced by the system. Accordingly, the second goal of this study is to determine how these biases can be attenuated, in particular by manipulating the kind of feedback produced by humans and systems during the interaction.

The remainder of this paper is organised as follows. Section 2 describes literature on human-human and human-system dialogue. The current study, which involved interactions between naïve participants and a simulated dialogue system, is described in Section 3. The results are reported in Section 4 and discussed in Section 5. Section 6 includes directions for future research.

2. Theoretical background: Dialogue memory in human-human interactions

Collaborative dialogue is an activity during which at least two partners interact in order to reach a common goal (Clark, 1992, 1996) and which might involve human partners only or human(s) and automated dialogue systems (see Klein, Feltovich, Bradshaw, & Woods, 2005).

Dialogue partners attempt to adapt to each other by favouring the production of easily understandable utterances not only at the beginning of the interaction, but also during the remainder of the dialogue. For instance, human partners talking about pictures of New York buildings adapt the references they use to designate the buildings depending on whether their partner knows New York well (in which case they might produce the reference “the Empire State Building”) or not (in which case they might produce the reference “the pointy building”) (Isaacs & Clark, 1987; see also Brennan & Clark, 1996; Clark & Wilkes-Gibbs,

1986; Nückles, Winter, Wittwer, Herbert, & Hübner, 2006). In a similar way, users interacting with a dialogue system about everyday objects reuse the same references to these objects as those previously used by the system, as they assume that the system should be capable of understanding them again (Bergmann et al., 2015; Branigan et al., 2011; Iio et al., 2015; for other examples, see also Branigan et al., 2003; Cavedon et al., 2015; Kiesler, 2005; Koulouri et al., 2015; Powers et al., 2005; Suzuki & Katagiri, 2007; Zoltan-Ford, 1991).

To determine what his or her partners are capable of understanding, each partner relies on the common ground, which consists in the knowledge and information that two dialogue partners share and are aware of sharing (in human dialogue) or the information which the user believes to be shared with the system (in human-system dialogue). Part of the common ground consists in the information produced earlier during the current interaction or during past interactions. Precisely, information is added to the common ground through a joint *contribution* process (Clark & Brennan, 1991; Clark & Schaefer, 1989; Clark & Wilkes-Gibbs, 1986; for a generalisation to human-system dialogue, see Brennan & Hulteen, 1995; Cahn & Brennan, 1999). One of the speakers starts by presenting a piece of information. For instance, Speaker A might say: "I would like to go to the cinema on Saturday." during an interaction with Speaker B. The latter then accepts this information, that is, he or she indicates that he or she believes that the information presented was understood well enough for current purposes. Acceptance is more or less explicit: Speaker B might accept the utterance produced by A by repeating it verbatim, by saying "okay" or by nodding his or her head. In any event, once presented and accepted, the information is added to the speakers' common ground (Clark & Brennan, 1991; McInnes & Attwater, 2004); in this example, this would imply that both A and B are aware that they both know that A would like to go to the cinema on Saturday. Either speaker may then resort to it for subsequent adaptation purposes.

The partners' capacity to remember *what* was said previously is therefore a central determinant of successful partner-adaptation.

Importantly, studies on human dialogue suggest that dialogue memory is more or less accurate depending on whether one needs to retrieve initially self- or partner-produced information from memory (Hjelmquist, 1984; Jarvella & Collas, 1974; Knutsen & Le Bigot, 2015; Stafford, Burggraf, & Sharkey, 1987; Stafford & Daly, 1984). For instance, Knutsen and Le Bigot (2015; see also Knutsen et al., in press) have recently shown that the distinction between self- and partner-production at the time of common ground construction directly affects dialogue memory. Indeed, after the end of an interaction, each partner remembers the information that he or she presented him- or herself better than the information presented by his or her partners; information accepted through verbatim repetition is also remembered better than information accepted implicitly (regardless of whether the acceptance was self- or partner-produced). Such memory biases have important consequences for subsequent partner-adaptation, as readily accessible information is more likely to be reused in the remainder of the interaction (Knutsen & Le Bigot, 2012, 2014; Knutsen et al., in press).

To date, these biases have exclusively been investigated in human-human dialogue. However, as mentioned already, similar processes are at play in humans engaged in human-human dialogue and in users engaged in human-system dialogue (e.g., Brennan, 1991; Powers et al., 2005), implying that users should also be subject to presentation and acceptance biases. This could have important consequences for human-system dialogue. Most users engage in this kind of dialogue in order to obtain pieces of information held by the system (e.g., the various stations at which a train calls or the departure time of a plane). If users' dialogue memory for human-system dialogue is subject to a self-presentation bias, this would imply that the information obtained from the system (i.e., system-presented information) would systematically be remembered less well than the information produced by

the user him- or herself (i.e., self-presented information). Furthermore, designers may rely on the fact that speakers tend to reuse words and structures previously mentioned by the system to ensure that users only produce words and structures that the system is capable of understanding (e.g., Koulouri et al., 2015; Zoltan-Ford, 1991). However, if the users' dialogue memory is biased towards remembering self-presented information better, then such convergence might not occur systematically, thus potentially impairing the interaction.

The acceptance bias might have important consequences for human-system dialogue as well. In the presentation-acceptance model, acceptance is more or less explicit (Clark & Brennan, 1991; Clark & Schaefer, 1989). When acceptance involves verbatim repetition of the presented reference, this reference becomes more readily accessible to both speakers (compared to references accepted through other means) (Knutsen & Le Bigot, 2014, 2015; Knutsen et al., in press). Such acceptance effect might be sufficient to attenuate the presentation effect from the point of view of the speaker performing the acceptance. For instance, if a system-presented reference is accepted through verbatim repetition by a user, this reference should benefit from a self-production effect (due to verbatim repetition at the time of acceptance) from the user's point of view, just like self-presented references. This should result in an increase in accessibility in memory of the system-presented reference, thus attenuating the strength of the presentation bias by reducing the difference in accessibility between self- and partner-presented references from the user's point of view. Importantly, there are both pros and cons associated with explicit acceptance in human-system dialogue. The main advantage associated with the user or the system repeating the information presented by the other partner is that it allows this partner to check that the information presented was understood correctly (Cahn & Brennan, 1999; Dybkjaer & Bernsen, 2001; Dybkjær & Bernsen, 2000). Furthermore, system explicit acceptance can increase user satisfaction, especially when the information repeated is important within the context of the

task framework (Stent, Dowding, Gawron, Bratt, & Moore, 1999). However, explicit repetition by the system is sometimes cumbersome and is not well adapted to all users (e.g., to users whose speech is typically well recognised; see Litman & Pan, 1999); in a similar way, explicit repetition by the user might feel unnatural and not always necessary. Furthermore, explicit repetition (by the system and/or by the user) decreases the efficiency of the interaction, as it increases the number of speech turns necessary to complete the task at hand (e.g., Wolters et al., 2009), potentially overloading the users' memory. The results of the current study are discussed below in light of these pros and cons.

In any event, finding presentation and acceptance biases in the current study would be in favour of the idea that similar psychological processes are at play in human-human and human-system dialogue (Brennan, 1991; Powers et al., 2005). It would also shed light on how users represent the content of past interactions with dialogue systems. Such knowledge would help understand how this kind of representation is used for subsequent adaptation purposes, as memory for past interactions is one of the sources of information used by dialogue partners to determine whether or not a piece of information belongs to the common ground (see Clark & Schaefer, 1989).

3. Method

The first goal of the current study was to determine whether the self-presentation bias occurs in human-system dialogue. The second goal was to determine whether the acceptance bias occurs in human-system dialogue as well, and whether it might contribute to reducing the self-presentation bias (i.e., the study sought to determine whether encouraging users to accept references presented by the system through verbatim repetition attenuates the users' tendency

to remember these references less well than self-presented ones). This section details the rationale for this study and the methodology employed.

3.1. Rationale of the experiment, cover story and operational hypotheses

The current experiment involved an interaction about the location of various types of restaurants between a naïve human participant and a simulated dialogue system. Leading human participants to believe that they are interacting with a genuine dialogue system when the system prompts are in fact controlled by a human confederate is usually referred to as the Wizard-of-Oz method (Cohen, Giangola, & Balogh, 2004; Fraser & Gilbert, 1991). This method can be used to study human behaviour in human-system-like interaction situations. Its main advantage is that it offers complete control over the content of the system prompts.

Each participant was led to believe that the experiment was part of a large-scale ergonomics project whose goal was to develop a dialogue system which would help users to locate restaurants in surrounding areas. He or she was told that the system was still in its phase test and that the goal of this experiment was to provide the system with opportunities to learn through genuine phone interaction with a human being. The experimenter then gave the participant a list of restaurants defined in terms of category (e.g., "Italian"), price (e.g., "20 euros") and location (e.g., "covert market") and explained that these were the only restaurants the system was currently capable of locating. The experimenter also explained that the system only understood queries using the syntax "Where is the [category] restaurant that costs [price]?" (e.g., "where is the Italian restaurant that costs 20 euros?") and answers using the syntax "it is next to [location]" (e.g., "it is next to the covert market"). The participant was told that his or her task was to produce queries and replies similar to the queries and replies the final system would have to manage.

The purpose of this experimental setup was to give the participant the opportunity to present references and to accept references presented by the system. The main manipulation concerned the kind of acceptance used by the participant and the system to accept the references presented. In one condition, the references were accepted through verbatim repetition; in the other condition, these were accepted implicitly. At the end of the experiment, the participant performed a memory (recognition) task during which he or she identified the references produced during the interaction with the system. The participant also completed a French translation of the SUS (System Usability Scale) (Brooke, 1996) in order to examine whether the kind of acceptance used affected the participants' perception of the system.

Three hypotheses were tested in this study. All three hypotheses concerned the participants' recognition of the references mentioned during the interaction. The first hypothesis (Presentation Hypothesis) was that self-presented target references are more likely to be recognised than partner-presented (i.e., system-presented) target references. The second hypothesis (Acceptance Hypothesis) was that target references accepted through verbatim repetition are more likely to be recognised during the memory task than target references accepted through another mean. Finally, we have suggested above that the presentation bias should be attenuated for the user when the reference presented by the system is then accepted explicitly by the user. Accordingly, the third hypothesis (Attenuation Hypothesis) was that the presentation bias (i.e., the participants' tendency to remember self-presented target references better and system-presented target references less well) is weaker when target references accepted through verbatim repetition than for target references accepted through another mean.

3.2. Participants

Fifty-two undergraduate students (41 female; mean age 20.2, *SD* = 2.3) took part in the experiment in exchange for partial course credit. All participants were native French speakers. They signed an informed consent form before the beginning of the experiment and were fully debriefed after the end of the experiment. At that point, the experimenter also explained about the simulated dialogue system and made sure that none of the participants had guessed that the system prompts were in fact controlled by a human confederate (no participant had suspected this).

3.3. Apparatus and materials

As specified above, the current study involved interacting with a dialogue system about the location of restaurants in a French town. Lists of restaurants, maps, a Wizard-of-Oz dialogue system and a usability scale were used to this end. These are described in more detail in this section.

3.3.1. "Category-location-price" items and lists

Twenty-four restaurant categories (which corresponded to 24 restaurant categories found on the French website http://www.linternaute.com), 24 prices (going from five euros to 62 euros) and 24 locations (which were randomly selected from the maps used by Knutsen & Le Bigot, 2012, 2014, 2015; what's more, the map used in the current study represented the same area as the map used by Knutsen & Le Bigot, 2012, Experiment 1) were randomly associated to create twenty-four "category-location-price" items included in the lists used by the participants to interact with the simulated dialogue system (see Appendix A for an example). Seven different lists were created for the purpose of the experiment. Each list featured all 24 items in a different (random) order. These seven lists were printed on A4 paper.

Four of these items ("Polish – 7 euros – Church", "Thai – 60 euros – Airport", "Sicilian – 62 euros – Administrative Centre" and "Ethiopian – 5 euros – Gym") were "error items". When these items were mentioned during the experiment, the system systematically committed a mistake (see below for more information about the features of these items); the purpose of these "error items" was to increase the credibility of the cover story used: the participants were led to believe that the system was still in its test phase, so it was likely to commit this kind of mistake.

3.3.2. Maps

A map featuring the locations included in the lists was created for the purpose of the experiment. It featured a total of 40 landmarks, 24 of which were the landmarks featured on the list and the remaining 16 were distractors (just like the 24 target landmarks used in this study, these 16 distractors were randomly taken from the materials used by Knutsen & Le Bigot, 2012, 2014, 2015). Two different versions of the map were created and printed on A4 sheets; the (random) position of each landmark varied across maps, ensuring that the participants' performance on the memory test was not due to the location of the landmarks on the map (see Appendix B for an example of a map). The participants were told that the map made the experiment more naturalistic, as end users would probably have at least some background knowledge of the town in which queries would be conducted. In reality, the aim of this was to make sure that the participants were familiar with the map used during the subsequent memory test, thus ensuring that the participants' performance was not simply due to them being unfamiliar with the map used.

3.3.3. Simulated dialogue system

The dialogue system included four components: a welcome component, an interaction component, a closing component and a help component. Because the experimental manipulation took place during the interaction component, only this component is detailed here; the other three components, whose sole purpose was to make the interaction more realistic, are presented in more detail in Appendix C.

The interaction component consisted in 24 trials, each corresponding to one of the items featured on the list. Each trial was divided into three parts (query, reply, feedback). The target reference was presented as part of the reply and accepted as part of the feedback. Two trial examples are provided in Figure 1.

Insert Figure 1 around here

All of the system prompts were pre-recorded by an artificial female voice in .mp3 format using the Voxygen Expressive Speech technology (http://www.voxygen.fr), a corpus-based concatenative synthesis. The voice used was "Agnès", which is described as "mature, very intelligible, institutional and natural". The pitch of this voice varies between 110Hz and 265Hz. The bitrate of the recordings was 64 kbps and the sampling frequency was 44,100 Hz.

An HTML interface was built containing hyperlinks to the recordings, each of which could be played by simply clicking on the corresponding hyperlink. This interface could be accessed from the confederate's computer, which was in a different room from the experimental room used by the participant. The interactions between the participants and the Wizard of Oz system were recorded using Audacity.

The participant and the confederate interacted over the phone. The Wizard of Oz interface was connected to a land phone; as for the participant, he or she used a mobile phone to call the system.

3.3.4. Usability scale

The SUS (System Usability Scale) is a questionnaire which is easy to administrate and which represents a valid means of quickly assessing the usability of a system (Bangor, Kortum, & Miller, 2008; Brooke, 1996). In this questionnaire, the participant is asked to indicate his or her agreement with ten statements by giving a number between 1 (strongly disagree) and 5 (strongly agree). These statements concern (1) the frequency with which the participant would like to use the system, (2*) its complexity, (3) it ease of use, (4*) whether they thought they would need technical support to use the system, (5) whether they found the various functions of the system well integrated, (6*) whether they thought there was too much inconsistency in the system, (7) whether they thought that most people could learn to use the system quickly, (8*) the system's cumbersomeness, (9) their confidence in using the system and (10*) whether they needed to learn a lot of things before they could start using the system (the asterisk denotes scales which were reversed when calculating each participant's final score, as described below). This scale was translated into French for the purposes of this study.

3.4. Task and procedure

At the beginning of the experiment, the participant was installed in a quiet experimental room and was explained the cover story by the experimenter.

The experiment was divided in three phases. At the beginning of the first phase (dialogue phase), the participant was given a list of 24 restaurants and a map of a French town in which these restaurants were supposedly located. The experimenter then dialled the number used to reach the system on a mobile phone which she then handed over to the participant.

This phase was divided in 24 trials. In half of these (i.e., trials in which the target reference was partner-presented), the participant was required to use the category and price information given in the list to formulate a query and to provide a feedback indicating whether or not the system's reply was correct. For instance, if the participant asked "where is the restaurant that costs 20 euros?", his or her task was then to make sure that the reply produced by the system corresponded to the location reported in the list (i.e., "it is next to the covert market") and to say whether this reply was correct or not. In the other half of trials (i.e., trials in which the target reference was self-presented), the participant was required to use the location information given in the list to answer the query formulated by the system; the system then produced a feedback indicating whether or not the participant's reply was correct. For instance, if the system asked "where is the restaurant that costs 20 euros?", the participant's task was to answer "it is next to the covert market" and to listen to the feedback produced by the system.

The main manipulation concerned the kind of feedback produced by the participants and the simulated system. In the "Acceptance through verbatim repetition" condition, the target reference presented during the preceding speech turn was repeated (e.g., "correct, it is next to the *town hall*"); in other words, the target reference was accepted through verbatim repetition in this condition. In the "Implicit acceptance" condition, the target reference presented during the preceding speech turn was not repeated (e.g., "correct"); in other words, the target reference was accepted implicitly in this condition. The type of feedback produced by the participant and the system was the same (i.e., if the participant was instructed to use verbatim repetition [or implicit acceptance] in his/her feedback, the system also used verbatim repetition [or implicit acceptance]). In both conditions, if the reply produced was incorrect, the partner simply said "incorrect". As mentioned above, the items discussed in four of the trials were "error items": the system systematically committed a mistake in these

trials. This mistake consisted in providing an incorrect location in trials where the query was produced by the participant or in indicating that the reply produced by the participant was incorrect even if this was not the case in trials where the query was produced by the system.

Queries were produced following the order in which the restaurants were featured on the list given to the participant. At the end of each trial, the system produced a transition prompt (either "it's now your turn to produce a query" or "it's now my turn to produce a query"), indicating to the participant which role he or she would play in the following trial. The participant and the system switched roles (i.e., asking a question or answering a question) after each trial.

At the end of the dialogue phase, the participant hung up the mobile phone and gave it back to the experimenter. He or she then performed a distraction task during which he or she counted backwards from 100 in threes for one minute before embarking on the second phase of the experiment (recognition phase). During this phase, the participant was asked to circle all of the landmarks mentioned during the interaction with the system on the map.

Finally, during the third phase (usability assessment phase), the participant completed a usability survey adapted from the SUS. Each of the statements was read out loud by the experimenter; the participant then indicated whether he or she agreed by giving a number between 1 and 5 (1 = strongly disagree, 5 = strongly agree).

The participant was fully debriefed after the end of the experiment. At this point, the experimenter made sure that he or she had not suspected that the system prompts were in fact controlled by a human being.

The experiment was not limited in time and usually lasted less than 20 minutes. A recap of the procedure can be found in Figure 2.

Insert Figure 2 around here

3.5. Experimental design

Two independent variables (IV) were used in this study. The Presentation IV referred to whether the target references had been presented by the participant or by the system. This was a within-participants IV with two modalities (self-presented, partner-presented; note that the modality labels reflect the participant's point of view). The Acceptance IV referred to whether these references were then accepted through verbatim repetition or implicitly by the other partner. This was a between-participants IV with two modalities (accepted through verbatim repetition, accepted implicitly).

3.6. Data coding and dependent variables

The purpose of the current study was to examine whether presentation and acceptance affected the participants' memory for the references to landmarks produced during the interaction. The participants were also asked to rate the usability of the system. Thus, two dependent variables were used in this study: landmark recognition and usability assessment. These two variables are described below.

3.6.1. Landmark recognition

Each landmark which had been referred to during the dialogue phase was coded depending on whether or not it was recognised by the participant during the second phase of the experiment (i.e., whether or not the participant had circled it on the map). This level of coding served as the binary outcome in the main statistical analysis.

3.6.2. Assessment of system usability

A usability score was calculated for each participant following the procedure described by Brooke (1996). For each participant, the score contributions for each item were summed (for items 1 [frequency of use], 3 [ease of use], 5 [integration of the various functions of the system], 7 [how quickly other people could learn to use the system] and 9 [confidence in using the system], the score contribution was the figure given by the participant minus one; for all other items, the score contribution was five minus the figure given by the participant). The sum of scores was then multiplied by 2.5. The figure obtained (one per participant; the SUS score could potentially range from 0 to 100) reflected how usable this participant found the dialogue system. This level of coding served as the numeric outcome in the analysis on usability.

4. Results

The analysis of the data from the recognition phase and the usability assessment phase were divided into three steps. First, a preliminary analysis assessed the false error rate during the recognition phase. Second, an analysis was conducted to determine whether landmark recognition depended on presentation and acceptance during the dialogue phase. This was the main analysis, as it intended to test the three hypotheses (Presentation, Acceptance and Attenuation hypotheses) presented above. Finally, one last analysis was conducted to determine whether Acceptance also affected the participants' perception of the simulated dialogue system.

4.1. Preliminary analysis of false alarm rate

The number of false alarms (i.e., the number of cases in which a participant circled a landmark which had not been referred to during the dialogue phase) was 25, implying that the

false alarm rate was 3% (25 distractors circled out of 832 [16 distractors * 52 participants] = 0.03). This fairly low false alarm rate allows discarding the possibility that the participants simply circled all of the landmarks featured on the map during the recognition phase.

4.2. Analysis 1: The influence of reference Presentation and Acceptance on subsequent landmark recognition

The recognition data were analysed using logistic mixed models. Logistic models in general are used in cases where the outcome variable is binary, which was the case here (a landmark was either recognised or not) (Jaeger, 2008). Logistic mixed models in particular are used in cases where there is more than one measure per participant (in this study, there was one measure per item from the dialogue phase). One of the indicators used in logistic regression is the odds ratio (OR), which compares the odds of two events occurring. For instance, in the current study, the odds of reusing self-presented references were compared with the odds of reusing partner-presented references. If partner-presented references were used as the reference category and that an odds ratio of 2 was found, this would mean that two self-presented references were recognised correctly for each partner-presented reference recognised correctly (see Jaccard, 2001)

As for mixed models, they allow introducing by-participants random intercepts and slopes (respectively accounting for potential variations across participants and for the fact that the participants might differ in their sensitivity to within-participants IVs) as well as by-items random intercepts and slopes (respectively accounting for potential variations across items and for the fact that the items might differ in their sensitivity to within-items IVs) (Barr, 2013; Barr, Levy, Scheepers, & Tily, 2013). The data were analysed using SAS 9.4 (GLIMMIX procedure).

The data from "error items" were removed from the analysis. Indeed, the mistakes committed by the system might have led the participants to pay more attention to the landmarks produced, thus increasing the accessibility in memory of corresponding references. The data from nine trials were also removed from the analysis due to technical issues (these were trials in which a participant failed to follow the instructions, e.g. trials in which the participant asked the system about a restaurant which was not in the list or trials in which the participant did not formulate a query, or answer the system's query, when prompted to do so). These data were not taken into account in the false alarm rate. As a result of this, the number of observations in each cell of the design was unbalanced, which was accounted for in the analysis by applying the Satterthwaite correction (Keselman, Algina, Kowalchuk, & Wolfinger, 1999; Satterthwaite, 1946).

The number of references recognised correctly by the participants is reported in Table 1.

Insert Table 1 around here

4.2.1. Model used and results of the main analysis

The statistical model used to analyse the data included Presentation, Acceptance and the interaction between these two factors as fixed effects and Recognition as the binary outcome. Following Barr et al. (2013), the first analysis conducted included the maximal random effects structure justified by the design (i.e., a random effects structure including by-participants and by-items random intercepts as well as by-participants random slopes corresponding to Presentation and by-items random slopes corresponding to Presentation, Acceptance and the interaction between these two factors). However, this analysis failed to converge due to the fact that the variance associated with some of the random effects was

equal to zero. Following Kiernan, Tao, and Gibbs (2012) these effects were identified (this is done automatically in SAS) and removed from the model; this had no effect on the outcome of the analysis. Specifically, the random effects causing convergence issues in the analysis were the by-participants random slopes corresponding to Presentation and the by-participant random slopes corresponding to the interaction between Presentation and Acceptance. These were removed from the final analysis.

The covariance parameter estimates and the results of the final analysis are reported in Appendix D. As expected, the effect of Presentation on Recognition was significant, $F(1, 51) = 9.88$, $p = .003$: participants were more likely to recognise self-presented references than partner-presented references, $OR = 1.62$, $CI_{.95} = 1.19, 2.20$. No significant effect of Acceptance on Recognition was found, $F(1, 36) < .001$, $p = .985$. However, the interaction between Presentation and Acceptance was significant, $F(1, 51) = 5.11$, $p = .028$. The difference between self- and partner-presented references was weaker when they had been accepted through verbatim repetition than when they had been accepted implicitly, as predicted, $b = - 0.69$, $p = .028$ (see Figure 3).

Insert Figure 3 around here

4.2.2. Pairwise comparisons

Additional pairwise comparisons (whose *p*-values were corrected using Sequential Bonferroni) revealed that the difference between self- and partner-presented references was significant for references accepted implicitly, $p = .002$, but that this difference was non-significant for references accepted through verbatim repetition, $p = 1.00$.

Importantly, Figure 2 might give the impression that the interaction was due to implicit acceptance increasing the accessibility in memory of self-presented references and

decreasing the accessibility of partner-presented references. However, the difference between self-presented references accepted through verbatim repetition and self-presented references accepted implicitly was not significant, $p = 1.00$, and the difference between partner-presented references accepted through verbatim repetition and partner-presented references accepted implicitly was not significant either, $p = 1.00$. Thus, this explanation was not supported by the data.

4.3. Analysis 2: Influence of Acceptance on usability

The purpose of the second analysis was to determine whether the kind of acceptance used by the participants and the system affected the participants' perception of the system. The SUS data were analysed using a one-way between-participants ANOVA including Acceptance as the IV and the participants' SUS scores as the numeric dependent variable. This analysis was conducted in SPSS 21.

Two preliminary analyses revealed that the normality assumption (Shapiro-Wilk test; verbatim acceptance condition: $p = .320$; implicit acceptance condition; $p = .113$) and the homogeneity of variance assumption (Levene's test, $p = .549$, based on means) were met. However, the main analysis revealed no significant effect of Acceptance on the participants' SUS scores, $F < 1$, $p = .749$ (see Table 2).

Insert Table 2 around here

5. Discussion

The purpose of this study was to examine users' dialogue memory for human-system interactions. When two human partners interact, they remember better self-presented

information and/or information accepted through verbatim repetition (Knutsen & Le Bigot, 2015; Knutsen et al., in press). Is this also the case when a human user interacts with a system? To answer this question, the current study examined participants' recognition of references to landmarks presented and accepted during an interaction with a simulated dialogue system.

First, the results confirmed that the participants remembered the references they had presented themselves better than the references that had been presented by the system, confirming that the self-presentation bias also occurs in human-system dialogue (Knutsen & Le Bigot, 2015; Knutsen et al., in press). This is also in line with the idea that similar psychological processes are at play when human beings interact with other humans or with automated dialogue systems (Brennan, 1991; Powers et al., 2005).

The self-presentation bias could potentially hinder human-system interaction. Users generally engage in interactions with dialogue systems in order to obtain information they currently do not have access to (e.g., a user might start interacting with a system in order to obtain the departure time of a flight). However, the self-presentation bias implies that the information presented by the system might be difficult to subsequently retrieve from memory, or at least more difficult to retrieve from memory than self-presented information. In a similar way, if a user produces an erroneous piece of information during the interaction and the system corrects him or her, the users might remember the incorrect information better than the correct information. Of course, other factors than presentation could affect the user's memory in this case as well – for instance, the system correcting him or her could lead the user to pay more attention to what the system says. Nonetheless, if the user believes that his or her flight is at 10 o'clock and the system informs him or her that the flight is in fact at 11 o'clock, the erroneous piece of information might remain particularly accessible in the user's memory due to self-presentation, potentially making this user more likely to arrive at the

airport at the wrong time. However, the current study also provides strong evidence that the self-presentation bias can be attenuated through acceptance in human-system dialogue. Specifically, the effect of acceptance was not statistically significant, failing to replicate a finding reported in several studies on human-human dialogue (Knutsen & Le Bigot, 2012, 2014, 2015; Knutsen et al., in press) and preventing us from concluding that the acceptance bias also occurs in human-system dialogue. This is potentially due to the fact that in previous human-human dialogue studies, dialogue partners were free to choose how they accepted the information they were presented (i.e., they could choose between accepting a piece of information through verbatim repetition, implicitly, or through any other mean). The fact that the participants were forced to use verbatim repetition or implicit acceptance in the current study might have led them to pay less attention to the feedbacks produced. Nonetheless, the expected Presentation x Acceptance interaction was statistically significant in this study, in line with the Attenuation Hypothesis (i.e., in line with the idea that the self-presentation bias is reduced when presented references are accepted through verbatim repetition in human-system dialogue). Dialogue system designers can therefore manipulate the nature of the feedback required by the system after the presentation of a piece of information to improve communication. This could be achieved by having the system explicitly ask the user to repeat the information presented. For instance, the system might say: "Your flight is scheduled at 11.45am. Could you please repeat this information?". Another possibility would be to train the users before they use the system (e.g., using a tutorial) to spontaneously produce such explicit feedback.

What's more, requiring explicit feedback from users may have additional benefits which are not directly related to information accessibility in memory. For instance, accepting information through verbatim repetition gives the system an opportunity to check the user's comprehension of the information presented. If the system produces the message: "Your

flight is at 11 o'clock" and that the user repeats: "At 10 o'clock", the system can infer that it was not understood correctly and point this out to the user (for similar ideas, see Cahn & Brennan, 1999; Dybkjaer & Bernsen, 2001; Dybkjær & Bernsen, 2000). Such miscomprehension could have gone unnoticed if the user had produced a less explicit feedback.

Importantly, we do not suggest that users should systematically be required to repeat system-presented information. Indeed, doing so could negatively affect the users' representation of the system for at least two reasons. First, it would systematically increase the number of speech turns necessary to perform the task underlying the interaction, thus potentially overloading the users with information (Wolters et al., 2009); second, the users could perceive such repetitions as useless in situations in which their next actions make it clear that they have understood the system correctly or in which potential user comprehension mistakes would be harmless (see Bernsen, Dybkjær, & Dybkjær, 1994). In the current study, no evidence was found that asking the participants to systematically provide explicit feedback affected their perception of the system's usability (i.e., there was no evidence that interacting with a system which systematically repeated the references to landmarks presented by the participant caused the latter to perceive the system more negatively). However, this lack of a significant difference might have been due to the experimental nature of the interaction (e.g., the participants in this study were not in a particular rush to complete the task, nor did they perceive the task as having high stakes). The users' perception of this kind of system might be quite different in real-life human-system interactions, where it should depend on the users' individual goals and constraints (e.g., it could depend on whether or not the users need to perform the task quickly). In this context, our recommendation is to encourage users to repeat important information only – that is, information whose memorization by the users is central to the task at hand.

6. Limitations, plans for future research and conclusion

The current study presents at least four limitations. First, the experiment conducted involved a simulated system rather than a real automated dialogue system. Although the Wizard of Oz technique can be used to examine the influence of participants' beliefs about the nature of their partner on their behaviour (Fraser & Gilbert, 1991), the next step of this work will consist in attempting to generalise the current results to genuine human-system dialogue. For instance, the comprehension capacities of a real dialogue system may differ greatly from those of a human confederate, potentially leading to more system comprehension mistakes which might cause users to pay more attention to the information and feedback produced by the system. Replicating the current findings using a genuine automated dialogue system will constitute even stronger evidence (1) that users are subject to memory biases in human-system dialogue and (2) that acceptance strategies may be used to attenuate these biases. Second, in this study, the participants were explicitly instructed to use either verbatim or implicit acceptance. In real-life interactions, however, speakers tend to use various kinds of acceptance strategies (Clark & Brennan, 1991; McInnes & Attwater, 2004), and the reasons which lead speakers to favour one kind of acceptance over another have seldom been investigated. Further research is needed to identify the determinants of information acceptance in order to determine how to lead users to favour verbatim acceptance when important information is presented by the system. Third, the participants in this study were mainly female psychology students. This raises the question of the generalisability of the results to other populations, as male participants and/or non-student participants could have behaved differently in the same experimental setting. Potential changes in the results could be due to variations in expertise with dialogue systems, for instance. Thus, future research should attempt to generalise the current results to a demographically broader sample. Finally,

in this experiment, the SUS questionnaire was administered by the experimenter, who read the questions out loud to the participant. This could have biased the participants' responses towards a more positive assessment of the system. To overcome this limitation, this kind of questionnaire should be administered in a different way (e.g., giving the participant the opportunity to complete it him- or herself) in future studies.

Despite these limitations, the results reported here confirm that presentation and acceptance play an important role in human-system dialogue. This raises a number of theoretical questions. In particular, research on human-human dialogue suggests that presentation and acceptance affect not only reference accessibility in memory after the end of an interaction, but also reuse throughout that interaction (self-presented and/or references accepted through verbatim repetition are more likely to be reused than any other reference; Knutsen & Le Bigot, 2012, 2014, 2015; Knutsen et al., in press). Future research should seek to determine whether this is also the case in human-system dialogue. This would contribute to a better understanding of how humans manage dialogue as they interact with a dialogue system. In addition, recall that dialogue memory plays a central role in subsequent partner-adaptation in human-human dialogue, as speakers routinely rely on what was said previously to adapt to each other (Brennan & Clark, 1996). If the current results extend to reuse, this would suggest that adaptation could be systematically biased towards the production of initially self-presented content. This could have negative consequences for the interaction if the system relies on linguistic convergence to ensure that users favour the production of words and structures that the system is capable of understanding (Koulouri et al., 2015; Zoltan-Ford, 1991).

To conclude, the current study suggests that human-system dialogue is subject to biases similar to those found in human-human dialogue (i.e., the self-presentation bias). This is in line with the idea that similar psychological processes are at play in both kinds of

dialogue. What's more, in human-system dialogue, the self-presentation bias can be attenuated by manipulating the kind of feedback produced by the user, potentially improving communication both directly, by increasing the accessibility in memory of partner-presented information, and indirectly, by providing the system with a means to verify that the information provided to the user was understood correctly.

Acknowledgements

This study was conducted as part of the first author's PhD, which was supported by the Direction Générale de l'Armement (DGA) and Région Poitou-Charentes. The authors would like to thank Voxygen for providing the voice synthesis technology used. The authors would also like to thank Yves Almécija for his technical assistance in setting up the Wizard of Oz system used in this study.

References

Asri, L. El, Lemmonier, R., Laroche, R., Pietquin, O., & Khouzaimi, H. (2014). NASTIA: Negotiating Appointment Setting Interface. In N. Calzorali, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, … S. Piperidis (Eds.), *Proceedings of the 9th International Conference on Language Resources and Evaluation* (pp. 266–271). Paris, France: European Languages Resources Association.

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the System Usability Scale. *International Journal of Human-Computer Interaction*, *24*, 574–594. http://doi.org/10.1080/10447310802205776

Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in Psychology*, *4*. http://doi.org/10.3389/fpsyg.2013.00328

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–278. http://doi.org/10.1016/j.jml.2012.11.001

Barrett, J., & Jiang, Y. (2012). *Apple iPhone Siri users* (Market Report). Dallas, Texas: Parks Associates.

Bergmann, K., Branigan, H. P., & Kopp, S. (2015). Exploring the alignment space - lexical and gestural alignment with real and virtual humans. *Frontiers in ICT*. http://doi.org/10.3389/fict.2015.00007

Bernsen, N. O., Dybkjær, L., & Dybkjær, H. (1994). A dedicated task-oriented dialogue theory in support of spoken language dialogue systems design. In *Proceedings of ICSLP ’94*.

Branigan, H. P., Pickering, M. J., Pearson, J., & McLean, J. F. (2010). Linguistic alignment between people and computers. *Journal of Pragmatics*, *42*, 2355–2368. http://doi.org/10.1016/j.pragma.2009.12.012

Branigan, H. P., Pickering, M. J., Pearson, J., McLean, J. F., & Brown, A. (2011). The role of beliefs in lexical alignment: Evidence from dialogs with humans and computers. *Cognition, 121,* 41–57. http://doi.org/10.1016/j.cognition.2011.05.011

Branigan, H. P., Pickering, M. J., Pearson, J., McLean, J. F., & Nass, C. I. (2003). Syntactic alignment between computers and people: The role of belief about mental states. In *Proceedings of the 25th Annual Meeting of the Cognitive Science Society* (pp. 186–191). Austin, TX: Cognitive Science Society.

Brennan, S. E. (1991). Conversation with and through computers. *User Modeling and User-Adapted Interaction, 1,* 67–86. http://doi.org/10.1007/BF00158952

Brennan, S. E. (1996). Lexical entrainment in spontaneous dialog. In *Proceedings of the 1996 International Symposium on Spoken Dialogue (ISSD-96)* (pp. 41–44). Tokyo, Japan: The Acoustical Society of Japan.

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 1482–1493. http://doi.org/10.1037/0278-7393.22.6.1482

Brennan, S. E., & Hulteen, E. A. (1995). Interaction and feedback in a spoken language system: A theoretical framework. *Knowledge-Based Systems, 8,* 143–151. http://doi.org/10.1016/0950-7051(95)98376-H

Brooke, J. (1996). SUS: A "quick and dirty" usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & A. L. McClelland (Eds.), *Usability evaluation in industry*. London, England: Taylor and Francis.

Cahn, J. E., & Brennan, S. E. (1999). A psychological model of grounding and repair in dialog. In *Proceedings of the AAAI Fall Symposium on Psychological Models of Communication in Collaborative Systems* (pp. 25–33). North Falmouth, MA: American Association for Artificial Intelligence.

Cavedon, L., Kroos, C., Herath, D. C., Burnham, D. K., Bishop, L., Leung, Y., & Stevens, C. J. (2015). "C'mon dude!": Users adapt their behaviour to a robotic agent with an attention model. *International Journal of Human-Computer Studies, 80*, 14–23. http://doi.org/10.1016/j.ijhcs.2015.02.012

Clark, H. H. (1992). *Arenas of language use*. Chicago, IL: University of Chicago Press.

Clark, H. H. (1996). *Using language*. Cambridge, MA: Cambridge University Press.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 127–149). Washington, DC: American Psychological Association.

Clark, H. H., & Schaefer, E. F. (1989). Contributing to discourse. *Cognitive Science, 13*, 259–294. http://doi.org/10.1207/s15516709cog1302_7

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition, 22*, 1–39. http://doi.org/10.1016/0010-0277(86)90010-7

Cohen, M. H., Giangola, J. P., & Balogh, J. (2004). *Voice user interface design*. Boston, MA: Addison-Wesley.

Dybkjær, L., & Bernsen, N. O. (2000). Usability issues in spoken dialogue systems. *Natural Language Engineering*, *6*, 243–271. http://doi.org/S1351324900002461

Dybkjaer, L., & Bernsen, N. O. (2001). Usability evaluation in spoken language dialogue systems. In P. Paroubek & D. G. Novick (Eds.), *Proceedings of the Workshop on Evaluation for Language and Dialogue Systems* (pp. 9–18). Toulouse, France: Morgan Kaufman.

Fraser, N. M., & Gilbert, G. N. (1991). Simulating speech systems. *Computer Speech & Language, 5*, 81–99. http://doi.org/10.1016/0885-2308(91)90019-M

Gibbs, R. W. (1986). Comprehension and memory for nonliteral utterances: The problem of sarcastic indirect requests. *Acta Psychologica, 62*, 41–57. http://doi.org/10.1016/0001-6918(86)90004-1

Grudin, J. (2005). Three faces of human-computer interaction. *IEEE Annals of the History of Computing, 27*, 46–62. http://doi.org/10.1109/MAHC.2005.67

Hjelmquist, E. (1984). Memory for conversations. *Discourse Processes, 7*, 321–336. http://doi.org/10.1080/01638538409544595

Iio, T., Shiomi, M., Shinozawa, K., Shimohara, K., Miki, M., & Hagita, N. (2015). Lexical entrainment in human robot interaction: Do humans use their vocabulary to robots? *International Journal of Social Robotics, 7*, 253–263. http://doi.org/10.1007/s12369-014-0255-x

Jaccard, J. (2001). *Interaction effects in logistic regression*. Thousand Oaks, CA: Sage.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*, 434–446. http://doi.org/10.1016/j.jml.2007.11.007

Jarvella, R. J., & Collas, J. G. (1974). Memory for the intentions of sentences. *Memory & Cognition, 2*, 185–188. http://doi.org/10.3758/BF03197513

Johnstone, A., Berry, U., Nguyen, T., & Asper, A. (1995). There was a long pause: Influencing turn-taking behaviour in human-human and human-computer spoken dialogues. *International Journal of Human-Computer Studies*, *42*, 383–411. http://doi.org/10.1006/ijhc.1995.1018

Keenan, J. M., MacWhinney, B., & Mayhew, D. (1977). Pragmatics in memory: A study of natural conversation. *Journal of Verbal Learning and Verbal Behavior, 16*, 549–560. http://doi.org/10.1016/S0022-5371(77)80018-2

Keselman, H. J., Algina, J., Kowalchuk, R. K., & Wolfinger, R. D. (1999). The analysis of repeated measurements: A comparison of mixed-model satterthwaite F tests and a nonpooled adjusted degrees of freedom multivariate test. *Communications in Statistics - Theory and Methods*, *28*, 2967–2999. http://doi.org/10.1080/03610929908832460

Kiernan, K., Tao, J., & Gibbs, P. (2012). Tips and strategies for mixed modelling with SAS/STAT procedures. Presented at the 2012 SAS Global Forum, Orlando, FL.

Kiesler, S. (2005). Fostering common ground in human-robot interaction. In *Proceedings of the IEEE International Workshop on Robot and Human Interactive Communication* (pp. 729–734). New York, NY: Institute of Electric and Electronics Engineers. http://doi.org/10.1109/ROMAN.2005.1513866

Klein, G., Feltovich, P. J., Bradshaw, J. M., & Woods, D. D. (2005). Common ground and coordination in joint activity. In W. B. Rouse & K. R. Boff (Eds.), *Organizational simulation* (pp. 139–184). Hoboken, NJ: John Wiley & Sons.

Knutsen, D., & Le Bigot, L. (2012). Managing dialogue: How information availability affects collaborative reference production. *Journal of Memory and Language*, *67*, 326–341. http://doi.org/10.1016/j.jml.2012.06.001

Knutsen, D., & Le Bigot, L. (2014). Capturing egocentric biases in reference reuse during collaborative dialogue. *Psychonomic Bulletin & Review*, *21*, 1590–1599. http://doi.org/10.3758/s13423-014-0620-7

Knutsen, D., & Le Bigot, L. (2015). The influence of reference acceptance and reuse on conversational memory traces. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*, 574–585. http://doi.org/10.1037/xlm0000036

Knutsen, D., Ros, C., & Le Bigot, L. (in press). Generating references in naturalistic face-to-face and phone mediated dialogue settings. *Topics in Cognitive Science*.

Koulouri, T., Lauria, S., & Macredie, R. D. (2015). Do (and say) as I say: Linguistic adaptation in human-computer dialogs. *Human-Computer Interaction*. http://doi.org/10.1080/07370024.2014.934180

Le Bigot, L., Caroux, L., Ros, C., Lacroix, A., & Botherel, V. (2013). Investigating memory constraints on recall of options in interactive voice response system messages. *Behaviour & Information Technology, 32*, 106–116. http://doi.org/10.1080/0144929X.2011.563800

Litman, D. J., & Pan, S. (1999). Empirically evaluating an adaptable spoken dialogue system. In *Proceedings of the 7th International Conference on User Modelling (UM'99)*. Secaucus, NJ: Springer-Verlag New York.

McInnes, F., & Attwater, D. (2004). Turn-taking and grounding in spoken telephone number transfers. *Speech Communication*, *43*, 205–223. http://doi.org/10.1016/j.specom.2004.04.001

Pasupathi, M., & Hoyt, T. (2010). Silence and the shaping of memory: How distracted listeners affect speakers' subsequent recall of a computer game experience. *Memory, 18*, 159–169. http://doi.org/10.1080/09658210902992917

Pieraccini, R., & Huerta, J. M. (2008). Where do we go from here? Research and commercial spoken dialogue systems. In L. Dybkjær & W. Minker (Eds.), *Recents Trends in Discourse and Dialogue* (pp. 1–24). Dordrecht, The Netherlands: Springer.

Powers, A., Kramer, A., Lim, S., Kuo, J., Lee, S., & Kiesler, S. (2005). Common ground in dialogue with a gendered humanoid robot. In *Proceedings of the IEEE International Workshop on Robot and Human Interaction*. New York, NY: Institute of Electric and Electronics Engineers.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics*, *2*, 110–114. http://doi.org/10.2307/3002019

Stafford, L., Burggraf, C. S., & Sharkey, W. F. (1987). Conversational memory: The effects of time, recall, mode, and memory expectancies on remembrances of natural conversations. *Human Communication Research, 14*, 203–229. http://doi.org/10.1111/j.1468-2958.1987.tb00127.x

Stafford, L., & Daly, J. A. (1984). Conversational memory: The effects of recall mode and memory expectancies on remembrances of natural conversations. *Human Communication Research, 10*, 379–402.

Stent, A., Dowding, J., Gawron, J. M., Bratt, E. O., & Moore, R. (1999). The CommandTalk Spoken Dialogue System. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*. Morgan Kaufmann.

Suzuki, N., & Katagiri, Y. (2007). Prosodic alignment in human–computer interaction. *Connection Science, 19*, 131–141. http://doi.org/10.1080/09540090701369125

van Lierop, K., Goudbeek, M., & Krahmer, E. (2012). Conceptual alignment in reference with artificial and human dialogue partners. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Meeting of the Cognitive Science Society* (pp. 1066–1071). Austin, TX: Cognitive Science Society.

Wolters, M., Georgila, K., Moore, J. D., Logie, R. H., MacPherson, S. E., & Watson, M. (2009). Reducing working memory load in spoken dialogue systems. *Interacting with Computers, 21*, 276–287. http://doi.org/10.1016/j.intcom.2009.05.009

Yoon, S. O., & Brown-Schmidt, S. (2012). Lexical differentiation in language production and comprehension. *Journal of Memory and Language, 69*, 397–416. http://doi.org/10.1016/j.jml.2013.05.005

Zhou, L. (2007). Natural language interface for information management on mobile devices. *Behaviour & Information Technology*, *26*, 197–207. http://doi.org/10.1080/01449290500402726

Zoltan-Ford, E. (1991). How to get people to say and type what computers can understand. *International Journal of Man-Machine Studies*, *34*, 527–547. http://doi.org/10.1016/0020-7373(91)90034-5

Appendix A: List example

This is an example of a list used by the participants to interact with the simulated dialogue system. Note that this is an English translation: the original lists used in the study were in French.

Category	Price	Location
French	22 euros	Regional education office
Tibetan	20 euros	Post office
Chinese	35 euros	Prefecture
Lebanese	52 euros	Museum
Cretan	17 euros	Cinema
Korean	10 euros	Town centre
African	37 euros	Roman arena
Spanish	55 euros	Train station
Portuguese	12 euros	Hospital
American	15 euros	Covert market
Japanese	32 euros	Congress centre
Sicilian	62 euros	Administrative centre
Creole	25 euros	Cathedral
Ethiopian	5 euros	Gym
Belgian	40 euros	Park
Moroccan	42 euros	Business school
Greek	50 euros	Music academy
Caribbean	27 euros	Theatre
Polish	7 euros	Church
Russian	47 euros	Library
Indian	57 euros	University
Mexican	30 euros	General council
Thai	60 euros	Airport
Italian	45 euros	Town hall

Appendix B: Example of a map used in the experiment

Appendix C: Structure of the Wizard of Oz system

During the first part of the interaction (welcome component), the participant listened to a welcome message providing him or her with brief instructions as to how to interact with the system. The closing prompt of this component was a message indicating who (the participant or the system) should produce the first query.

The main experimental manipulation was included in the interaction component. In this component, the participant and the system took it in turns to produce queries about the location of restaurants on the map given to the participant. An additional hypertext link was used by the human confederate to request additional information in cases where the participant's query only included one piece of information (i.e., the participant's query only included the category of the restaurant or its price). A second additional link was used by the confederate in cases where she did not understand the query produced by the participant (because of the quality of the call, or because the participant did not follow the instructions given by the experimenter). At the end of each trial, a system prompt informed the participant either that it was his or her turn to produce a query or that it was the system's turn to produce a query.

The closing component ended the interaction after the 24 trials had been completed by the participant.

Finally, the help component was used in cases where the participant mentioned that he or she needed help interacting with the system. It is noteworthy that this component was hardly ever used by the participants, who usually asked their questions directly to the experimenter present in the experimental room.

The structure of the system is presented in detail in Table A1.

Table A1

Detail of the Structure of the Wizard of Oz System (P = Participant, S = System)

Component	Content (English translation)
Welcome	Welcome to the “Restos par téléphone" platform. This platform allows you to interact with the restaurant search system. To produce a query, you can speak normally. For instance, at any time, you can obtain help by saying “I want help”. - *Trials in which the first query was produced by the system*: It’s now my turn to produce a query. - *Trials in which the first query was produced by the participant*: It’s now your turn to produce a query.
Interaction (x 24)	<u>Verbatim acceptance condition</u> *Query produced by the system :* - S: Where is the Chinese restaurant that costs 35 euros? - P: It’s next to the prefecture. - S: Correct, it’s next to the prefecture. *Query produced by the participant:* - P: Where is the Chinese restaurant that costs 35 euros? - S: It’s next to the prefecture. - P: Correct, it’s next to the prefecture. <u>Implicit acceptance condition</u> *Query produced by the system :* - S: Where is the Chinese restaurant that costs 35 euros? - P: It’s next to the prefecture. - S: Correct. - *Query produced by the participant:* - P: Where is the Chinese restaurant that costs 35 euros? - S: It’s next to the prefecture. - P: Correct. *Cases in which an incomplete query was produced by the participant:* - You are looking for a Chinese restaurant. Can you specify the price? - You are looking for a restaurant that costs 35 euros. Can you specify the category? *Cases where a problem occurs:* I didn’t understand what you said, please repeat. *Transition to a trial where the query was produced by the system:* It’s now my turn to produce a query. *Transition to a trial where the query was produced by the participant:* It’s now your turn to produce a query.
End	Thank you for using “Restos par téléphone”, see you soon.
Help	The “Restos par téléphone” test platform allows you to interact with the system which will eventually allow users to find restaurants in their town on the basis of different criteria. The aim of this platform is to reinforce the system’s database. To interact with the system, you can speak normally.

Appendix D: Results of the mixed model analysis (final model)

Covariance parameter estimates			
Covariance parameter	Subject	Estimate	Standard error
Intercept	Participant	0.31	0.14
Presentation	Participant	0.16	0.12
Intercept	Item	0.14	0.08
Acceptance	Item	0.02	0.06

Fixed effects				
Effect	Estimate	Standard error	DF	*p*
Intercept	-0.58	0.21	78.48	0.007
Presentation: self	0.83	0.22	51.57	< .001
Presentation: partner	0			
Acceptance: verbatim	0.34	0.27	68.92	0.211
Acceptance: implicit	0			
Interaction: Verbatim – self	-0.69	0.31	51.09	0.028
Interaction: Verbatim - other	0			
Interaction: Implicit - self	0			
Interaction: Implicit – other	0			

Test of fixed effects				
Effect	Num DF	Den DF	*F*	*p*
Presentation	1	51.15	9.88	0.003
Acceptance	1	35.74	< .001	0.985
Interaction	1	51.09	5.11	0.028

Figure captions

Figure 1. Trial examples. The target reference is italicized in both trials. In the first example (upper panel), the target reference is presented by the participant and accepted through verbatim repetition by the system. In the second example (lower panel), the target reference is presented by the system and accepted through verbatim repetition by the participant.

Figure 2. Recap of the procedure used in the study. The arrow represents the time course of the experiment.

Figure 3. Proportion of references recognised as a function of Presentation and Acceptance. These proportions were obtained by dividing the number of landmarks recognised in each cell of the design by the total number of measures in each cell of the design (self-presented, verbatim: 258; self-presented, implicit: 260; partner-presented, verbatim: 256, partner-presented, implicit: 257).

Figure 1

Partner	Prompt type	Content
System	Query	Where is the Italian restaurant which costs 45 euros?
Participant	Reply	It is next to *the town hall*.
System	Feedback	Correct, it is next to *the town hall*.

Partner	Prompt type	Content
Participant	Query	Where is the French restaurant which costs 22 euros?
System	Reply	It is next to *the regional education office*.
Participant	Feedback	Correct, it is next to *the regional education office*.

Figure 2

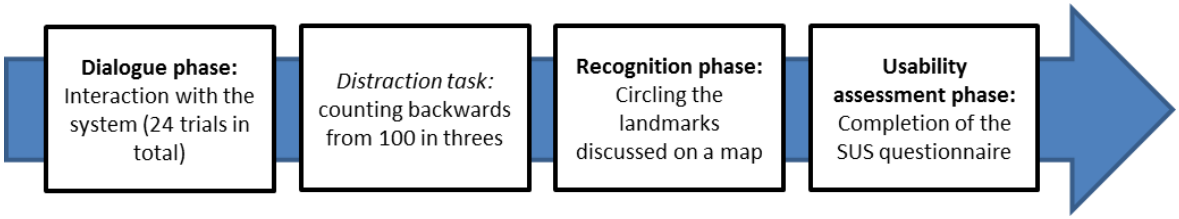
Dialogue phase: Interaction with the system (24 trials in total)
Distraction task: counting backwards from 100 in threes
Recognition phase: Circling the landmarks discussed on a map
Usability assessment phase: Completion of the SUS questionnaire

Figure 3

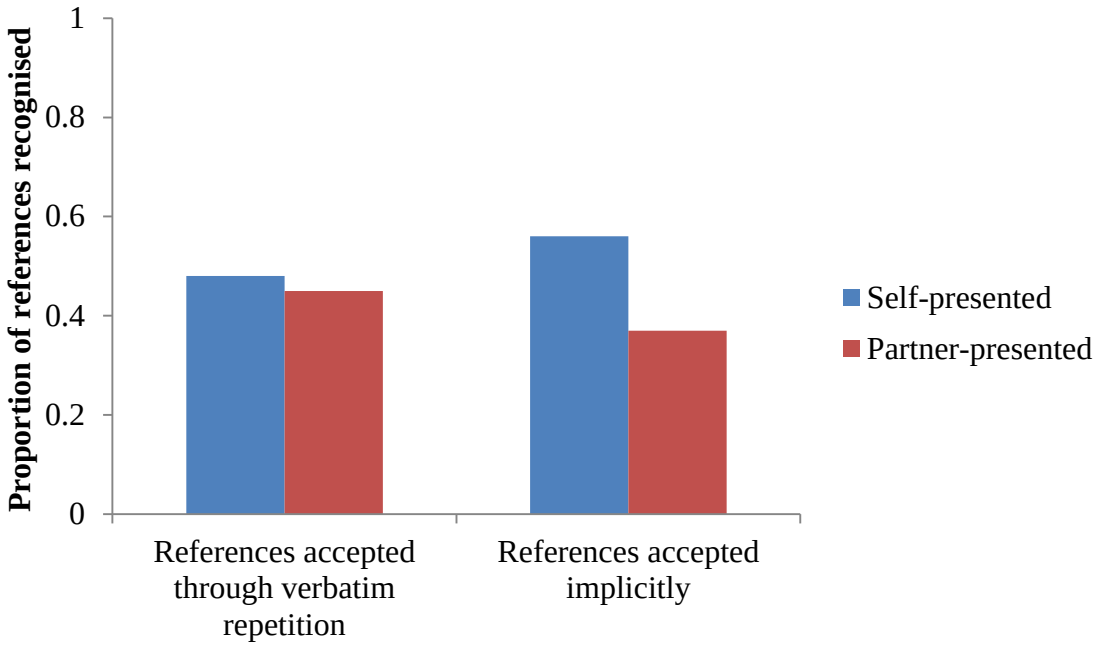
Proportion of references recognised
1
0.8
0.6
0.4
0.2
0
References accepted through verbatim repetition
References accepted implicitly
Self-presented
Partner-presented

Table 1

Number of Landmarks Recognised during the Second Phase of the Experiment (Recognition Phase) as a Function of Presentation and Acceptance

	Self-presented references	Partner-presented references	**Total**
Accepted through verbatim repetition	123	114	**237**
Accepted implicitly	145	95	**240**
Total	**268**	**209**	**477**

Note. In this table, the "total" figure (i.e., 477) is the total number of landmarks correctly recognised by all of the participants during the recognition phase, regardless of who had initially presented the corresponding references and of how these were accepted.

Table 2

Average SUS Scores as a function of Acceptance

	Verbatim acceptance	Implicit acceptance
Average SUS score	73.58 (14.53)	75.06 (13.31)

Note. Standard deviations are reported in brackets.