Do (and Say) as I Say: Linguistic Adaptation in Human-Computer Dialogues

THEODORA KOULOURI, STANISLAO LAURIA AND ROBERT D. MACREDIE
Department of Computer Science, Brunel University, UK

Abstract

There is strong research evidence showing that people naturally align to each other's vocabulary, sentence structure and acoustic features in dialogue, yet little is known about how the alignment mechanism operates in the interaction between users and computer systems let alone how it may be exploited to improve the efficiency of the interaction. This paper provides an account of lexical alignment in human-computer dialogues, based on empirical data collected in a simulated human-computer interaction scenario. The results indicate that alignment is present, resulting in the gradual reduction and stabilisation of the vocabulary-in-use, and that it is also reciprocal. Further, the results suggest that when system and user errors occur, the development of alignment is temporarily disrupted and users tend to introduce novel words to the dialogue. The results also indicate that alignment in human-computer interaction may have a strong strategic component, and is used as a resource to compensate for less optimal (visually impoverished) interaction conditions. Moreover, lower alignment is associated with less successful interaction, as measured by user perceptions. The paper distils the results of the study into design recommendations for human-computer dialogue systems and uses them to inform a model of dialogue management that supports and exploits alignment through mechanisms for in-use adaptation of the system's grammar and lexicon.

1 Introduction

There has been significant and sustained research over the last three decades into the design of natural language user interfaces, embedded in dialogue systems, robots and embodied conversational agents, to support goal-oriented use of computer systems. Despite widespread predictions of success, these systems have yet to enable effective, efficient and natural interactions with the user. This failure has been at least partly attributed to insufficient understanding about how users will address the system or, indeed, what people really do when they communicate. Similarly, relatively little is known about the design and nature of the computer as an interlocutor itself (Porzel, 2006). Therefore, insights derived from empirical studies of goal-oriented human communication have the potential to be of immediate relevance for the design of natural language user interfaces to computer systems.

Empirical models of human communication on which we can draw in understanding how to model and inform the design of user-system interaction emphasise that language is dynamic, adaptable to the context of use and emerges as a function of inter-individual processes (Clark, 1996; Pickering & Garrod, 2004). In particular, it is well-established that speakers adapt to the perceived needs and abilities of the addressee. For instance, an individual will speak in different ways to a young child, a colleague or someone from a different country. However, in the context of human-computer communication, forming assumptions about what a system can do and understand is problematic for most people. In turn, forming assumptions about how users will ‘talk to’ the system is also likely to be problematic for system developers. The potential for variability in how users will communicate with a system is enormous and has been dubbed ‘The Vocabulary Problem’. The extent of the problem was measured in the well-known study by Furnas, Landauer, Gomez, & Dumais et al. (1987), in which participants were asked to name objects for a computer to understand in five scenarios. The probability of two people using the same word to refer to an object ranged from 7% to 18%, indicating that the limited vocabulary of a system is unlikely to match the one utilised by the user. For instance, even in the restricted domain of route instructions, there are myriad ways to formulate the same command; the route instruction “take the second turn on the right” is pragmatically identical to “go straight ahead until you pass a junction; do not take this turn, go straight on until there is another junction on your right. Turn there”. This highlights the ability to predict and constrain user input as a key factor in the success of the system, in terms of enabling efficient and natural user-system interaction. Moreover, the content and structure of communication is largely dictated by the affordances and constraints of the interaction situation (Clark & Brennan, 1991). For instance, when the interlocutors are collocated and share visual space, utterances such as “turn here” are highly more likely than any of the aforementioned instructions. However, if the interaction setting precludes visibility or co-temporality, elaborate instructions like the previous ones are necessary to achieve the same level of understanding. Taken together, it is important to consider natural communication mechanisms and how they are influenced by the interaction situation when designing systems.

It has been observed that dialogue is largely repetitive; that is, speakers in dyads progressively use the same expressions. This natural phenomenon has been referred to as ‘adaptation’ (Brown & Dell, 1987), ‘entrainment’ (Brennan and Clark, 1991), ‘accommodation’ (Giles, Coupland, & Coupland, 1991), ‘convergence’ (Brennan, 1996) and ‘alignment’ (Pickering & Garrod, 2004). This paper will adopt the term ‘alignment’ as it is part of a complete framework of language use, the Interactive Alignment Model (Pickering & Garrod, 2004). According to this model, successful communication is the result of a process of alignment across all linguistic levels, such that speakers converge in how they understand and use sounds (phonetics), language structure (syntax), word meanings (semantics) and contextual information (pragmatics). Pickering and Garrod (2004) proposed that alignment operates as follows: interlocutors initially start by using different referring expressions and, as the dialogue progresses, the most frequently used words, syntactic structures and situation structures become increasingly likely to be reused, inhibiting the other competing expressions.

Although alignment is a prominent and well-documented phenomenon in human communication, it has received little attention in the context of human-computer interaction. This is particularly surprising given that alignment, as a mechanism that promotes language re-use, can be of practical relevance to the ‘Vocabulary Problem’. Specifically, it is argued in this paper that alignment can be exploited not only to support successful and natural interaction but, more importantly, to predict and constrain the variability of user input.

Having identified the importance to the field of better understanding alignment, this paper uses the dialogue paradigm to identify and categorise the occurrence of alignment in users’ interactions with computer systems. It sets out to elucidate the characteristics of alignment in problem-free communication as well as in cases of user error, system error and non-understanding. The paper starts with a description of alignment in human-human interaction and then discusses the existing literature on alignment in human-computer interaction. From this analysis, a number of research hypotheses are framed which are subsequently tested through a study of simulated human-system dialogues in two different visual co-presence conditions. Synthesising existing findings with its results, the study aims empirically to demonstrate the practical implications of alignment and provide design recommendations relevant to the development of computer systems with natural language interfaces. The paper concludes by presenting a general model for the integration of alignment in dialogue-based human-computer interaction.

2 Alignment

Alignment is argued to be a basic interactive mechanism that takes place in dialogues at all levels – phonetic, phonologic, lexical, syntactic, semantic and pragmatic – and that makes communication between people ‘easy’, efficient and effective (Pickering & Garrod, 2004; Garrod & Pickering, 2004). The evidence for this comes from multiple data-driven studies which show that alignment occurs at the phonetic and phonological levels with participants converging in terms of pronunciation (Pardo, 2006). In respect of lexical alignment, dialogue is full of repetition of the same words (Tannen, 1989); interlocutors align in terms of vocabulary in the sense that they use the same referring expressions (Garrod & Anderson, 1987); and when interlocutors refer to the same object they tend to re-use a previously-used term, even when simpler terms are available (Brennan & Clark, 1996). In terms of syntax, speakers will select a specific syntactic structure (such as either 'give the apple to Jim' or 'give Jim the apple') based on that which their interlocutors have been using (Branigan, Pickering, & Cleland, 2000). At the situational (or pragmatic) level, interlocutors align on reference frames, such that if one speaker uses an egocentric frame of reference (e.g., using 'to the left', signifying his/her own left), the other speaker will do the same (Schober, 1993).

The phenomenon, described as part of Pickering and Garrod's (2004) Interactive Alignment Model, develops through two processes. First, alignment occurs as a result of the local, between-speakers priming mechanism ('input-output matching') at the same linguistic level (for instance, lexical, where speakers repeat each other's lexical choices). Subsequently, alignment at one level leads to further alignment at other levels, such that the re-use of a particular lexical item will activate a particular situation model (that is, the information relevant for the situation under discussion). From this perspective, since successful communication is seen as alignment of the interlocutors' situational models, communication success largely results from linguistic alignment (Pickering & Garrod, 2006). Eventually, the repeated use of the same syntactic, phonetic and lexical expression to refer to the same object results in the development of the chunking of those expressions into 'dialogue routines' which, over time, optimise and stabilise interaction. With respect to 'dialogue routinisation', the Collaborative Model (by Clark and colleagues; see Clark, 1996) seems to coincide with the Interactive Alignment account; in particular, Brennan and Clark (1996) propose that when interlocutors use the same expression to refer to an object, they enter into a tacit 'conceptual pact' in which they agree to keep referring to the same object in the same way. However, as explained below, the Interactive Alignment Model assumes that routinisation is automatic, whereas the Collaborative Model views this phenomenon as a result of partner-specific common ground.

Finally, there have been several explanations of why this phenomenon occurs. These include the social explanation, which argues that people who align linguistically with their partners expect to and may be positively perceived (see, for example, Giles, Coupland, & Coupland's, 1991, 'Communication Accommodation Theory' that proposes various factors behind convergence in speech patterns such as 'an individual's desire for social approval', attraction, power relations and social norms), and the 'audience design' explanation of the Collaborative Model, which argues that by choosing the same referring expressions interlocutors maximise their chances of successful communication (Brennan & Clark, 1996). This also resonates with the Interactive Alignment Model (Pickering & Garrod, 2004) which argues that alignment will result in communicative success. Yet, the accounts diverge in terms of whether the mechanism of alignment is automatic or strategic. In particular, Pickering and Garrod (2004) argue that alignment is a process that invariably occurs owing to mechanisms within the human processing system and is the basis of communication success. Work within the Collaborative Model (Isaacs & Clark, 1987; Clark & Murphy, 1982), however, assumes that alignment is mediated by explicit modelling of the interlocutor and context, which is updated on a turn-by-turn basis through feedback, in order to increase the likelihood of communication success. Pickering and Garrod

(2004) also recognise that ‘audience design’ may occur, but maintain that it is a one-off, optional decision, occurring at the beginning of the dialogue (p. 11; p. 48).

Having argued that alignment is pervasive in human communication, there remains the question of whether this mechanism also operates in the communication between a human and a computer system and, if it does, in what ways. If alignment is an automatic process (following the Interactive Alignment Model), then it should present similar patterns as are seen in alignment in human-human interaction. If it is a strategic process (following the ‘audience design’ explanation of the Collaborative Model), it should manifest in different ways. If alignment has a ‘social’ dimension, it is less clear how, or indeed if, alignment will occur, since one party in the dialogue is non-human.

There is a corpus of research looking at aspects of human-computer alignment which may be relevant here. A large segment of this research is dedicated to the study of alignment at the phonological/acoustic level, and shows that people tend to adjust their speaking rate (Bell, Gustafson, & Heldner, 2003), amplitude and pause frequency (Oviatt, Darves & Coulston , 2004; Suzuki and Katagiri, 2007) to that of the computer with which they interact. Moving beyond acoustic features as the focus, Branigan, Pickering, Pearson, McLean, and Nass (2003) and Cowan, Beale, and Branigan (2011) have investigated syntactic alignment between a human and a real or simulated computer in a picture-naming task, demonstrating evidence of alignment beyond the phonological level. From the perspective of dialogue system development and the ‘Vocabulary Problem’, however, alignment in terms of vocabulary seems to have more practical significance.

Pioneering work on lexical alignment in Human-Computer Interaction was conducted by Brennan (1996) who aimed to address the question of whether people adopt the same lexical terms used by the computer to the same extent as they do when interacting with other humans. Wizard-of-Oz experiments were conducted in relation to a database query task, and the results showed that when the ‘system’ responded using a different term to that originally used by the human user, the user tended to accept and subsequently use the system’s term. The rate of alignment with the computer (or ‘convergence’) was found to be comparable to the alignment rate with humans. This finding supports the hypothesis that alignment is a basic, automatic mechanism operating in all contexts of language use.

A series of studies by Branigan and her colleagues also focus on lexical alignment in HCI (see Branigan and Pearson (2006) and Branigan, Pickering, Pearson, and McLean (2010) for an overview). In Branigan et al. (2004), users were told that they would interact with a computer program or a human (via a computer) in an object-naming and selecting task, though the interlocutor was a computer program in both conditions. In the study, the users saw two objects on the screen (for instance, a bench and an apple). The objects could be referred to in two ways; for instance, the bench could be referred to accurately as ‘bench’ or less accurately as ‘seat’ (accuracy refers to preferred or dispreferred synonyms based on a pre-test conducted as part of their study). In both conditions, the computer would name one of the objects using the more or less accurate term and the user would select the named object. Subsequently, the roles were reversed; presented with the same pair of objects, the user named one of them and could see the computer’s selection of it. The researchers measured whether the user would choose the less accurate term if the computer had done so. The same experimental setup was again deployed in Pearson, Hu, Branigan, Pickering, and Nass (2006). This study involved users completing the task with a computer, but they were made to believe that they would interact with either the 'basic' or the 'advanced' version of the system, whereas in reality both versions were the same.

The findings of these studies show that lexical alignment is prevalent in both human-computer and human-human interaction, with users in both studies using the less accurate term when it was used by their interlocutor (human or computer). On first consideration, this may suggest that alignment is an automatic process, a perspective supported by Branigan, Pickering, Pearson, McLean, and Nass's (2003) study in syntactic alignment which observed similar rates of alignment for both computer and 'human' addressees, leading to the conclusion that alignment is an automatic imitation mechanism that does not involve any decision or strategic component. Oviatt, Darves and Coulston's (2004) study with children also reported that acoustic and prosodic adaptation to the speech synthesis system was bidirectional, rapid and re-adaptable, which may also suggest automaticity. However, in Branigan et al.'s (2004) study, lexical alignment was considerably greater where the user was interacting with the computer compared to when their interlocutor was (what s/he thought was) a human, possibly because the former was perceived as being more 'error-prone'. The explanation for this is that speakers align their linguistic behaviour according to the perceived, rather than actual, capabilities of the system. This is confirmed by observations from Pearson, Hu, Branigan, Pickering, and Nass's (2006) follow-up study which showed that users aligned more to the 'basic' version of the system (than to the 'advanced' one). As the authors point out, this indicates that alignment has a strategic dimension, as users aligned more in order to maximise the likelihood of successful communication (Branigan & Pearson, 2006). In summary, alignment between humans may be equally mediated by automatic priming processes, a social and a strategic component. Yet, human-computer interaction appears to involve a stronger strategic component, which is specifically clear in the case of lexical alignment.

2.1 The effect of visual feedback on alignment

How visual feedback influences goal-oriented interaction has attracted interest across many disciplines. Such research is necessary for understanding phenomena in normal human communication. It also informs the development of computer systems that share the same visual or physical space with human users. Moreover, computer-mediated communication (CMC) and computer-supported cooperative work (CSCW) technologies may integrate video or support sharing visual perspective and, therefore, better awareness of the role of visual information as a conversational resource can lead to improved designs. Relevant literature in task-oriented CMC and human communication (Gergle, Kraut & Fussell, 2013; Kraut, Gergle & Fussell, 2002; Clark & Krych, 2004; Gergle, Kraut, & Fussell, 2004; Kraut, Fussell, & Siegel, 2003; Brennan, 2005) shows that visual information (termed 'shared workspace' or 'visual co-presence') offers several advantages for the accomplishment of the task, namely: it affords direct observation of task status; it provides visible feedback on the addressee's actions; and it ensures a joint focus of attention and common reference frame which augments the interlocutors' common ground. These studies involved dyads of participants, with one person providing instructions to his/her partner on how to complete a task. They compared the condition in which the instructor was able to observe the physical actions, movements and relevant shared objects in the environment with a language-only condition. Their results also indicate that sharing visual information has a profound effect on coordination patterns and communication content. For example, in visual co-presence, speakers may produce linguistic shortcuts such as ‘turn here’ or ‘take this road’ instead of more complex constructs such as ‘take the third road to your left’. Similarly, their addressees may demonstrate understanding without having explicitly to state it but through performing the action (since visual evidence is stronger than linguistic). These phenomena have been largely interpreted through the concepts of grounding and common ground, as discussed within the Collaborative Model.

The findings from human communication outlined here give rise to rich questions with regards to how visual feedback affects the interaction with a computer system. Specific to the central aim of this paper, it would be interesting to identify how visual feedback influences the coordination mechanism of alignment between a human and a computer system. In addition to the practical importance, exploring whether alignment is stronger or weaker depending on the interaction condition may have implications for theoretical models of communication. As shown in the previous section, findings remain inconclusive regarding whether alignment is an automatic, ‘post-conscious’ (Bargh, 1989) process or a strategy that interlocutors ‘intentionally’ employ to maximise the probability for communication success. Therefore, if it is found that alignment is consistent across both conditions of presence and absence of visual feedback, it may suggest that it is an automatic mechanism that ordinarily occurs irrespective of situation. On the other hand, if alignment is stronger or weaker in one condition, it could hint at the existence of a strategic component.

3 Research aim and hypotheses

The studies outlined in the previous section provide strong evidence regarding the presence of alignment in HCI. However, four possible limitations have been identified. First, the studies employed tasks and scenarios (e.g., object-naming) that were restricted and only weakly related to real-life applications. Second, they failed to assess the fundamental characteristic of alignment; in particular, that alignment is mutual. Instead, they focused on the ‘one-way’ alignment of user to system. It would be interesting to see whether user alignment varies depending on whether the system is also primed to repeat user’s expressions. Third, alignment was measured in interactions with a system that were completed in two utterances. Yet, alignment operates and develops over the full course of a dialogue (as shown from the original ‘maze game’ experiments by Garrod and Anderson (1987), in which pairs produced spatial descriptions guiding each other in a maze and found that, over time, they converged on similar spatial descriptions). Fourth, these studies provide evidence of the local priming mechanism of alignment (‘input-output matching’), with less scope for the global, longer-lasting alignment that persists throughout the dialogue (relating to ‘dialogue routines’). As a result, questions remain with regards to whether and how alignment occurs and develops in human-computer dialogues.

Motivated by these studies and in an attempt to address the noted limitations, this paper sets out to identify and describe alignment in the domain of human-computer interaction, with the aim of informing the development of practical, goal-oriented dialogue systems. To this end, the study formulates research hypotheses and tests them through analysis of experimental data from a dialogue study. The hypotheses are given below.

> *Hypothesis 1: Alignment occurs in the interaction between a human user and a computer system.*

> *Hypothesis 2: Alignment occurs as a mutual phenomenon.*

As outlined in section 2.1, previous studies have provided substantial evidence regarding the effect of visual feedback on task-oriented communication. Given its scope, this study seeks to identify how visual feedback influences the processes of alignment. In particular, the third research hypothesis focuses on whether the strength of alignment is different across two conditions of (i) absence and (ii) presence of visual feedback.

> *Hypothesis 3: Visual feedback influences alignment between a user and a system.*

The fourth research hypothesis is concerned with miscommunication. In goal-oriented human communication, instances in which the hearer fails correctly to interpret an utterance are natural and ubiquitous. Similarly, speakers commonly produce not only underspecified and vague utterances, but also inaccurate ones. For systems with natural language interfaces, miscommunication is more prevalent owing to challenges with automatic speech-recognition technologies. This is aggravated by misplaced assumptions by the user regarding the functional and linguistic capabilities of the system. Therefore, the scope and frequency as well as costs (in the case of systems, such as robots, that operate in the same environment as humans, where potential hazards are involved) of miscommunication make it an essential part of system design (McTear, 2008).

Given the objectives of the paper, it is important to understand the behaviour of users when miscommunication is detected. Miscommunication appears to be the basis of linguistic change, as it is at this point when speakers need to consciously reformulate their utterances – to be more compatible with what the ‘hearer’ can understand. Therefore, it is expected that miscommunication will disrupt lexical alignment, leading to the fourth research hypothesis below. Within the same problem domain, it is practically relevant to continue the investigation to find out whether users will attempt to recover from an error by using vocabulary that ‘worked’ earlier in the dialogue, or they will use an entirely novel expression.

> *Hypothesis 4: Miscommunication locally disrupts the process of alignment in human-computer communication.*

As noted in section 2, the main premise of studies adopting the Interactive Alignment Model is that alignment underlies successful communication. There is also evidence that alignment has a social dimension, leading people to align their verbal and non-verbal behaviour to express affiliation (Giles, Coupland, & Coupland, 1991), and that this behaviour is perceived favourably by peers. Although it is a contentious issue whether the same social norms persist in people's interactions with computers (see, for example, Nass & Moon, 2000), research has shown that users rated more positively systems that imitated their head movements (Bailenson & Yee, 2005), personality attributes (Moon & Nass, 1996) and acoustic and prosodic features (Nass & Lee, 2001; Ward & Nakawaga, 2002). There do not, though, seem to be any similar findings in relation to lexical and syntactic alignment in task-oriented interactions. Therefore, the fifth research hypothesis deals with the relationship between alignment and user evaluation of interaction success.

Hypothesis 5: Lower alignment is linked to lower user perceptions of interaction success.

While the aforementioned studies explored original territory and provided new ideas and novel data on the operation of alignment in HCI, there was no focused attempt to use their findings to frame specific recommendations for interactive systems. The paper will address this shortcoming, drawing on the findings related to the five research hypotheses to distil guidelines relevant to the development of practical, goal-oriented dialogues with computer-based systems. It will also seek to contribute to the limited work on dialogue models that leverage the effects of the mechanism by aiming to describe elements of a theoretically- and empirically-motivated dialogue model that supports and exploits alignment.

4 Methodology

The study essentially explores whether it is possible to limit and predict the range of utterances that the user can potentially employ to interact with a computer-based system, by taking advantage of the two mechanisms of alignment that naturally occur in interactions: input/output matching; and routinisation. The context used to explore the research hypotheses is the investigation of route instructions produced in real-time dialogue with a computer-controlled robot within a restricted spatial network. The remainder of this section describes the development of, and rationale behind, the methodology used to address the research hypotheses.

4.1 The experimental method

Since human-human interaction differs from human-computer interaction (Amalberti, Carbonell, & Falzon, 1993; Fraser & Gilbert, 1991), data and ideas to inform the design of computer-based dialogue systems should be derived from interactions with such systems, rather than directly from studies of human-human interaction. This requires that a dialogue system already exists or that one is simulated. A commonly-employed approach that uses a simulated system is the Wizard of Oz (WOz) method (Fraser & Gilbert, 1991) where two people interact, one of whom is made to believe that he/she is interacting with a system rather than a person. The 'wizard' in a WOz experiment is the experimenter or a single, trained confederate. However, this approach will inevitably offer one (expert and possibly biased) interpretation of the instructions, inhibiting effects of interaction and individual differences in language interpretation and strategy. To address this, in the WOz study reported in this paper the wizards were also naive participants who were given no dialogue script or guidelines on what to say.

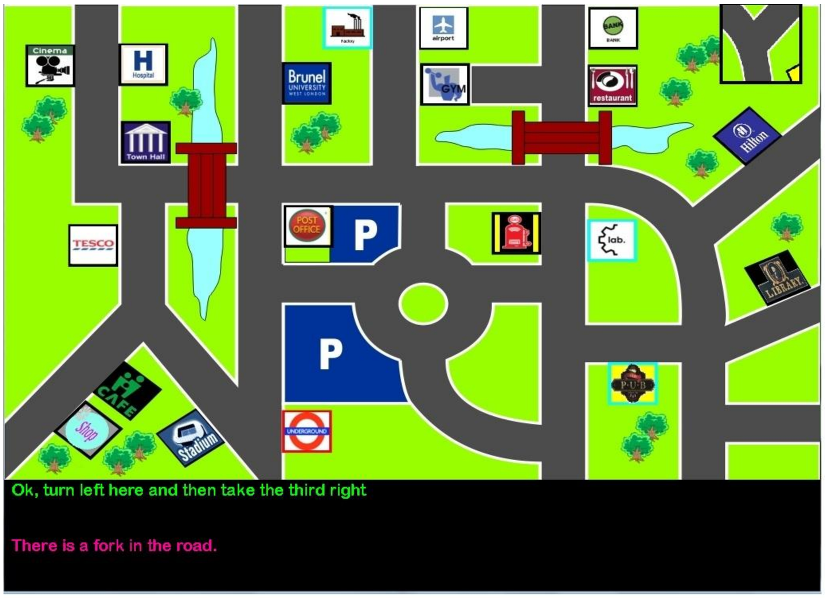
The study was designed to elicit spontaneously-generated route instructions. The experimental technique involved dyads of participants (instructors and followers) collaborating in an urban navigation scenario, with the instructors being under the impression that they were conversing with a software agent (simulating a robot follower). Given the focus of the research on alignment, both instructors and 'robots'/followers were subjects in the study. A system was developed to support synchronous text communication and execution of route instructions between the paired participants. To implement the experimental conditions aiming to assess whether alignment is modulated by the presence/absence of visual feedback (research hypothesis 3), the system could enable or restrict visual access to the actions of the robot.

The domain used in the experiment was pedestrian navigation in a simulated town. Each participant had two overt sources of information: what was on his/her map; and what the other pair member said. Thus, the participants were given the opportunity to interact with each other in a relatively natural manner, while the information available to them was controlled at any given point in the dialogue. The user/instructor (hereafter the user) had to guide the 'robot'/follower (hereafter the 'robot') to six designated locations in the town. The cooperative nature of the task lay in two additional characteristics. First, in each pairing, only the user knew the destinations and had a global view of the environment, so the 'robot' had to rely on the user's instructions and location descriptions. Second, the user needed the 'robot's' descriptions to determine its exact position and perspective. The details of the setup and the system used to support the simulation are provided in the following sections.

4.1.1 The system

The experiment relied on a custom-built system which supported the interactive simulation and enabled real-time, direct text communication between the user and 'robot' in a pair. The system connected two interfaces over a Local Area Network using TCP/IP as the communication protocol, kept a log of the dialogues and also recorded the coordinates of the current position of the 'robot' at the moment at which messages were transmitted. Thus, it was possible to analyse the descriptions against a matching record of the 'robot's' position and reproduce its path with temporal and spatial accuracy. The interfaces consisted of a graphical display and an instant messaging facility (the dialogue box). The dialogue box displayed each participant's messages (in green) in the upper part of the dialogue box; the messages sent by the other participant in the pair were displayed (in magenta) in the lower part of the dialogue box.

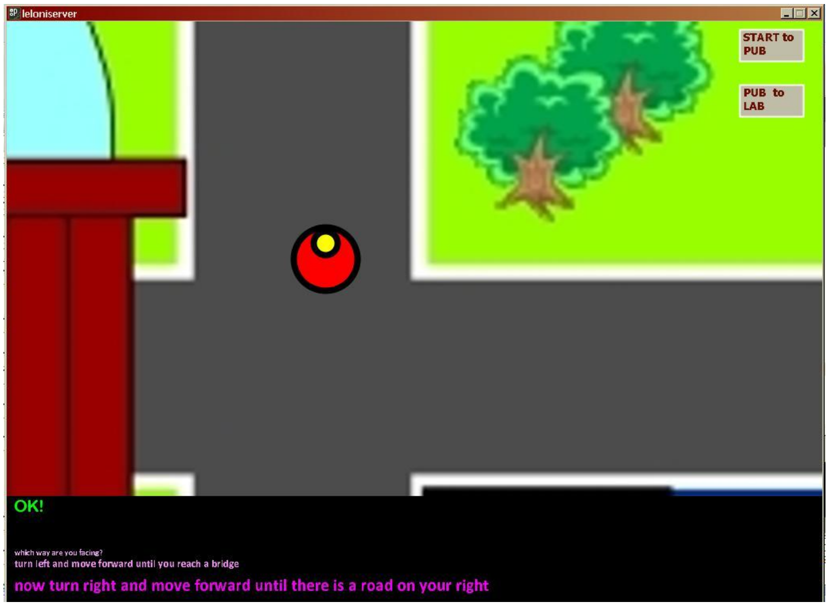
The interface seen by the user displayed the full map of the simulated town. The destination location was shown in red and the tasks that had been completed were shown in blue. In order to examine the nature of alignment through the presence/absence of visual feedback (research hypothesis 3), there were two variants of the user's screen. In the first, called the 'Monitor Condition', a small 'monitor' was displayed in the upper right corner of the screen showing the 'robot's' immediate locality, but not the robot itself (see Figure 1). This meant that the user shared the same visual space as the ‘robot’ and could see the area changing as the ‘robot’ was moving. In the ‘No Monitor Condition’, this feature was disabled so that the user had no direct visual information relating to the ‘robot’s’ position in the environment.



Ok, turn left here and then take the third right

There is a fork in the road.

Figure 1. The interface of the user/instructor as presented in the Monitor condition. The monitor window can be seen in the upper right corner. In the No Monitor condition, this feature was removed.



OK!

which way are you facing?
turn left and move forward until you reach a bridge
now turn right and move forward until there is a road on your right

Figure 2. The interface of the ‘robot’/follower.

The ‘robot’s’ interface displayed a fraction of the overall environment map, showing only the surroundings of the robot’s current position (see Figure 2). The ‘robot’ (signified by a red circle with a yellow ‘face’) was operated by the follower using the arrow keys on the keyboard. The dialogue box also displayed a history of the user’s previous messages to the ‘robot’. To simulate the ability of the ‘robot’ to learn routes, after each task was completed a button for the completed route appeared on the robot’s/follower’s screen. If the ‘robot’ was then instructed to go to a previously visited destination, the follower could press the corresponding button and the ‘robot’ would automatically execute the move. In the example provided in Figure 2, the ‘robot’ has ‘learnt’ two routes: (i) from the ‘start’ to the ‘pub’; and (ii) from the ‘pub’ to the ‘lab’.

4.1.2 Participants

64 participants (32 males and 32 females), recruited from undergraduate and postgraduate students of various departments at a UK university, were randomly allocated to the two roles (user or ‘robot’) and to each of the experimental conditions (‘Monitor’ or ‘No Monitor’). Each was paid £10 for participating in the experiment. Previous experience in using computers was necessary but no specific computer expertise or other skill was required to take part.

4.1.3 Procedure

Users and ‘robots’ were seated in separate rooms equipped with desktop PCs, on which the respective interfaces were displayed. Participants received verbal and written instructions related to the task from their role perspective. The participants that were assigned to be ‘robots’ were fully informed about the experimental setup and that they were to pretend to be robots. No examples or instructions were provided on how to communicate or complete the task. The ‘robots’ were also given a brief demonstration of, and time to familiarise themselves with, the operation of the interface. In brief, the training of the ‘robots’ in terms of communication style followed the guidelines set in Amalberti, Carbonell, and Falzon (1993): natural language should be used, there were no constraints in comprehension and production and no dialogue script, but ‘robots’ could only produce task-related utterances, and the use of slang words was not permitted (abbreviations and misspellings were automatically corrected). The users were told that they would interact directly with a robot, which for practical reasons was a computer-based, simulated version of the actual robot. The users were told that robots were proficient in understanding and producing spatial language. They were given no other examples of, or instructions about, how to interact with the robot.

Each pair attempted six tasks, presented in the same order; the user navigated the ‘robot’ from the starting point (bottom right of the map) to six designated locations (pub, lab, factory, tube, Tesco, shop). At the end of the experiment, the users were debriefed and the full nature of the experimental setup was disclosed and explained. Before this disclosure, questioning was used to determine whether users had become aware that the experiment was a simulation. Though relevant literature suggested that participants in WOz studies are easily convinced (Fraser & Gilbert, 1991), the experimenters were prepared to discard the data if any user expressed doubt over the simulation. However, all users confirmed their belief in the setup and expressed surprise on being told during the debriefing that they had been interacting with a human acting as a ‘robot’. During the interviews, no user expressed that they were impressed with the (linguistic and functional) capabilities of the robot. This may be due to the fact that users have no experience of interacting with real robotic systems, which may lead to inflated or no *a priori* assumptions about what a robot can do.

There is an interesting body of research focusing on users' perceptions of systems' capabilities. The study by Amalberti, Carbonell, and Falzon (1993), for example, presented an experiment in which two groups of users interacted with the same human experimenter; one group was told that they would talk to a human, and the other group that they would interact with a dialogue system. The human experimenter followed the same guidelines as the 'robots' in the study reported in this paper. The results showed that users approached the roles in the interaction differently, and tended to rely less on the problem-solving capacity of the 'computer' compared to the human interlocutor. Interestingly, any linguistic differences tended to disappear as subjects gained familiarity with the system. Along the same lines, research by Levin and colleagues (Levin, Killingsworth, Saylor, Gordon & Kawamura, 2013; Levin, Killingsworth & Saylor, 2008) demonstrates that people are willing to attribute human-like cognitive characteristics such as intentionality to robots more than they do with computers, but only when users are given time to observe intentional behaviour by the robot. However, robots cannot be perceived as fully intentional. For the study reported in this paper, the post-task interviews and relevant literature findings give confidence that any significant effects yielded by the data are not a result of language adaptation by the users arising from them realising that they were instructing another person.

4.2 Data analysis approach

The study yielded a corpus of 184 dialogues, which comprised 3,876 turns (messages sent) by the participants (2,125 user turns and 1,751 'robot' turns). The users produced 1051 turns that included 1,660 different instructions. First, all utterances in the corpus were analysed in terms of their components, following the Communication of Route Knowledge (CORK) framework developed by Vanetti and Allen (1988). Utterances could contain references to environmental features – (i) Landmarks, (ii) Pathways, (iii) Choice Points and (iv) Destination – and could incorporate delimiters, which fall into four categories: (i) Distance designations; (ii) Direction designations; (iii) Relational terms; and (iv) Modifiers.

Identification and analysis of the levels of alignment and miscommunication were performed with respect to the two visual information conditions ('Monitor' and 'No Monitor'), to address the first four research hypotheses. Data related to research hypothesis 5 was gathered through a user questionnaire (see section 4.2.4).

4.2.1 Annotation of alignment

The analysis with respect to lexical alignment basically investigated whether speakers used the same words as their partner. Following the Interactive Alignment Model of human communication (see section 2) and addressing the limitations of related work in HCI (see section 3), it was necessary to capture alignment in the dialogue both 'locally', as priming, and 'globally', as lexical innovation. So, first, alignment was measured by looking at the adjacency pairs in the dialogue and comparing the two utterances (what the Interactive Alignment Model terms 'input/output matching'). An adjacency pair is a sequence of two *related* utterances by *two different* speakers, such that the second utterance is a response to the first – for instance, paired responses like a question followed by an answer, or an offer followed by acceptance or rejection (Levinson, 1983, p.303). So, a turn was a 'match' if it contained the same component as the turn to which it was a response. For each matching component in an utterance, a score of 1 was given. If no component matched, the turn was a 'mismatch' and a score of 0 was given. The annotation of alignment on the adjacency pair level is exemplified through two dialogue excerpts, shown in Tables 1 and 2.

In the first example, the user's utterance matches the previous utterance by the 'robot', repeating the modifier, 'bendy', and the pathway reference, 'road'. Thus, it is marked as containing '2' matches. The aligned components are shown in bold in Table 1.

Utterance	Match
R: I am at the junction by the bridge, facing the ***bendy road***.	
U: Go into the ***bendy road***.	2

Table 1. First dialogue example; the user response repeats two components of the robot utterance. R and U denote 'robot' and user, respectively.

In the second example (see Table 2), the user first produces an instruction which does not match the previous utterance. This is immediately reformulated to repeat the exact expression used by the 'robot', 'at y-shaped junction', containing '2' matches.

Utterance	Match
R: I am at ***y-shaped junction.***	
U: make a right.	0
U: make a right at ***y-shaped junction***.	2

Table 2. Second dialogue example; the two consecutive user responses repeat 0 and two components of the robot utterance, respectively.

Second, lexical innovation, the rate of unique words introduced over the course of the dialogue, was used as an indicator of global alignment (following the approach of Mills, 2007). When interlocutors introduce new expressions instead of re-using those that have already occurred in the dialogue (as the Interactive Alignment Model postulates), alignment is low. Lexical innovation was calculated by comparing every constituent word in an utterance to the previous words in the dialogue. For example, an utterance such as 'turn left' leads to a backwards search in the dialogue for the previous occurrence of 'turn', adding '1' to the alignment score if not found and '0' if found, before moving on to the next word. Lexical innovation was also used to capture alignment achieved by the end of the dialogue and was measured by the ratio of unique words produced in undertaking the final task of the session. Simply put, the lower the ratio of unique words towards the end of the dialogue, the higher the level of alignment ultimately achieved.

4.2.2 Annotation of miscommunication

Specifically relevant to research hypothesis 4, the logged interactions were annotated in order to detect and classify interaction problems. In dialogue studies, miscommunication is defined as encompassing two forms of problems, misunderstandings and non-understandings (Hirst, McRoy, Heeman, Edmonds, & Horton, 1994). A misunderstanding occurs when the addressee obtains an interpretation that s/he believes is correct and complete, but not the one that the speaker intended her/him to obtain. Misunderstandings are only noticed when the addressee acts upon them (Hirst, McRoy, Heeman, Edmonds, & Horton, 1994). In this study, misunderstandings corresponded to execution errors, which refer to instances in which the ‘robot’ failed to understand the instruction and deviated from the described route. The system logged and time-stamped messages and ‘robot’ coordinates so that an execution could be matched with the instruction that produced it. An example of an execution error is provided in the dialogue excerpt in Figure 3. Figure 3 also illustrates the route that the user described and the ‘robot’ followed during this interaction.

Utterance
U: Walk past the lab. Then turn left. After you get to the post office turn in the road off the post office. Keep going until you see the Underground station **[1]**
U: Where are you now? **[2]**
R: Brunel University is on my right **[3]**

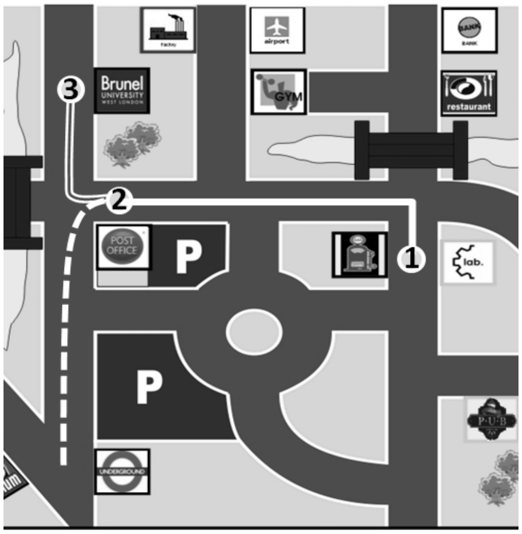


Figure 3. An excerpt of a dialogue containing an execution error. The number inside the square brackets denotes the position of the ‘robot’ on the map at the time that the utterance was sent. The map on the right shows the ‘robot’s’ execution of the instructions in this dialogue: the solid line illustrates the accurately executed route; the dashed line represents the route that the instructor described but the ‘robot’ failed to execute; the double line shows the deviation from the intended route; the numbers along the executed route indicate the position of the ‘robot’ when the utterances were sent.

A non-understanding occurs when the hearer obtains an uncertain interpretation of an utterance, no interpretation or more than one (Hirst, McRoy, Heeman, Edmonds, & Horton, 1994). Instances of non-understandings are immediately recognised, as the hearers are aware of them and articulate them. The analysis measured the utterances by the ‘robot’ that expressed non-understanding. These responses could be formed explicitly, as in statements like “I don’t understand", or as clarification requests (Gabsdil, 2003) (for example, "Back to the bridge or back to the factory?" after the user instruction "Go back to the last location.").

These two forms of miscommunication are normally attributed to the addressee, who, in this scenario, is the 'robot'. However, the source of execution errors was not only the incorrect interpretation of an instruction; they also occurred as a result of inaccurate instructions. Therefore, the analysis of miscommunication also extends to 'user errors' (Oulasvirta, Engelbrecht, Jameson, & Möller, 2006). In the study's dialogue corpus, incorrect instructions occurred mainly because of unintended mistakes or misconceptions regarding the position and orientation of the 'robot'.

Figure 1 (in section 4.1.1) shows a screenshot of an interaction and serves to exemplify an incorrect instruction due to a mistake in the spatial direction. The destination of the particular interaction was the Tube. As can be seen from the small window in the top right corner of the user's monitor and the 'robot's' message in the dialogue box ("There is a fork in the road"), the robot is on the y-junction beside the Lab. The next instruction from the user is "Ok, turn left here and then take the third *right*" which is incorrect, having confused 'left' with 'right'. The 'robot' accurately executes the incorrect instruction and arrives at Brunel University. As such, this miscommunication incident was tagged as 'incorrect instruction' and not 'execution error'.

Table 3 summarises the study's alignment and miscommunication measures and provides short definitions.

Measure	**Definition**
Alignment	
Match/Mismatch	A turn is a match (or mismatch) if it repeats (or not) the same component as the turn to which it responds. The number of repeated components in a turn was also counted.
Lexical Innovation	The rate of unique words introduced over the course of the dialogue; that is, the number of new words in each turn was counted.
Miscommunication	
Execution Errors	The instances in which the 'robot' deviated from the described route.
Non-understandings	The utterances by the 'robot' that expressed non-understanding, either explicitly or as clarification requests.
Incorrect Instructions	An incorrect instruction by the user.

Table 3. The study's list of measures and their definitions.

4.2.3 Reliability of annotation

Lexical innovation was automatically calculated. The rest of the measures were manually annotated. The manual annotation was performed by cross-referencing the utterances with the system logs of the robot actions and position at the time each message was sent or received. As explained above, the annotation involved little subjective judgement. The annotation process was performed in two stages. During the first stage, 25% of the corpus (48 dialogues, 933 turns, from both conditions) was coded by two annotators: an expert annotator and an annotator with no prior knowledge of discourse analysis or experience in dialogue data annotation, who received a training session before undertaking the analysis. The annotators coded the same 25% of the corpus, and worked independently. The consistency of the annotation was calculated by a series of Cohen's Kappa. The Kappa values obtained for the four measure categories (Match/Mismatch, Execution Errors, Non-understandings and Incorrect Instructions) were .961, .842, .886 and .816, respectively, showing a generally high level of agreement between the annotators (values above .70 are normally considered satisfactory (Lazar, Feng, & Hochheiser, 2010, p. 298)). The few items where disagreement occurred were discussed between the annotators. In the second stage of the annotation, only the expert annotator annotated the remaining 75% of the corpus, because of the high level of inter-annotator agreement from stage 1.

4.2.4 User perceptions of the interaction

A simple questionnaire was designed to collect data on user perceptions, based on the related studies by Williams and Young (2004) and Skantze (2005). After the completion of each of the six tasks, the users were asked to complete a questionnaire in which they rated their agreement with five declarative statements of opinion. The questionnaire used a Likert scale with seven levels of agreement: *strongly disagree; disagree; slightly disagree; neutral; slightly agree; agree; and strongly agree.* The items probed five different aspects of the user's experience of their interaction with the 'robot': perceived task completion (Statement 1: "I did well in completing the task"); execution accuracy (Statement 2: "The system was accurate"); ease of use (Statement 3: "The system was easy to use"); helpfulness of the system (Statement 4: "The system was helpful"); and overall satisfaction (Statement 5: "I am generally satisfied with this interaction"). The responses were mapped to integer values between one and seven (with seven representing the highest level of agreement). The scores associated with each statement were summed for all six tasks, which resulted in a cumulative score for each statement ranging from 6 to 42.

5 Results

This section reports the results of the analysis in relation to the focus of each of the study's five research hypotheses.

5.1 Evidence of alignment

To test the first research hypothesis, evidence of alignment between user and 'robot' was sought. The rate of lexical innovation was determined by the number of new words introduced as the dialogue progressed. Figure 4 shows the number of new words plotted against the utterance number (averaged for all pairs). The graph demonstrates a decrease of innovation over time and shows that the vocabulary utilised by the participants becomes relatively stable after approximately 70 turns. This finding fits the basic predictions offered by the Interactive Alignment Model which suggests that participants will come to rely on previously used expressions as dialogues progress. Confirming the first research hypothesis, the decrease in the rate of lexical innovation that occurs early in the dialogue hints at a rapid development of alignment.

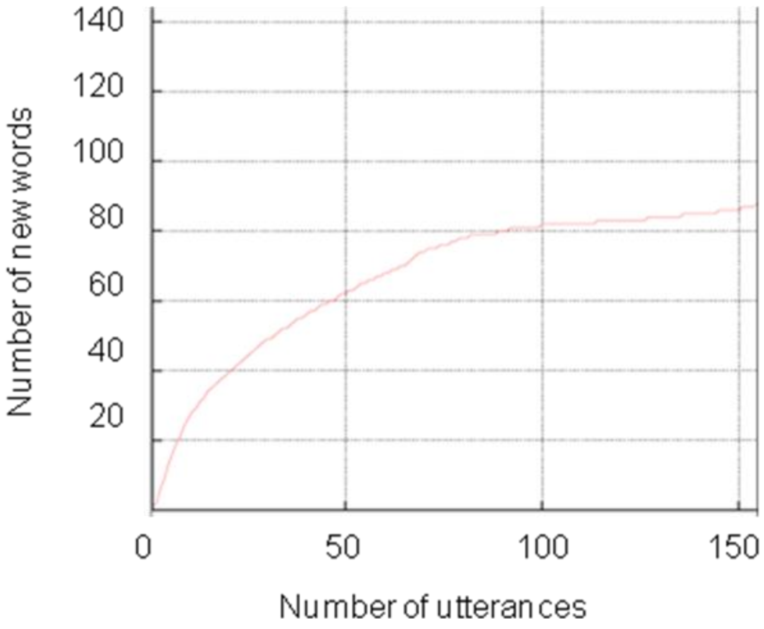


Figure 4. Lexical innovation over time

Lexical innovation was also measured by the ratio of unique words produced in undertaking the final task of the session. Not surprisingly, there was a significant negative correlation between match scores for users and 'robots' and the ratio of unique words in the final task ($r_{(32)}$ = -.53, p = 0.002). That is, 'robots' and users that were aligning to each other on the adjacency pair level were also more likely to conclude the dialogue with a more concise vocabulary. This finding also serves to validate the fitness of lexical innovation as a measure of alignment.

5.2 Evidence of the mutuality of alignment

The analysis in relation to lexical innovation pointed to the existence of alignment. Additional evidence was required to determine whether both interlocutors coordinate their lexical choices, and therefore whether, as research hypothesis 2 stated, alignment is a mutual phenomenon.

Correlational analysis showed that user match scores and 'robot' match scores were positively and strongly related ($r_{(32)}$ = .82, p = .001). The computation of r-squared indicated that 68% of the variability in the user match scores could be directly predicted by the variability in 'robot' match scores. Therefore, as the 'robot' 'match' scores increased the user 'match' scores were also very likely to increase. This finding provides evidence that alignment is not merely present but also mutual and conditional: if one speaker uses aligned responses, their partner is more likely to do so at a similar rate. The scattergram in Figure 5 illustrates that the data points are reasonably well-distributed along the regression line, in a linear relationship with no outliers. Similarly, there is a positive correlation between the 'mismatch' scores of users and 'robots', with the 'mismatch' scores of users rising when the 'mismatch' scores for 'robots' rise ($r_{(32)}$ = .42, p = .017).

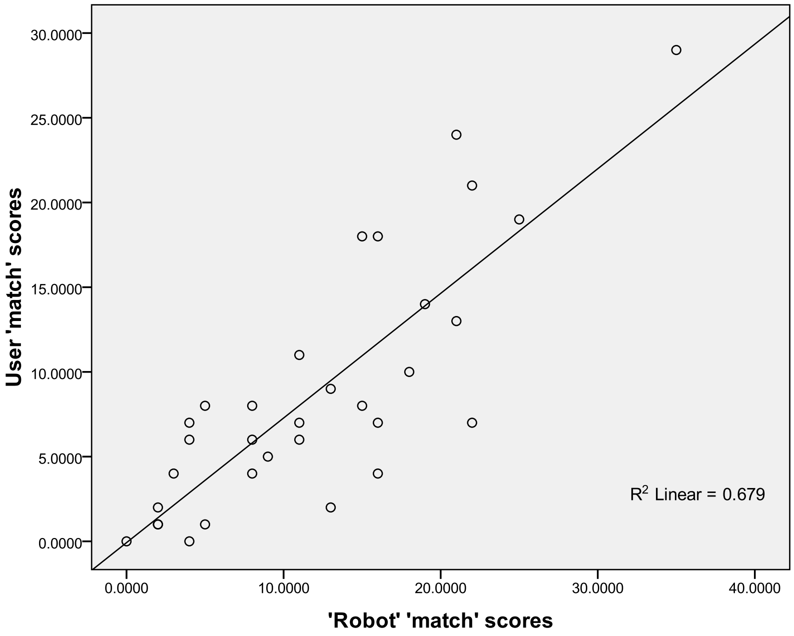


Figure 5. Scattergram showing the relationship between match scores by users and 'robots'.

5.3 The effect of visual feedback on alignment

Relevant to the third research hypothesis, the analysis sought to discover whether the levels of alignment varied with the absence of visual feedback.

A mixed ANOVA design was employed to explore the effect of visual feedback. The within-subjects factor corresponded to 'match' and 'mismatch' scores of a pair, and the between-subject factor was Monitoring (Monitor and No Monitor). The investigation began by looking at whether there was a difference between the 'match' and 'mismatch' scores of the pairs; in other words, whether interlocutors aligned to each other. The means of the scores seemed to suggest a preference for aligned responses (a mean of 18.4 'matches' as opposed to 15.3 'mismatches' for each pair). However, the ANOVA showed that the difference between 'match'/'mismatch' scores was only marginal ($F_{(1,30)}$ = 2.75, p = .052), and the post–hoc paired t-test confirmed the absence of significant effect ($t_{(15)}$ = 1.91, p = .065). The analysis determined a significant effect for Monitoring ($F_{(1,30)}$ = 5.78, p = .023). The significant interaction clarified the effect ($F_{(1,30)}$ =

4.85, p = .035, η^2 = .125). Inspection of error bar charts and Bonferroni-corrected post-hoc t-tests (significance value set to p < .0125) verified that only the 'match' scores in the Monitor condition accounted for the observed difference. In particular, the 'match' scores of pairs were significantly higher in the No Monitor condition (*M* = 4.73, *SD* = 2.34) compared to the Monitor condition (*M* = 2.48, *SD* = 2.31) ($t_{(30)}$ = -2.74, p = .010, d = 0.97). Similarly, 'match' scores in the No Monitor condition was markedly higher compared to 'mismatch' scores in both conditions. This result suggests that in the absence of visual feedback participants relied more heavily on alignment as a mechanism/strategy to ensure dialogue success.

Next, the analysis considered a speaker effect, and, thus, a mixed ANOVA was performed on the 'match' scores of the user and 'robot' as the within-subjects factor. Most importantly, the analysis reiterated that 'match' scores of both speakers were significantly higher in the No Monitor condition ($F_{(1,30)}$ = 7.50, p = .01, η^2 = .25). This parallel increase demonstrates that it is not the scores of *one* of the participants that account for the previous observation; rather, both 'robots' and users aligned more when visual information was not available. Finally, the analysis measured lexical innovation in the final task to assess alignment. The t-test revealed reliable differences between the Monitor and No Monitor conditions ($t_{(30)}$ = 2.87, p = .007, d = 1.06). In particular, in the Monitor condition the final task contained 21.1% new words (*SD* = 0.049), which dropped to 17.1% in the No Monitor condition (*SD* = 0.027). This finding provides further evidence that alignment is higher when users do not have access to visual information.

5.4 Miscommunication and alignment

This subsection presents the analysis related to the fourth research hypothesis; the effect of miscommunication on alignment was explored through lexical innovation.

First, lexical innovation in the final task was considered using the measure of the ratio of unique words. The analysis revealed that there was a positive relationship between the number of incorrect instructions and the ratio of new words, suggesting that pairs concluded the dialogue being less aligned when more incorrect instructions had been given ($r_{(32)}$ = 0.41, p = .021).

As a result, a chi-square analysis was performed to clarify the link between lexical innovation and miscommunication. This analysis considered the number of new words contained in an utterance immediately after a (i) non-problematic and (ii) problematic utterance (that is, a dialogue turn marked as a non-understanding, an incorrect instruction or in which an execution error occurred; a combined measure was used since the nature and cause of miscommunication was not the focus of this analysis). All utterances were grouped based on whether or not they contained new words and whether or not they followed a problematic utterance.

Chi-square tests, using both linear and standard Pearson's chi-square for completeness, were performed and showed an association between the number of new words in an utterance and the occurrence of miscommunication ($\chi^2_{(1)}$ = 18.52, p < .001). The linear-by-linear association (calculated using Pearson's *r*) confirmed the result (M^2= 18.52, p < .001) and the phi coefficient was equal to .068. The odds ratio was 1.78, indicating that the odds of novel words being used were 1.78 times higher after miscommunication than after a non-problematic utterance.

So far, this section has shown that novel vocabulary is more likely to be input by the user when s/he detects miscommunication, whereas in problem-free communication, vocabulary from the preceding dialogue is reiterated. The results in section 5.3 (high ‘match’ scores and low lexical innovation) suggested that alignment increased when users did not have visual access to the ‘robot’s’ actions. Therefore, it was necessary to tease apart the effect of visual information, and refine our observations on how miscommunication shapes the development of alignment.

Again, chi-square analysis was carried out to discover whether there was a significant relationship between the three variables: number of new words in an utterance (0 or 1 to many), type of previous utterance (non-problematic or problematic) and visual information (Monitor or No Monitor condition). The resulting test indicated a significant association between occurrence of miscommunication and lexical innovation, but only in the No Monitor condition ($\chi^2_{(1)} = 15.71$, $p < .001$), and was confirmed by the linear chi-square ($M^2 = 15.70$). Under both conditions, only around 34% of the utterances contained new words when communication was smooth. However, when a problem occurred, this figure climbed to 54% in the No Monitor condition. The odds ratio indicated that, if visual information was withheld, new words were 2.33 times more likely to be introduced after miscommunication. Figure 6 illustrates that the probability of introducing new words is elevated after miscommunication, whereas it is most likely that users draw their vocabulary from the preceding dialogue in cases where the communication is problem-free. The number of utterances with new words also rose, to 44%, in the Monitor condition, but failed to yield a significant result ($\chi^2_{(1)} = 1.78$, $p = .182$). The results of both data sets (Monitor and No Monitor) are shown in Figure 6, indicating more pronounced differences in the No Monitor condition (the right graph).

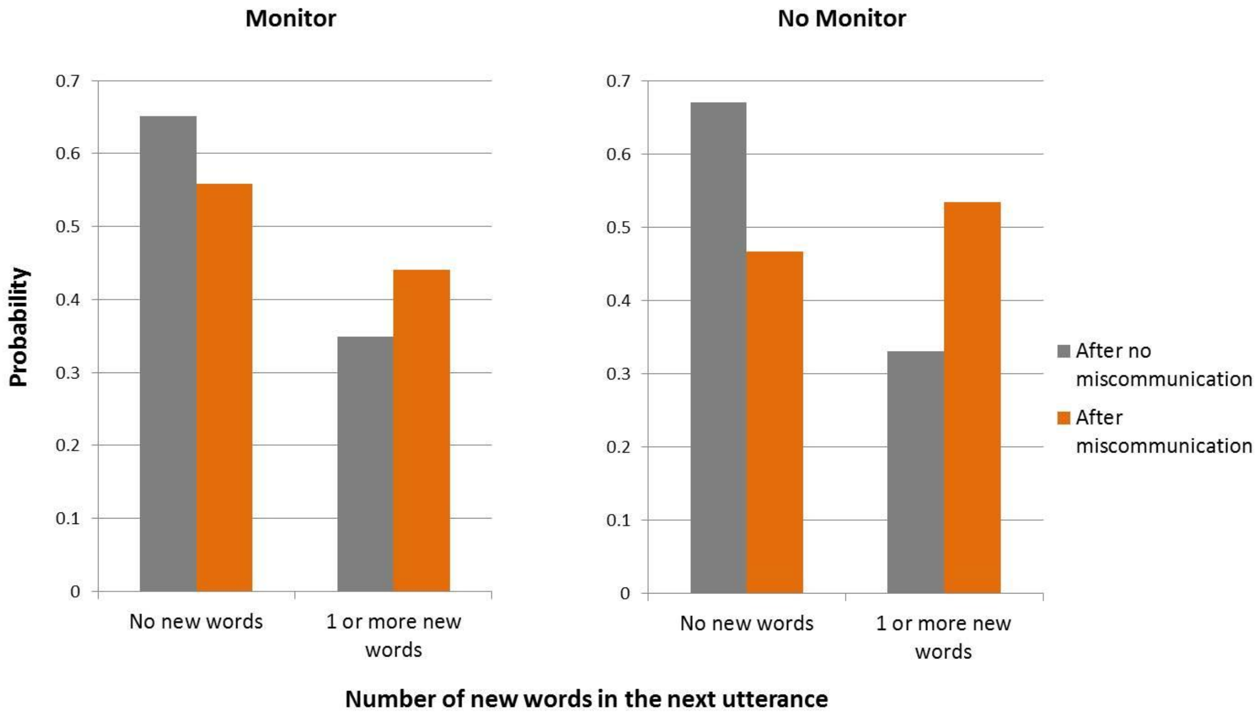


Figure 6. Probability of occurrence of new words after non-problematic and problematic utterances in the Monitor condition (the left graph) and No Monitor condition (the right graph). Probabilities are calculated as the ratio of actual count over total number of utterances.

Taken together, the results confirm that the development of alignment is locally disrupted by the occurrence of miscommunication. Users do not tend to resort to expressions that were previously used and successfully understood, but instead tend to introduce novel words. Further, this effect was pronounced in the condition in which users had no visual information (the No Monitor condition), providing further evidence of the strategic nature of alignment.

5.5 Alignment and user perceptions of interaction success

The final research hypothesis looked at user perceived interaction success. As described in section 4.2.4, users completed a seven-point Likert-scale questionnaire in which they rated their agreement with five statements. The values for each statement were summed for all six tasks and correlational analysis was performed for lexical innovation (ratio of unique words in the final task). Though the use of parametric or non-parametric tests on rating scores has been a controversial issue, Likert scale data are commonly and legitimately treated as if they were interval (Gravetter & Forzano, 2012, p. 92; Norman, 2010). Employing such an approach has been recommended by HCI practitioners and applied statisticians (Sauro & Lewis, 2012, p. 243-246; Lewis, 1993) and was therefore adopted in this study. The analysis revealed a significant negative correlation between user experience of task success ("*I did well in completing the task*") and lexical innovation ($r_{(30)}$ = - .47, p = .013). That is, users perceived that the interaction was less successful when alignment was weaker. The analysis failed to reveal significant relationships between the other statements. Yet, as expected, all statements were negatively correlated with higher frequency of non-understandings and execution errors. These results are summarised in the correlation matrix provided in Table 4.

	Statement 1	Statement 2	Statement 3	Statement 4	Statement 5
Lexical innovation	**$r_{(30)}$ = - .47, p = .013**	$r_{(30)}$ = - .29, p = .127	$r_{(30)}$ = - .12, p = .532	$r_{(30)}$ = - .19, p = .308	$r_{(30)}$ = - .16, p = .396
Execution Errors and Non-understandings	**$r_{(30)}$ = -.49, p = .015**	**$r_{(30)}$ = -.62, p = .001**	**$r_{(30)}$ = -.52, p = .003**	**$r_{(30)}$ = -.51, p = .004**	**$r_{(30)}$ = -.72, p = .001**

Table 4. Correlation matrix showing the correlations between the questionnaire statements and lexical innovation/execution errors and non-understandings. The significant correlations appear in bold font.

5.6 Summary of results

Table 5 lists the five research hypotheses tested in the study and summarises the respective outcomes, with the right-hand column giving the number of the sub-section where the relevant results were presented. The results reported in these sub-sections provide insights into the local and global processes of alignment in a user's dialogue with a system. First, the stabilisation of working vocabulary early in the interaction reveals the operation of alignment between speakers that settle on a set of grounded expressions for dealing with the ensuing dialogue. Second, the analysis of the experimental data confirmed that the magnitude of alignment is reciprocal, with interlocutors aligning to each other at similar rates. Third, analysis of data from two different visual co-presence conditions produced evidence that may also indicate that alignment in human-computer dialogues has a strategic component. That is, in the absence of visual evidence of understanding, correct execution and joint reference, speakers tended to adapt their linguistic choices more strongly, possibly in an effort to compensate for the lack of this resource and in an attempt to enhance (the impoverished) communication. Fourth, the development of alignment is locally disrupted by the occurrence of miscommunication such that novel words are introduced instead of falling back on previously used vocabulary. Users and 'robots' converged in shorter vocabularies when user errors were lower. Yet, while the lack of visual feedback promoted alignment, when miscommunication occurred users were considerably less likely to draw from the grounded expressions. Finally, analysis of the user perception data revealed that users rated their performance less favourably when alignment was weaker. The next section of this paper will discuss the implications of these results for the development of natural language interfaces to computer systems.

Research Hypothesis	High-level Result	Sub-section
1. *Alignment occurs in the interaction between a human user and a*	Confirmed; vocabulary stabilised early in the dialogue suggesting the operation of	5.1

computer system.	alignment.	
2. *Alignment occurs as a mutual phenomenon.*	Confirmed; 'robots' and users aligned to each other at similar rates.	5.2
3. *Visual feedback influences alignment between a user and a system.*	Confirmed; 'robots' and users aligned more strongly in the absence of visual feedback.	5.3
4. Miscommunication locally disrupts the process of alignment in human-computer communication.	Confirmed; the development of alignment was *locally* disrupted; new vocabulary was introduced after miscommunication.	5.4
5. Lower alignment is linked to lower user perception of interaction success.	Confirmed; lower task success perceptions are associated with higher final lexical innovation.	5.5

Table 5. List of research hypotheses and respective results.

6 Discussion

There are at least three important reasons for seeking to better understand and characterise alignment in human-computer dialogues. First, better understanding of processes that play a part in the interaction between users and computer systems may help to inform more naturalistic system designs. Second, if alignment is indeed a precondition for communicative success, systems that do not support this mechanism are destined to fail. Third, alignment may help 'prime' desirable user input and inhibit out-of grammar words. These issues are discussed in the following sub-sections where the findings from this study are translated into design recommendations which are subsequently used to inform the development of a framework of dialogue management that incorporates linguistic alignment.

6.1 Alignment in human-computer communication develops early and reciprocally

Section 5.2 reported a one-to-one coupling of user and 'robot' inputs at the adjacency pair level. The analysis demonstrated a trend, according to which the more aligned one participant is, the more aligned their partner will be. Hence, it is likely that a computer dialogue system which consistently matches the input of the user will trigger similar user tactics. In turn, as these expressions become grounded, the use of different lexical items by the user may well be more inhibited. In addition to local priming, the analysis in section 5.1 demonstrated its operation over the course of the dialogue: the interlocutors, although presented with different landmarks and environment configurations during the session, began to rely more and more on previously-used expressions. This led to a small-sized working vocabulary that peaked and stabilised after only 70 dialogue turns. As such, speakers simply drew from the preceding dialogue to formulate future utterances. Taken together, these observations provide strong evidence that alignment operates in human-computer dialogues through both local priming and longer-lasting alignment of vocabulary.

In summary, there is symmetry in the linguistic input and output of system and user which gains stability over time. That is, the user aligns with the system and the system aligns with the user at the utterance pair level, which eventually results in a relatively stable set of expressions that are being re-used. As such, alignment appears instrumental in addressing the ‘Vocabulary Problem’, allowing prediction and constraint of the linguistic input of the user. These observations suggest that, through their output, dialogue managers should seek to prime users such that they are more likely to input in-grammar terms and structures. Production and interpretation are coupled processes, so system prompts should contain no syntactic or lexical items that the system itself cannot interpret. In addition to this, specific design issues arise with regards to how the system’s dialogue manager supports lexical alignment in order to restrict the vocabulary in use, and these will be considered in section 6.5 as part of a proposed dialogue model.

6.2 Lack of alignment is linked to lower user perceptions of task success.

Previous work in human communication emphasises that linguistic alignment is the basis of stable, successful communication (Pickering & Garrod, 2004; 2006). Reitter and Moore’s (2007a) findings support this, reporting a strong correlation of task success and long-term alignment of syntactic structures, though no effect was found for local priming, and concluding that lexical and syntactic alignment is a reliable predictor of task success, and that “successful dialogue requires syntactic alignment” between human interlocutors in a spatial task (Reitter & Moore, 2007b, p.1). The question that naturally follows from the analysis of relevant work in human communication, and which motivated this study, is whether alignment is also a precondition for successful communication with computer systems. The results presented in section 5.5 suggest that it is, demonstrating a link between lower perceived task success and lower lexical alignment achieved by the end of the dialogue. While there is literature that reports that systems that aligned to their users in terms of prosodic or other paralinguistic elements are rated more positively (e.g., Nass & Lee, 2001; Bailenson & Yee, 2005), to our knowledge no other study has presented evidence that interactions are perceived to be more successful when systems align to users.

Taken together, while the results of this study are correlational, in view of a strong basis of previous empirical and theoretical findings, they argue for a potential effect of alignment on perceived communication success. In effect, they reverse the priorities, bringing the role of system-generated responses into the foreground, and suggest that alignment by the computer system is of key importance to the success of the interaction. As such, though important, system prompts designed to prime the user to provide desirable input (as recommended in section 6.1) may not suffice to yield effective interactions. Rather, it is suggested that alignment can be instrumental in interaction success if the system is also primed to repeat user output. This suggests that, through their output, dialogue managers should seek to repeat user outputs to promote alignment. This recommendation will be revisited in section 6.5 to explore its place in the development of a dialogue management model.

While interesting for the purposes of this exploratory study, these results remain preliminary, given that they were produced by correlational analyses. On the basis of the results, it is possible to argue for an association, but it remains unknown whether low success perception is because of low alignment. To give evidence of causation, it would be necessary to replicate this study using appropriate experimental manipulations in order to test the directional hypothesis that *'aligned robot responses increase user satisfaction'*. This could be achieved by the replication of the study involving two groups of trained 'robots' instructed to either systematically repeat the same lexical items as the user or use different forms, and measuring the effect in terms of user perceptions.

6.3 The effect of visual feedback on alignment in HCI

Studies by Brennan, and Branigan and her colleagues (discussed in section 2) have demonstrated strong presence of linguistic alignment in HCI which suggests that it is an automatic mechanism that invariably manifests in communication. Later research has added that it is also a strategy that is consciously-employed based on the speaker's beliefs about the linguistic competence of the interlocutor (for example, in the case where users aligned more to 'basic' computers than to 'advanced' ones and more to computers than to human partners (see Pearson, Hu, Branigan, Pickering, & Nass, 2006)). As one explanation, Branigan, Pickering, Pearson, and McLean (2010) have suggested that since computers are perceived as less competent interlocutors, alignment is more prevalent in HCI than human-human interaction, and has a stronger strategic component. Unifying this body of results, Branigan, Pickering, Pearson, McLean, & Brown (2011) concluded that lexical alignment is mediated by beliefs about interlocutors, and that speakers align more strongly when they believe that this will facilitate interaction success.

It is difficult to interpret the findings of the present study to contribute to the debate around the nature of alignment. Yet, from a different standpoint, they reiterate the conclusions offered by Branigan and her colleagues. The analysis presented in section 5.3 showed that the extent of alignment in HCI was determined by the interaction condition; in particular, alignment was prevalent when visual feedback was absent, and yet comparatively scarce in the condition of visual co-presence. When users could not readily establish joint reference, monitor task status or have instantaneous evidence of the system's understanding and execution, speakers aligned more strongly. Therefore, the results of this study add weight to those previous findings that argue that alignment in HCI is used when communication success appears to be at risk and as a 'safeguard' against a perceived elevated likelihood of miscommunication.

From a wider practical perspective, awareness of how visual information affects collaboration and communication patterns is important for the design of CMC, CSCW systems and agents in situated interactions. Previous studies in CMC have discussed how visual information (particularly of the work area) increases awareness of the current state of the task and facilitates conversation and grounding, such that interlocutors can use linguistic shortcuts and simpler language (see, for instance, Gergle, Kraut & Fussell, 2013). It was found that it profoundly changes the structure and content of dialogue, since utterances may be substituted or complemented by actions and gestures[1]. Inspection of the dialogue corpus of the present study reiterates these observations and extends them to the domain of human-computer dialogues; when visual feedback is withheld, interlocutors tend to use more explicit and longer utterances (i.e., more words). Adding to this, the results of the analysis showed that users and 'robots' also systematically repeat each other's words. This is exemplified in the dialogue excerpt in Table 6 below.

Utterance
U: Now turn to your right, walk straight ahead until you reach the ***road junction***
R: Yes, I am at the ***road junction*** now
U: Once you are at this ***road junction***, please turn to your left and walk straight ahead until you reach a right turning in the road
R: Ok, I can see a ***right turn***
U: Good, please turn right at this ***right turn*** and follow the road until you reach a ***roundabout***
R: Ok
R: I am in the ***roundabout*** now
U: Good, you are at the ***roundabout.***
U: Is there a ***car park on your left***?
R: Yes
R: The ***car park*** is ***on my left*** now
U: Please turn to your left and take the first left exit off the roundabout, keeping ***the car park on your left***

Table 6. Dialogue excerpt from the No Monitor condition.

Qualitative examination of the dialogues also revealed an interesting phenomenon; in a few exchanges in the No Monitor condition, users were inclined to repeat even the erroneous or idiosyncratic vocabulary of the 'robots' (as illustrated in the example in Table 7). This observation appears to echo the findings from the aforementioned studies by Branigan and colleagues, which indicated that human users tended to repeat the term that the computer used, even if it was less accurate or normal. As noted in the previous section, it may be worthwhile to explore the validity and extent of this phenomenon and its specific impact on aspects of interaction success using a controlled experimental setup, in which 'robots' systematically produce such terms.

U: turn left
R: Go pass the bridge?
U: go forward
R: ***Go pass*** the town hall?

[1] As expected, in the Monitor condition, many responses by the robot were carried out through a physical action rather than verbal means. It should be clarified that this was not annotated as a mismatch.

*U: **go pass*** Tesco

Table 7. Dialogue excerpt from the No Monitor condition.

This study, then, has illustrated the impact of visual feedback on the patterns of communication when other parameters are kept the same. The results suggest that users align more strongly to systems when visual feedback is not possible, increasing the necessity to implement linguistic alignment capabilities in the dialogue manager of systems that are not physically or visually co-present with their users.

6.4 The effect of miscommunication

Miscommunication is a natural and ubiquitous phenomenon within communication, both between humans and, perhaps even more so, in computer-based dialogue systems. In interaction with such systems, miscommunication manifests as user errors, system errors and non-understandings. The ability to predict what users will do in terms of linguistic choices after the occurrence of errors is a matter of enormous practical significance. Addressing the fourth research hypothesis, section 5.4 explored how users reacted when they detected miscommunication.

It was found that, after miscommunication, users were more likely to use new words, whereas successful utterances were typically followed by responses that exclusively reused lexical items from the dialogue history. A simple explanation of this phenomenon is that, as the dialogue progresses, interlocutors build up a body of aligned expressions that seems to be mutually intelligible and that functions successfully. When miscommunication occurs, interlocutors lose confidence in the efficiency of these expressions and the interaction as a whole and introduce new expressions. This user behaviour was more pronounced when visual feedback was absent. This is likely to be because visual evidence offers a more effective and economic way of grounding compared to verbal-only evidence (Brennan, 2005). Thus, it can be argued that the status of lexical items that are grounded under a visual co-presence condition is less susceptible to the impact of miscommunication.

Two specific recommendations can be drawn from these findings. First, as suggested in section 6.3, dialogue managers should account for different interaction conditions of visual and verbal-only feedback. In particular, when miscommunication is detected in visual co-presence conditions the system should adhere to the vocabulary established in the course of the dialogue. In verbal-only conditions, the system should anticipate novel words in the user input, and ‘expect’ departure from those previously recorded in the dialogue history.

The second recommendation concerns the miscommunication (or error) handling functionalities of the dialogue manager. The efficiency of dialogue systems is often compromised by their inability to detect speech recognition and language understanding errors. In turn, it has been found that humans do not typically provide explicit cues that a misunderstanding has occurred, but prefer implicit strategies such as reformulating their statements or even moving on (Skantze, 2005; Koulouri & Lauria, 2009; Bohus & Rudnicky, 2005). Therefore, the detection of out-of-vocabulary words may be used by the dialogue system as an indicator that an error has occurred.

Along with those from sections 6.1-6.3, these recommendations will be incorporated in the dialogue model discussed in the following section.

As a final note, it is interesting to refer to studies in human communication that have proposed that, despite a local disruption, miscommunication may in fact accelerate *global semantic* alignment (Healey and Mills, 2006; Mills, 2007); by interactively repairing problems, interlocutors are able to converge more on their semantic models. Along the same lines, Martinovsky and Traum (2003) suggested that through miscommunication, interlocutors gain awareness of the state and capabilities of each other. In this light, miscommunication between humans and computers is not seen as a pathological phenomenon that should be prevented, but as a key component of longer-term successful interactions. This observation acquires increased significance, given the general ignorance of users with regards to computers as interlocutors. Thus, it is argued that research efforts should be redirected from trying to eradicate all possible errors to designing dialogue managers with the capability to detect errors and to make the corresponding update. This insight served as the initial motivation for the model presented in the next section.

6.5 Towards an alignment-driven approach to dialogue management

Dialogue systems are typically based on modular pipeline architectures. Depending on the application domain, a basic architecture consists of modules for natural language understanding (including components such as speech recognition and language parsing), natural language generation (including speech synthesis) and dialogue management. In the case of spoken dialogue systems, the speech signal is captured and the speech recogniser produces a hypothesis which is passed to the natural language understanding (NLU) component. Speech recognition and NLU typically use language modelling to predict the next word given the identities of the previous words. The NLU component parses this input and submits a semantic representation to the dialogue manager, which determines the next system action, based on the dialogue history and other knowledge sources. This action is forwarded to the natural language generation (NLG) component which creates a system response. The speech synthesiser outputs the response. Text-based dialogue systems omit speech recognisers and synthesisers but use the rest of the core architecture. The NLU and NLG components typically use static data from application-specific grammars and lexicons – the set of allowed structures and words (sometimes collectively referred to as grammar). The dialogue manager also makes use of the same linguistic resources. Figure 7 summarises the interactions between the modules in such an architecture.

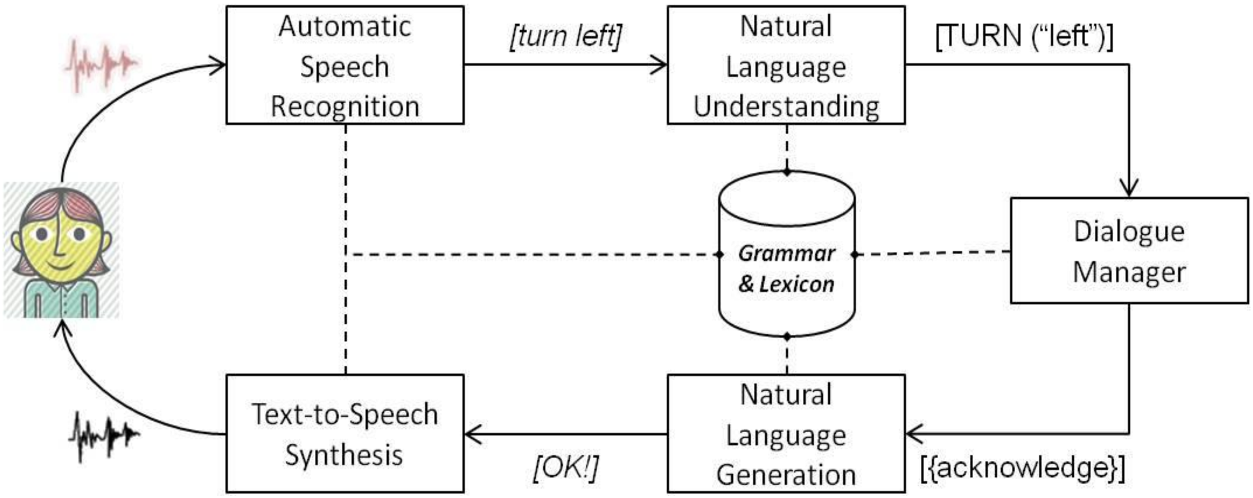


Figure 7. A typical dialogue system architecture. Text-based dialogue systems omit the speech recognition and synthesis modules.

After this brief review of the architecture and technologies of a typical dialogue system, this paper concludes by incorporating the recommendations detailed in the previous sub-sections into a high-level dialogue model for task-oriented interactions with a computer-based system. In particular, the model focuses on the dialogue manager's interaction with the grammar for determining the content of the next system action and adapting the lexicon, as a result of the processes of linguistic alignment on which the study reported in this paper focused.

The dialogue manager's operation based on the proposed model will be illustrated though a simplified dialogue example from a human-robot supervised navigation scenario. The dialogue example corresponds to a task completed within one transaction: a user utterance instructing the robot to turn left at a junction, and the robot executing the instruction. Based on empirically collected data, the environmental feature, *junction*, was more or less accurately referred to as 'v-shaped junction', 'three-way junction', 'y-junction', 'intersection', 'crossroad', 'cross junction', 'fork' and 't-junction' by different users (as observed in this study). At the beginning of the dialogue, the grammar contains all possible synonymous lexical items. A weighting feature is assigned to each lexical item, indicating its frequency of use in the dialogue. Thus, all lexical items begin by having equivalent weightings.

The user initiates the interaction using the instruction "*turn left at the fork*". At this point, there are three communication outcomes: (i) correct understanding; (ii) non-understanding; or (iii) misunderstanding. In the cases of correct understanding and non-understanding, the system gives positive or negative evidence of understanding, respectively.

First, in the case of correct understanding, the dialogue manager triggers a verbal acknowledgement followed by the physical action of the system. The execution is based on particular expressions that referred to actions and objects in the interaction situation. If the understanding was indeed correct, as evidenced by the user acknowledging successful execution, the expression is taken to be conceptually-equivalent for both user and system to refer to the relevant actions and objects. As such, the dialogue manager should perform two *grammar* *updates,* which reinforce the use of this lexical item in subsequent similar situations: (i) the expression should be mapped to a particular situation (object or action); and (ii) the expression's weighting should be increased, meaning that it will subsequently be favoured over synonymous expressions in the grammar.

Then, following the basic 'input/output alignment' principle in the Interactive Alignment Model and the recommendations in sections 6.1 and 6.2, the system should immediately repeat the expression by generating a verbal acknowledgement which reinforces the expression used (i.e., *"I have turned left at the fork*"). This system output, in turn, should further prime the user to re-use the expression to refer to this object, inhibiting the use of any alternative term. This will eventually lead to the particular expression becoming 'fixed', and routine for this dialogue (Pickering & Garrod, 2004). As described in section 2, 'routines' (following the Interactive Alignment Model) or 'conceptual pacts' (following the Collaborative Model) are linguistic constructs that are agreed between the interlocutors to refer to an entity in the situation model. Following the process described so far, as the dialogue progresses the working grammar will be gradually reduced in variation and size, with some expressions being dispreferred and others being favoured until, ideally, the grammar becomes stabilised and only consists of dialogue routines.

Second, in the case in which the instruction is not understood, the dialogue manager will implement the strategy specified in the error recovery module of the dialogue manager (strategies include asking the user to repeat or rephrase the problematic utterance, or, if the system has advanced inferential capabilities, asking task-level reformulations, such as "turn left after the bridge?" (see Bohus & Rudnicky, 2005; Gabsdil, 2003)). The results in this study suggest that when miscommunication occurs, users lose confidence in the efficiency of established dialogue routines and introduce new expressions (see sub-section 6.4). Therefore, in case of non-understanding, the initial system response should be not to increase the weighting of any expressions used.

Similarly, no grammar update is performed in cases of misunderstanding (execution errors in the user/robot scenario from the study in this paper).

The observation that users tend to introduce new lexical items when they perceive an error may also be translated into a guideline for late error detection. Drawing on the recommendation framed in section 6.4, it is suggested that late error detection approaches should include monitoring for the presence of alternative lexical items in user turns (that is, words that currently hold lower weightings in the grammar compared to the most frequently-used expression for a situation) as a valid negative cue to detect errors (in combination with other typically used cues such as longer turns, word order, rejections etc.).

In summary, this section has outlined a high-level model to illustrate how linguistic alignment can be supported by the dialogue manager. The dialogue manager performs two types of update as a function of the usage of an expression over the course of a single dialogue: it creates an association between the lexical item and a referent; and changes its weighting within the lexicon. Possible benefits of the suggested approach include: enhanced recognition accuracy, owing to rescoring of word probabilities based on their weightings; improved intelligibility of system generated output, owing to it consisting of recurring words; and user interaction with the system that is more natural and cognitively easy. Although this study and framework focus on lexical alignment, alignment is expected to operate in comparable ways across all other linguistic levels. Therefore, it could be extended to apply to, for example, syntactic structures.

7 Summary and future work

The nature of this research and the hypotheses it aimed to address motivated the experimental approach. Our methodological decisions, however, encompass potential limitations, which, in turn, lead to reflections on ways in which this work could be advanced. First, a potential criticism is whether the validity and extensibility of the experimental results are limited owing to the differences between text and spoken utterances as modalities. Studies (Brennan, 1996; Branigan, Pickering, Pearson, McLean, & Brown, 2011) have confirmed that modality had no effect on alignment, but, even if this were not the case, the study would be useful given the immediate practical relevance that any findings would offer for text-based interaction with a computer system or computer-mediated communication between people. Second, valid questions may arise about whether the lack of trained confederate(s) and dialogue script limit the generalizability of the results to HCI. Indeed, in order to move closer to the state-of-the-art of agent technologies, future work should replicate the experiment using a 'typical' WOz setup, in which 'robots' are trained to use either the same or a different lexical item at given points in the developing dialogue.

Speakers tend to repeat their own and each other's linguistic choices in dialogue, a phenomenon which arguably underlies communication success. However, with its occurrence and effects in human-computer interaction remaining ill-defined, the practical benefits of alignment have been unexploited. The study reported in this paper has drawn on the Interactive Alignment Model and existing work in HCI in order to investigate alignment in task-oriented dialogues with computer systems. The experimental data, obtained from naturalistic human-robot navigation dialogues, have helped to address important questions about the operation and role of alignment in the effectiveness and success of the interaction. In addition, the analysis has led to design guidelines which were subsequently used in the development of a simple alignment-driven approach for dialogue management. It is hoped that the model presented in this paper will serve as a starting point for exploring the potential of alignment within computer-based dialogue models and system implementations. Building on this work, our future research aims to capture syntactic alignment and integrate it into the proposed model, and to develop a platform for the implementation and evaluation of the model.

References

Amalberti, R., Carbonell, N., & Falzon, P. (1993). User representations of computer systems in human–computer speech interaction. *International Journal of Man–Machine Studies, 38*(4), 547–566.

Bailenson, J. N., & Yee, N. (2005). Digital chameleons—automatic assimilation of nonverbal gestures in immersive virtual environments. *Psychological Science, 16*(10), 814–819.

Bargh, J. A. (1989). Conditional automaticity: varieties of automatic influence on social perception and cognition. In J. S. Uleman, J. S. & J.A Bargh, (Eds.), *Unintended thoughts*, (3–51). Guilford Press.

Bell, L., Gustafson, J., & Heldner, M. (2003). Prosodic adaptation in human–computer interaction. *Proceedings of the International Congress of Phonetic Sciences*. 2453–2456.

Bohus, D. & Rudnicky, A. I. (2005). Sorry, I didn't catch that! – an investigation of non-understanding errors and recovery strategies. *Proceedings of SIGdial2005*. Lisbon, Portugal.

Branigan, H.P & Pearson J. (2006). Alignment in Human-Computer Interaction. In K. Fischer (Ed.) *Proceedings of the Workshop on How People Talk to Computers, Robots, and Other Artificial Communication Partners* (140-156). Delmenhorst, Germany: HWK.

Branigan, H. P., Pickering, M., & Cleland, A. A. (2000). Syntactic co-ordination in dialogue. *Cognition, 75*(B), 13–25.

Branigan, H.P., Pickering, M.J., Pearson, J. & McLean, J.F. (2010). Linguistic alignment between humans and computers. *Journal of Pragmatics, 42*, 2355–2368.

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*(1), 41-57.

Branigan, H.P., Pickering, M.J., Pearson, J.M., McLean, J.F., & Nass, C. (2003) Syntactic alignment between computers and people: the role of belief about mental states. *Proceedings of the Twenty-fifth Annual Conference of the Cognitive Science Society*. (186-191). Mahwah: Erlbaum.

Branigan, H.P., Pickering, M.J., Pearson, J., McLean, J.F., Nass, C.I., & Hu, J. (2004). Beliefs about mental states in lexical and syntactic alignment: Evidence from human-computer dialogs. *Proceedings of the CUNY Conference on Human Sentence Processing.*

Brennan, S. E. (1996). Lexical entrainment in spontaneous dialog. *Proceedings of the International Symposium on Spoken Dialogue*, 41–44.

Brennan, S. E. (2005). How conversation is shaped by visual and spoken evidence. In J. C. Trueswell & M. K. Tanenhaus (Eds.), *Approaches to studying world-situated language use: Bridging the language-as-product and language-as-action traditions.* (95-129). Cambridge, MA: MIT Press.

Brennan, S. & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *22*, 1482–1493.

Brown, Paula M., and Gary S. Dell. (1987): Adapting production to comprehension: The explicit mention of instruments. *Cognitive Psychology*. *19*(4). 441-472.

Clark, E. V. (1993). *The Lexicon in Acquisition*. Cambridge: Cambridge University Press.

Clark, H. H. (1996). *Using Language*. New York, US: Cambridge University Press.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. *Perspectives on socially shared cognition, 13*, 127-149.

Clark, H.H., & Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Memory & Language J., 50*, 62-81.

Clark, H. H., & Murphy, G. L. (1982). Audience design in meaning and reference. *Language and comprehension, 9*, 287-299.

Cohen, M.H., Giangola, J.P. & Balough, J. (2004). *Voice User Interface Design*. Addison Wesley.

Cowan, B., Beale, R., & Branigan, H.P. (2011). Investigating syntactic alignment in spoken natural language human-computer communication. *Proceedings of the ACM CHI conference on Human Factors in computing systems,* Vancouver.

Fraser, N.M., & Gilbert, G.N. (1991). Simulating speech systems. *Computer Speech and Language, 5*, 81–99.

Furnas, G. W., Landauer, T. K., Gomez, L. M., & Dumais, S. T. (1987). The vocabulary problem in human-system communication. *Communications of the ACM, 30*(11), 964-971.

Gabsdil, M. (2003). Clarification in spoken dialogue systems. *Proceedings of the 2003 AAAI Spring Symposium Workshop on Natural Language Generation in Spoken and Written Dialogue*, Stanford, USA.

Garcia, A., & Jacobs, J. B. (1998). The interactional organization of computer mediated communication in the college classroom. *Qualitative Sociology*, *21*(3), 299-317.

Garrod, S. and Anderson, A. (1987). Saying what you mean in dialogue: a study in conceptual and semantic co-ordination. *Cognition*, *27*, 181–218.

Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy? *Trends in Cognitive Sciences, 8*, 8-11.

Gergle, D., Kraut, R. E., & Fussell, S. R. (2013). Using visual information for grounding and awareness in collaborative tasks. *Human-Computer Interaction.28* (1).

Gergle, D., Kraut, R. E., & Fussell, S. R. (2004). Language efficiency and visual technology: minimizing collaborative effort with visual information. Journal of Language & Social Psychology, 23(4), 491-517. Thousand Oaks, CA: Sage Publications.

Giles, H., Coupland, N., & Coupland, J. (1991). Accommodation theory: communication, context, and consequence. In H. Giles, J. Coupland, & N. Coupland (Eds.), *Contexts of Accommodation: Developments in Applied Sociolinguistics.* (1-68). Cambridge: Cambridge University Press.

Gravetter, F. J. & Forzano, L. B. (2012). *Research methods for the behavioral sciences* (4th ed.). Belmond, CA,USA: Cengage Learning.

Healey, P. G., & Mills, G. (2006). Participation, precedence and co-ordination in dialogue. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 1470-1475).

Hirst, G., McRoy, S., Heeman, P., Edmonds, P. & Horton, D. (1994). Repairing conversational misunderstandings and non-understandings. *Speech Communication, 15*(3-4), 213-229.

Isaacs, E. A., & Clark, H. H. (1987). References in conversation between experts and novices. *Journal of Experimental Psychology: General, 116*(1), 26.

Koulouri, T. & Lauria, S. (2009). Exploring miscommunication and collaborative behaviour in Human-Robot Interaction. In M. Purver (Ed.). *Proceedings of SIGdial09.* Association for Computational Linguistics, (111-119). Stroudsburg, PA, USA.

Kraut, R. E., Fussell, S. R., & Siegel, J. (2003). Visual information as a conversational resource in collaborative physical tasks. *Human Computer Interaction*, *18,* 13-49.

Kraut, R. E., Gergle, D., & Fussell, S. R. (2002, November). The use of visual information in shared visual spaces: Informing the development of virtual co-presence. In *Proceedings of the 2002 ACM conference on Computer supported cooperative work* (pp. 31-40). ACM.

Lazar, J., Feng, J. H., & Hochheiser, H. (2010). *Research methods in human-computer interaction*. Wiley.

Levin, D. T., Killingsworth, S. S., & Saylor, M. M. (2008). Concepts about the capabilities of computers and robots: A test of the scope of adults' theory of mind. In *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction* (pp. 57-64). ACM.

Levin, D. T., Killingsworth, S. S., Saylor, M. M., Gordon, S. M., & Kawamura, K. (2013). Tests of concepts about different kinds of minds: Predictions about the behavior of computers, robots, and people. *Human–Computer Interaction*, 28, 161-191.

Levinson, S. C. (1983). *Pragmatics.* Cambridge, England: Cambridge University.

Lewis, J. R. (1993). Multipoint scales: mean and median differences and observed significance levels. *International Journal of Human-Computer Interaction, 5*(4), 383-392.

Martinovsky, B., & Traum, D. (2006). The error is the clue: Breakdown in human-machine interaction. In *Proceedings of the ISCA Workshop on Error Handling in Dialogue Systems.*

McTear, M. (2008). *Handling miscommunication in spoken dialogue systems: why bother?* In L. Dybkjaer , L & W. Minker (Eds.), *Recent Trends in Discourse and Dialogue.* (101 – 122). Springer.

Mills, G.J. (2007). *Semantic co-ordination in dialogue: the role of direct interaction.* (Doctoral dissertation, Queen Mary University, London, UK). Retrieved from http://www.dcs.qmul.ac.uk/tech_reports/RR-07-07.pdf.

Moon, Y., & Nass, C. I. (1996). How ‘‘real’’ are computer personalities? Psychological responses to personality types in human–computer interaction. *Communication Research, 23*(6), 651–674.

Nass, C. I., & Lee, K. M. (2001). Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied, 7*(3), 171–181.

Nass, C. I., & Moon, Y. (2000). Machines and mindlessness: social responses to computers. *Journal of Social Issues, 56*(1), 81–103.

Nass, C. I. , Moon, Y., & Carney, P. (1999). Are people polite to computers? Responses to computer-based interviewing systems. *Journal of Applied Social Psychology, 29*(5), 1093–1110.

Norman, G. (2010) Likert scales, levels of measurement and the “laws” of statistics. *Adv Health Sci Educ., 15*, 625–632.

Oulasvirta, A., Engelbrecht, K-P., Jameson, A., & Möller, S. (2006). The relationship between user errors and perceived usability of a spoken dialogue system. *The 2nd ISCA/DEGA Tutorial & Research Workshop on Perceptual Quality of Systems.* Berlin, Germany.

Oviatt, S. L., Darves, C., & Coulston, R. (2004). Toward adaptive conversational interfaces: modeling speech convergence with animated personas. *ACM Transactions on Computer–Human Interaction, 11*(3), 300–328.

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America, 119*, 2382-2393.

Pearson, J., Hu, J., Branigan, H.P., Pickering, M.J., & Nass, C.I. (2006). Adaptive language behavior in HCI: how expectations and beliefs about a system affect users' word choice. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* 1177–1180. New York: ACM.

Pickering, M. J. & Garrod, S. (2004). Towards a mechanistic psychology of dialogue. *Behavioural and Brain Sciences*, *27*(2), 169–190.

Pickering, M.J., & Garrod, S. (2006). Alignment as the basis for successful communication. *Research on Language and Computation, 4*, 203-228.

Porzel, R. (2006). How people (should) talk to computers. In K. Fischer (Ed.) *Proceedings of the Workshop on How People Talk to Computers, Robots, and Other Artificial Communication Partners* (7-37). Delmenhorst, Germany: HWK.

Reitter, D. & Moore, J.D. (2007a). Predicting success in dialogue. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 808–815. Prague, Czech Republic: Association for Computational Linguistics.

Reitter, D. & Moore, J.D. (2007b). Successful dialogue requires syntactic alignment. *20th Annual CUNY Conference on Human Sentence Processing.*

Sauro, J., & Lewis, J. R. (2012). *Practical statistics for user research.* Burlington, MA: Morgan Kaufmann.

Schober, M. F. (1993). Spatial perspective-taking in conversation. *Cognition*, *47*(1), 1–24.

Schober, M. F. & Clark, H. H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology*, *21*, 211–232.

Skantze G. (2005). Exploring human error recovery strategies: implications for spoken dialogue systems. *Speech Communication, 45*(3), 207-359.

Suzuki, N., & Katagiri, Y. (2007). Prosodic alignment in human–computer interaction. *Connection Science*, *19*(2), 131–141.

Tannen, Deborah L. (1987). Repetition in conversation: toward a poetics of talk. *Language*, *63*(3), 574–605.

Vanetti, E. J. & Allen, G. L. (1988). Communicating environmental knowledge: The impact of verbal and spatial abilities on the production and comprehension of route directions. *Environment and Behavior*, *20*, 667-682.

Ward, N., & Nakagawa, S. (2002). Automatic user-adaptive speaking rate selection for information delivery. *Proceedings of the International Conference on Spoken Language Processing*, 549–552.

Williams, J.D., & Young, S. (2004). Characterizing task-oriented dialog using a simulated ASR channel. *Proceedings of the International Conference on Spoken Language Processing*, Jeju, South Korea.