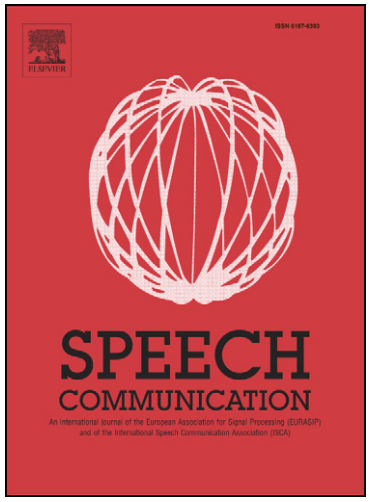
Towards human-like spoken dialogue systems

Jens Edlund, Joakim Gustafson, Mattias Heldner, Anna Hjalmarsson



Towards human-like spoken dialogue systems

Jens Edlund, Joakim Gustafson, Mattias Heldner, Anna Hjalmarsson

PII: S0167-6393(08)00054-X
DOI: 10.1016/j.specom.2008.04.002
Reference: SPECOM 1707

To appear in: *Speech Communication*

Received Date: 31 August 2007
Revised Date: 5 April 2008
Accepted Date: 6 April 2008

Please cite this article as: Edlund, J., Gustafson, J., Heldner, M., Hjalmarsson, A., Towards human-like spoken dialogue systems, *Speech Communication* (2008), doi: 10.1016/j.specom.2008.04.002

Abstract:

This paper presents an overview of methods that can be used to collect and analyse data on user responses to spoken dialogue system components intended to increase human-likeness, and to evaluate how well the components succeed in reaching that goal. *Wizard-of-Oz variations*, *human-human data manipulation*, and *micro-domains* are discussed in this context, as is the use of third-party reviewers to get a measure of the degree of human-likeness. We also present the *two-way mimicry target*, a model for measuring how well a human-computer dialogue *mimics* or *replicates* some aspect of human-human dialogue, including human flaws and inconsistencies. Although we have added a measure of innovation, none of the techniques is new in its entirety. Taken together and described from a human-likeness perspective, however, they form a set of tools that may widen the path towards human-like spoken dialogue systems.

Introduction

The evaluation and development of spoken dialogue systems is a complex undertaking, and much effort is expended on making it manageable. Research and industry endeavours in the area often seek to compare versions of existing systems, or to compare component technologies, in order to find the best methods – where "best" is defined as *most efficient*. Sometimes, *user satisfaction* is used as an alternative, more human centred metric, but as the systems under scrutiny are often designed to help users perform some task, user satisfaction and efficiency are highly correlated. Much effort is also spent on minimising the cost of evaluation, for example by designing evaluation methods that will generalise over systems and that may be re-used (e.g. Dybkjær et al., 2004; Walker et al., 2000; see also Möller et al., 2007 for an overview); by automating the evaluations (e.g. Bechet et al., 2004; Glass et al., 2000); or by utilising simulations instead of users in order make low cost repeat studies (e.g. Georgila et al., 2006; Schatzmann et al., 2005).

In this paper, we look at the particular issues involved in evaluating and developing *human-like spoken dialogue systems* – systems that aim to mimic human conversation to the greatest extent. The discussion is limited to collaborative spoken dialogue, although the reasoning may hold for a wider set of interaction types, for example text chats. Discussing human-like spoken dialogue systems implicitly requires that we formulate what "human-like" means, and the next two sections provide background on the concept of human-likeness. The first section proposes an analysis of how users perceive spoken dialogue systems in terms of other, more familiar things. The following section is a brief overview of pros and cons of striving for human-likeness in spoken dialogue systems. The following three sections deal with, in turn, how "increased human-likeness" can be understood, how to gather the experimental data needed to evaluate a component intended to increase human-likeness, and how to analyse that data.

Two faces of spoken dialogue systems

Spoken dialogue system research is often guided by a wish to achieve more natural interaction. The term "natural" is somewhat problematic and rarely defined, but is generally taken to mean something like "more like human-human interaction". For example, Jokinen (2003) talks about "computers that mimic human interaction" and Boyce (2000) says "the act of using natural speech as input mechanism makes the computer seem more human-like". We will use "human-like" to mean "more like human-human interaction".

There is little reason to assume that every spoken dialogue system becomes *better* by adding a component that is human-like, however. To evaluate this, criteria for "better" must be agreed upon, but this is the source of controversy with respect to interfaces in general as well as to spoken dialogue systems. There is a long-standing debate within human-machine interaction (HMI) between proponents of *direct manipulation interfaces* and *interface agents* (Shneiderman & Maes, 1997). A discussion more directly aimed at speech technology is that of *tool-like* versus *anthropomorphic* interfaces (see Qvarfordt, 2004 for an overview).

In Edlund et al. (2006), we tentatively proposed a different analysis in which such controversies in part disappear, in that the criteria for "better" is made dependent on the kind of system one wants to create. This analysis is elaborated in the following.

Interfaces and interlocutors

We argue that users of spoken dialogue systems perceive the systems largely *metaphorically* – within a conceptual space in which events and objects are interpreted. Others have made

similar claims, see for example the HMI concept of *mental models* in Norman (1983), popularised in Norman (1998), and the discussion on conversational dialogue in Allen et al. (2001). The thought that users interpret spoken dialogue systems metaphorically is not far-fetched: metaphors help us understand something that is odd to us in terms of something that is not. Talking computers are not yet commonplace, so interpreting them in terms of something that is more familiar to us makes sense. It is worth noting that metaphors may be seen from a designer as well as a user perspective. In the first case, the metaphor is what the designer intends for the user to perceive, similar to Norman's *design models*. In the second case, it is the image in the light of which the user actually understands the system, similar to the mental models described by Norman and others.

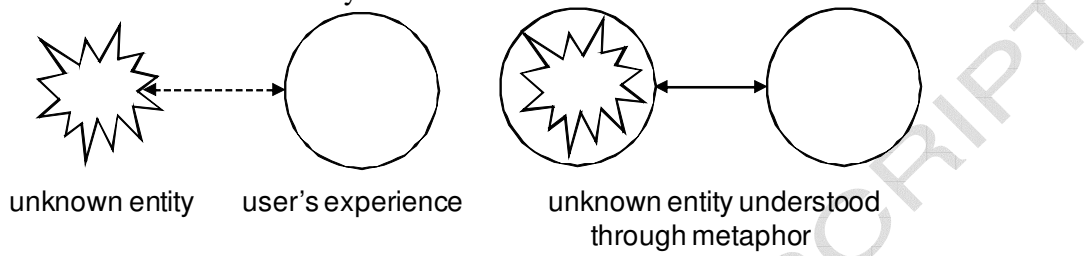


Figure 1: understanding the unfamiliar through the familiar – schematic illustration adapted by permission from an original by Jeffrey J. Morgan

Naturally, it is entirely possible to learn how to interact with a computer without explicit use of metaphors, for example simply by learning instructions one by one. Metaphors can help, however, in that they allow us to quickly learn what to expect, and to pick up a whole behaviour package, as illustrated in Figure 1. In order to get the most out of the metaphor, one should take advantage of a metaphor that is already available to the user. Based on observations, we propose that two metaphors commonly used by people in spoken human-machine interaction are *the human metaphor* and *the interface metaphor*.

Within the interface metaphor, the spoken dialogue system is perceived as a machine interface – often a computer interface. Speech is used to accomplish what would have otherwise been accomplished by some other means of input, such as a keyboard or a mouse. If a user perceiving a train ticket booking system in this manner says "Bath", it is equivalent to choosing Bath in for example a web form. The interface metaphor lies close at hand because many of the spoken dialogue systems that users are in contact with today basically provide alternatives to existing interfaces.

Within the human metaphor, on the other hand, the computer is perceived as an interlocutor: a being with human-like conversational abilities. Speech is not a substitute, nor is it one of many alternatives – it is the primary means of communication. Much like a television set, which is not part of the film it shows, the computer is not itself represented in this metaphor. The human metaphor is plausible – after all, speech is strongly associated with human-human interaction and if something speaks, it is reasonable to perceive it as human-like.

Supplying final proof that the proposed analysis holds is difficult, and the experiment achieving this has yet to be designed. There is evidence to be found, however, in that many observations are consistent with the analysis:

- Within the industry, users talking in a brief, semaphoric style are appreciated by system designers, since current systems can handle their behaviour. Conversely, commercial systems stand little chance with users who speak freely and at length, and these users are often viewed as a lost cause.
- Riccardi & Gorin (2000) investigated human responses to machine greeting prompts and noted a bimodal distribution of the durations of the responses. One mode was very similar to human-human interaction, the other corresponded to what they call *menu-speak*. This

speaking style has been labelled differently by other authors – Martinovsky & Traum (2003), for example, calls it *machine talk*.

- When characterising human-computer dialogue, Fischer speaks about *players* and *non-players*. The former "take up the system's cues and pretend to have a conversation with it", whereas the latter "only minimally reacts to interpersonal information provided by the system or even refuse communication at that level" (Fischer, 2006a) and "approach [the talking computer] as an instruction-obeying device" (Fischer, 2006b).
- In several of our systems, such as Waxholm (Blomberg et al., 1993), August (Gustafson & Bell, 2000) and AdApt (Gustafson et al., 2000), we have noted that some users happily respond to greetings and social system utterances, whereas others never respond to them at all. Fischer (2006b) also mentions that players open with a greeting, as opposed to non-players.
- Users of spoken dialogue systems designed for conversational speech sometimes use command-like language that the system is ill equipped to understand, or simply remain silent. We found that out of eight users interacting with a partially simulated system, six used feedback such as "uh-huh" at every opportunity. The remaining two, however, did not use feedback of that sort at all (Skantze et al., 2006a).
- During the development of the TeliaSonera customer care call centre, we observed that the open prompt "How may I help you?" is occasionally met with silence by users who know they are talking to a machine, suggesting that they are expecting the counterpart to a DTMF (Touch-Tone) navigation system, in which case there would be a list of alternatives. If, on the other hand, the voice is viewed as a human operator, users should have no difficulty formulating their concerns.

Users applying the interface metaphor seem to draw their knowledge of how to use spoken dialogue systems more or less directly from experiences with web or DTMF interfaces, and they show signs of not understanding more human-like features. Conversely, users applying a human metaphor view the system as a person and have expectations of its conversational abilities that are based on that.

Design implications

We discuss *two* metaphors here because they fit the observations at hand, but the idea that people understand spoken dialogue systems metaphorically is not restricted to those two metaphors. Users might for example perceive interface metaphor systems that replicate a web form differently from those that replicate DTMF choices over the telephone. Furthermore, it is appealing to envision a metaphor that lies between a human and a machine – *the android metaphor*, perhaps. Through such a metaphor, a system could be expected to behave human-like in some respects, say turn-taking and error handling, but not in other respects – it may for example not show or interpret emotion.

Regarding spoken dialogue system design, the analysis in terms of metaphors raises three concerns: we must carefully and knowingly choose metaphor; we must display the metaphor clearly to the user; and finally we must make the system internally coherent with the metaphor.

A suitable image. The choice of metaphor should not be taken lightly, as it affects the users' experience of a system and their expectations of what it can achieve. It is unlikely that there are set user types with predetermined and fixed ideas of how to understand spoken dialogue systems. Instead, users can be guided towards viewing a system in the light of one metaphor rather than another, and the same user can understand one system according to one metaphor and another system according to another. In many cases, either metaphor can be used to perform a given task. Travel booking, for example, can be achieved with an interface metaphor system functioning much like a voice controlled web form on the one hand, and

with a human metaphor system behaving like a travel sales person on the other. In other cases, however, our choice of metaphor can be guided by the task. Crucially for the account presented here, the choice also has effects on how the system is best evaluated.

Flaunting it. Displaying the metaphor clearly is important. If users can comprehend more than one metaphor, they can also be made aware that there are different types of spoken dialogue systems which are accessible through different metaphors. Evidence suggests that users' behaviour is influenced by their experience with a system as well as by their initial expectations of it. Amalberti et al. (1993) found differences between a group that thought they were talking to a computer and one that thought they were talking to a human over a noisy channel (which was in reality the case for both groups). However, the differences they were mostly present at the beginning of the interactions and dissipated with time.

Another example is that users presented with well-timed feedback such as "uh-huh" and "ok" generally speak less command-like, more in a manner consistent with human-human dialogue, as observed by ourselves (e.g. Gustafson et al., in press) and Gorin and colleagues, who noted a very long tail in the histogram of *utterance length* (measured either in words or seconds) in caller-operator dialogues (the data collection is described in Riccardi & Gorin, 2000). Some utterances were several minutes long, and upon closer inspection, it turned out that these utterances were interspersed with backchannel feedback by the operator. When similar interactions were recorded with pre-recorded prompts (and no backchannel feedback), the long tail disappeared (A. Gorin, personal communication, February 2nd 2006).

If a user habitually interprets speaking computers according to a specific metaphor, she will need to be given a reason to change this. Thus, clearly displaying through what metaphor users may best understand a system is helpful, and the display should come early on to avoid perpetuating less-than-optimal behaviours. Clearly, there is more than one way to display a metaphor. Visual cues such as rotating hourglasses, progress bars and blinking LEDs, and acoustic cues such as *earcons*, all point towards the interface metaphor. Conversely, the use of an *embodied conversational agent* (ECA) points towards a human metaphor (Cassell et al., 1999). The design of greetings is another example (see e.g. Balentine & Morgan, 2001; Boyce, 1999). "Hello!" suggests a system based on the human metaphor, whereas "This is the NN automated travel booking system. Where do you want to go from?" suggests an interface metaphor.

Keeping it real. Systems should be internally coherent with the chosen metaphor. A user will feel more comfortable with a system that can be coherently understood within one metaphor, rather than one that lends images arbitrarily from two or more. What constitutes coherence is dependent on the chosen metaphor. Human-like system behaviour is coherent with the human metaphor and clearly desirable if we want our users to understand the system that way, but it makes less sense if the spoken dialogue system is to be understood in terms of the interface metaphor. Instead, an interface metaphor system is coherent if it behaves like some corresponding interface – a web form or a DTMF menu.

At this point, let us briefly return to the somewhat problematic and often undefined concept of naturalness in spoken dialogue systems, and suggest that it can be understood in the following terms: If "internally coherent" means that the target of the metaphor (i.e. the spoken dialogue system) successfully mimics the behaviour one would expect from the source of the metaphor (e.g. a human interlocutor), and is less similar to some other (strong) source (e.g. a web interface), then "natural" can perhaps be understood as "internally coherent with a human metaphor".

Opting for human-likeness

Before turning our attention entirely to the design and evaluation of human-like spoken dialogue systems, let us apply the metaphor distinction to current spoken dialogue systems,

and pre-emptively address some objections to the endeavour of increasing human-likeness in spoken dialogue systems.

Human-likeness in current spoken dialogue systems

A number of oft-quoted benefits of spoken dialogue systems are consistent with the interface metaphor (some are even expressed in terms of comparisons with other interfaces): (1) they *work in hands-free and eyes-free situations* (Berry et al., 1998; Julia et al., 1998; Julia & Cheyer, 1998); (2) *when other interfaces are inconvenient or when our resources are occupied with other things* (Martin, 1976; Nye, 1982); (3) *when disabilities render other interfaces useless* (Damper, 1984); and (4) *with commonly available hardware*, such as a telephone. Similarly, some of the domains that have been exploited by spoken dialogue system designers are well suited for the interface metaphor, since speech is indeed an alternative interface in these domains: (a) *information retrieval systems*, such as train time table information or directory inquiries (e.g. Aust et al., 1995; Blomberg et al., 1993; Peckham, 1991; Seneff et al., 1998; Zue et al., 2000); (b) *ordering and transactions* , such as ticket booking (e.g. Boye et al., 1999; Wahlster et al., 2001); (c) *command control systems*, such as smart homes ("turn the radio off") or voice command shortcuts ("save") (e.g. Bolt, 1980; Rayner et al., 2001); and (d) *dictation*, for example Nuance Dragon Dictate and Windows Vista (for a discussion of the field, see Leijten & van Waes, 2001).

On the other hand we have a number of equally oft-quoted benefits of spoken dialogue systems that are more consistent with a human metaphor and often expressed in terms of human behaviour: speech is supposedly (5) *easy-to-use* since we already know how to talk; it is (6) *flexible*, and human dialogue is *responsive and fast* (Chapanis, 1981; Goodwin, 1981); human conversations (7) come with *extremely resilient error handling* and feature *fast and continuous validation* (Brennan, 1996; Brennan & Hulteen, 1995; Clark, 1994; Nakatani & Hirschberg, 1994); and human-human dialogue is (8) *largely social* and, importantly, *enjoyable* (Cassell & Bickmore, 2002; Isbister, 2006). There is a corresponding range of domains, predominantly used for research systems, which draw upon these features. Examples include (e) *games and entertainment* (e.g. Gustafson et al., 2005; Hey, you, Pikachu![1], Seaman[2]); (f) *coordinated collaboration*, with tasks involving control or over-view of complex situations (Rich et al., 2001; Traum & Rickel, 2001); (g) *computer mediated human-human communication*, systems that support people in their communication with each other and often present themselves as an *avatar*, a representation of a human (Agelfors et al., 2006; Edlund & Hjalmarsson, 2005; Wahlster, 2000); (h) *expert systems*, systems that diagnose and help people reason, as in computer support situations and help desks (Allen et al., 1995; Allen et al., 2001; Bohus & Rudnicky, 2002; Cassell et al., 1999; Smith et al., 1992); (i) *learning and training systems* that help people learn and practise new skills (Hjalmarsson et al., 2007; Johnson et al., 2000; Lester & Stone, 1997; Lester et al., 1999); and (j) *guiding, browsing and presentation*, such as a city guide or a narrator (Cassell et al., 2002; Gustafson & Bell, 2000; Gustafson et al., 2000; Gustafson & Sjölander, 2002; Oviatt, 2000; Skantze et al., 2006b).

The listings are not exhaustive, but serve to illustrate the following: Current speech applications generally exploit 1-4 above well enough in domains a-d. There are guidelines to speech interface design with a more or less clear focus on these items (e.g. Balentine &

[1] Nintendo, *Hey, you, Pikachu!/Pikachu Genki Dechu: A N64 game published by Nintendo*, Available from http://www.heyyoupikachu.com, 1999.

[2] SEGA, *Seaman a dreamcast computer game published by SEGA*, Available from http://www.sega.com/games/dreamcast/post_dreamcastgame.jhtml?PRODID=194, 2000.

Morgan, 2001; Reeves et al., 2004; Rudnicky, 1996; and see Möller et al., 2007 which provides an overview). The benefits in 5-8 – the ones we associate with the human metaphor – reflect salient aspects of human-human communication whose potential is rarely exploited in commercial systems. It is not obvious that systems such as those in a-d would become "better" – more efficient, or more appreciated by their users – if they exploited 5-8, but many systems in domains e-j attempt to draw upon these features.

Objections to striving for increased human-likeness

Feasibility. Can we ever hope to achieve human-likeness to a degree that it is of use to us? Importantly, there is no need to design spoken dialogue systems that are virtual copies of humans, or even systems that behave like real humans – a prospect which is severely questioned, not least within the AI society (see Larsson, 2005 for a discussion from a dialogue research perspective). Instead, the aim is a system that can be understood through a human metaphor or as Cassell (2007) puts it "a machine that acts human enough that we respond to it as we respond to another human" –a much more feasible goal.

Humans are often quite willing participants, and the human metaphor allows us to draw on this by borrowing from other areas. For example, the gaming, film and fiction industries rely heavily on *willing suspension of disbelief* – the ability to ignore minor inconsistencies in order to enjoy a work of fiction (e.g. Hayes-Roth, 2004). Users may be quite willing to ignore a measure of inconsistencies as long as the sequence of events as a whole makes sense. One way of using this is by explaining inconsistencies *in-character* – within the story, or within the metaphor. Human features may be used to explain short-comings of a system, for example by portraying the character represented by the spoken dialogue system as being stupid, arrogant, preoccupied, uninterested, flimsy, etc (see e.g. Gustafson et al., 2005). One could also claim that the character is far away and experiencing line noise, or that the conversation is being translated for the character, in order to explain misunderstandings. Naturally, we are not suggesting that these systems be built with smoke and mirrors alone. In order to make the metaphor internally coherent and the systems believable, many aspects of them need to be improved.

Utility. Given that human-like spoken dialogue systems are feasible, the question remains whether they are useful. One answer is that achieving human-likeness in implemented computer-human dialogues will provide valuable insights about how humans communicate with each other – analogous to how cognitive science aims to explain cognition by means of explicit computational models. Another straightforward answer has already been hinted at – some of the promises held by the idea of talking computers (e.g. robustness, flexibility, responsivity) are strongly linked to human-like dialogue, and there are a number of applications with the primary, perhaps even solitary, goal of human-like and social behaviour, for example characters in games and entertainment. There is also a growing interest for human-likeness in spoken dialogue systems amongst researchers (e.g. Philips, 2006, keynote speech, Interspeech, Pittsburgh, PA, US; Zue, 2007), and many researchers have made a case for anthropomorphism in spoken dialogue systems. To the extent that human-like behaviour in a spoken dialogue system strengthens the human metaphor, their arguments are valid for human-likeness as well, as anthropomorphism can be seen as applying a human metaphor – Laurel (1990), for example, says that anthropomorphism is "not the same thing as relating to other people, but is rather the application of a metaphor with all its concomitant selectivity".

Uncanny valley. As our spoken dialogue systems become more and more human-like, they are heading straight for the *uncanny valley* (Mori, 1970) – a point where human-likeness supposedly stops being attractive to users, and instead becomes eerie and unnerving. It is however hard to see this threat as sufficient cause to refrain from studying human-likeness. For one thing, no reports of uncanny valley effects from users actually interacting with

human-like spoken dialogue systems are known to us, possibly because spoken dialogue systems aiming at human-likeness are yet too immature. If this is the case, we will cross the uncanny valley when we come to it.

There are, however, studies where users are asked their opinion in the matter. Dautenhahn et al. (2005) reports that 71% of respondents to a questionnaire in a user study on robot companions stated that they would like the robot to communicate in a more human-like manner, whereas only 36% wanted it to behave more human-like, and 29% wished for it to appear more human-like. As there are no fully functional human-like systems to demonstrate, such responses are by necessity elicited by posing hypothetical questions. This is difficult, as evidenced by Saini et al. (2005), who had two groups of users test different versions of a talking robot cat, one with more social intelligence and one more neutral. The groups were then asked if they would have preferred a system with the characteristics of the other, to them unseen, system. Both groups responded that they clearly preferred the system they had tested. The authors concluded that it is difficult to imagine and judge something one has not experienced.

Symmetry. One might object that if spoken dialogue systems behave in a more human-like manner, users will expect them to have more human capabilities in general: human understanding and knowledge of the world, for example. Although this is true, the principle of symmetry – that a spoken dialogue system should be able to understand what it can produce – is designed to avoid communication breakdown in task oriented systems. We shall see in the following that there may be situations where the context frees us from the principle. Finally, allowing the fact that we are not currently able to achieve some aspects of a topic (e.g. understanding in a human-like manner) to stop us from researching the topic (i.e. human-likeness) would only serve to perpetuate our ignorance.

Towards human-likeness

Whether one subscribes to a metaphoric view or not, one may opt to aim for spoken dialogue systems with increased human-likeness, begging the question how to proceed. What technology must be developed, and how is it evaluated to ensure that it brings human-likeness and not something else?

Evaluation target

The first question is what to evaluate against. As the aim is increased human-likeness, we can simply assume that this is a valuable goal, and evaluate *increased human-likeness* directly.

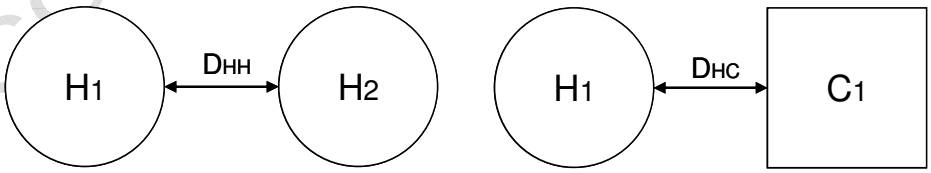


Figure 2: human-human interaction (D_{HH}) and human-computer interaction (D_{HC})

The question is how well a component mimics or replicates an aspect of human-human dialogue, or a phenomenon present in it. Figure 2 shows schematically D_{HH} a human-human dialogue between H_1 and H_2, and D_{HC} a human-computer dialogue between H_1 and C_1. We propose to evaluate human-likeness in human-computer interaction using a *two-way mimicry*

target: the general goal is to design C_1 so that (some aspect of) D_{HC} mimics, or resembles as much as possible, (some aspect of) D_{HH}. The goal can be subdivided:

1. C_1's behaviour in D_{HC} should resemble H_2's behaviour in D_{HH} – the machine should behave like a human interlocutor in a human-human conversation;
2. and H_1's behaviour in D_{HC} should resemble H_1's behaviour in D_{HH} – the human speaking to the computer should behave similarly as to when speaking to another human.

This subdivision places the focus on the computer's behaviour, and views the human interlocutor's behaviour as a result of this, whereas spoken dialogue systems are otherwise often categorised according to the restrictedness of the human interlocutor's behaviour with little mention of the spoken dialogue system's behaviour (Porzel, 2006).

A major difference between human-likeness evaluation and traditional evaluation lies in how the comparison of parameters is done: In ASR evaluations, a lower WER is better, given the same context. With the two-way mimicry target, or any human-likeness target, the comparison operand is "more similar to", so that if for example pause length is *distributed* in a certain way in human-human dialogue given some context, the most human-like system would elicit the same *pause length distribution* on the system side as well as the user side, given the same context, rather than the fastest system. An example of this can be seen in Figure 4.

Evaluation context

The process of evaluating isolated features is sometimes called micro-evaluation, particularly in the embodied conversational agent (ECA) community (e.g. Ruttkay & Pelachaud, 2004). Conversely, the evaluation of systems as a whole is called macro-evaluation. The steps introduced below are all micro-evaluation in this scheme: although the evaluation context is broadened with each step, the effect of the isolated component is nevertheless what is evaluated. Although macro-evaluation of human-like systems is beyond the scope of this article, it is worth mentioning that the metrics we describe here could for example be used instead of or in parallel with the cost measures in PARADISE (Walker et al., 2000), with a resulting system-wide mapping between human-likeness, task success, and user satisfaction.

Developing and evaluating a component for increased human-likeness can be described as a multi-step process, where the component is tested in an increasingly broad context. Here, *component* does not refer to soft- or hardware of any particular size or scope, but is used loosely and without reference to system architecture. It should be taken to mean a part of a spoken dialogue system with a well-defined task.

Candidate selection. The first step, then, is to identify candidate features for the sought-after effect. This can be done by studying the literature if the phenomenon has been researched, or by exploratory data studies if it has not. In order to find candidates for human-likeness, it is generally best to use human-human dialogue, although in some cases Wizard-of-Oz data collections (Wooffitt et al., 1997) and similar methods may be useful to get around sparse data and avoid "nonchalant treatment of phenomena in language and speech that are known or assumed to have low frequencies of occurrence" (Möbius, 2005). What, then, are the good candidates? The broad answer is *anything that is observable in human-human communication*, but they can be exemplified by phenomena others have looked at: Kawamoto et al. (2004) mentions grunts and back-channel feedback, use of prosody to indicate utterance type and emotion, incremental understanding and interruptability, and facial animation with lip synchronisation; Porzel (2006) focuses on turn-taking issues; to mention but a few.

The selected candidate can then be tested for *perception*, *understanding* and *response*. These are, roughly speaking, of increasing complexity, which is reflected in the effort it takes to perform them. They are also hierarchical: what is understood is also perceived, and what we

respond (appropriately) to is also understood. In other words, tests for response implicitly test perception and understanding as well, but as response tests are considerably more expensive, it is prudent to test candidates using perception and understanding tests first. In the following, the first two are only described briefly, while response is discussed in more detail in subsequent sections.

Perception. The easiest tests to perform are simple perception tests to see if subjects can perceive the phenomenon. If, to take an example from studies of turn taking and prosody, a mid level tone allegedly coincides with turn keeping and a downwards pitch movement with turn yielding, a first experimental study could test that subjects can actually differentiate between the two (as realised in the stimuli), or what the JND (just noticeable difference) is.

Understanding. Once it is established that a feature can be perceived and that a distinction can be made, experiments to find out *how* it is perceived are needed. The fact that a listener is able to perceive the difference between two stimuli does not prove that the difference means anything in particular. Examples of how to map stimuli include asking subjects to freely describe what they believe a stimulus means or answering multiple choice questions. In many cases, the message we want the stimuli to convey (e.g. a desire to keep the floor or to provide positive feedback) can be paraphrased into a concrete verbal request, such as "hang on, let me finish" or "yes I agree". Such paraphrases can be exploited to make more straightforward multiple choice questions along the lines of "Do you think what you just heard is equivalent to utterance (a), (b) or (c) below?"

Response. When it is shown that users can perceive and understand a feature in the way it was intended, it is time to test the pragmatics, thus finding out whether subjects respond accordingly – something which standard listening or production test are not likely to achieve. Candidates may be evaluated for perception and understanding using simple perception tests – given the illustration in Figure 2, it is enough to investigate whether C1 in D_{HC} behaves like H2 in D_{HH}, but pragmatic evaluation also involves the user responses - H_2's behaviour in D_{HC}.

Eliciting pragmatic evidence

In the following, we will discuss techniques to elicit the pragmatic data needed for evaluating user responses against human-likeness gold standards. The methods we discuss – *Wizard-of-Oz variations*, *human-human data manipulation*, and *micro-domains* – are all commonly used to collect data in order to build models of human-computer dialogue, and references to such usage are provided for completeness. Note that when these methods are used in the traditional sense, it is important to ensure that the data is representative for computer-directed speech, as illustrated by $D_{H'C}$ in Figure 3. This contrasts with the two-way mimicry target, which tests how close we can get to the situation in $D_{HC'}$.

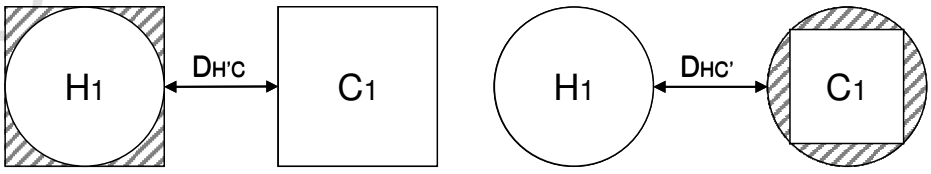


Figure 3: human-computer interaction on the computer's terms ($D_{H'C}$) and human-computer interaction on human terms ($D_{HC'}$)

Wizard-of-Oz variations

This section describes a number of variations of the Wizard-of-Oz data collection paradigm, in which people (i.e. wizards) simulate a system unbeknownst to the subjects (see Wooffitt et al., 1997 for a thorough discussion). Although a number of critiques have been put forth against the Wizard-of-Oz methodology (for an overview, see Eklund, 2004), it is often used for human-computer data collection when building a fully functional system is impractical or too expensive. The paradigm has been used for initial data collection in the development of a great number of spoken dialogue systems, including early systems such as Circuit Fix It Shop (Moody, 1988), ATIS (Hemphill et al., 1990), SUNDIAL (Peckham, 1991), The Philips train timetable system (Aust et al., 1995), Waxholm (Bertenstam et al., 1995), MASK (Life et al., 1996), and AdApt (Gustafson et al., 2000).

In these data collections, wizards are often required to produce output that resembles what can be expected from a talking computer, so that the data is representative for human-computer interactions (Dahlbäck et al., 1993), as illustrated in $D_{H'C}$ (Figure 3). Techniques to achieve this include instructing the wizard to misunderstand (Bernsen et al., 1998); using text filters that introduce insertion and deletion errors (Fraser & Gilbert, 1991); using a wizard to determine what utterances the system should even attempt to understand (Peissner et al., 2001); training wizards to speak without disfluencies and removing samples from the speech signal (Lathrop et al., 2004); providing the wizards with a limited selection of pre-recorded system prompts (Wirén et al., 2007); and limiting the wizard to making choices within a strictly defined dialogue model (Tenbrink & Hui, 2007).

As noted above, the requirements when using the Wizard-of-Oz paradigm to test human-likeness methods stand in stark contrast to this. When the aim is to model how humans behave when talking to a system that is as human-like as possible, as seen in $D_{HC'}$ (Figure 3), instructing the wizard to behave in a computer-like manner is moot.

Constraining the wizard. Designing a wizard-of-Oz system for testing human-likeness is no easy task, however. Wizards, being humans, are capable of participating in a real human-human dialogue in the most natural manner imaginable almost subconsciously. Using this capability to control a system is a different matter, however: the wizard must make conscious choices for the system, for example by pressing buttons in a computer interface. Wizards, then, should be given as much programmatic assistance as possible. It is often prudent to keep the task of the wizard simple and straightforward. Programmatic constraints can also make the wizards' task less ambiguous (by removing fuzzy requirement such as behaving "like a computer would") and can alleviate the problem of wizards being asked to take on roles in which they are not really representing themselves, noted by Allwood & Haglund (1992). Well-construed programmatic support also makes it possible to use more than one wizard to simulate a system, as suggested by Salber & Coutaz (1993), who talk about using one wizard for handling the input, one for handling the task level decisions and one for formulating the answers – although they also suggested measures for keeping the dialogue more "machine-like". Multiple wizards were used in the NICE Project (Gustafson et al., 2005) where the system's decision-making was supervised by wizards. The domain was quite open, and the wizard sometimes had to choose from a rather large number of possible alternative actions when correcting the system. In order to keep the interaction fast and responsive, a second wizard controlled more rapid and semantically light-weight backchannels and feedback utterances allowing the system to keep a straight face during delays. Dialogue 1 shows an example from such a recording: an interaction between the fairy-tale character Cloddy Hans and a user. The contributions labelled *Cloddy* are monitored by the primary wizard, and the ones labelled *Cloddy (secondary)* are initiated by the secondary wizard in order to buy time when the primary wizard runs into trouble. The second Cloddy (secondary) contribution actually covers for a restart of a failing system component. A wizard setup that is fast,

responsive and easy to use allows us to test methods in a context that is close to a functioning human-like spoken dialogue system, where the wizard emulates components we are yet unable to build.

Dialogue 1: dialogue example from Nice data collection	
Cloddy	du kan prata med mig och be mig göra saker *you can talk to me and ask me to do things*
User	här kan du hämta den där boken som ligger på hyllan *here can you fetch that book on the shelf*
Cloddy (secondary)	Mmm
Cloddy	vill du att jag ska ta boken *do you want me to take the book*
User	Mmm
Cloddy	om du säger det så *if you say so*
User	Ja *yes*
Cloddy	jag ska gå bort till hyllan *i'll go over to the shelf*
User	Ok
Cloddy	[action:go to shelf]
Cloddy	jag ska ta boken *i'll take the book*
User	Ok
System	[system failure]
User	Hallå *hello*
Cloddy (secondary)	mhm [during component restart by main wizard]
Cloddy	[action:pickup book]
User	Mmm
Cloddy (secondary)	Mm
Cloddy	här står alla trollformler man behöver *every spell you need is in here*

Wizard-as-component. We can also restrict the wizards' influence to simulate a "perfect" component with limited scope, for example to decide whether an ASR hypothesis is correct or faulty. This provides data reflecting how users would respond to a system featuring such a component. Skantze et al. (2006a) provides an example where a fully functional system was used to gather information on subjects' responses to brief grounding utterances in the form of colour words with different prosodic realisation (see Dialogue 2). The wizard that was utilised was given a double task: (1) recognise any mention of the colours red, green, blue or yellow; and (2) identify the end of user utterances. The tasks had different rationales. By letting the wizard spot colour words, the data could be analysed automatically and immediately, as it was manually verified on-line (1). The benefit of using brief grounding utterances may be countered by long response times, so delays were kept at a minimum by letting the wizard identify places for the system to speak (2).

Dialogue 2: dialogue example from colour feedback experiments	
System	and what colour is this
System	[presents visual colour stimuli]
User	that looks like red to me
System	Red
User	yes red

Using wizards to aid a spoken dialogue system with tasks beyond its capabilities can also have a positive side effect: If the wizards are allowed to use whatever means they are given to the best of their ability and any restrictions imposed on them are encoded in the software, then the wizards' actions represent the target the component designer should aim at – an idea akin to Paek (2001), who suggests using human wizards' behaviour as a gold standard for dialogue components.

Wizard-as-Subject. As hinted at above, Wizard-of-Oz experiments can be set up with the purpose of studying the actions of *the wizards*. This is exemplified by the development of a Swedish natural language call routing system at TeliaSonera (Wirén et al., 2007, Boye & Wirén, 2007). To overcome the lack of realism in traditional Wizard-of-Oz collections, a method which was coined *in-service Wizard-of-Oz data collection* was introduced, in which the wizards were real customer care operators and the callers were real customers with real problems. Having actual customer care operators act as wizards provided valuable feedback on dialogue and prompt design. The wizards used a prompt piano – a set of keys corresponding to pre-recorded prompts: an *initial open prompt* designed to inform callers that they were talking to a machine, but could express themselves freely, and prompts to engage in a system-driven, menu-based dialogue in order to collect additional pieces of information needed for routing the call. The wizards' choices were carefully logged to model the final production system. We will call the corpus collected in this design *in-service Woz I* (ISWOZ-I).

In problematic cases, the wizards could route failing calls back to themselves, taking over the

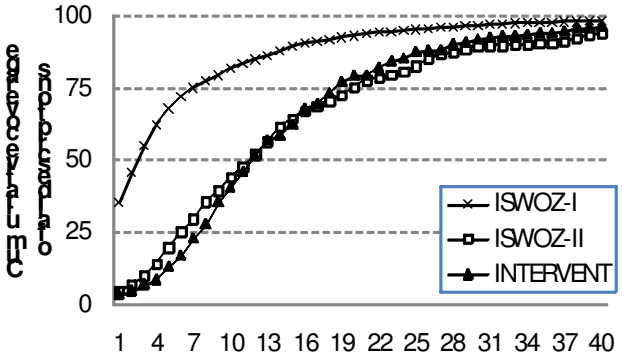


Figure 4. *Distribution of utterance lengths (in number of words) in the two Wizard-ofOz collections and in the human-human intervention dialogues.*

call in the role of the operator. In Gustafson et al. (in press), we used a corpus of such calls (INTERVENT) to gain insights about how they were resolved. The analysis revealed that the problems were not insurmountable – operators succeeded without exception in collecting the information needed to route these calls, and a set of specific reasons causing the dialogues to fail were discernible. Based on these observations, we redesigned the prompt piano to facilitate the observed human-human dialogue behaviour. The in-service setup with service agents acting as Wizards-of-Oz was kept and used to collect a new corpus with the new prompt piano, ISWOZ-II. The resulting dialogues were generally successful. Figure 4 shows that the length distribution of callers' descriptions in ISWOZ-II is similar to the human-human dialogues in INTERVENT, and dissimilar to the dialogues in ISWOZ-I, exemplifying humanlikeness evaluation: given the criteria for humanlikeness presented previously, this shows that with respect to user utterance length, ISWOZ-II is the more human-like dialogue. A setup similar to the in-service design is used in a later study by Porzel (2006), who calls the method WOT (Wizard and Operator Test). A human acts as wizard by operating a speech

synthesis interface. An obvious system breakdown is then simulated, after which the wizard breaks in and finishes the remaining tasks. This way, human-computer (wizard) and human-human data are collected from the same dialogue.

A general prerequisite for Wizard-of-Oz studies is that users be under the impression that they are talking to a computer (Fraser & Gilbert, 1991, Wooffitt et al., 1997), as is the case in the studies above. The aim of studying human-like behaviour, however, may void this prerequisite. Skantze (2003), for example, uses a modified Wizard-of-Oz configuration where subjects interacting with each other are openly informed of the experiment setup, but where the speech in one direction is passed through an ASR, and in the other direction is filtered through a vocoder, with the goal of investigating and modelling how people deal with a signal ridden with typical ASR errors. In a similar setup, Stuttle et al. (2004) used a human typist to simulate the ASR component, then added controlled errors and asked a wizard to act the system's part by talking freely in order to investigate how the wizard handled the errors. This type of experiment, where the line between wizard and subject is blurred, brings us to the next data collection method: *human-human data manipulation*.

Human-human data manipulation

Unconstrained human-human interactions are, naturally, the most human-like interaction data available. In order to test the effects of particular methods, or what happens when a particular feature is manipulated, however, a certain measure of control is unavoidable. This section describes data collections that somehow manipulate human-human data to get results that would otherwise not be possible. The technique can be divided into *off-line* and *on-line manipulation*, depending on when the manipulation is done: off-line manipulation takes place in a post-recording step, whilst on-line manipulation takes place during the interaction.

Like Wizard-of-Oz experiments, data manipulation has been used to record data for modelling human-computer interaction. Examples include Wizard-of-Oz setups where the wizard talks to the user through a Vocoder, as in Dybkjaer et al. (1993), and the many data driven user simulations used to test systems is another. Again, the aim is commonly to gather data that is representative for computer-directed speech, whereas our goal is to experiment with making computer-human dialogue as similar to human-human dialogue as possible.

Off-line data manipulation. Human-likeness experiments involving off-line data manipulation can be exemplified by listening tests where subjects are asked to judge recordings where for example the pitch, voice, or lexical information in spoken utterances has been altered. The basic idea is to record some interaction, manipulate the recording somehow, and ask subjects to judge it in some manner. A drawback with this is that it more or less excludes the use of the participants themselves as judges. From a dialogue perspective, an obvious drawback is that manipulations of one speaker's speech will not affect the other participant's behaviour in the recording – something highly likely to occur had the speech been manipulated *during* the interaction. The same problem haunts many tests on pre-recorded spoken dialogue system data, where post factum changes to a contribution in the middle of a recorded interaction are assumed to yield new versions of the entire interaction from which valid conclusions can be drawn.

The method is useful for initial tests of human-likeness, however. In Hjalmarsson & Edlund (in press), we studied human-like language generation by investigating how a dialogue system displaying complex human behaviours, such as fragmental utterances and human-like grounding, is perceived by non-participating listeners. These behaviours are frequent in human-human conversation, but typically not present in synthesised system output. Based on recordings of human-human dialogues, two versions of simulated human-machine dialogue were created by replacing one of the speakers with a synthesised voice. In one of the versions (UNCONSTRAINED), the original choice of words was kept intact, as were features such as

hesitations, filled pauses, and fragmental utterances. In the other version (CONSTRAINED), a set of constraint rules were applied before synthesising, similar to the distillation described by Jönsson & Dahlbäck (2000). The rules left timing intact, while removing disfluencies and reducing lexical variation. The dialogues were then be compared to each other in tests. Dialogue 3 shows an example.

Dialogue 3. A CONSTRAINED and an UNCONSTRAINED utterance.	
System (UNCONSTRAINED)	**mm she she has a dinner on friday mm but she is available on saturday and sunday and on thursday as well**
System (CONSTRAINED)	**anna is available for dinner on thursday saturday and Sunday**

THE UNCONSTRAINED dialogues were rated more HUMAN-LIKE, POLITE and INTELLIGENT. For the two remaining dimensions, EFFICIENCY and UNDERSTANDING, there was no significant difference, neither was there any preference for a particular version. When EFFICIENCY was checked for correlations with the other dimensions, the only correlation found was with the rating of UNDERSTANDING. The results imply that the specific manners of human communication have no negative effect on perceived efficiency or level of understanding, but that they cause the dialogues to be perceived as more human-like, more intelligent and more polite.

On-line data manipulation. In on-line manipulation of human-human data, some part of the interaction is doctored on-line and in real-time as the interaction proceeds. This gives a strong advantage to the off-line counterpart, as the responses recorded are the participants' actual responses to the manipulated interaction. The technique suffers from the difficulties involved in manipulating spoken interaction in this manner, and the manipulation will unfailingly introduce latency. Exactly how sensitive human-human dialogue is to latency varies: a sixth of a second had a negative impact when Reeves & Nass (1996) shifted the synchrony of video and audio, and Kitawaki & Itoh (1991) showed that the minimum detectable delay can vary from 90ms to over a second, while in practice, round trip delays of 500ms and more have ill effects for communication.

Examples of on-line manipulation data collections include Purver et al. (2003) and Healey & Mills (2006), both of which used text chats and no speech, Schlangen & Fernández (2007), who manipulated the speech channel in dialogue by blurring speech segments to elicit clarification requests. Gratch et al. (2006) had pairs of interlocutors recorded in a setting where a listener can hear the speaker's real voice while watching what they are told are graphic representations of the speaker and the speaker's gestures on monitors. In reality, the visual representation on the monitor is controlled by other parameters. A similar setup was used in Edlund & Beskow (2007), where user behaviour is shown to vary with turn-taking gestures in a talking head without the subjects being explicitly aware of this.

On-line data manipulation is particularly powerful because it captures one participant's reactions *as if the manipulated data had actually been produced by the other participant*. Other benefits include that the data collection can be made symmetrical, so that data from both participants' perspective can be exploited.

Micro-domains

Interspeech 2009 hosts the Loebner prize, which for the first time contains a spoken class (Roger Moore, presentation at Interspeech 2007) for anyone who can build a speaking computer that will make a human believe she is talking to another human. It is probably safe to say that no-one believes the prize will be won. Turing said "Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's? If this were then subjected to an appropriate course of education one would obtain the adult brain" (Turing, 1950). Playing the devil's advocate, we could argue that should a believable child be built, the Turing test would already be passed. The idea of building a spoken dialogue system that can be taken for a human by *some person*, under *some set of circumstances* leads us to what we will call micro-domains.

Given sufficient control of the context, making a person believe she is talking to a human is trivial – some would say that the early and text based ELIZA (Weizenbaum, 1966) came close, and many of us have been tricked into believing we are talking to a person when it is in fact an answer phone. If the dialogue situation is sufficiently predictable, understanding or even perceiving user responses may not be necessary to support the illusion – modelling someone taking telephone messages, for example, can be done without any real understanding. Similarly, people who simply do not care or do not listen can be modelled easily – like someone delivering a furious ear-bashing or shouting out a fire alarm. The examples may seem trite, but micro-domains can be very useful for gathering data on how human-like features are received. Nigel Ward's humming machine (Ward & Tsukahara, 2000) is an example of a machine that can potentially make a human believe she is talking to another human, if presented in an appropriate context – like a telephone conversation where one person does most of the talking. Apart from the novelty value, however, such a machine goes a long way to test models of where to produce backchannels – the reason the machine was designed in the first place.

Another example of a workable micro-domain is that of the instructor or narrator. In certain contexts, for example when introducing film or a play, the speaker is not expected to listen to others, merely to make sure that the audience is paying attention, and the need for semantic capabilities is small. Given sufficient understanding of grounding, attention and interaction control, a narrating system may perform quite close to a human. Several research labs, including ourselves, are working on such systems with the intention to use the final system to test methods to model attention, interruption, etc. Examples include the animated interactive fiction system described in Piesk & Trogemann (1997), the virtual child Sam (Cassell et al., 2000), and the story-telling robot described in Mutlu et al. (2006).

Data collections in micro-domains have in common with most Wizard-of-Oz data collections that they rely on the user being deceived in some way. Naturally, this can be quite problematic ethically, and great care has to be taken to ensure that such data collections are acceptable, legally, morally, and scientifically.

Analysing the experiment data

Having collected data on how users respond to a component, we are left with the task of evaluating the data. To reiterate, with reference to Figure 2, the human-like criteria is that C_1's behaviour in D_{HC} should resemble H_2's behaviour in D_{HH} and H_1's behaviour in D_{HC} should resemble H_1's behaviour in D_{HH}, as illustrated by $D_{HC'}$ in Figure 3. There are several approaches to the task of measuring this.

Automatic comparisons

If data from similar human-human dialogues are available, features that can be operationalised and automatically extracted can be compared automatically, which is appealing since it makes the evaluation less subjective and easier to repeat. Examples of such features include different durations, such as the lengths of pauses, overlapping speech and utterances, which can be accessed using speech activity detection; prosodic features, such as intensity and pitch; and turn-taking patterns, which can be accessed from SAD decisions and an interaction model (e.g. Brady, 1968). The comparison in itself can be made complex, so that also the context in which different features occur matters, but in the simplest case, the overall distribution of a feature is compared with that of the same feature in human-human dialogue, as illustrated in Figure 4.

Naturally, quite a few features that can only be reliably accessed by manual annotation or a combination of manual annotation and automatic extraction can be compared in the same manner. Examples include vocabulary; syntactic structure; use of communicative acts; displays of emotion; and use of filled pauses, hesitations, clarification requests. In other cases, particularly during the early steps of development, but also when evaluating more complex response behaviour, automatic measures will not do, and we must resort to subjective measures.

Human judgement

The same subjective techniques as those used for manual annotation of training data can be used to analyse human-likeness data, with the same measures to maintain their validity in spite of their subjectiveness. In the following, we discuss two approaches to judging or labelling data, *reviewing* and *participant studies*, and argue that the former is more valuable for human-likeness studies.

Reviewing vs. participant studies. In *reviewing*, judgement is not passed by a participant in the interaction, but by an external person – a reviewer – with the specific task of judging some aspect of the interaction. In a *participant study*, a participant of the interaction is asked to judge some aspect of it.

A key argument for participant studies is that participants alone know what they perceive, but there is a rather large body of research suggesting that participants are sometimes "unaware of the existence of a stimulus that importantly influenced a response", "unaware of the existence of the response", and "unaware that the stimulus has affected the response", as Nisbett & Wilson (1977) put it in a review of the field. Although perhaps the strongest lesson to be learned from this is that it is preferable to measure *the effects* of stimuli and events in subjects rather than asking the subjects how they perceive them, they also imply that participants may not have that much of an advantage to onlookers when it comes to judging the interaction they are involved in.

A clear drawback of participant studies is that they do not permit immediate responses without disturbing the flow of the interaction. The PARADISE evaluation scheme (Walker et al., 2000), for example, calls for a success/failure judgement to be made by a participant after each information transaction – something that could easily become very disruptive. Still, the responses need to be collected during or shortly after the interaction occurs, lest it turn into a memory test. Conversely, the perhaps most compelling advantage of reviewing is that it decouples the judgement from the interaction with respect to time, participants and location, which allows researchers to do a number of things that would otherwise not be possible:

- Run evaluations at any time after the interaction was recorded.
- Run as many tests with as many reviewers that can be afforded.

- Get multiple judgements for each data point by using a group of reviewers rather than one participant.
- Compare judgements from reviewers of different background.
- Using a reviewer (group of reviewers) for each factor (set of factors) makes it possible to test any number of factors, whereas a participant only can answer so many questions.
- Check results by running the same test on different data – perhaps even on data recorded elsewhere or for different purposes.
- Make comparative studies between interactions with different participants.
- Manipulate the amount of context given to the reviewers, or the context itself, for example to examine what parts of an interaction provide necessary and sufficient conditions for a decision.
- Manipulate the recorded data itself before reviewing, as exemplified in the section on *off-line data manipulation* above.
- Check the significance of trends by re-running the test with identical design and stimuli, but with new reviewers.

Some of these items involve running repeated tests, which can invalidate the significance of the results, so precautions have to be taken.

The field abounds with examples of judgements by a single or a few people of unmanipulated data which is presented in a chunk – standard transcription and labelling are good examples, as are *eavesdropper* and *overhearer* evaluations (e.g. Whittaker & Walker, 2006), where a reviewer is asked to listen in on a recorded dialogue and pass judgement on some aspect. We conclude the discussion on analysis by presenting two variations on reviewing that are less commonly used.

Plenary experiments. Reviewing allows us make plenary experiments where a great many people pass judgement simultaneously. This can be done with the use of paper forms, a method used in (Granström et al. (2002)), but is more efficient with an audience response system (ARS), as was recently done to allow an audience of 60 to judge song synthesis (Special session for song synthesis, Interspeech 2007). Ideally, one would use ARS in the manner it is used in Hollywood screenings, where subjects adjust a lever continuously to indicate what they think of some aspect of the film they are shown. This method records judgement of continuous sequences of events, such as dialogue flow or the quality of synthesized speech. An example of such a system is Audience Studies Institute *mini-theatre research service*, which currently uses audiences of 25 (Vane et al., 1994, chapter 6).

Human computation. Reviewing also allows for different media to be used. A highly interesting example is human computation, a method first proposed by Luis von Ahn and Laura Dabbish (von Ahn & Dabbish, 2004), in which people are encouraged to do labelling tasks embedded in some entertaining environment such as a game. Von Ahn and Dabbish first introduced the method in *the ESP Game*, and it has since been used for a number of tasks (e.g. von Ahn et al., 2006; Brockett & Dolan, 2005; Paek et al., 2007). The idea is to draw on the tremendous pool of human resources available through the Internet, and to do so by exchanging human computation for entertainment. A typical Human Computation setup contains a game idea, which is what provides the participants with motivation for doing the task; a validation method with the main purpose of avoiding spam and vandalism – typically points are given to users who give the same answer to the same question or something similar; methods to eliminate cheating; methods to deliver the data to be annotated to the participants and to retrieve the annotation data. The setup holds a promise of almost uncountable numbers of human annotators, something that may be needed both to annotate training material and to evaluate the results of machine learning methods.

Future work

It would be especially gratifying to make use of known human-human dialogue phenomena in this type of evaluation. The Lombard reflex can serve as an example. It is known that under noisy conditions, speakers change their voice in a number of ways: they raise their voice and their pitch, etc. If noise is added to a human-computer dialogue, a human-like computer would be expected to exhibit the same changes (so that C_1's behaviour in D_{HC} should resemble H_2's behaviour in D_{HH}). This could be easily tested automatically. Accommodation (or entrainment, alignment, etc.), the propensity of interlocutors to make their behaviour similar to that of their conversational partner is another phenomenon that could be utilised – it suggests that the more human-like the computer in a human-computer dialogue behaves, the more human-like the behaviour of the human, as well. In the case of the Lombard reflex, this could mean that a human talking to a machine may not exhibit the Lombard reflex to the full extent if she does not interpret the situation as a human-human dialogue.

Conclusion

This paper has presented an overview of methods to collect data on how users respond to techniques intended to increase human-likeness in spoken dialogue systems and to analyse the results. Some of these represent fairly traditional ways of accomplishing this, such as Wizard-of-Oz studies with the subjects being the objects of study; or having subjects judge manipulated interactions off-line. Other methods have added a measure of innovation, including studying the wizards as subjects, allowing the wizards to enter the conversation in the event of insufficient wizard interfaces, and to redesign the wizard interfaces based on such interventions. Other more novel twists include manipulating human-human dialogues on-line, effectively treating both participants as subjects and recording continuous judgments of some parameter by a panel of reviewers equipped with different kinds of audience response systems. Taken together these techniques form a set of tools that may widen the path towards human-like spoken dialogue systems.

Acknowledgement

The call routing experiments took place at TeliaSonera, Sweden. They would not have been possible hadn't it been for the ASR 90 200 pilot team that provided the data collection tools for the skilled wizards. Note also that Anders Lindström at TeliaSonera took part in designing the second experiment. Finally, our heartfelt thanks to everybody in the research group at Speech, Music and Hearing at KTH, the two anonymous reviewers, and to colleagues everywhere for valuable input and support.
Part of the work presented here was funded by the Swedish research council projects #2006-2172 (Vad gör tal till samtal/What makes speech special) and #2007-6431 (GENDIAL), and the European Commission's Sixth Framework Program project IP-035147 (MonAMI).

References

Agelfors, E., Beskow, J., Karlsson, I., Kewley, J., Salvi, G., & Thomas, N. (2006). User Evaluation of the SYNFACE Talking Head Telephone. *Lecture Notes in Computer Science, 4061*, 579-586.
Allen, J. F., Byron, D., Dzikovska, M., Ferguson, G., Galescu, L., & Stent, A. (2001). Towards conversational human-computer interaction. *AI Magazine, 22*(4), 27-37.
Allen, J., Schubert, L., Ferguson, G., Heeman, P., Hwang, C., Kato, T., Light, M., Martin, N., Miller, B., Poesio, M., & Traum, D. (1995). The TRAINS Project: A case study in defining a conversational planning agent. *J. Exp. Techn. Artif. Intelligence, 7*, 7-47.
Allwood, J., & Haglund, B. (1992). *Communicative Activity Analysis of a Wizard of Oz Experiment*. Technical Report, Göteborg University, Gothenburg, Sweden.

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.
Aust, H., Oerder, M., Seide, F., & Steinbiss, V. (1995). The Philips automatic train timetable information system. *Speech Communication, 17*, 249-262.
Balentine, B., & Morgan, D. P. (2001). *How to build a speech recognition application: a style guide for telephony dialogues.* San Ramon CA: Enterprise Integration Group.
Bechet, F., Riccardi, G., & Hakkani-Tur, D. (2004). Mining Spoken Dialogue Corpora for System Evaluation and Modeling. In *the 2004 Conference on Empirical Methods in Natural Language Processing* (pp. 134-141). Barcelona, Spain.
Bernsen, N., Dybkjaer, H., & Dybkjaer, L. (1998). *Designing interactive speech systems.* London: Springer-Verlag.
Berry, G. A., Pavlovic, V. I., & Huang, T. S. (1998). A Multimodal Human–Computer Interface for the Control of a Virtual Environment. In *AAAI Workshop on Intelligent Environments*. Stanford, CA, US: Stanford University.
Bertenstam, J., Beskow, J., Blomberg, M., Carlson, R., Elenius, K., Granström, B., Gustafson, J., Hunnicutt, S., Högberg, J., Lindell, R., Neovius, L., Nord, L., de Serpa-Leitao, A., & Ström, N. (1995). The Waxholm system - a progress report. In Dalsgaard, P. (Ed.), *Proc of ESCA Workshop on Spoken Dialogue Systems* (pp. 281-284). Vigs¿, Denmark.
Blomberg, M., Carlson, R., Elenius, K., Gustafson, J., Granström, B., Hunnicutt, S., Lindell, R., & Neovius, L. (1993). An experimental dialogue system: WAXHOLM. In *Proceedings Eurospeech '93* (pp. 1867-1870). Berlin.
Bohus, D., & Rudnicky, A. (2002). LARRI: A Language-based Maintenance and Repair Assistant. In *Proc. of the ISCA Workshop Multi-Modal Dialogue in Mobile Environments*. Kloster Irsee, Germany.
Bolt, R. (1980). Put that there: Voice and gesture at the graphics interface. *Comp. Graph., 14*, 262-270.
Boyce, S. J. (1999). Spoken natural language dialogue systems: user interface issues for the future. In Gardner-Bonneau, D. (Ed.), *Human factors and voice interactive systems* (pp. 37-61). Kluwer Academic Publishers.
Boyce, S. J. (2000). Natural spoken dialogue systems for telephony applications. *Communications of the ACM, 43*(9), 29-34.
Boye, J., & Wirén, M. (2007). Multi-slot semantics for natural-language call routing systems. In *Proceedings of the Workshop on Bridging the Gap: Academic and Industrial Research in Dialog Technologies* (pp. 68–75).
Boye, J., Wiren, M., Rayner, M., Lewin, I., Carter, D., & Becket, R. (1999). Language-Processing Strategies and Mixed-Initiative Dialogues. In *Proc. of IJCAI-99 Workshop on Knowledge and Reasoning in Practical Dialogue Systems*. Stockholm.
Brady, P. T. (1968). A statistical analysis of on-off patterns in 16 conversations. *The Bell System Technical Journal, 47*, 73-91.
Brennan, S. E. (1996). Lexical entrainment in spontaneous dialog. In *Proceedings of ISSD* (pp. 41-44).
Brennan, S., & Hulteen, E. (1995). Interaction and feedback in a spoken language system: a theoretical framework. *Knowledge-Based Systems, 8*, 143-151.
Brockett, C., & Dolan, W. (2005). Echo Chamber: A game for eliciting a colloquial paraphrase corpus. In *AAAI 2005 Spring Symposium, Knowledge*.
Cassell, J., & Bickmore, T. (2002). Negotiated Collusion: Modeling Social Language and its Relationship Effects in Intelligent Agents. *User Model. Adapt. Interf., 12*, 1-44.
Cassell, J., Ananny, M., Basu, A., Bickmore, T., Chong, P., Mellis, D., Ryokai, K., Vilhjalmsson, H., Smith, J., & Yan, H. (2000). Shared Reality: Physical Collaboration with a Virtual Peer. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI),* (pp. 259-260).
Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjálmsson, H., & Yan, H. (1999). Embodiment in Conversational Interfaces: Rea. In *CHI'99* (pp. 520-527). Pittsburgh, PA, US.
Cassell, J., Stocky, T., Bickmore, T., Gao, Y., Nakano, Y., & Ryokai, K. (2002). MACK: Media lab Autonomous Conversational Kiosk. In *Proceedings of Imagina02*. Monte Carlo.
Cassell, J. (2007). Body language: lessons from the near-human. In Riskin, J. (Ed.), *Genesis Redux: Essays on the history and philosophy of artificial life* (pp. 346-374). University of Chicago Press.
Chapanis, A. (1981). Interactive human communication: Some lessons learned from laboratory experiments. In Shackel, B. (Ed.), *Man-Computer Interaction: Human Factors Aspects of Computers and People*. Rockville, Maryland, US: Sijthoff & Noordhoff.
Clark, H. H. (1994). Managing problems in speaking. *Speech Communication, 15*, 243-250.
Dahlbäck, N., Jönsson, A., & Ahrenberg, L. (1993). Wizard of Oz studies – why and how. In *Proceedings from the 1993 International Workshop on Intelligent User Interfaces* (pp. 193-200).
Damper, R. (1984). Voice-input aids for the physically handicapped. *Int. J. Man–Machine Stud., 21*, 541-553.

Dautenhahn, K., Woods, S., Kaouri, C., Walters, M., Koay, K., & Werry, I. (2005). What is a robot companion - friend, assistant or butler?. In *Proceedings of Int Conf on Intelligent Robots and Systems (IROS 2005)* (pp. 1192-1197).

Dybkjaer, H., Bernsen N., ., & Dybkjaer, L. (1993). Wizard-of-Oz and the trade-off between naturalness and recognizer constraints. In *Proceedings of the 3rd European Conference on Speech Communication and Technology, Berlin*.

Dybkjær, L., Bernsen, N. O., & Minker, W. (2004). Evaluation and usability of multimodal spoken dialogue systems. *Speech Communication, 43*, 33-54.

Edlund, J., & Beskow, J. (2007). Pushy versus meek – using avatars to influence turn-taking behaviour. In *Proceedings of Interspeech 2007*. Antwerp, Belgium.

Edlund, J., & Hjalmarsson, A. (2005). Applications of distributed dialogue systems: the KTH Connector. In *Proceedings of ISCA Tutorial and Research Workshop on Applied Spoken Language Interaction in Distributed Environments (ASIDE 2005)*. Aalborg, Denmark.

Edlund, J., Heldner, M., & Gustafson, J. (2006). Two faces of spoken dialogue systems. In *Interspeech 2006 - ICSLP Satellite Workshop Dialogue on Dialogues: Multidisciplinary Evaluation of Advanced Speech-based Interactive Systems*. Pittsburgh PA, USA.

Eklund, R. (2004). *Disfluency in Swedish human–human and human–machine travel booking dialogues*. Doctoral dissertation, Linköping University, Linköping, Sweden.

Fischer, K. (2006a). The role of users' preconceptions in talking to computers and robots. In *Proceedings of Workshop on 'How People Talk to Computers, Robots, and Other Artificial Communication Partners', Hansewissenschaftskolleg, Delmenhorst.* (pp. 112-130).

Fischer, K. (2006b). *What Computer Talk Is and Is not: Human-Computer Conversation as Intercultural Communication.* Saarbrücken: AQ.

Fraser, N. M., & Gilbert, G. N. (1991). Simulating speech systems. *Computer Speech and Language, 5*(1), 81-99.

Georgila, K., Henderson, J., & Lemon, O. (2006). User simulation for spoken dialogue systems: learning and evaluation. In *the 9th International Conference on Spoken Language Processing (Interspeech - ICSLP)*. Pittsburgh, PA, US.

Glass, J., Polifroni, J., Seneff, S., & Zue, V. (2000). Data collection and performance evaluation of spoken dialogue systems: the MIT experience. In *ICSLP-2000* (pp. 931-934).

Goodwin, C. (1981). *Conversational organization: Interaction between speakers and hearers.* New York: Academic Press.

Granström, B., House, D., & Swerts, M. G. (2002). Multimodal feedback cues in human-machine interactions. In Bel, B., & Marlien, I. (Eds.), *Proc of the Speech Prosody 2002 Conference* (pp. 347-350). Aix-en-Provence: Laboratoire Parole et Langage.

Gratch, J., Okhmatovskaia, A., Lamothe, F., Marsella, S., Morales, M., van der Werf, R. J., & Morency, L-P. (2006). Virtual rapport. In *Proceedings of 6th International Conference on Intelligent Virtual Agents*. Marina del Rey, CA, US.

Gustafson, J., & Bell, L. (2000). Speech Technology on Trial: Experiences from the August System.. *Natural Language Engineering, 6*(Special issue on Best Practice in Spoken Dialogue Systems).

Gustafson, J., & Sjölander, K. (2002). Voice transformations for improving children's speech recognition in a publicly available dialogue system. In *Proc of ICSLP 2002* (pp. 297-300). Denver, Colorado, USA.

Gustafson, J., Bell, L., Beskow, J., Boye, J., Carlson, R., Edlund, J., Granström, B., House, D., & Wirén, M. (2000). AdApt - a multimodal conversational dialogue system in an apartment domain. In *Proc. of ICSLP 2000, 6th Intl Conf on Spoken Language Processing* (pp. 134-137). Beijing.

Gustafson, J., Boye, J., Fredriksson, M., Johannesson, L., & Königsmann, J. (2005). Providing computer game characters with conversational abilities. In *Proceedings of Intelligent Virtual Agent (IVA05)*. Kos, Greece.

Gustafson, J., Heldner, M., & Edlund, J. (in press). Potential benefits of human-like dialogue behaviour in the call routing domain. To be published in *Proceedings of the 4th IEEE Workshop on Perception and Interactive Technologies for Speech-Based Systems*. Kloster Irsee, Germany.

Hayes-Roth, B. (2004). What Makes Characters Seem Life-Like?. In Prendinger, H., & Ishizuka, M. (Eds.), *Life-Like Characters: Tools, Affective Functions, and Applications* (pp. 447-462). Germany: Springer.

Healey, P. G. T., & Mills, G. (2006). Clarifying spatial descriptions:Local and global effects on semantic coordination. In *brandial'06, the 10th International Workshop on the Semantics and Pragmatics of Dialogue (SemDial10)*.

Hemphill, C., Godfrey, J., & Doddington, G. (1990). The ATIS Spoken Language Systems, Pilot Corpus. In *Proceedings of 3rd DARPA Workshop on Speech and Natural Language, Hidden Valley, PA* (pp. 102-108).

Hjalmarsson, A., & Edlund, J. (in press). Human-likeness in utterance generation: effects of variability. To be published in *Proceedings of the 4th IEEE Workshop on Perception and Interactive Technologies for Speech-Based Systems*. Kloster Irsee, Germany.

Hjalmarsson, A., Wik, P., & Brusk, J. (2007). Dealing with DEAL: a dialogue system for conversation training. In *Proceedings of SigDial* (pp. 132-135). Antwerp, Belgium.
Isbister, K. (2006). *Better game charachters by design: a psychological approach.* Elsevier.
Johnson, W., Rickel, J., & Lester, J. (2000). Animated Pedagogical Agents: Face-to-Face Interaction in Interactive Learning Environments. *Int. J. Artif. Int. Educ., 11*, 47-78.
Jokinen, K. (2003). Natural interaction in spoken dialogue systems. In *Proceedings of the Workshop "Ontologies and Multilinguality in User Interfaces" at HCI International 2003* (pp. 730-734). Crete, Greece.
Julia, L., & Cheyer, A. (1998). Cooperative Agents and Recognition Systems (CARS) for Drivers and Passengers. In *OzCHI '98* (pp. 32-38). Adelaide, Australia.
Julia, L., Cheyer, A., Dowding, J., Bratt, H., Gawron, J. M., Bratt, E., & Moore, R. (1998). How Natural Inputs Aid Interaction in Graphical Simulations?. In *VSMM '98* (pp. 466–468). Gifu, Japan.
Jönsson, A., & Dahlbäck, N. (2000). Distilling dialogues - a method using natural dialogue corpora for dialogue systems development. In *Proceedings of the 6th Applied Natural Language Processing Conference* (pp. 44-51). Seattle.
Kawamoto, S-i., Shimodaira, H., Nitta, T., Nishimoto, T., Nakamura, S., Itou, K., Morishima, S., Yotsukura, T., Kai, A., Lee, A., Yamashita, Y., Kobayashi, T., Tokuda, K., Hirose, K., Minematsu, N., Yamada, A., Den, Y., Utsuro, T., & Sagayama, S. (2004). Galatea: Open-source Software for Developing Anthropomorphic Spoken Dialog Agents. In Prendinger, H., & Ishizuka, M. (Eds.), *Life-Like Characters: Tools, Affective Functions, and Applications* (pp. 187-212). Germany: Springer.
Kitawaki, N., & Itoh, K. (1991). Pure delay effects on speech quality in telecommunications. *IEEE Journal on Selected Areas in Communications, 9*(4), 586-593.
Larsson, S. (2005). Dialogue systems: simulations or interfaces?. In Gardent, C., & Gaiffe, B. (Eds.), *Proceedings of the ninth workshop on the semantics and pragmatics of dialogue*.
Lathrop, B., Cheng, H., Weng, F., Mishra, R., Chen, J., Bratt, H., Cavedon, L., Bergmann, C., Hand-Bender, T., Pon-Barry, H., Bei, B., Raya, M., & Shriberg, L. (2004). A Wizard of Oz framework for collecting spoken human-computer dialogs: An experiment procedure for the design and testing of natural language in-vehicle technology systems. In *Proceedings of the 12th International Congress on Intelligent Transportation Systems, San Francisco CA, USA*.
Laurel, B. (1990). Interface agents: metaphors with character. In Laurel, B. (Ed.), *The art of human-computer interface design* (pp. 355-365). Addison Wesley.
Leijten, M., & van Waes, L. (2001). The influence of voice recognition on the writing process: cognitive and stylistic effects of speech technology on writing business texts. In *Proc. of Human-Computer Interaction: INTERACT '01.* (pp. 783–784). Tokyo, Japan.
Lester, J., & Stone, B. (1997). Increasing believability in animated pedagogical agents. In Johnson, W. L., & Hayes–Roth, B. (Eds.), *Proc. of the First International Conference on Autonomous Agents* (pp. 16-21). Marina del Rey, CA, US.
Lester, J., Voerman, J., Towns, S., & Callaway, C. (1999). Deictic Believability: Coordinated Gesture, Locomotion, and Speech in Lifelike Pedagogical Agents. *Applied Artificial Intelligence, 13*, 383–414.
Life, A., Salter, I., Temem, J., Bernard, F., Rosset, S., Bennacef, S., & Lamel, L. (1996). Data Collection for the MASK Kiosk: WOZ vs. Prototype System. In *Proceedings of ICSLP '96, 3–6 October, Philadelphia, PA*.
Martin, T. B. (1976). Practical applications of voice input to machines. *IEEE, 64*, 487–501.
Martinovsky, B., & Traum, D. (2003). The error is the clue: breakdown in human-machin einteraction. In *Proceedings of the ISCA Tutorial and Research Workshop Error Handling in Spoken Dialogue Systems*. Château d'Oex, Vaud, Switzerland.
Moody, T. (1988). *The Effects of Restricted Vocabulary Size on Voice Interactive Discourse Structure*. Doctoral dissertation, North Carolina State University..
Mori, M. (1970). The Uncanny Valley. *Energy, 7*(4), 33-35.
Mutlu, B., Hodgins, J., & Forlizzi, J. (2006). A Storytelling Robot: Modeling and Evaluation of Human-like Gaze Behavior. In *Proceedings 2006 IEEE-RAS International Conference on Humanoid Robots (HUMANOIDS'06)*. Genova, Italy.
Möbius, B. (2005). Rare events and closed domains: two delicate concepts in speech synthesis. *International Journal of Speech Technology, 6*(1), 57-71.
Möller, S., Smeele, P., Boland, H., & Krebber, J. (2007). Evaluating spoken dialogue systems according to de-facto standards: A case study. *Computer Speech & Language, 21*(1), 26-53.
Nakatani, C., & Hirschberg, J. (1994). A corpus-based study of repair cues in spontaneous speech. *JASA, 95*, 1603–1616.
Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we know: Verbal reports on mental processes. *Psychological Review, 84*(3), 231-259.
Norman, D. A. (1983). Some observations on mental models. In Gentner, D., & Stevens, A. L. (Eds.), *Mental models* (pp. 7-14). Lawrence Erlbaum Associates.

Norman, D. A. (1998). *The design of everyday things.* MIT Press.
Nye, J. (1982). Human factors analysis of speech recognition systems. *Speech Technology, 1*, 50-57.
Oviatt, S. (2000). Talking To Thimble Jellies: Children's Conversational Speech with Animated Characters. In *Proc. of ICSLP '00* (pp. 877–880). Beijing, China.
Paek, T., Ju, Y-C., & Meek, C. (2007). People Watcher: a game for eliciting human-transcribed data for automated directory assistance. In *Proceedings of Interspeech 2007* (pp. 1322-1325). Antwerp, Belgium.
Paek, T. (2001). Empirical methods for evaluating dialog systems. In *ACL 2001 Workshop on Evaluation Methodologies for Language and Dialogue Systems*.
Peckham, J. (1991). Speech Understanding and Dialogue over the Telephone: An Overview of the ESPRIT SUNDIAL Project. In *Proc. of the DARPA Speech and Natural Language Workshop* (pp. 14-27). Pacific Grove, CA, US.
Peissner, M., Heidmann, F., & Ziegler, J. (2001). Simulating recognition errors in speech user interface prototyping. In *Usability evaluation and interface design: Proceedings of HCI International* (pp. 233–237). Mahwah, NJ, US: Lawrence Erlbaum.
Piesk, J., & Trogemann, G. (1997). Animated interactive fiction: storytelling by a conversational virtual actor. In *Proceedings of VSMM'97* (pp. 100-108). Geneva, Switzerland: IEEE Computer Society Press.
Porzel, R. (2006). How Computers (Should) Talk to Humans. In *Proceedings of Workshop on 'How People Talk to Computers, Robots, and Other Artificial Communication Partners'*. Hansewissenschaftskolleg, Delmenhorst. Germany.
Purver, M., Healey, P. G., King, J., Ginzburg, J., & Mills, G. J. (2003). Answering clarification questions. In *the 4th SIGdial Workshop on Discourse and Dialogue*. Sapporo, Japan.
Qvarfordt, P. (2004). *Eyes on multimodal interaction*. Doctoral dissertation, Linköping University.
Rayner, M., Lewin, I., Gorrell, G., & Boye, J. (2001). Plug and Play Speech Understanding. In *Proc. 2nd SIGdial Workshop on Discourse and Dialogue*. Aalborg, Denmark.
Reeves, B., & Nass, C. (1996). *The Media Equation.* Stanford, CA, US: CSLI Publications.
Reeves, L. M., Lai, J., Larson, J., Oviatt, S., Balaji, T. S., Buisine, S., Collings, P., Cohen, P., Kraal, B., Martin, J-C., McTear, M., Raman, T. V., Stanney, K. M., Su, H., & Wang, Q. Y. (2004). Guidelines for multimodal user interface design. *Communications of the ACM, 47*(1), 57-59.
Riccardi, G., & Gorin, A. L. (2000). Stochastic Language Adaptation over Time and State in Natural Spoken Dialogue Systems. *IEEE Trans. Speech Audio Proc., 8*, 3-10.
Rich, C., Sidner, C., & Lesh, N. (2001). COLLAGEN: Applying Collaborative Discourse Theory to Human–Computer Interaction. *Artificial Intelligence Magazine, 22*, 15-25.
Rudnicky, A. (1996). *Speech interface guidelines.* Unpublished manuscript.
Ruttkay, Z., & Pelachaud, C. (2004). *From Brows till Trust: Evaluating Embodied Conversational Agents.*
Saini, P., de Ruyter, B., Markopoulos, P., & van Breemen, A. (2005). Benefits of Social Intelligence in Home Dialogue Systems. In *Proceedings of Interact 2005 - Communicating Naturally through Computers, Rome, Italy, 2005.*.
Salber, D., & Coutaz, J. (1993). Applying the Wizard of Oz Technique to the Study of Multimodal Systems. In Bass, L., Gornostaev, J., & Unger, C. (Eds.), *Proceedings of Human-Computer Interaction, 3rd International Conference EWHCI '93,* (pp. 219-230). Springer Verlag.
Schatzmann, J., Georgila, K., & Young, S. (2005). Quantitative evaluation of user simulation techniques for spoken dialogue systems. In *6th SIGdial Workshop on Discourse and Dialogue*. Lisbon, Portugal.
Schlangen, D., & Fernández, R. (2007). Speaking through a noisy channel: experiments on inducing clarification behaviour in human-human dialogue. In *Proceedings of Interspeech 2007*. Antwerp, Belgium.
Seneff, S., Hurley, E., Lau, R., Pao, C., Schmid, P., & Zue, V. (1998). Galaxy-II: A reference architecture for conversational system development. In *Proc. of ICSLP '98* (pp. 931-934). Sydney, Australia.
Shneiderman, B., & Maes, P. (1997). Direct manipulation vs. interface agents: Excerpts from debates at IUI'97 and CHI'97. *Interactions, 4*, 42-61.
Skantze, G., Edlund, J., & Carlson, R. (2006b). Talking with Higgins: Research challenges in a spoken dialogue system. In André, E., Dybkjaer, L., Minker, W., Neumann, H., & Weber, M. (Eds.), *Perception and Interactive Technologies* (pp. 193-196). Berlin/Heidelberg: Springer.
Skantze, G., House, D., & Edlund, J. (2006a). User responses to prosodic variation in fragmentary grounding utterances in dialogue. In *Proceedings of Interspeech 2006 - ICSLP* (pp. 2002-2005). Pittsburgh PA, USA.
Skantze, G. (2003). Exploring human error handling strategies: implications for spoken dialogue systems. In *Proceedings of ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems* (pp. 71-76). Chateau-d'Oex-Vaud, Switzerland.
Smith, R. W., Hipp, D. R., & Biermann, A. W. (1992). A Dialog Control Algorithm and its Performance. In *Proceedings of the 3rd Conference on Applied Natural Language Processing* (pp. 9-16). Trento, Italy.
Stuttle, M., Williams, J. D., & Young, S. (2004). A framework for dialogue data collection with a simulated ASR-channel. In *Proceedings of ICSLP*. Jeju, South Korea.

Tenbrink, T., & Hui, S. (2007). Negotiating spatial goals with a wheelchair. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue* (pp. 103-110). Antwerp, Belgium.
Traum, D., & Rickel, J. (2001). Embodied Agents for Multi-party Dialogue in Immersive Virtual Worlds. In *Proc. of IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems* (pp. 766-773). Seattle, WA, US.
Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*, 433-460.
Vane, E. T., Schafer, L., & Gross, S. (1994). *Programming for TV, Radio, and Cable.* Focal Press.
von Ahn, L., & Dabbish, L. (2004). Labeling images with a computer game. In *ACM Conference on Human Factors in Computing Systems, CHI 2004* (pp. 319-326). Vienna, Austria.
von Ahn, L., Ginosar, S., Kedia. M., Liu, R., & Blum, M. (2006). Improving accessibility of the web with a computer game. In *Proceedings of the SIGCHI conference on Human Factors in computing systems* (pp. 79-82). Montréal, Québec, Canada.
Wahlster, W., Reithinger, N., & Blocher, A. (2001). SmartKom: Towards Multimodal Dialogues with Anthropomorphic Interface Agents. In *Proc. of MTI Status Conference* (pp. 26-27). Berlin, Germany.
Wahlster, W. (2000). *Verbmobil: Foundations of Speech-to-Speech Translation.* Berlin: Springer.
Walker, M., Kamm, C., & Litman, D. (2000). Towards developing general models of usability with PARADISE. *Natural Language Engineering, 6*, 363-377.
Ward, N., & Tsukahara, W. (2000). Prosodic features which cue back-channel responses in English and Japanese. *Journal of Pragmatics, 32*, 1177-1207.
Weizenbaum, J. (1966). ELIZA - A computer program for the study of natural language communication between man and machine. *Communications of the Association for Computing Machinery, 9*, 36-45.
Whittaker, S., & Walker, M. (2006). Evaluating Dialogue Strategies in Multimodal Dialogue Systems. In Minker, W., Bühler, D., & Dybkjær, L. (Eds.), *Spoken Multimodal Human-Computer Dialogue in Mobile Environments*. Springer Netherlands.
Wirén, M., Eklund, R., Engberg, F., & Westermark, J. (2007). Experiences of an in-service Wizard-of-Oz data collection for the deployment of a call-routing application. In *Bridging the Gap: Academic and Industrial Research in Dialog Technology, HLT–NAACL Workshop*. Rochester, New York, USA.
Wooffitt, R., Fraser, N. M., Gilber, N., & McGlashan, S. (1997). *Humans, computers and wizards.* London and New York: Routledge.
Zue, V., Seneff, S., Glass, J., Polifroni, J., Pao, C., Hazen, T., & Hetherington, L. (2000). Jupiter: A telephone-based conversational interface for weather information. *IEEE Trans. Acoust. Speech Sign. Proc., 8*, 85-96.
Zue, V. (2007). On organic systems. In *Proceedings of Interspeech 2007* (pp. 1-8). Antwerp, Belgium.