Cross-Cultural Assessment of Automatically Generated Multimodal Referring Expressions in a Virtual World

Ielka van der Sluis[a], Saturnino Luz[b], Werner Breitfuß[c], Mitsuru Ishizuka[c], Helmut Prendinger[d]

[a] *i.f.van.der.sluis@rug.nl*
Department of Communication and Information Sciences
University of Groningen, the Netherlands
[b] *luzs@scss.tcd.ie*
Department of Computer Science
Trinity College Dublin, Ireland
[c] *werner,ishizuka@mi.ci.i.u-tokyo.ac.jp*
University of Tokyo,
7-3-1 Hongo, Bunkyo-ku,
Tokyo, 113-8656, Japan
[d] *helmut@nii.ac.jp*
National Institute of Informatics,
2-1-2 Hitotsubashi, Chiyoda-ku,
Tokyo, 101-8430, Japan

Abstract

This paper presents an assessment of automatically generated multimodal referring expressions as produced by embodied conversational agents in a virtual world. The algorithm used for this purpose employs general principles of human motor control and cooperativity in dialogues that can be parametrised so as to vary the precision of the pointing gestures and the amount of linguistic information included in the referring expressions. The study assessed how native speakers of English and Japanese perceived three different algorithmic outputs for multimodal referring behaviour in terms of understandability, human-likeness and a social practice (selling). Results show that users generally prefer mobile agents that are economical in their linguistic descriptions to stationary verbose agents. They also show the need for further calibration of the algorithm to accommodate the differences between the two groups. In addition to the detailed description of the set up and results of the study, the paper discusses implications for the design and use of agents, methodological issues that arose while conducting the cross-cultural study and directions for future work.

Preprint submitted to Elsevier *May 3, 2012*

Van der Sluis, I., Luz, S., Breitfuss, W., Ishizuka, M., and Prendinger, H. Cross-cultural assessment of automatically generated multimodal referring expressions in a virtual world. *International Journal of Human-Computer Studies 70*, 9 (2012), 611–629.

Keywords: perception of multimodal referring expressions, virtual worlds, dialogue, translation, cross-cultural differences.

1. Introduction

Research in Human Computer Interaction (HCI) shows an increased interest in developing interfaces that closely mimic human communication. The development of "embodied conversational agents" (ECAs) or life-like characters with appropriate verbal and non-verbal behaviour with regard to a concrete spatial domain clearly fits this interest (Prendinger and Ishizuka, 2004; Kopp et al., 2003; Cassell et al., 2002; André et al., 2000). In the last decade ECAs have increasingly been used in areas including product presentation and sale (Eichner et al., 2007; Prendinger et al., 2007), training (Kenny et al., 2007a,b; Traum et al., 2005), education (Nakasone et al., 2011) and, of course, in Entertainment and games (see Rist et al., 2003, for an overview). Currently, however, the ability of an ECA to interact with human users is very limited using mainly stationary agents with which interactions mostly rely on pre-scripted system output, whereby the manual generation of human-like and convincing agent behaviour is a cumbersome task.

An issue addressed in many systems is that of identifying a certain object in a visual context accessible to both user and system. This can be done by an ECA that points to the target object while uttering a linguistic referring expression. In the past decades many algorithms have been proposed with which such multimodal referring expressions (MREs) can be produced (Claassen, 1992; Reithinger, 1992; André and Rist, 1993; Huls et al., 1995; Lester et al., 1999; Kranstedt et al., 2006). The work presented in this paper employs one such algorithm (Van der Sluis and Krahmer, 2007). Based on universal principles of human behaviour and human motor control, this algorithm combines linguistic properties and pointing gestures in a compositional way. In contrast to other approaches, the algorithm is flexible in that it can include pointing gestures that vary in their level of precision, allowing an agent to move closer to objects in its environment or identify them from a distance similar to the way in which humans use their environment.

This paper presents a carefully designed evaluation method to assess the quality of automatically generated MREs by ECAs in a virtual environment as a first attempt to calibrate the output of this MRE algorithm. To the best of our knowledge this is the first study to assess, user perception of MREs as used by agents that are allowed to move and interact dynamically in the virtual space they inhabit. Our assessment of the perception of automatically produced MREs by

Van der Sluis, I., Luz, S., Breitfuss, W., Ishizuka, M., and Prendinger, H. Cross-cultural assessment of automatically generated multimodal referring expressions in a virtual world. *International Journal of Human-Computer Studies 70*, 9 (2012), 611–629.

humans was carried out through a cross-cultural study in Dublin (Ireland) and Tokyo (Japan). The need to approach the development and design of ECAs from a cross-cultural perspective has been acknowledged by Payr and Trappl (2004), but to date only few systems have actually addressed this issue, and a Western perspective prevails (Rehm et al., 2009a; De Rosis et al., 2004).

The results presented below indicate that, both in Tokyo and in Dublin, participants prefer a mobile agent that uses short linguistic descriptions over a stationary and more verbose agent, even if the agent's movement is not smooth at all times. Further fine tuning is, however, required to accommodate the variations that surfaced between the preferences of our two experimental groups. In addition to presenting detailed results and discussing their implications for the design and use of ECAs, we also discuss the methodological difficulties encountered in conducting the study and lessons learnt. Finally, we also offer a number of directions for future cross-cultural studies in dynamic settings.

The paper is structured as follows. Section 2 provides some background on the generation of MREs, ECAs and virtual worlds, scripted dialogue and cross-cultural studies. Section 3.1 presents our hypotheses. Section 3 describes the cross-cultural study. Section 4 closes the paper with a discussion of the findings, conclusions and future work.

2. Background

National and linguistic boundaries are regarded as important factors in HCI design (Fernandes, 1994). These factors are often said to determine cultural differences which can be systematically studied from a managerial or system design perspective (Hofstede, 1983; Bourges-Waldegg and Scrivener, 1998). While this position is not entirely uncontentious, given the complexity of the concept of culture (Kroeber and Kluckhohn, 1952), it has been adopted in empirical comparisons of user interface elements and designs involving user groups across national boundaries (e.g. Noiwan and Norcio (2006); Dong and Salvendy (1999)). This is the perspective adopted in this paper. Furthermore, we follow Traum (2009) and adopt the following definition: '[a] culture is a set of rules (i.e. normative, communicative and inferential) that a group has common knowledge of and orientation towards'. This is applicable both to national groups and their social practices Cassell (2009). The rules in question refer, in this study, to the use of MREs. A preliminary study served to elicit relevant factors and differences Van der Sluis and Luz (2011). In this section we review prior work in MRE generation and related cross-cultural studies. The process of "enculturating" the materials used in

Van der Sluis, I., Luz, S., Breitfuss, W., Ishizuka, M., and Prendinger, H. Cross-cultural assessment of automatically generated multimodal referring expressions in a virtual world. *International Journal of Human-Computer Studies 70*, 9 (2012), 611–629.

this study is discussed in Section 3.5.

2.1. Generating Multimodal Referring Expressions

Generating referring expressions is a central task in Natural Language Generation (NLG), and various algorithms which automatically produce referring expressions have been developed (Dale and Reiter, 1995; Krahmer et al., 2003; Jordan and Walker, 2005; Van Deemter and Krahmer, 2006; Van Deemter, 2006; Funakoshi et al., 2006). Most of these algorithms assume that both speaker and addressee have access to the same information. This information can be represented by a knowledge base that contains the objects and their properties present in the domain of conversation. A typical algorithm takes as input a single object (the target) and a set of objects (the distractors) from which the target object needs to be distinguished (borrowing terminology from Dale and Reiter, 1995). The task of the algorithm then is to determine which set of properties is needed to single out the target from the distractors. Research in this area has aimed for human-likeness, that is matching human produced REs that are not only distinguishing, but minimal (i.e., uniquely describing the target object) or overspecified (i.e. including more properties than strictly necessary to identify the target) depending on what human speakers produce.

In human communication, referring expressions which include pointing gestures are commonly used (Beun and Cremers, 1998). Speakers can indicate objects uniquely by using pointing gestures from a close distance, but they also can use pointing gestures from a greater distance which are less precise in that they not unambiguously single out the intended referent (Kranstedt et al., 2005). Similarly, an MRE algorithm should be able to produce not only referring expressions that include precise pointing gestures, but also ones that include underspecified pointing gestures with distractor objects in their scope. The algorithm by Van der Sluis and Krahmer (2007) approaches the generation of MREs as a compositional task in which language and gestures can be combined in a flexible way. To mimic the way in which speakers combine gesture and language, for instance for use by ECAs (Byron, 2003), the algorithm needs sophisticated specifications.

Clark and Wilkes-Gibbs (1986) noted that, in cooperative dialogue, a speaker tries to minimise both her own and the hearer's effort. Consequently, a speaker's goal is to make object identification by the hearer as easy as possible by providing enough but not too much information. At the same time, the speaker wants to minimise her own effort in producing the referring expression. Besides balancing the amount of information, this universal principle also determines the kind of information that is used; in some cases a pointing gesture is the optimal way to

Van der Sluis, I., Luz, S., Breitfuss, W., Ishizuka, M., and Prendinger, H. Cross-cultural assessment of automatically generated multimodal referring expressions in a virtual world. *International Journal of Human-Computer Studies 70*, 9 (2012), 611–629.

refer to an object, whereas in others a linguistic description, or a combination of the two, is more appropriate. Speakers integrate their use of pointing gestures and linguistic material in a compositional way (Lücking et al., 2004; Mc Neill, 2000; Kita, 1990; De Ruiter, 2000; De Ruiter et al., 2012), they do not use precise pointing gestures all the time, but frequently decide to point at a target object from a distance.

Similarly, the algorithm by Van der Sluis and Krahmer (2007) co-relates speech and gesture based on a trade-off between the effort it takes to produce a pointing gesture, based on Fitts's Law (Fitts, 1954), and the effort it takes to produce a linguistic description taking human preference into account (Pechmann, 1989; Dale and Reiter, 1995). Thus, generation of distinguishing MREs can be linked to a notion of effort (i.e. balancing the kind of information necessary for identification and the cost of delivering that information). A detailed description of the algorithm can be found in (Van der Sluis and Krahmer, 2007). The output of this MRE algorithm depends on a function through which the costs of producing linguistic descriptions and pointing gestures can be varied. The algorithm defines the cost of pointing gestures employing Fitts's Law (Fitts, 1954) and thus depends on the distance to the target and the size of the target. The cost of the linguistic descriptions are calculated by summing over the cost of the linguistic properties, which could vary, for instance, according to whether these properties are absolute (such as 'colour' and 'type') or relative, that is, whether they require inspection of other objects in the domain (such as 'size'). Therefore, the amount of linguistic information and the type of pointing gesture included in the referring expression can differ.

An assumption that pervades the literature reviewed in this section is that MRE generation behaviour is, at a basic level, universal and that cultural and linguistic differences can be handled (i.e. accounted for theoretically and implemented in practice) through appropriate parametrisation. This paper tests a particular embodiment of this assumption empirically by contrasting user perceptions (in Ireland and Japan) of three paradigmatic parametrisations described in Section 3.4.

2.2. Cross-Cultural Studies

Adapting an ECA's communicative behaviour to particular cultural traits is expected to make the agent seem more convincing, believable and efficient (Rehm et al., 2009a; Aylett et al., 2009). There seems to be no standard approach to the design of culture specific agents. The work by Jan et al. (2006), for instance, models a set of cultural parameters that define an agent's behaviour according to reviewed literature and tests the realism of the output on people from different

Van der Sluis, I., Luz, S., Breitfuss, W., Ishizuka, M., and Prendinger, H. Cross-cultural assessment of automatically generated multimodal referring expressions in a virtual world. *International Journal of Human-Computer Studies 70*, 9 (2012), 611–629.

cultural backgrounds. In the Cube-G project (Nakano and Rehm, 2009), a corpus elicited from lab studies in which participants are interacting with professional actors is used to train a system to mimic human communicative behaviour in a number of social settings. Yet another approach models ECAs based on observations of people in their natural environment and tests the result in comparable situations (Cassell et al., 2009; Huang et al., 2009). Similarly, the work presented in this paper assesses the output of an algorithm that is based on empirical evidence from human communication. However, in our case we test three types of algorithmic output embedded in a setting that is culturally and linguistically adapted to the preferences and background of two groups of participants.

Current work in the area of enculturated ECAs addresses overlaps and pauses in speech (Endrass et al., 2009), gaze and turn-taking in dialogue (Jan et al., 2006), gesture and posture (Rehm et al., 2009b) and linguistic communication, including linguistic codes (Cassell, 2009). Interestingly, as far as we know, this work does not include the use of MREs. Moreover, we are not aware of any references to work carried out in cross-cultural contexts in which agents are allowed to freely use the virtual space they inhabit to point out particular objects (cross-cultural work in this area appears to be restricted to stationary agents). We rely on cognitive science research as a background for cross-cultural perception and production of MREs in human communication. Nisbett and Miyamoto (2005) argue that perception is culture dependent. In Western cultures, they claim, people organise objects using rules and classification methods, focusing on salient objects independent of their context. In contrast, people from East Asian cultures focus more on the relationships between objects and the context in which they appear, grouping objects based on family resemblance rather than category membership (see Masuda and Nisbett, 2001; Kitayama et al., 2003; Fernald and Morikawa, 1993). Although we may infer from this that the perception and production of MREs differ depending on the cultural background and setting of the speakers, we decided to use linguistic object descriptions composed only of properties that are commonly considered in the literature on generating referring expressions as discussed in Section 3.3.

Pointing gestures are unique to human behaviour and almost inevitable in human communication (Kita, 1993; Kendon, 1994). In this paper we limit ourselves to pointing gestures used to indicate objects, locations or directions (i.e. concrete pointing gestures (Mc Neill, 2000). Although such pointing gestures can be achieved with various body parts, we consider pointing gestures that are produced solely by the hand. Such pointing gestures appear in multiple forms with different interpretations which depend on the cultural background and context of the

Van der Sluis, I., Luz, S., Breitfuss, W., Ishizuka, M., and Prendinger, H. Cross-cultural assessment of automatically generated multimodal referring expressions in a virtual world. *International Journal of Human-Computer Studies 70*, 9 (2012), 611–629.

speaker (Calbris, 1990; Kendon and Versante, 2003; Wilkins, 2003). Although differences in the semantic coordination between linguistic and gestural representations in English and Japanese have been found in lab-based studies with respect to descriptions of events (cf. Kita, 1997; Kita and Özürek, 2003; Özürek et al., 2005), to our knowledge, no studies have been conducted to investigate the use of pointing gestures in the English and Japanese language with respect to particular social contexts in terms of dynamic characteristics and frequency of pointing.

3. A Cross-cultural Study to Assess MRE generation

Three dialogues between two ECAs, a furniture seller and a potential buyer, set in a furniture shop built for this experiment in the Second Life®(SL) environment[1] were presented to subjects in Tokyo and Dublin. These presentations were designed so as to cover three paradigmatic cases of language and gesture combinations through inclusion of five MREs produced by the Van der Sluis and Krahmer (2007) algorithm, and therefore included referring expressions ranging over two extremes with respect to linguistic and pointing information. These cases can be characterised as follows:

POINT: all referents were indicated by precise pointing gestures and linguistic descriptions such as 'this one', for instance, corresponding to a setting of parameter that assigns high cost to linguistic properties and low cost to pointing gestures. Thus the agents move through the shop and close to the furniture items under discussion so as to indicate them uniquely (Figure 1(a)).

LANGUAGE: all available linguistic properties were included in the descriptions which were accompanied by imprecise pointing gestures, corresponding to a parametrisation that assigns low cost to linguistic properties and high cost to pointing gestures. As a result, the agents remain stationary in the front of the shop as depicted in Figure 2 from where the seller agent produce very detailed linguistic descriptions of objects while pointing in their direction (Figure 1(b)).

MIXED: this corresponds to a parametrisation between the above described presentation. Referring expressions consisted of pointing gestures and linguistic descriptions such that the seller ECA moved to point to objects near the

[1] http://secondlife.com/

Van der Sluis, I., Luz, S., Breitfuss, W., Ishizuka, M., and Prendinger, H. Cross-cultural assessment of automatically generated multimodal referring expressions in a virtual world. *International Journal of Human-Computer Studies 70*, 9 (2012), 611–629.

front of the shop and referred to objects further far away (i.e. those in the back of the shop) by using all available linguistic properties as well as an imprecise pointing gesture. Objects in the middle of the shop were referred to through descriptions containing the most preferred linguistic properties (Dale and Reiter, 1995) as well as a pointing gesture which varied according to location. Thus, two of the five targets (a singleton and a set containing a few objects) were identified with precise pointing gestures, two targets (both singletons) were identified with imprecise pointing gestures and one singular target was identified with a slightly imprecise pointing gesture.

(a) (b)

Figure 1: (a) The Irish furniture Seller points at the red chair from a close distance while uttering 'This one', and (b) the Japanese furniture Seller points at the red chair from further away while uttering 'The large red chair in the front of the shop'.

3.1. Hypotheses

In order to test the parametrisations above, we evaluated the corresponding MREs with respect to human-likeness, understandability and social practice. Understandability was defined as the clarity with which objects were referred to through the MREs used by the furniture seller in the dialogue. Human-likeness can be influenced by various factors such as turn-taking behaviour, level of politeness, speech quality etc. In this study, however, all factors (except the MREs) were constant. Therefore human-likeness in this context must necessarily indicate how similar to human behaviour the participants judge the MREs produced by the furniture seller agent (i.e., the way the furniture seller agents moves and speaks) to be. In terms of the human-likeness and understandability criteria we expect that:

Van der Sluis, I., Luz, S., Breitfuss, W., Ishizuka, M., and Prendinger, H. Cross-cultural assessment of automatically generated multimodal referring expressions in a virtual world. *International Journal of Human-Computer Studies 70*, 9 (2012), 611–629.

Figure 2: The agents in the furniture shop.

HH, Human-likeness: the furniture seller ECA will be perceived as most human-like in the MIXED presentation. In the MIXED presentation the ECA balances his efforts according to the physical context by moving close to items in the front to point at them unambiguously, and uses more elaborate linguistic descriptions for the items that are located further away.

Symbolically: MIXED $\prec$ LANGUAGE, POINT[2];

HU, Understandability: the seller ECA will be perceived as most understandable in the POINT presentation. In the POINT presentation the agent points at each item unambiguously from a close distance.

Symbolically: POINT $\prec$ MIXED, LANGUAGE;

As regards "social practice", the study is restricted to user preferences for a particular "style of selling" and the assumption behind it is that participants will equate the different presentation styles to the ECA's salesmanship. The hypothesis in this case is the following:

HS, Social Practice: the seller's style of selling in the POINT presentation will be preferred by the participants. In the POINT version the agent makes a greater physical effort for the sake of the customer by pointing at each item from a close distance. It is expected that participants will appreciate this effort when they are asked to imagine themselves in the role of a customer.

Symbolically: POINT $\prec$ MIXED, LANGUAGE).

[2]In this paper the expression "A $\prec$ B" stands for "A is preferred to B"; the comma denotes no specific preference.

Van der Sluis, I., Luz, S., Breitfuss, W., Ishizuka, M., and Prendinger, H. Cross-cultural assessment of automatically generated multimodal referring expressions in a virtual world. *International Journal of Human-Computer Studies 70*, 9 (2012), 611–629.

Analysis of variance, supplemented by post-hoc tests considering gender and familiarity with Second Life where appropriate, will be employed to assess the above hypotheses. Language use is gender-dependent, specially in Japanese (see Section 3.5). In testing the above described hypotheses, as well as the assumptions on MRE generation behaviour discussed below, we would also like to quantify the effect this variable has (if any) on the participant's judgements. Similarly, we conjecture that familiarity (or not) with the virtual environment might lead to differences, and therefore we also tested with respect to this variable. Rejection of the null hypotheses (that there are no differences in the presentation types) will be sought at the $p < 0.05$ level (Cairns, 2007). Where the difference between means is significant and conforms to the ordering described in the hypothesis statements above we say that the hypothesis has been "confirmed" or, equivalently, that the assessment "agrees with the hypothesis".

As regards the assumption of universality of MRE generation behaviour discussed in 2.1, we seek to reject the (null) hypothesis that Tokyo and Dublin participants agree in their assessments of MRE types. Where this null hypothesis is rejected at the above mentioned level, the preferred styles are taken to correspond to the preferences of users for ECA interaction in the respective cultural setting (in this narrowly defined sense).

3.2. Participants

A total of 30 Japanese and 30 English native speakers participated in the main study, 15 males and 15 females in each group. The native speakers of English participated at Trinity College Dublin and the native speakers of Japanese participated at the National Institute of Informatics at Tokyo. They were paid 10 Euros and 1,000 Yen, respectively, for their participation. The average age of the Dublin group was 21.13 (std. 1.83). To the statement 'I am familiar with virtual worlds', 13 said 'yes' and 17 said 'no'. None of the participants in the Dublin group visited Second Life regularly. To the statement 'I like Second Life', 24 said 'don't know', 5 said 'no' and 1 responded 'yes'. The average age in the Tokyo group was 22.07 (std. 1.99), 10 indicated that they were not familiar with virtual worlds, 20 said that they knew about them. Four participants in the Tokyo group indicated that they visited Second Life regularly, 26 did not. To the statement 'I like Second Life', 22 said 'don't know', 7 said 'no' and 1 responded 'yes'. Participants' comments on virtual world in general are presented in Table 6 in the Appendix.

Van der Sluis, I., Luz, S., Breitfuss, W., Ishizuka, M., and Prendinger, H. Cross-cultural assessment of automatically generated multimodal referring expressions in a virtual world. *International Journal of Human-Computer Studies 70*, 9 (2012), 611–629.

3.3. Setting

We employed a Fully Generated Scripted Dialogue (FGSD) approach (André et al., 2000; Williams et al., 2007) to evaluate the output of the Van der Sluis and Krahmer algorithm. With FGSD entire dialogues are produced by one generator. Initially, scripted dialogues made heavy use of canned text, but recently this approach has been integrated with Natural Language Generation techniques (Van Deemter et al., 2008; Piwek, 2008). FGSD allows us to produce dialogues, without implementing a full natural language interpretation module. However, more importantly, using FGSD ensure that the study participants were all presented with the same materials, which would have been impossible if participants had been allowed to interact with the agent directly.

The furniture domain was chosen because detailed data on how speakers refer to furniture items in terms of their cardinality and their 'colour', 'size' and 'location' properties are available through the COCONUT corpus (Di Eugenio et al., 2000) and the TUNA corpus (Gatt et al., 2007), and we hoped that these knowledge sources would help us to construct believable REs. A setting was built in Second Life®(SL) specifically for this study. The online virtual world enabled us to choose a specific domain of conversation in which all objects and their properties are known. This allows for complete semantic and pragmatic transparency, which is important for a content determination task like the generation of referring expressions. The domain and the dialogue were designed as to allow for the assessment of five carefully chosen MREs that reflect current research issues like cardinality, location and vague properties like 'size'. An existing gesture generation system was employed to control and animate agents in SL which automatically adds nonverbal behaviour to a dialogue and produces a play-able script using MPML3D (Prendinger et al., 2010; Breitfuß et al., 2008). We extended this system by adding three levels of pointing gestures, precise, imprecise and very imprecise, as suggested in (Van der Sluis and Krahmer, 2007). The English voices were created with Loquendo[3]. For the Japanese voices NeoSpeech™[4] was used. It is conceivable that the differences between the synthesised voices used for the materials presented to the Japanese and the English speaking participants in the study might have influenced their respective results. However, using synthesised speech ensured that, within the two language groups, the voice quality was uniform across all conditions, which would have been harder to control had natural

[3]http://www.loquendo.com/

[4]http://www.neospeech.com/

Van der Sluis, I., Luz, S., Breitfuss, W., Ishizuka, M., and Prendinger, H. Cross-cultural assessment of automatically generated multimodal referring expressions in a virtual world. *International Journal of Human-Computer Studies 70*, 9 (2012), 611–629.

human voice-overs been used.

Figure 3: The furniture shop.

The virtual shop (Figure 3) was populated with two ECAs. It contained 53 objects, 16 of which are actually referred to in the dialogues, (i.e. a large red office chair, a large blue desk, a small blue desk, a set of 6 large red chairs and a set of 7 small green chairs). The other items in the shop were used as distractor objects. The target objects were distributed so that there were objects to refer to in the front (i.e. the singular red chair), in the middle (i.e. the sets of chairs) and in the back of the shop (i.e. the desks). As the ECAs by default were located in the front of the shop, it was relatively easy for them to walk over to and point precisely at the objects located in the front of the shop. Pointing precisely at the objects located in the middle of the shop would require some more effort, and walking to the back of the shop to precisely point out the objects located far away would take even more time and effort.

The dialogue used was originally written in English and consists of 19 utterances with 5 first-mention references to furniture items (3 singletons and 2 sets containing multiple objects). It features a conversation between a female agent purchasing furniture for her office, and a male shop-owner guiding her through the store while describing some furniture items. Hence, the quality of the MREs produced by the algorithm was assessed with respect to the behaviour of the furniture seller. For validation purposes, the dialogue was read by 3 native speakers of English and improved based on their feedback. The dialogue was used as a

Van der Sluis, I., Luz, S., Breitfuss, W., Ishizuka, M., and Prendinger, H. Cross-cultural assessment of automatically generated multimodal referring expressions in a virtual world. *International Journal of Human-Computer Studies 70*, 9 (2012), 611–629.

template in which the five referring expressions were varied so as to produce the three presentation type define above (i.e. POINT $\prec$ MIXED, LANGUAGE). The referring expressions used to fill out the slots can be automatically reproduced with the MRE algorithm (Van der Sluis and Krahmer, 2007).

3.4. Experiment Design

A pilot study was conducted prior to the large scale study reported in this paper in order to test its setup and fine-tune the methods employed. The pilot had ten participants recruited among colleagues in our respective departments. For further details on this study the reader is referred to Breitfuß et al. (2009).

For the large scale study, two groups of participants (one based in Dublin and the other based in Tokyo) were asked to judge the above described presentations in terms of human-likeness, understandability and social practice. Judgements were collected with questionnaires that are presented in section 3.6.Hypotheses on human-likeness (HH) and understandability (HU) were tested between subjects as well as within subjects. The hypothesis on social practice (HS) is only tested between subjects. The experiment has therefore a two by three design, with *MRE type* as a within subjects variable and *Language/Culture* as a between subjects variable. Table 1 summarises the experimental design.

Table 1: Overview of the experimental design with *language/culture* as between subjects and *MRE quality* as within subjects variables.

		Language/Culture	
		Japanese (Tokyo)	*English (Dublin)*
MRE type	POINT	I	II
	LANGUAGE	III	IV
	MIXED	V	VI

3.5. Translation and Localisation

The next step in the preparation of the cross-cultural study was to translate and localise the English presentations to the Japanese language and culture. In close collaboration with one of the authors of this paper and guided by contextualisations of the dialogue in SL, our professional Japan-based translator took meticulous care in the translation of the dialogue and the referring expressions therein. The goal of the translation was to produce a Japanese dialogue in which the referring expressions were as close to the English originals as possible. The

Van der Sluis, I., Luz, S., Breitfuss, W., Ishizuka, M., and Prendinger, H. Cross-cultural assessment of automatically generated multimodal referring expressions in a virtual world. *International Journal of Human-Computer Studies 70*, 9 (2012), 611–629.

scenario, however, was adapted to the Japanese style such that the Japanese audience could easily conceive the situation. With respect to the MREs, difficulties arose mainly in translating the locative expressions and the use of demonstratives (see Van der Sluis et al., 2009; Van der Sluis and Luz, 2011, for details). The dialogue was first checked with a native speaker of Japanese based in Japan, who was impressed with its quality. In addition, we performed two data elicitation studies to further validate the work of our translator (Van der Sluis and Luz, 2011).

In general, the English dialogue was very verbose compared to an equivalent Japanese dialogue. The Japanese language (especially colloquial language) has a great tendency to omit, abbreviate and to positively use 'silence', or in other words, to trust in the addressee's ability to comprehend the implications of the unspoken words. The translation also recovered some differences in the attitudes expressed in the dialogue. The beginning of the dialogue, for instance, was altered considerably in the Japanese version, both in speech and in gestures. In the Dublin version, the furniture seller opens the dialogue with 'Hi, how can I help you?' accompanied with no particular gesture. In the Japanese version, the furniture seller says: 'Irasshai-mase' meaning 'Welcome to our shop' accompanied with a bow of 30 degrees. A literal translation of 'how can I help you', was considered to be too assertive. Overall, our translator chose translations that felt most natural in the given context and which preserved the flow of the dialogue as well as the MRE output as much as possible.

The language used in Japanese dialogues is extremely dependent on the relationship between the dialogue partners, their gender, age and social standing. Likewise, the virtual furniture shop, its standing and size, influences the attitude of the agents that populate it, and the phrases and words they use. In order to maintain naturalness with respect to these contextual factors, we used localised agents in the Tokyo as well as in the Dublin set up, and designed their physical appearances so as to suggest an age of about 25 to 35 years. The relationship between the agents was defined as a 'shop owner-office lady' relationship and the kind of furniture shop as an average middle-end shop. All this implied that the kind of Japanese language used by the agents should be a socially polite form whose modern usage, especially among the younger generation of Japan, does not include much difference between the masculine and feminine form.

In the Japanese presentations, the size of the non-deictic gestures was adapted to occupy less space (Nakano and Rehm, 2009) and the pointing gestures were made using the whole hand instead of with just the index finger. The virtual furniture shop that was built for the study displayed the furniture in a relatively large space, which made it easy for the agents to move around. In Dublin as well as

Van der Sluis, I., Luz, S., Breitfuss, W., Ishizuka, M., and Prendinger, H. Cross-cultural assessment of automatically generated multimodal referring expressions in a virtual world. *International Journal of Human-Computer Studies 70*, 9 (2012), 611–629.

in Tokyo, the setting may not reflect a typical furniture store. However, we decided to keep the space for the agents so that they were able to move around without stumbling over objects. The appearance of the ECAs were localised in terms of their clothing, hair colour and body size. The facial features were deliberately kept vague to prevent distractions due to misaligned lip-syncing and expressions of emotions and personality.

To check the set up and materials, a small pilot study was conducted in Tokyo with three native speakers of Japanese. No problems were reported, the presentations were understandable and considered in agreement with the Japanese style.

3.6. Materials

Materials consisted of four videos of presentations in SL[5], one introduction and three presentations featuring the dialogues with the algorithmic outputs described in Section 3.3. The material was presented to participants on a $50''$ LCD screen using a standard media player. To minimise any distractions, participants sat in a secluded space, about a meter's distance from the screen and used a headset while watching the videos. Apart from the debriefing which was done by the experimenter in person, the introduction, consent form, instructions and questionnaires were handed to participants on paper. The materials for the Dublin group were written in English, the Tokyo group received the translations in the Japanese language. For both languages, all materials were checked and validated. The questionnaires were validated in several steps: first, for content through presentation to several experts in an iterative process during the questionnaire's development, then by initial application to a small group, still in the development phase, and finally by a pilot study, reported elsewhere (Breitfuß et al., 2009). Since, to the best of our knowledge, no other questionnaires exist which assess MRE generation behaviour, it was impossible to assess validity in terms of correlation with other similar instruments.

Three questionnaires were used, which we will refer to as *A*, *B* and *C*. Questionnaire *A* aimed at obtaining a baseline and contained ten questions about the agents, the setting and the conversation plus some general questions about the participant's background. Some questions were open and some used a Likert scale that ranged from one ('strongly agree') to seven ('strongly disagree'). The data resulting from questionnaire *A* were analysed with t-tests at the $p < .05$ level. Questionnaire *B* was used for evaluating the three presentations and consisted of

[5]The presentations can be viewed http://isluis.home.xs4all.nl/videos/

Van der Sluis, I., Luz, S., Breitfuss, W., Ishizuka, M., and Prendinger, H. Cross-cultural assessment of automatically generated multimodal referring expressions in a virtual world. *International Journal of Human-Computer Studies 70*, 9 (2012), 611–629.

four sections addressing the interaction between the agents, the agents themselves, their role-play and the conversation. In total there were twenty-one questions, of which three were relevant to the issues investigated in this paper. For all questions, questionnaire *B* used the same Likert scale as *A* and was handed to the participants to collect their assessment of the three presentations. To test our hypotheses on human-likeness (HH) and understandability (HU) we used three statements in this questionnaire (i.e. S1, S2 and S4 as specified below) and analysed the data collected with these statements through ANOVA tests with repeated measures at the $p < .05$ level. In terms of the variables used in Table 1, S1, S2 and S4 were compared within subjects I, III, and V for the Tokyo group and II, IV and VI for the Dublin group. The between subjects comparison for the two groups was carried out for S1, S2 and S4 (I, III, and V versus II, IV and VI) and an interaction was computed between groups over all variables within the groups. Questionnaire *C* compared the three presentations directly and was handed to the participants after they had seen all three presentations. We used two statements in this questionnaire to gain further evidence for our hypotheses for human-likeness (HH) and understandability (HU) (respectively S3 and S5, as specified below). One statement in questionnaire *C* was used to test our hypothesis on social practice (HS) (S6, see below). Possible answers were ('Presentation 1', 'Presentation 2', 'Presentation 3', 'Don't know', 'No Difference'). The data obtained with questionnaire *C* were analysed using the frequencies with which participants in our study selected one of the five answers to the statements. In terms of the variables used in Tabel 1, the frequencies collected for the statements S3, S5 and S6 were compared between groups I, III, and V for the Tokyo group versus II, IV and VI for the Dublin group.

All questionnaires allowed participants to enter free comments, which participants were explicitly asked to supply at the start of the session. The Tokyo group wrote 202 comments, with a total word count of 3328, whereas the Dublin group wrote 314 comments with a total word count of 6594. The Japanese comments were automatically (machine) translated into English. Where the translations were difficult to understand, the comments were then further translated manually by a native speaker of Japanese. A selection of comments collected in various stages of the experiment is presented in the Appendix.

3.7. Procedure

At the start of the experiment participants were handed a welcome sheet describing the purpose of the study, in which they were introduced to the study with the following text:

Van der Sluis, I., Luz, S., Breitfuss, W., Ishizuka, M., and Prendinger, H. Cross-cultural assessment of automatically generated multimodal referring expressions in a virtual world. *International Journal of Human-Computer Studies 70*, 9 (2012), 611–629.

> 'In this study we want to investigate users' impressions of a dialogue between our two animated 3D agents. They will be auditioning for a part in a play.'

The participants signed a consent form which stated that their participation in the study was voluntary and anonymous and that their data would be used for research purposes only and would be treated with full confidentiality. Subsequently, participants received a more detailed instruction to the experiment and were asked to put on the headset and to watch a short introductory video in which the agents introduced themselves. After the introduction participants filled out questionnaire A. Then a further instruction prepared them for our three presentations in which the auditioning part was further specified as follows:

> 'In each presentation, the male character, Wena, will audition for the part of furniture seller and the female character, Haruna, will play the supporting part of someone who wants to buy furniture.'

Participants were instructed to carefully observe especially the behaviour of the male agent and prepare for the questionnaires that would follow the presentations. The order in which participants saw the presentations was randomised and after each video, participants filled out questionnaire *B* about the presentation they had just seen. After participants had seen all three presentations they were asked to fill out questionnaire C. Finally, participants were debriefed by the experimenter and paid.

3.8. Results

Baseline Descriptives

Participants in our experiment (N= 30) filled out questionnaire A after they had seen the introductory presentation and before they saw the three presentations featuring the dialogues. The descriptives (means and standard deviations) and the results of an independent samples t-test are shown in Table 2, no values were missing. The data tell us that both the Dublin and the Tokyo group were positive about their familiarity with the setting. The Dublin group generally agreed with the statements we asked them to rate, the Tokyo group was significantly less positive than the Dublin group and rated the statements between 3 and 4 on our 7-point scale. The quality of the voices used for the agents was positively received by both groups (i.e. rates between 2 and 3). They only differed with respect to the pleasantness of the voice of the female agent. Despite the fact that we used different speech synthesisers in the two groups, the Tokyo and Dublin groups rated the

Van der Sluis, I., Luz, S., Breitfuss, W., Ishizuka, M., and Prendinger, H. Cross-cultural assessment of automatically generated multimodal referring expressions in a virtual world. *International Journal of Human-Computer Studies 70*, 9 (2012), 611–629.

Table 2: Baseline means and standard deviation within brackets for the *Dublin* and *Tokyo* group (N = 30) rated on a scale one ('strongly agree') to seven ('strongly disagree') as collected with Questionnaire A (where ... stands for 'a virtual agent to move/speak') and independent samples t-test results, where * denotes significant difference at the $p < .05$ level.

	Dublin	*Tokyo*	*t-test*
Familiarity:			
I am now familiar with the furniture shop	2.73 (1.23)	3.97(1.21)	3.90 *
I am now familiar with the male agent and his role	2.00 (0.70)	3.57(1.25)	6.00 *
I am now familiar with the female agent and her role	2.24 (0.70)	3.80(1.32)	5.11 *
Quality:			
The male agent has a pleasant voice	3.93 (1.75)	3.10(1.47)	1.97
The female agent has a pleasant voice	3.79 (1.74)	2.77(1.40)	2.50 *
The male agent speaks clearly	2.79 (1.15)	2.80(1.47)	.020
The female agent speaks clearly	3.10 (0.28)	2.57(1.33)	1.45
Human-likeness:			
The male agent moved in a human-like manner (S1)	5.00 (1.70)	4.03(1.65)	2.23 *
The female agent spoke in a human-like manner	4.87 (1.63)	3.77(1.61)	2.62 *
The male agent spoke in a human-like manner (S2)	5.30 (1.42)	4.40(1.38)	2.49 *
The female agent moved in a human-like manner	5.17 (1.53)	3.90(1.35)	3.40 *
Expectations:			
The female agent moved better than I expected...	4.63 (1.54)	3.00(0.98)	4.89 *
The male agent spoke better than I expected...	4.90 (1.45)	3.43(1.28)	4.16 *
The female agent moved better than I expected...	5.13 (1.25)	3.80(1.06)	4.44 *
The male agent moved better than I expected...	5.27 (1.08)	3.67(1.21)	5.40 *
Understandability:			
I found the conversation easy to follow	2.72 (1.53)	3.10(1.56)	0.93

voice quality almost the same. The ratings for the human-likeness of the agents significantly differ between the two groups. The Tokyo group was more positive than the Dublin group (4 vs. 5). In the statements about their expectations participants were asked to compare what was presented with what they expected from virtual agents in terms of the manner in which they would speak and move. The Tokyo group judged the agents more similar to what they expected (3 to 4) than the Dublin group (6). As regards understandability, both groups responded positively to the statement 'The conversation was easy to follow' (3).

Comments, see Table 7, also show that compared to the Dublin group, Tokyo group was generally less negative, not only in the overall number of negative comments but also in the sense that their verbalisations were less strong ('a little strange' vs. 'jerky') and usually included some positive feedback as well. We

Van der Sluis, I., Luz, S., Breitfuss, W., Ishizuka, M., and Prendinger, H. Cross-cultural assessment of automatically generated multimodal referring expressions in a virtual world. *International Journal of Human-Computer Studies 70*, 9 (2012), 611–629.

checked for differences in the baseline data within the Tokyo group where 4 participants had indicated that they visited Second Life regularly, but did not find any.

The following sections present an analysis of results for presentation style preferences in Dublin and Tokyo, including an assessment of differences by gender.

HH - Human-likeness: MIXED $\prec$ LANGUAGE, POINT

To test for effects on human-likeness of the MREs we asked participants to rate to which extent they agreed with the following statements[6]:

S1: The male agent spoke in a human-like manner.

S2: The male agent moved in a human-like manner.

On the responses to these questions in baseline questionnaire A, which the participants filled out with respect to the introductory presentation, we performed a t-test for independent samples and found significant differences between the responses of the Dublin and the Tokyo group (S1: t(1,58)= 2.23, $p < .05$, Cohen's d = 0.58 and S2: t(1,58)= 2.49, $p < .05$, Cohen's d = 0.64). The means and standard deviations, illustrated in Table 2, show that Dublin group generally disagreed (5) with the two statements, while the Tokyo group on average answered the two questions neutrally (4). To test if participants agreed on the human-likeness of the male ECA in each of the presentations, ANOVAs with repeated measures were performed on the data elicited with S1 (spoke in a human-like manner) and S2 (moved in a human-like manner) in questionnaire B, which the participants filled out after each of the three presentations (POINT, LANGUAGE and MIXED). The means and standard deviations for these data are presented in Table 3. In all cases the independent variables were MRE type and cultural setting (as outlined in Table 1) and the dependent variable was the rating given. Mauchly's test indicated that the assumption of sphericity was not violated for MRE type ($W(2) = 0.96, p = 0.35$). For statement S1 (Questionnaire B), whether the male agent spoke in a human-like manner, we found an effect between groups ($F(1,58) = 12.44,\ p < .05,\ \eta^2 = 0.13$), which tells us that the Tokyo group found the manner in which the agent spoke significantly more human-like than the Dublin group. In fact, the Tokyo group agreed with the statement (3) when

[6]Based on the data from the pilot study, we split the statement 'The male agent behaved naturally' into S1 and S2.

Van der Sluis, I., Luz, S., Breitfuss, W., Ishizuka, M., and Prendinger, H. Cross-cultural assessment of automatically generated multimodal referring expressions in a virtual world. *International Journal of Human-Computer Studies 70*, 9 (2012), 611–629.

judging the POINT and the LANGUAGE presentations, where the Dublin group disagreed (5). The two groups judged the MIXED presentation similarly (4). Post-hoc Tukey HSD tests indicated a significant difference for the POINT presentation between the two groups ($p < .05$). There was also an interaction ($F(1,58) = 14.44$, $p < .05$, $\eta^2 = 0.05$), which tells us that there was a significant difference between how the Dublin group and the Tokyo group rated the human-likeness of the three presentations.Overall, the Dublin group found that the agent spoke most human-like in the MIXED presentation and the least human-like in the POINT presentation, while the Tokyo group found him most human-like in the POINT presentation and least human-like in the MIXED presentation. Post-hoc pairwise, paired t-tests with Holm correction for multiple testing showed no significance for the Tokyo group. However, the same test showed significant differences between POINT and LANGUAGE ($p = 0.002$) and between POINT and MIXED presentations ($p = 0.003$) for the Dublin group.

Table 3: Means with standard deviation within brackets for the statements in Questionnaire B about the human-likeness of the manner in which the male agent spoke (*S1*) and moved (*S2*) and understandability of the MREs (*S4*) in the POINT, LANGUAGE and MIXED presentations, where '*' indicate a significant difference at the $p < .05$ level for post-hoc Tukey HSD tests.

	POINT			LANGUAGE		MIXED		
	Tokyo	*Dublin*		*Tokyo*	*Dublin*	*Tokyo*	*Dublin*	
S1	3.37 (1.65)	5.30(1.29)	*	3.53(1.28)	4.57(1.61)	4.00(1.37)	4.40(1.65)	
S2	3.39 (1.29)	5.57(1.41)	*	4.07(1.20)	4.60(1.57)	3.77(1.36)	5.67(1.49)	*
S4	2.30 (1.24)	1.43 (.57)		3.17(1.51)	2.50(1.76)	2.20(1.22)	2.20(1.16)	

For statement S2, we also found an effect between groups ($F(1,58) = 29.03, p < .05, \eta^2 = 0.2$), which indicates that the Dublin and the Tokyo group differed significantly in their judgements about the human-likeness of the manner in which the agent moved. Post-hoc Tukey HSD tests resulted in significant differences between the two groups when comparing the MIXED and the POINT presentations at the $p < .05$ level. Similarly an interaction effect was found ($F(1,116) = 5.4, p < .05, \eta^2 = 0.04$) which prompted post-hoc analysis of MRE types within groups. The Tokyo group rated the agent's movement in the POINT presentation most human-like (3) and his movement in the LANGUAGE presentation the least human-like (4). As before, however, pairwise t-tests showed that the differences were non-significant. In contrast, the Dublin group preferred the LANGUAGE pre-

Van der Sluis, I., Luz, S., Breitfuss, W., Ishizuka, M., and Prendinger, H. Cross-cultural assessment of automatically generated multimodal referring expressions in a virtual world. *International Journal of Human-Computer Studies 70*, 9 (2012), 611–629.

sentation (4 to 5), and rated the POINT and the MIXED presentations similarly (i.e. 5 to 6). Pairwise t-tests with with Holm correction showed significant differences between POINT and LANGUAGE ($p < .05$) and between LANGUAGE and MIXED presentations ($p < .05$).

We proceeded with a check for gender effects within Tokyo and the Dublin group for statements S1 (spoke human-like) and S2 (moved human-like). Table 11 in the Appendix of this paper presents the results for the Tokyo group. In general, females in the Tokyo group found that the agent spoke and moved more human-like than the males, however, there were no significant differences. Table 12 in the Appendix presents the results for the Dublin group by gender. In the Dublin group we found a between subject effect for S1 ($F(1,28) = 11.92, p < .05, \eta^2 = 0.1$), which indicates a significant difference between the male and female participants from Dublin in the way they rated the manner in which the agent spoke. Female participants judged the MIXED and the POINT presentation similarly human-like, whereas male participants judged the MIXED presentation less human-like than the POINT presentation.

After they had seen the three presentations, we asked participants to compare the presentations directly in questionnaire C. Participants were asked to indicate:

S3: The conversation between the agents was most human-like in:

In this case, participants could choose ('Presentation 1', 'Presentation 2', 'Presentation 3', 'No difference', 'Don't Know')[7]. The resulting data presented in Table 4 shows that the Tokyo group preferred the POINT presentation with the MIXED coming a close second. The participants in the Dublin group were more divided, but most also preferred the POINT presentation. The differences in preferences between the two cultural groups are not significant according to Pearson's χ^2 test: $\chi^2(4) = 3.78, p = 0.4, \phi = 0.25$). For further insight into the reasons for these preferences, we refer to the comments listed in Table 8 in the Appendix.

In summary, as regards HH one can conclude from the statistically significant results presented above that (i) the Dublin and the Tokyo group do not agree on the human-likeness of the presentations, and (ii) that the Dublin group generally rated the more verbose presentations (MIXED and LANGUAGE) more human-like, thus partially confirming HH for that group. In direct comparison, however, the majority of participants in both groups preferred the least verbose presentation (POINT).

[7]Note that the order in which participants saw the presentations was randomised.

Van der Sluis, I., Luz, S., Breitfuss, W., Ishizuka, M., and Prendinger, H. Cross-cultural assessment of automatically generated multimodal referring expressions in a virtual world. *International Journal of Human-Computer Studies 70*, 9 (2012), 611–629.

Table 4: Judgements of participants with respect to statements on human-likeness (*S3*) and the understandability (*S5*) of the conversation and the preferred style of selling (*S6*) obtained with the comparisons in Questionnaire C.

		POINT	LANGUAGE	MIXED	*No Difference*	*Don't know*
S3	*Tokyo*	12 (40%)	5 (16.67%)	10 (33.33%)	2 (6.67%)	1 (3.33%)
	Dublin	13 (43.33%)	8 (26.67%)	9 (30%)	0	0
S5	*Tokyo*	14 (46.67%)	4 (13.33%)	11 (36.67%)	1 (3.33%)	0
	Dublin	9 (30%)	6 (20%)	7 (23.33%)	7 (23.33%)	1 (3.33%)
S6	*Tokyo*	12 (40%)	3 (10%)	13 (43%)	1 (3.33%)	1 (3.33%)
	Dublin	17 (56.67%)	6 (20%)	7 (23.33%)	0	0

HU - Understandability: POINT $\prec$ MIXED, LANGUAGE

To find out if participants found the MREs used in the presentations understandable we asked after each presentation in questionnaire *B* to what extent they agreed with the statement:

S4: It was always clear to me which item was under discussion.

Table 3 summarises the results for S4. ANOVA was used to test the results. We found a between subjects effect ($F(1,58) = 4.87, p < 0.05, \eta^2 = 0.04$), which shows that the Dublin group found the presentations significantly more understandable than the Tokyo group. This is especially true for the POINT and the LANGUAGE presentations, as the MIXED presentation was rated similarly between groups.Unlike the previous cases, Mauchly's test showed that sphericity was violated for both the within-subjects variable MRE type ($W(2) = 0.74, p < 0.05$) and it's interaction with cultural setting ($W(2) = 0.74, p < 0.05$). Applying the Greenhouse-Geisser correction ($\varepsilon = 0.792$) we found a within subjects effect ($F(1.58, 91.86) = 11.12, p < 0.05, \eta^2 = 0.09$). Pairwise, post-hoc, paired t-tests indicated that the MREs in the POINT presentation were considered more understandable than LANGUAGE ($p < 0.01$) and MIXED presentations ($p < 0.01$) by the Dublin group. The Tokyo group considered the MREs in the MIXED presentation more understandable than the LANGUAGE presentation ($p < 0.01$), but no significant difference was found between MIXED and POINT ($p = 0,66$). When corrected for sphericity, the interaction between the explanatory variables was not found to be significant ($F(1.58, 91.86) = 2.37, p = 0.11$).

We checked our data for statement S4 on effects of gender. In the Appendix Table 11 presents the results for the Tokyo group and Table 12 presents the results for the Dublin group. For the Dublin group we found a between subjects

Van der Sluis, I., Luz, S., Breitfuss, W., Ishizuka, M., and Prendinger, H. Cross-cultural assessment of automatically generated multimodal referring expressions in a virtual world. *International Journal of Human-Computer Studies 70*, 9 (2012), 611–629.

effect ($F(1,28) = 4.68 p < .05, \eta^2 = 0.14$), which shows that, in general, female participants found the presentations significantly more understandable than male participants (females avg = 1.76, sd = .95; males avg = 2.33, sd = 1.27).The same effect was found for the Tokyo group ($F(1,28) = 4.814, p < .05, \eta^2 = 0.15$, females avg = 2.18, sd = 1.20; males avg = 2.93, sd = 1.34).

We also checked the understandability of the MREs with statement S5 in the comparison questionnaire *C,* where participants were asked to indicate:

S5: I found the conversation most easy to follow in:

As with S3 (The conversation between the agents was most human-like in:), participants could choose 'Presentation 1', 'Presentation 2', 'Presentation 3', 'No difference' or 'Don't Know'. The resulting data presented in Table 4 shows that the Tokyo group had a clear preference for the POINT presentation with the MIXED presentation in second place. The participants in Dublin group were more divided, but most prefer the POINT presentation. Pearson's χ^2 test did not reveal the slight differences in choices by the Dublin and Tokyo groups to be statistically significant ($\chi^2(4) = 7.88, p = 0.1, \phi = 0.36$).Table 9 in the Appendix presents the participants' comments on the understandability of the MREs produced by the furniture seller.

HS - Social Practice: POINT $\prec$ MIXED, LANGUAGE

In the context of the furniture shop and the roles of the agents in it, we expected participants both in Dublin and in Tokyo to prefer the furniture seller that shows all relevant sale items from nearby instead of from a distance. In other words, we expected that, when imagining themselves in the role of customers, the participants would prefer looking closely at the items on sale and therefore appreciate the agent's effort in pointing at the items from a close distance. To test this hypothesis, participants were asked to indicate in questionnaire C:

S6 'If I were a buyer, I would prefer to deal with the male agent from: ...'

The possible answers were one of the presentations (1, 2, or 3), 'no difference' and 'don't know'. Results, presented in Table 4, show that the Dublin group had a clear preference for the POINT presentation. Participants in the Tokyo group are much more divided in their judgements, most of them prefer the MIXED presentation but the POINT presentation is almost equally preferred. Comparison between the overall distributions of preferences between the Dublin and Tokyo groups though Pearson's χ^2 test, however, showed no statistical significance ($\chi^2(4) = 5.32, p = 0.26, \phi = 0.29$).The participants' comments about the furniture seller's social practice are presented in Table 10 in the Appendix.

Van der Sluis, I., Luz, S., Breitfuss, W., Ishizuka, M., and Prendinger, H. Cross-cultural assessment of automatically generated multimodal referring expressions in a virtual world. *International Journal of Human-Computer Studies 70*, 9 (2012), 611–629.

4. Discussion

Table 5 gives an overview of the hypotheses we tested with our study. From the results of our ANOVA tests for statements S1, S2 and S4 from questionnaire B, we can infer a threefold rejection of the null hypotheses with respect to cultural differences. The Dublin group and the Tokyo group do not agree in their judgements of the presentations in terms of human-likeness (HH), understandability (HU) or social setting (HS). On the other hand, direct comparison of participant's ratings after they had seen all three presentations in terms of human-likeness (S3) and understandability (S5), indicates that both groups preferred the POINT presentation. However, both groups vary in their judgements for the MIXED and POINT presentations depending on which of the three criteria they were asked to consider.Neither group liked the LANGUAGE presentation (i.e. the most verbose, least mobile presentation). The null hypothesis on social practice (HS), that the Dublin and Tokyo groups would prefer the same selling style was rejected by direct comparison in questionnaire C with statement S6.

Table 5: Overview of the hypotheses (*HH*, *HU* and *HS*) tested within subjects (*D*ublin and *T*okyo) and between subjects (*T* vs *D*) and the *Inter*action between the two groups and the three presentations, with their outcomes (MIXED, POINT or LANGUAGE) for each of the statements (*S#*). Where applicable, boldface indicates agreement with the hypothesis. Significant differences at the levels $p < 0.05$ are denoted '*'.

	Dublin	*Tokyo*	*T vs D*	*Inter.*
HH, Human-likeness: (MIXED $\prec$ LANGUAGE, POINT)				
Spoke in human-like manner (S1,QB)	**MIXED**	POINT	*	*
Moved in human-like manner (S2,QB)	LANGUAGE	POINT	*	
Conversation was most human-like in: (S3,QC)	POINT	POINT		
HU, Understandability: (POINT $\prec$ LANGUAGE, MIXED)				
Item under discussion was known (S4,QB)	**POINT**	MIXED	*	*
Conversation was most easy to follow in: (S5,QC)	**POINT**	**POINT**		
HS, Social Practice: (POINT $\prec$ LANGUAGE, MIXED)				
Preferred selling style: (S6,QC)	**POINT**	MIXED		

Within the two groups, the participants from Dublin matched our expectations best confirming our understandability (HU) and our social practice hypothesis (HS). They found the presentation in which the agent identified each furniture item at a close distance easiest to follow as well as the preferred style of selling furniture. The Dublin group also partly confirmed our human-likeness hypothesis

Van der Sluis, I., Luz, S., Breitfuss, W., Ishizuka, M., and Prendinger, H. Cross-cultural assessment of automatically generated multimodal referring expressions in a virtual world. *International Journal of Human-Computer Studies 70*, 9 (2012), 611–629.

as they judged the way the agent spoke in the MIXED presentation the most human-like. The Tokyo group only partly confirmed our understandability hypothesis by expressing a preference for the POINT presentation, when asked to compare which of the three presentations they found the easiest to follow in statement S5. In terms of human-likeness the Tokyo group definitely preferred the POINT presentation to the MIXED one. The Tokyo group preferred the MIXED presentation with respect to understandability (S4) and social practice (S6), indicating that the furniture seller agent should balance his efforts in indicating sale items and only move to a particular item when it is not too far away. With the benefit of hindsight we have identified certain shortcomings of the study which need to be stated. Recall that we substituted the single statement about the 'naturalness of the agent's behaviour' by two statements about the human-likeness of the way in which the agent spoke and the human-likeness of the way the agent moved. It is possible that, because we used more statements in the questionnaires to check the human-likeness of the presentations, it became more difficult to find unanimous results for this aspect. Another possibility is that the statements were still too imprecise. In other words, the way in which the agent moves can have different meanings, such as movement in terms of the gestures used while talking, or movement of the agent around the shop. Similarly, the way in which the agent spoke, could have been interpreted as referring to 'what the agent said' or referring to 'the way the agent said it'.

In addition, we know from their comments that some participants found it difficult to tell which of the three styles of selling they preferred. The data gathered to test our hypothesis on social practice (HS) show that the Dublin group definitely preferred the POINT version, where the Tokyo group, although more divided, preferred the MIXED version. This is interesting because it addresses both the physical distance from which they were asked to view the presentation as well as our use of scripted dialogue. Recall that the experiment introduction and also the instructions told the participants that they were about to watch two virtual agents auditioning for a part in a play. In our setting, participants were watching a presentation and although the camera moved along with the agents through the shop, participants did not have an active part in the play. Apart from the fact that there was a physical distance between the participants and what was happening in the presentation, participants had no personal interest in the furniture. As a result, it may have been the case that participants rated the statements in the questionnaires differently depending on whether they thought the goal was to comprehend the dialogue, or whether they tried to imagine themselves in the shoes of the customer. It would be interesting to investigate the perception of the output of the

Van der Sluis, I., Luz, S., Breitfuss, W., Ishizuka, M., and Prendinger, H. Cross-cultural assessment of automatically generated multimodal referring expressions in a virtual world. *International Journal of Human-Computer Studies 70*, 9 (2012), 611–629.

algorithm in other settings than the sales setting.

Among the factors that may explain our findings at a more basic level are the different ways in which people from Western and Asian cultural backgrounds perceive objects. Nisbett and Miyamoto (2005), for instance, report that while Westerners tend to focus on salient objects when interpreting a scene, Asians tend to focus on the whole visual context. This broadly agrees with our results in terms of understandability, where Tokyo participants strongly dispreferred the LANGUAGE presentation while Dublin participants were more divided as to the presentation style they preferred. Still, the two groups do agree, that the LANGUAGE presentation was the least preferred presentation for all criteria. The differences between ratings for the POINT presentation and the MIXED presentation, however, were small in many cases, indicating that finer-grained calibration exercises to optimise the output of the MRE algorithm are in order.

5. Conclusions

5.1. Summary

In this paper we presented our approach to evaluate the output of an existing algorithm for the generation of multimodal referring expressions for which we employed a scripted dialogue presented by two ECAs acting as a seller and a buyer in a virtual furniture shop in Second Life. The study aimed to test algorithmic output which allows for a flexible composition of MREs in terms of the precision of the pointing gesture and the number of properties to be included in the linguistic description. Depending on the knowledge representation scheme chosen, the complexity of the objects and the amount of objects in the domain the output variability could be infinite. We assessed three types of referring behaviour by the seller in a relatively complex, but life-like domain, where all MREs produced by the seller always included a pointing gesture. Two of these types were extreme in the sense that the MREs either derived their distinguishing quality from the information included in the pointing gesture (i.e. the 'POINT presentation') or from the information included in the linguistic description (i.e. the 'LANGUAGE presentation'). The third presentation was a mixture of the two extremes in which the MREs were dependent on the distance between the agents and the target objects. The study was conducted cross-culturally, involving translation and localisation of the materials to accommodate participants in Dublin and Tokyo. From the results reported in this paper we conclude that both in Dublin and Tokyo, finer-grained calibration exercises for the use of MREs by ECAs should range from 'POINT' to 'MIXED', that is, a mobile agent is preferred over a stationary agent which relies

Van der Sluis, I., Luz, S., Breitfuss, W., Ishizuka, M., and Prendinger, H. Cross-cultural assessment of automatically generated multimodal referring expressions in a virtual world. *International Journal of Human-Computer Studies 70*, 9 (2012), 611–629.

on distinguishing linguistic descriptions, even if its movements appear clumsy or imperfect. With the range from POINT to MIXED MREs, the most preferred output will most likely vary across cultures.

The use of ECAs and, more generally, anthropomorphic components in user interfaces is still a matter of debate Shneiderman and Maes (1997); Yee et al. (2007); von der Pütten et al. (2010). While most studies in this area have focused on the effectiveness of ECAs or on the effects of factors such as realism and socio-emotional assumptions, little research has been done on the aspects of production of multimodal references by ECAs addressed in this paper. Since, as with other research on ECAs, the study of MRE is fraught with methodological difficulties, we thought a discussion of methodological issues we encountered in carrying out this study might be useful to others embarking on similar research. Therefore such discussion is presented next, followed by a discussion on the implications of the findings presented here for the design of user interface ECAs.

5.2. Methodological Issues

As far as we know, no other examples exist of evaluations of multimodal referring expressions produced by ECAs that are able to move freely through a virtual space. Perception experiments on referring expressions address the issue from a natural language generation perspective and usually bear little relation to real life settings such as the one employed in this study. The main tasks in setting up a study of this sort are to design the virtual setting, to compose the dialogues, to decide on the appearances of the ECAs, and to elaborate the questionnaires to be used in the evaluation. Each of these tasks present certain issues. From their (free-text) comments we gather that participants were not generally positive about the presentations due to their "artificiality" and technical problems with the virtual environment in which they were set. Although the chosen scenario relates closely to real life (e.g., most people are familiar with a furniture shop), the issues above were compounded by the fact that the participants found it hard to engage with the presentations.

As regards cross-cultural differences, we highlighted the careful way in which the localisation of the set up to the Japanese language and culture was conducted. We are confident that the thorough assessment of the Japanese dialogue and the MREs therein (cf. Van der Sluis and Luz, 2011) allowed us to test materials that were recognisable and acceptable for each of the two groups involved in our study. However, it is well known that there are strong differences in politeness and social rules between the two cultures we considered i.e., it may have been that the participants in the Tokyo group were more polite towards the experimenter than the

Van der Sluis, I., Luz, S., Breitfuss, W., Ishizuka, M., and Prendinger, H. Cross-cultural assessment of automatically generated multimodal referring expressions in a virtual world. *International Journal of Human-Computer Studies 70*, 9 (2012), 611–629.

participants in the Dublin group. We conducted sentiment analyses on the comments written by our participants, across culture as well as gender, but we did not find any remarkable differences. Although our evaluation design did not control for personality effects, our study was specifically designed to check for gender effects, employing equal numbers of male and female participants at both Tokyo and Dublin. The effects of gender found in the Dublin group could be related to a difference in the familiarity with virtual worlds and ECAs between the male and female participants. The focus on the behaviour of the male ECA may also have contributed to these effects.

Although many evaluations of ECAs have been performed, systematic studies on specific aspects of interaction are scarce (Ruttkay and Pelachaud, 2004; Dehn and Van Mulken, 2000). In this study, we made a great effort to ensure that the materials tested were equivalent between the cultures under consideration and with respect to the conditions tested within the two groups. The translation and localisation of the materials in terms of the dialogue and the appearance of the ECAs ensured a clear focus on the referring behaviour of the furniture seller. This behaviour was assessed through questionnaires. The use of questionnaires has been criticised as not suitable to assess the notion of "presence" (roughly defined as the "sense of being there") in virtual reality environments (Slater, 2004). Whatever the merits of this criticism as applied to the notion of presence, we consider that assessing the referring behaviour of a particular ECA in a virtual environment is a considerably more manageable (and better defined) empirical question than assessing whether one gets a sense of being there by interacting in a virtual world. Therefore Slater's criticism of the use of questionnaires does not apply to the research question investigated in this paper. The participant's judgements regarding referring behaviour were elicited through the use of scripted dialogue as well as the way the study was framed to the participants (i.e. an ECA auditioning for a part in a play). A possible caveat is that even though we randomised the order in which the presentations were shown to each participant, participants were asked to judge three relatively similar presentations with the same questionnaire and may therefore have suffered from repeated exposure effects.

To assess content determination algorithms like the MRE algorithm, one could imagine other types of evaluation. For instance, one could imagine a set up in which participants are asked to judge the behaviour of real life actors instead of artificial agents. Such a scenario, may help to overcome distractors like the quality of the text to speech system or the virtual ECA's motor control. Another approach could involve asking participants to interact with the agent directly, instead of watching an interaction between two agents in real life or in a virtual world. Such

Van der Sluis, I., Luz, S., Breitfuss, W., Ishizuka, M., and Prendinger, H. Cross-cultural assessment of automatically generated multimodal referring expressions in a virtual world. *International Journal of Human-Computer Studies 70*, 9 (2012), 611–629.

a set up would allow not only for task-based evaluations with which the quality of the MREs could be assessed in terms of the efficiency and success with which a participant is able to carry out a particular assignment, but it would also allow for linguistic analyses in which the quality of MREs could be assessed in terms of for instance the length and number of clarification dialogues and the alignment of MREs. Obviously, such methods would make assessment less controlled and subject to more influencing factors such as gesturing and other nonverbal behaviour, like personality, emotional state and idiosyncratic features of individuals. Nevertheless, such factors would inform the judgement of multimodal behaviour and it would be interesting to investigate them further.

5.3. Some implications for the Design and Use of ECAs

This investigation on three possible parametrisations of a MRE generation algorithm indicated that ECAs should use language sparingly when producing descriptions of objects. They should also be allowed to move and gesture freely on the screen, if their behaviour is to be considered human-like by the user and their referring expressions correctly understood as picking out the right objects. This finding, which appears to be robust (at least across the two different cultural settings assessed in this study), is in remarkable contrast with current practice in interface agent design where agents are still predominantly represented as stationary "talking heads" Yee et al. (2007) who take rather verbose dialogue turns.

Finding the right balance between verbosity on the one hand, and mobility and pointing gesture precision on the other is unlikely to suffice, by itself, in making ECA behaviour seem acceptably natural in the eyes of the user. However, this issue should certainly be addressed in the design of user interfaces containing ECAs, along with phenomena such as "uncanny valley" effects and considerations on the advantages or disadvantages of anthropomorphism. From a practical perspective, designers wishing to use ECAs in user interfaces are faced with the question of whether to focus implementation efforts on graphical realism or on behavioural naturalness (even if the agents are to be represented by cartoon-like characters, an often employed technique Yee et al. (2007)). In the study presented in this paper, participant comments indicated that they were initially sensitive to technical problems inherent to Second Life (e.g. 'The men's shoes disappeared', 'The male agent "moonwalked" across the room in [MIXED version],'The woman also merged with the table when she walked towards it which looked very strange', 'The male had a strange blue transparent ring around his hand and fell through the floor at one point when walking'. However, such negative impressions tend to become less salient as the user gets accustomed with the system and engages in the

Van der Sluis, I., Luz, S., Breitfuss, W., Ishizuka, M., and Prendinger, H. Cross-cultural assessment of automatically generated multimodal referring expressions in a virtual world. *International Journal of Human-Computer Studies 70*, 9 (2012), 611–629.

task. This can be seen in the fact that, compared to the participants in the Dublin group, the Tokyo participants were less negative in their overall judgements of the agent's appearance and actions at the baseline level possibly because they were more familiar with virtual worlds ('I was now accustomed to the strange movement of the character'). In both groups, however, the judgements became more positive when they were presented with combinations of language, movement and gestures which indicated the object under discussion more precisely. This suggests that design effort directed at balancing correctly the elements of MRE generation should take priority over improvements in graphical realism.

The results presented above also suggest that different cultural (linguistic) settings might require these factors to be balanced differently for each setting. The localisation process involved should, therefore, extend far beyond translation (Van der Sluis and Luz, 2011). More research is needed to further explore the space of parameters for MRE generation algorithms.

5.4. Future Work

The work presented in this paper opens up various directions for future work. One direction relates to the type of linguistic descriptions people use to identify objects in real-life settings. Current work on the generation of referring expressions revolves around lab-based production experiments that often seem unrelated to real life settings. While we have access to detailed data from corpora (Di Eugenio et al., 2000; Van Deemter et al., To Appear) on how speakers refer to pieces of furniture, the way in which speakers refer to furniture items in the particular seller-buyer relationship in a furniture shop may differ. An issue that deserves further investigation is the extent to which corpora obtained through such lab-based studies can be useful to real life applications. The relevance or salience of the attributes used in a linguistic description may be dependent on a (scenario-specific) utility function, which is likely to go beyond the usual 'colour', 'size' and 'location' representations that are commonly used in algorithms that generate referring expressions. Participants commented, for instance, that '[he] should have invited her to sit in chairs (sic)'. See also Van der Sluis and Luz (2011) for a discussion.

With respect to the social practice we took as a starting point in our study, it would be necessary to investigate protocols for shopkeepers and find out more about cultural differences in their verbal and non-verbal behaviour in the sales context. In general, although many studies provide evidence that people do use pointing gestures, not much is known about when people perform pointing gestures as opposed to linguistic descriptions and what factors play a role in these

Van der Sluis, I., Luz, S., Breitfuss, W., Ishizuka, M., and Prendinger, H. Cross-cultural assessment of automatically generated multimodal referring expressions in a virtual world. *International Journal of Human-Computer Studies 70*, 9 (2012), 611–629.

decisions. Focused studies on human MRE production behaviour in particular social practices would allow us to formulate more precise hypotheses than the ones used as starting points for the work presented in this paper, and would inform the parametrisation of the MRE algorithm proposed by Van der Sluis and Krahmer in terms of the cost of pointing gestures and linguistic descriptions to be generated in a particular sales setting.

In our translation and localisation process of the scripted dialogue and especially the MREs, it became clear that the current approach to referring expression generation (of which the Van der Sluis and Krahmer algorithm is an example) has a strong bias towards the English language and parametrisation was not as straightforward as hoped for. Although a general conclusion based on our findings would be that pointing gestures should not be costlier than linguistic properties (i.e. participants preferred mobile agents over stationary agents), we cannot report much about the type of linguistic descriptions that should accompany these pointing gestures. Therefore, future work should most certainly include the design of MRE algorithms that allow adaptation to languages other than English.

From a methodological point of view, it would be of interest to investigate whether the findings presented in this paper can be reproduced when employing more interactive and/or task-based studies with ECAs as described above. Although such studies would make it more difficult to control the variables involved, they would certainly ensure a higher level of involvement of the participants and allow for evaluation metrics other than questionnaires. In particular, our finding that participants generally prefer mobile agents over stationary ones despite some technical imperfections, implies the need for future cross-cultural studies in more dynamic settings.

6. Acknowledgements

This research is partly funded SFI grant (07/CE/I1142) for the Centre for Next Generation Localisation (www.cngl.ie) at Trinity College Dublin, and the National Institute of Informatics. Thanks are due to Mr Fukunaga, Mr Suzuki and Ms. Junko Nagai at Dai Nippon Printing and Kumiko Campbell and Naoto Nishio and Tsuyoshi Okita for help with the translation and localisation of the experiment materials. We also like to thank Prof. Nobuhiro Furuyama and his group at National Institute of Informatics, Prof. Takenobu Tokunaga and his group at the Tokyo Institute of Technology, Tokyo and Prof. Yukiko Nakano and Afia Akhter at Seikei University, Tokyo for good discussions and valuable comments.

Van der Sluis, I., Luz, S., Breitfuss, W., Ishizuka, M., and Prendinger, H. Cross-cultural assessment of automatically generated multimodal referring expressions in a virtual world. *International Journal of Human-Computer Studies 70*, 9 (2012), 611–629.

References

André, E., Rist, T., 1993. The design of illustrated documents as a planning task, in: Maybury, M. (Ed.), Intelligent Multimedia Interfaces. AAAI Press, pp. 94–116.

André, E., Rist, T., Mulken, S., Van Klesen, M., 2000. The automated design of believable dialogues for animated presentation teams, in: Embodied Conversational Agents. MIT Press.

Aylett, R., Vannini, N., André, E., Paiva, A., Enz, S., Hall, L., 2009. But that was in another country: agents and intercultural empathy, in: Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems, pp. 329–336.

Beun, R., Cremers, A., 1998. Object reference in a shared domain of conversation. Pragmatics & Cognition 6, 121–152.

Bourges-Waldegg, P., Scrivener, S.A., 1998. Meaning, the central issue in cross-cultural HCI design. Interacting with Computers 9, 287–309.

Breitfuß, W., Prendinger, H., Ishizuka, M., 2008. Automatic generation of gaze and gestures for dialogues between embodied conversational agents. Int'l J of Semantic Computing 2, 71–90.

Breitfuß, W., Van der Sluis, I., Luz, S., Prendinger, H., Ishizuka, M., 2009. Evaluating an algorithm for the generation of multimodal referring expressions in a virtual world: a pilot study, in: Proceedings of the 9th International Conference on Intelligent Virtual Agents, IVA 2009, Springer. pp. 71–90.

Byron, D., 2003. Understanding referring expressions in situated language, some challenges for real world agents, in: Proceedings of the 1st International Workshop on Language Understanding and Agents for the Real World, Hokkaido University.

Cairns, P., 2007. Hci ...not as it should be: inferential statistics in hci research, in: Proceedings of HCI 2007, vol 1 BCS, pp. 195–201.

Calbris, G., 1990. The Semiotics of French Gesture. Indiana University Press, Bloomington.

Van der Sluis, I., Luz, S., Breitfuss, W., Ishizuka, M., and Prendinger, H. Cross-cultural assessment of automatically generated multimodal referring expressions in a virtual world. *International Journal of Human-Computer Studies 70*, 9 (2012), 611–629.

Cassell, J., 2009. Social practice: Becoming enculturated in human-computer interaction. HCI 7, 303–313.

Cassell, J., Geraghty, K., Gonzalez, B., Borland, J., 2009. Modeling culturally authentic style shifting with virtual peers, in: Proc. of ICMI-MLMI.

Cassell, J., Stock, T., Bickmore, T., Gao, Y., Nakano, Y., Ryokai, K., Tversky, D., Vaucelle, C., Vilhjalmsson, H., 2002. Mack: Media lab autonomous conversational kiosk, in: Proceedings of IMAGINA'02.

Claassen, W., 1992. Generating referring expressions in a multimodal environment, in: Dale, R., Hovy, E., Rösner, D., Stock, O. (Eds.), Aspects of Automated Natural Language Generation, Lecture Notes in Artificial Intelligence. Springer Verlag, Berlin. volume 587, pp. 263–276.

Clark, H., Wilkes-Gibbs, D., 1986. Referring as a collaborative process. Cognition 22, 1–39.

Dale, R., Reiter, E., 1995. Computational interpretations of the Gricean maxims in the generation of referring expressions. Cognitive Science 18, 233–263.

Van Deemter, K., 2006. Generating referring expressions that involve gradable properties. Computational Linguistics 32, 195–222.

Van Deemter, K., Gatt, A., Van der Sluis, I., Power, R., To Appear. Generation of referring expressions: Assessing the incremental algorithm. Cognitive Science .

Van Deemter, K., Krahmer, E., 2006. Graphs and Booleans: on the generation of referring expressions, in: Bunt, H., Muskens, R. (Eds.), Computing Meaning, Studies in Linguistics and Philosophy. Kluwer, Dordrecht. volume 3.

Van Deemter, K., Krenn, B., Piwek, P., Klesen, M., Schroeder, M., Baumann, S., 2008. Full generated scripted dialogue for embodied agents. AI Journal 172, 10.

Dehn, D., Van Mulken, S., 2000. The impact of animated interface agents: a review of empirical research. Int. J. Human-Computer Studies 52, 1–22.

Dong, J., Salvendy, G., 1999. Designing menus for the chinese population: Horizontal or vertical? Behaviour & Information Technology 18, 467–471.

Van der Sluis, I., Luz, S., Breitfuss, W., Ishizuka, M., and Prendinger, H. Cross-cultural assessment of automatically generated multimodal referring expressions in a virtual world. *International Journal of Human-Computer Studies 70*, 9 (2012), 611–629.

Eichner, T., Prendinger, H., André, E., Ishizuka, M., 2007. Attentive presentation agents, in: Proc 7th Int'l Conf on Int'l Virtual Agents (IVA'07), pp. 283–295.

Endrass, B., Rehm, M., André, E., 2009. Culture-specific communication management for virtual agents, in: 8th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2009), pp. 281–287.

Di Eugenio, B., Jordan, P., Thomason, R., Moore, J., 2000. The agreement process: an empirical investigation of human-human computer-mediated collaborative dialogues. Intl. Journ. Human-Comp. Studies 6, 1017–1076.

Fernald, A., Morikawa, H., 1993. Common themes and cultural variations in japanese and american mothers speech to infants. Child Dev. 64, 637–656.

Fernandes, T., 1994. Global interface design, in: Conference companion on Human factors in computing systems, ACM, New York, NY, USA. pp. 373–374.

Fitts, P., 1954. The information capacity of the human motor system in controlling amplitude of movement. Journal of Experimental Psychology 47, 381–391.

Funakoshi, K., Watanabe, S., Tokunaga, T., 2006. Group-based generation of referring expressions, in: Proc. of the INLG-06, pp. 73–80.

Gatt, A., Van der Sluis, I., Van Deemter, K., 2007. Evaluating algorithms for the generation of referring expressions using a balanced corpus, in: Proc. of the ENLG-07.

Hofstede, G., 1983. National cultures in four dimensions: A Research-Based theory of cultural differences among nations. International Studies of Management & Organization 13, 46–74.

Huang, H., Cerekovic, A., Pandzic, I., Nakano, Y., Nishida, T., 2009. Toward a multi-culture adaptive virtual tour guide agent with a modular approach. AI and Society 24, 225–235.

Huls, C., Claassen, W., Bos, E., 1995. Automatic referent resolution of deictic and anaphoric expressions. Computational Linguistics 21, 59–79.

Jan, D., Martinovski, D.H.B., Novick, D., Traum, D., 2006. A computational model of culture-specific conversational behavior, in: Pelachaud, C., Martin, J., Andre, E., Chollet, G., Karpouzis, K., Pele, D. (Eds.), IVA 2007. LNCS (LNAI), vol. 4722. Springer, Heidelberg, pp. 45–56.

Van der Sluis, I., Luz, S., Breitfuss, W., Ishizuka, M., and Prendinger, H. Cross-cultural assessment of automatically generated multimodal referring expressions in a virtual world. *International Journal of Human-Computer Studies 70*, 9 (2012), 611–629.

Jordan, P., Walker, M., 2005. Learning content selection rules for generating object descriptions in dialogue. Journal of Artificial Intelligence Research 24, 157–194.

Kendon, A., 1994. Do gestures communicate? A review. Research on Language and Social Interaction 27.

Kendon, A., Versante, L., 2003. Pointing by the hand in neapolitan, in: Kita, S. (Ed.), Pointing, where Language, Culture and Cognition Meet. Lawrence Erlbaum Associates Publishers, Manwah, New Jersey, London, pp. 109–137.

Kenny, P., Hartholt, A., Gratch, J., Swartout, W., Traum, D.R., Marsella, S., Piepol, D., 2007a. Building interactive virtual humans for training environments, in: Proceedings of I/ITSEC.

Kenny, P., Parsons, T., Gratch, J., Leuski, A., Rizzo, A., 2007b. Virtual patients for clinical therapist skills training, in: Proceedings of the 7th International Conference on Intelligent Virtual Agents, Paris, France. pp. 197–210.

Kita, S., 1990. The Temporal Relationship between Gesture and Speech: A study of Japanese-English Bilinguals. MS, Department of Psychology, University of Chicago.

Kita, S., 1993. Language and Thought Interface: A Study of Sponteneous Gestures and Japanese Mimetics. Ph.D. thesis. University of Chicago.

Kita, S., 1997. Two-dimensional semantic analysis of japanese mimetics. Linguistics 35, 379–416.

Kita, S., Özürek, A., 2003. What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. Journal of Memory and Language 48, 16–32.

Kitayama, S., Duffy, S., Kawamura, T., Larsen, J., 2003. Perceiving an object and its context in different cultures: A cultural look at the new look. Psychol. Sci. 14, 201–206.

Kopp, S., Jung, B., Lessmann, N., Wachsmuth, I., 2003. Max - a multimodal assistant in virtual reality construction. KI Kunstliche Intelligenz 5, 11–17.

Krahmer, E., van Erk, S., Verleg, A., 2003. Graph-based generation of referring expressions. Computational Linguistics 29, 53–72.

Van der Sluis, I., Luz, S., Breitfuss, W., Ishizuka, M., and Prendinger, H. Cross-cultural assessment of automatically generated multimodal referring expressions in a virtual world. *International Journal of Human-Computer Studies 70*, 9 (2012), 611–629.

Kranstedt, A., Lücking, A., Pfeiffer, T., Rieser, H., Wachsmuth, I., 2005. Deixis: How to determine demonstrated objects, in: Presented at the 5th International Gesture Workshop (GW'05), Ile de Berder, France.

Kranstedt, A., Lücking, A., Pfeiffer, T., Rieser, H., Wachsmuth, I., 2006. Deictic object reference in task-oriented dialogue, in: Rickheit, G., Wachsmuth, I. (Eds.), Situated Communication. Mouton de Gruyter.

Kroeber, A., Kluckhohn, C., 1952. Culture: A Critical Review of Concepts and Definitions. Random House.

Lester, J., Voerman, J., Towns, S., Callaway, C., 1999. Deictic believability: Coordinating gesture, locomotion and speech in lifelike pedagogical agents. Applied Artificial Intelligence 13, 383–414.

Lücking, A., Rieser, H., Stegmann, J., 2004. Statistical support for the study of structures in multimodal dialogue: Inter-rater agreement and synchronization, in: Proceedings of the 8th Workshop on the Semantics and Pragmatics of Dialogue, (Catalog'04), Barcelona, Spain. pp. 93–100.

Masuda, T., Nisbett, R., 2001. Attending holistically vs. analytically: Comparing the context sensitivity of japanese and americans. J. Pers. Soc. Psychol. 81, 922–934.

Nakano, Y., Rehm, M., 2009. Multimodal corpus analysis as a method for ensuring cultural usability of embodied conversational agents, in: Kurosu, M. (Ed.), Human Centered Design, Springer. pp. 521–530.

Nakasone, A., Tang, S., Shigematsu, M., Heinecke, B., Fujimoto, S., Prendinger, H., 2011. Openbiosafetylab: A virtual world based biosafety training application for medical students, in: Proc 8th Int'l Conf on Information Technology: New Generations (ITNG'11), IEEE CPS.

Mc Neill, D., 2000. Language and Gesture. Cambridge University Press.

Nisbett, R., Miyamoto, Y., 2005. The influence of culture: holistic versus analytic perception. TRENDS in Cognitive Sciences 9, 383–414.

Noiwan, J., Norcio, A.F., 2006. Cultural differences on attention and perceived usability: Investigating color combinations of animated graphics. International Journal of Human-Computer Studies 64, 103–122.

Van der Sluis, I., Luz, S., Breitfuss, W., Ishizuka, M., and Prendinger, H. Cross-cultural assessment of automatically generated multimodal referring expressions in a virtual world. *International Journal of Human-Computer Studies 70*, 9 (2012), 611–629.

Özürek, A., Kita, S., Allen, S., Furman, R., Brown, A., 2005. How does linguistic framing of events influence co-speech gestures? insights from crosslinguistic variations and similarities. Journal of Memory and Language 5, 219–240.

Payr, S., Trappl, R., 2004. Agent Culture: Human- Agent Interaction in a Multicultural World. London: Lawrence Erlbaum Associates.

Pechmann, T., 1989. Incremental speech production and referential overspecification. Linguistics 27, 98–110.

Piwek, P., 2008. Presenting arguments as fictive dialogue, in: Proc. of the CMNA-08.

Prendinger, H., Eichner, T., André, E., Ishizuka, M., 2007. Gaze-based infotainment agents, in: Proc ACM SIGCHI Int'l Conf on Advances in Computer Entertainment Technology (ACE'07).

Prendinger, H., Ishizuka, M. (Eds.), 2004. Life-Like Characters. Tools, Affective Functions, and Applications. Cognitive Technologies, Springer Verlag, Berlin Heidelberg.

Prendinger, H., Ullrich, S., Nakasone, A., Ishizuka, M., 2010. Mpml3d: Scripting agents for the 3D Internet. IEEE Transactions on Visualization and Computer Graphics .

von der Pütten, A.M., Krämer, N.C., Gratch, J., Kang, S.H., 2010. 'it doesn't matter what you are!' Explaining social effects of agents and avatars. Computers in Human Behavior 26, 1641–1650.

Rehm, M., Andre, E., Nakano, Y., 2009a. Some pitfalls for developing enculturated conversational agents. LNCS: Human-Computer Interaction. Ambient, Ubiquitous and Intelligent Interaction 5612, 340–348.

Rehm, M., Nakano, Y., Andre, E., Nishida, T., Bee, N., Endrass, B., Wissner, M., Lipi, A., Huang, H., 2009b. From observation to simulation: generating culture-specific behavior for interactive systems. AI and Society 24, 267–280.

Reithinger, N., 1992. The performance of an incremental generation component for multi-modal dialog contributions, in: Dale, R., Hovy, E., Rösner, D., Stock, O. (Eds.), Aspects of Automated Natural Language Generation, Lecture Notes in Artificial Intelligence. Springer Verlag, Berlin. volume 587, pp. 263–276.

Van der Sluis, I., Luz, S., Breitfuss, W., Ishizuka, M., and Prendinger, H. Cross-cultural assessment of automatically generated multimodal referring expressions in a virtual world. *International Journal of Human-Computer Studies 70*, 9 (2012), 611–629.

Rist, T., André, E., Baldes, S., Gebhard, P., Klesen, M., Kipp, M., Schmitt, M., 2003. A review on the development of embodied presentation agents and their application fields, in: Prendinger, H., Ishizuka, M. (Eds.), Life-like Characters. Tools, Affective Functions and Applications. Springer, Berlin, pp. 377–404.

De Rosis, F., Pelachaud, C., Poggi, I., 2004. Transcultural believability in embodied agents: a matter of consistent adaptation, in: Payr, S., Trappl, R. (Eds.), Agent Culture: Designing HumanAgent Interaction in a Multicultural World. Laurence Erlbaum Associates, Mahwah.

De Ruiter, J., 2000. The production of gesture and speech, in: Mac Neill, D. (Ed.), Language and Gesture. Cambridge University Press.

De Ruiter, J., Bangerter, A., Dings, P., 2012. The interplay between gesture and speech in the production of referring expressions: Investigating the tradeoff hypothesis. Topics in Cognitive Science 4, 232–248.

Ruttkay, Z., Pelachaud, C., 2004. From Brows to Trust: Evaluating Embodied Conversational Agents. Kluwer.

Shneiderman, B., Maes, P., 1997. Direct manipulation vs interface agents. interactions 4, 42–61.

Slater, M., 2004. Slater, m.: How colorful was your day? why questionnaires cannot assess presence in virtual environments. Presence-Teleoperators and Virtual Environments 13, 484 –493.

Van der Sluis, I., Krahmer, E., 2007. Generating multimodal referring expressions. Discourse Processes 44, 145–174.

Van der Sluis, I., Luz, S., 2011. Issues in translating and producing Japanese referring expressions for dialogues. Linguistic Issues in Language Technology 5, 1–46.

Van der Sluis, I., Nagai, J., Luz, S., 2009. Producing referring expressions in dialogue: Insights from a translation exercise, in: Proc. of preCogsci 2009: Production of Referring Expressions: Bridging the gap between computational and empirical approaches to reference. At CogSci'09, Amsterdam, The Netherlands.

Van der Sluis, I., Luz, S., Breitfuss, W., Ishizuka, M., and Prendinger, H. Cross-cultural assessment of automatically generated multimodal referring expressions in a virtual world. *International Journal of Human-Computer Studies 70*, 9 (2012), 611–629.

Traum, D., 2009. Models of culture for virtual human conversation. LNCS: Universal Access in Human-Computer Interaction. Applications and Services 5616, 434–440.

Traum, D., Swartout, W., Marsella, S., Gratch, J., 2005. Virtual humans for non-team interaction training, in: Proceedings of the AAMAS Workshop on Creating Bonds with Embodied Conversational Agents.

Wilkins, D., 2003. Why pointing with the index finger is not a universal (in socio-cultural terms), in: Kita, S. (Ed.), Pointing, where Language, Culture and Cognition Meet. Lawrence Erlbaum Associates Publishers, Manwah, New Jersey, London, pp. 109–137.

Williams, S., Piwek, P., Power, R., 2007. Generating monologue and dialogue to present personalised medical information to patients, in: Proceedings of ENLG-07, pp. 167–]170.

Yee, N., Bailenson, J.N., Rickertsen, K., 2007. A meta-analysis of the impact of the inclusion and realism of human-like faces on user experiences in interfaces, in: Proceedings of the SIGCHI conference on Human factors in computing systems, ACM, New York, NY, USA. pp. 1–10.

Van der Sluis, I., Luz, S., Breitfuss, W., Ishizuka, M., and Prendinger, H. Cross-cultural assessment of automatically generated multimodal referring expressions in a virtual world. *International Journal of Human-Computer Studies 70*, 9 (2012), 611–629.

Appendix: Participants' Comments and Results by Gender

Table 6: Comments from participants in the Dublin and the Tokyo groups about virtual worlds in general for Questionnaire A.

	Familiarity and likability of virtual worlds in general
Dublin	'I realise that it is an upcoming field of technology. I am just not very familiar with it as of yet.', 'Any simulated environment is simply more engaging fiction.', 'Fun until you realise it's taking over your actual life!'
Tokyo	'Interesting', 'I enjoy the real world; I do not feel a need for virtual reality. (So far)'

Table 7: Comments from participants in the Dublin and Tokyo groups about the introductory presentation for Questionnaire A.

	The setting
Dublin	'well-set', 'realistic' to 'badly designed', 'fairly messy', 'somewhat unusual'
Tokyo	'Very pretty', 'Was a beautiful space' to 'Materials and furniture felt unnatural.', 'the furniture does not feel close to the characters.', 'I do not know enough; you can not determine the status of good or bad.'
	Understandability
Dublin	'Clear concise.', 'I thought it was well put together.', 'While the characters spoke each word clearly the intonation/phrasing made the conversation difficult to understand.'
Tokyo	'Flow was smooth.', 'I found it hard to grasp the situation.'
	Human-likeness
Dublin	'The voices seemed a little bit warbled (like a vocoder effect)', 'They were definitely robotic in their movements but not overly so; about what I would expect from computer characters', 'Seem pleasant', 'The characters appeared quite human-like', 'The movements were jerky', 'robotic'
Tokyo	'The characters do not speak like native Japanese.', 'The male character is behaving proud which makes him impressive', 'The women looks very professional', 'Movement was not very natural', 'Practiced unnatural. Overall lack of spontaneity', 'No facial expressions; they would make them look more like real human'

Van der Sluis, I., Luz, S., Breitfuss, W., Ishizuka, M., and Prendinger, H. Cross-cultural assessment of automatically generated multimodal referring expressions in a virtual world. *International Journal of Human-Computer Studies 70*, 9 (2012), 611–629.

Table 8: Comments from participants in the Dublin and the Tokyo groups about the human-likeness of the furniture seller agent in the three presentations for Questionnaire B.

	POINT *presentation*
Dublin	'Speech was stilted; pauses in speech were too long.', 'The walking was not very realistic; strides were too big.', 'The movement distracts me and makes it harder to follow.', 'The movement put me off.', 'I think it would be good if they still communicated as they were walking. More lifelike.', 'He moved to the products he was describing instead of pointing at them which was good.'.
Tokyo	'Moving combination is not smooth sometimes; but the motion of person is natural.', 'A bit unnatural but no problem.', 'Many pauses.', 'The pointing of the male character to the items was good and logical'
	LANGUAGE *presentation*
Dublin	'Didn't move to items; direction of pointing was hard to gauge.', 'It seemed better this time. Clearer because he wasn't moving around all the time so it was easier to focus on what he was selling.'
Tokyo	'The conversation between the two people feels unnatural. The intonation of words here and there were strange.', 'I am a bit skeptical about the quality of the communication.'
	MIXED *presentation*
Dublin	'Less movement/walking; pointing out objects in background more realistic; bringing items into conversation in a more realistic manner.', 'This was the most realistic so far in terms of moving around the shop; the seller showed his wares; but also didn't waste time by moving too much.', 'When the agents were walking across the room they seemed to spin around and glide backwards which wasn't very human-like.'
Tokyo	'The male shows some good gestures which enhance their communication.', 'Very good; but sometimes strange when you hear the monotonous intonation.' and 'There was some strange walk backwards to move forward.'

Van der Sluis, I., Luz, S., Breitfuss, W., Ishizuka, M., and Prendinger, H. Cross-cultural assessment of automatically generated multimodal referring expressions in a virtual world. *International Journal of Human-Computer Studies 70*, 9 (2012), 611–629.

Table 9: Comments from participants in the Dublin and the Tokyo groups about the understandability of the furniture seller agent in the three presentations for Questionnaire B.

	POINT *presentation*
Dublin	'Because the persons moved to the items under discussion it was easy to understand what they were talking about.', 'In previous presentations; where it had been unclear as to which blue desk the subject of the conversation; this was no longer a problem; as the male seller physically moved to the object and faced it in close proximity; leaving no doubt in my mind.'
Tokyo	'This version was easy to understand.'
	LANGUAGE *presentation*
Dublin	'The pointing from afar was not as clear as when they walked up to the items.', 'It seemed better this time. Clearer because he wasn't moving around all the time so it was easier to focus on what he was selling.'
Tokyo	'This video is very difficult to understand.', 'I want to move around the room', 'Just standing in place may make it difficult to understand what they are talking about.'
	MIXED *presentation*
Dublin	'The male agent pointed at the first blue desk. This was confusing as the small blue desk stood out more than the big one.'
Tokyo	'The conversation is always logical and clear'

Table 10: Comments from participants in the Dublin and the Tokyo groups about the social practice of the furniture seller agent in the three presentations for Questionnaire B.

	POINT *presentation*
Dublin	'I would prefer to deal with dialogue 2 (i.e. POINT presentation) as he brought the woman to the item and showed it to her up close so she could evaluate it; he should have offered to let her try it out though'
Tokyo	'By approaching each product description; the clerk comes across as more helpful and caring.'
	LANGUAGE *presentation*
Dublin	'He didn't show the customer the products well; pointing from far away is less engaging than walking around.', 'I understood what furniture the male was talking about but as the salesman; he should show them off instead of pointing them out.'
Tokyo	'Because this conversation took place at one spot' I felt as I was put on a little fun.'
	MIXED *presentation*
Dublin	'Moving around is important but so is not taking too much time to go around.'
Tokyo	'Exchange seemed somehow business-like.'

Van der Sluis, I., Luz, S., Breitfuss, W., Ishizuka, M., and Prendinger, H. Cross-cultural assessment of automatically generated multimodal referring expressions in a virtual world. *International Journal of Human-Computer Studies 70*, 9 (2012), 611–629.

Table 11: Means with standard deviation within brackets for the statements in Questionnaire B about the human-likeness of the manner in which the male agent spoke (*S1*) and moved (*S2*) and understandability of the MREs (*S4*) in the POINT, LANGUAGE and MIXED presentations for the Tokyo group by gender.

	POINT		LANGUAGE		MIXED	
	Females	*Males*	*Females*	*Males*	*Females*	Males
S1	3.00 (1.65)	5.73(1.62)	3.33(1.34)	3.37(1.22)	3.87(1.55)	4.13(1.18)
S2	3.49 (1.24)	4.47(1.13)	3.37(1.22)	4.40(1.12)	3.80(1.42)	3.73(1.33)
S4	1.73 (.96)	2.87(1.25)	2.80(1.32)	3.53(1.64)	2.00(1.31)	2.40(1.12)

Table 12: Means with standard deviation within brackets for the statements in Questionnaire B about the human-likeness of the manner in which the male agent spoke (*S1*) and moved (*S2*) and understandability of the MREs (*S4*) in the POINT, LANGUAGE and MIXED presentations for the Dublin group by gender.

	POINT		LANGUAGE		MIXED	
	Females	*Males*	*Females*	*Males*	*Females*	Males
S1	5.33 (1.11)	5.27(1.49)	4.67(1.68)	4.47(1.60)	4.67(1.68)	4.13(1.64)
S2	5.33 (1.72)	5.80(1.01)	4.87(1.41)	4.33(1.72)	5.80(1.37)	5.53(1.64)
S4	1.20 (.41)	1.67(.62)	2.27(1.75)	2.73(1.79)	1.80(.68)	2.60(1.40)

Van der Sluis, I., Luz, S., Breitfuss, W., Ishizuka, M., and Prendinger, H. Cross-cultural assessment of automatically generated multimodal referring expressions in a virtual world. *International Journal of Human-Computer Studies 70*, 9 (2012), 611–629.