Perception of Visual Scene and Intonation Patterns of Robot Utterances*

Ivana Kruijff-Korbayová
German Research Center for Artificial Intelligence
DFKI GmbH
Saarbrücken, Germany
ivana.kruijff@dfki.de

Raveesh Meena
German Research Center for Artificial Intelligence
DFKI GmbH
Saarbrücken, Germany
raveesh.meena@dfki.de

Pirita Pyykkönen
Department of Computational Linguistics
Saarland University
Saarbrücken, Germany
pirita@coli.uni-sb.de

ABSTRACT

It is established that assigning intonation to dialogue system output in a way that reflects relationships between entities in the discourse context can enhance the acceptability of system utterances. Previous research has concentrated on the role of linguistic context in processing; dialogue *situatedness* and hence the role of visual context in determining accent placement has not been studied. In this paper, we present an experimental study addressing the influence of visual context on the perception of nuclear accent placement in synthesized clarification requests. We found that utterances with nuclear accent placement licensed by the visual scene are perceived as appropriate more often then utterances with nuclear accent placement not licensed by the visual scene.

Categories and Subject Descriptors

[**HRI Communication**]

General Terms

Experimentation, Situatedness, Intonation

Keywords

conveying intentions, dialogue, language processing, situated awareness, enabling technologies, experimental methods

1. INTRODUCTION

Since the pioneering work of Pierrehumbert and Hirschberg [7] it is generally accepted that speakers choose particular *intonation tunes* to convey relationships between their utterance, the currently perceived *beliefs* of the hearer(s), and anticipated contributions of subsequent utterances. These relationships are conveyed compositionally via the selection of *pitch accents*, *phrase accents*, and *boundary tones* that make up tunes. Consequently, when generating natural spoken system output in dialogue systems or situated human-robot interaction the contextual appropriateness of its *intonation* needs to be modeled (cf. [2, 8, 5, 12]).

It is established that pitch accents make the individual words with which they are associated *salient*. The accented item is rendered salient not only phonologically but also from an informational standpoint: the assignment of nuclear accent marks contrast between the intended referent and contextually relevant alternative(s), e.g., [7, 10].

Although it is generally assumed that both linguistic and situational (visual) context influence the surface realization of utterances, discussions of contrast and placement of nuclear accent in the literature usually concern only discourse context established linguistically (i.e., by preceding utterances). For example, consider the following two possible realizations of the utterance *"That is a red box"* with different placement of nuclear accent:[1]

(1) That is a RED box
H* LL%

(2) That is a red BOX
H* LL%

(1) but not (2) is appropriate in a context where one or more boxes have been mentioned, and it is the red color which distinguishes the intended referent from the rest. On the other hand, (2) is appropriate when no boxes have been mentioned yet.

With respect to visual context, the presence of multiple objects in the visual scene and hence the availability of competing visual properties should similarly affect the use of contrast and placement of nuclear accent in situated dialogue. In this paper we present the results of an experiment designed to investigate this hypothesis. We concentrate on clarification requests of the form illustrated in (3) and (4).

(3) R: Is that a RED box?
L* HH%

*Supported by EU FP7 Project 'CogX'(FP7-ICT-215181).

[1]The words printed in SMALL CAPITALS indicate the alignment of the nuclear accent in the intonation contour. The description of the intonation contour using ToBI shown beneath the utterances follows [6].

(4) R: Is that a red BOX?
L* HH%

In Section 2.1, we explain the goal of the experiment in detail. In Section 2.2 we describe our methodology, in Section 2.3 we present the results. In Section 3 we discuss our findings and conclude.

2. THE EXPERIMENT

2.1 Goal

In this experiment, we investigated whether visual scenes influence the perception of nuclear accents in utterance comprehension.

An accent in an utterance is found to mark contrast with respect to competing alternatives available due to their prior mention, or pragmatic accommodability [7, 10]. However, less is known whether a listener's perception of the intonation tune of an utterance is also influenced by the content of the current visual scene in situated dialogue. It might well be that appropriate and inappropriate nuclear accent placements are governed by the presence of competing alternatives in the visual context. Psycholinguistic studies suggest that while language directs people's visual attention to mentioned objects on the scene, also the information in the scenes can constrain and alter linguistic comprehension processes [1, 3, 11].

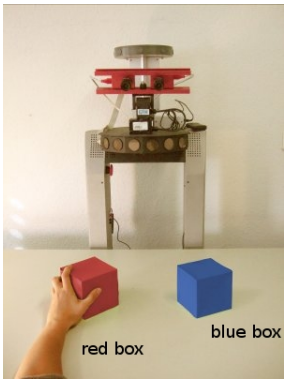
Therefore, the observations on contrast and placement of nuclear accent might also apply to visual context. It could thus be hypothesized that the presence of multiple objects in the visual scene, and the availability of competing visual properties govern the use of contrast and placement of nuclear accent in robot utterances.

Along the lines sketched above, (3) but not (4) would thus be appropriate in the visual context of Fig. 1(a),[2] where the presence of a red and a blue box licences the use of contrast on the *color* property for distinguishing the intended box from the other one.[3]

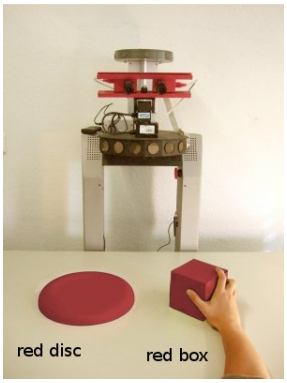
The nuclear accent placement in (4) is licensed in the visual context of Fig. 1(b) where both objects have red color and the object type is a distinguishing property. The accent placement in (3) is not licensed in the visual context of 1(b).

Using visual scenes as in Fig. 1 and clarification request utterances like (3) and (4), we have setup an experiment to test the following hypothesis:

If human language comprehension is sensitive to the relationship between visual context and nuclear accent placement then variations in the placement of nuclear accent in an utterance can be perceived. A preference of one pattern of accent placement over another provides evidence in support of the role of visual context in determining the appropriate intonation of an utterance.



(a) A visual context that is *congruent* for the accent placement in (3), but *non-congruent* for the accent placement in (4).



(b) A visual context that is *congruent* for the accent placement in (4), but *non-congruent* for the accent placement in (3).

Figure 1: Various visual contexts in a situated human-robot dialogue.

2.2 Methodology

2.2.1 Participants

Thirty-one students and researchers participated in the experiment.[4]

All participants were offered a sum of 5 Euros or an Amazon Gift Card worth 5 Euros for their successful completion of the experiment. Additionally, three participants were drawn for a prize gift voucher worth 20 Euros each.

2.2.2 Material and Design

The stimuli consisted of a visual scene and a corresponding audio. The visual scene was a picture of a PeopleBot standing at a table with one object already present and another one being introduced by a human (therefore held by a hand, as in Fig. 1). The audio consisted of the robot's clarification request about this visual scene followed by a human response 'Yes' or 'No', depending upon the correctness of the robot's request.

[2]The text labels on the objects were not present in the original pictures. We added them for presentation purposes in this paper, because the colors are not sufficiently distinguishable in black-and-white print.

[3]The placement of nuclear accent in (3) is also appropriate when the robot is uncertain about the color of the box it intends to refer to. The accent placement in (4) is also appropriate when the robot is uncertain about the type of the object being introduced. We do not address uncertainty as a factor in accent placement in this paper.

[4]Twenty-one accessed an online version of the experiment and the remaining ten undertook the experiment in our lab. Twelve were confirmed native English speakers; the rest claimed to speak US-English. Various psycholinguistic findings [4] reveal that L2 speakers of English are equally sensitive to intonational variations. However, their interpretation of tunes may vary with the individual's experience with the L2 language. Because the language background was potentially critical, we compared the results of natives and non-natives in the same analyses as presented in Section 2.3 below. Since natives and non-natives did not differ in their responses, the data was collapsed over these two groups in the analyses reported in this paper.

The audio files were synthesized using the MARY text-to-speech synthesizer (TTS) [9].[5] The input to the TTS was provided in MaryXML format indicating the type and location of nuclear accent and intonational boundary type. An MBROLA[6] US-English male speaker voice was used for synthesizing the robot's clarification requests. The human responses of 'Yes' and 'No' were also synthesized using MARY TTS, albeit with a US-English female speaker unit selection based voice.

For the visual stimuli, two (not necessarily different) objects were paired in a picture (of 300x400 pixels). The pairing of objects was done so that each object occurs as an object that is already present on the table, and as an object that is being introduced (held by a hand). We used sixteen object-type pairs and twelve color pairs. The twelve color pairs for each of the sixteen object pairs result in a total of 12x16=192 unique pictures for the visual scenes. The object being introduced was randomly held in the left-hand or right-hand to avoid visual saturation, e.g., Fig. 1(a). and Fig. 1(b).

We used a 2x2x2 design with three factors of two levels each, i.e., intonation (accent placement on color vs. type of the object), visual context (congruent vs. non-congruent) and human response ('Yes' vs. 'No').

Intonation. Two types of nuclear accent placement were chosen – either on the *color* or on the *type* property of the object. We labeled the intonation contour resulting from nuclear accent placement on the color property of the object as *tune A* (as in (3)) and the one resulting from nuclear accent placement on the type property of the object as *tune B* (as in (4)).

Visual Context. Based on the presence or absence of competing object properties (color or type) in a scene the nuclear accent placement in an utterance was labeled either *congruent* (C), i.e., licensed by the visual scene, or *non-congruent* (NC), i.e., not licensed by the visual context. An accent on color was labeled congruent (from the robot's viewpoint) if and only if the other object in the visual scene had a different color; otherwise, accent on type was labeled congruent.

For example, the combination of accent placement in (3) and the visual scene in Fig. 1(a) correspond to a congruent experimental condition. On the other hand, the combination of accent placement in (4) and the visual scene in Fig. 1(a) correspond to a non-congruent condition.

Response. The human's response 'Yes' or 'No' indicated to the robot whether its perception about the target object in the scene as expressed in the clarification request is correct or not.

The addition of this condition should decrease bias in a subject's judgement due to (in)correctness of the robot's clarifications. Introducing correct and incorrect hypotheses enabled us to check whether our setup worked to make the subject concentrate on the realization of an utterance and not on its correctness.

For the convenience of referring to various combinations of the three experimental conditions in this paper we represent them as A-C-YES, A-C-NO, B-C-YES, B-C-NO, A-NC-YES, A-NC-NO, B-NC-YES and B-NC-NO. In the appendix we provide examples of the visual stimuli corresponding to each of the experimental conditions combination.

Clarification requests of the form "Is that a `color type`" were chosen for the robot's utterances, e.g., (3) and (4). The color and type values were selected so that they were monosyllabic words, to maintain uniformity and avoid other sources of prosodic variation in the clarification request than the accent placement. We used the following eight object types: *ball*, *box*, *disc*, *heart*, *ring*, *sphere*, *star* and *wedge*. Each type appeared in six colors: *black*, *blue*, *brown*, *green*, *pink* and *red*. Using these eight object types and the six colors, we designed 8x6=48 stimuli in the aforementioned form.

These items were distributed across 8 lists so that each subject encountered each item in only one condition; all subjects received equally many items in each condition. An additional 48 clarification requests were added to each list as fillers. Two additional nuclear accent placements were used in the fillers to overcome auditory saturation due to tunes A and B in the experimental stimuli. These filler tunes exhibit accent placement on either the referential expression "that" or the verbal head "is". We label them as *tune C* and *tune D*, respectively. Table 1 summarizes the experimental and filler tunes and their corresponding intonation contours.

Table 1: Experimental (Exp) and filler (Fill) intonation tunes.

Tune	Type	Nuclear accent placement
A	Exp	Is that a RED box? L* HH%
B	Exp	Is that a red BOX? L* HH%
C	Fill	Is THAT a red box? L* HH%
D	Fill	Is that a red box? L* HH%

2.2.3 Predictions

We predicted that if comprehension is sensitive to the relationship of visual context and the nuclear accent placement then there should be a difference in the judgement of the appropriateness of utterances, namely, the utterances corresponding to the *congruent* condition would be judged more appropriate than the utterances in a *non-congruent* condition.

We expected that this outcome of the subjective judgements would hold irrespective of the type of intonational tune of the utterance. That is, if comprehension is only sensitive to the relationship of visual context and the nuclear accent placement then both the accent placement in tune A and tune B would be perceived more appropriate in congruent

[5] `mary.dfki.de`
[6] `http://tcts.fpms.ac.be/synthesis/`

conditions than in non-congruent ones.

2.2.4 Procedure

The experiment was implemented using the WebExp[7] system for conducting psychological experiments over the World Wide Web.[8]

On arrival at the Web-Experiment page the participants first read instructions: They were informed that they will see scenes with a robot with one object already on the table that the robot knows about, and then another object being presented by a human. The robot asks a question to verify whether it recognized correctly the *type* and the *color* of the object being shown. Since its recognition capacity is imperfect, it may make a mistake. The human responds to the robot with a 'Yes' or a 'No'. Their task is to evaluate whether the robot asked the question in a way appropriate to the current scene, irrespective of whether it recognized the object (its type and color) correctly or not. In addition, they were instructed to answer whether simple math calculations shown betwen the robot trials were correct.

After the instructions, participant information, i.e., age, gender, mother tongue, English they speak (US, UK, etc.), educational background, and their past experience with spoken language interfaces was collected.

Subjects were then automatically assigned one of the 8 stimuli lists. First, they went through a practice session consisting of 6 stimuli to get familiar with the presentation style of the stimuli and their tasks. Next, they entered the main session. In the practice and the main session, the presentation and the evaluation of the stimuli proceeded in three steps.

In the first step, the visual scene was shown to the subject, and after a 1500ms picture preview the corresponding audio stimuli of the robot's clarification request followed by the audio of the human user's response was played. The picture preview allowed the subject to inspect the scene before the audio was played. In the absence of a visual preview, linking the attention captured by the visual scene with the audio stimulus from the clarification would have been a challenging task for the subject: The sentence would be over before the participants would have started to pay attention to the spoken stimuli. Once the audio stopped playing, the visual scene disappeared after a delay of 1s. This delay was added to give the subject some time for linking the dialogue with the visual scene.

In the second step, the subject was asked for their judgement of the robot's utterance: *"Your evaluation of how appropriately the question was asked."* The subject indicated their judgement by selecting a response on a 5-point scale between good and bad.

In the third step, the subject judged the correctness of a simple math calculation task. An audio with the ticking of a clock was also played until the subject responded. The purpose of the calculation task and the clock audio was to interrupt the subject's visual and audio stimulation, due to the preceding stimuli presentation, before proceeding to the next one. Once the subject responded to the calculation task, the next stimulus was presented as just described.

The experiment took around 20-25 minutes to complete.

[7]http://www.webexp.info/

[8]Those participants who took part in the experiment in our lab were simply directed to the Web-Experiment main page and were asked to follow the instructions there.

2.3 Results

We report analysis results based on the evaluation of the robot's clarification requests from all participants.

The congruent stimuli were expected to be more acceptable then the non-congruent ones. The score count in Fig. 2 and 3 suggest, that the distribution of subjective judgement for congruent and non-congruent stimuli is very similar at both higher and lower ends of the scores. This coarse-grained analysis suggests that congruent stimuli were not considered more acceptable than the non-congruent ones.

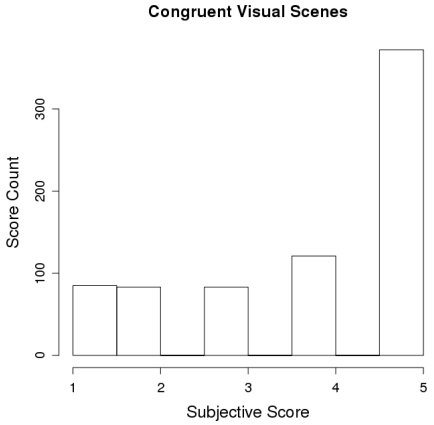


Figure 2: Distribution of subjective scores for congruent stimuli.

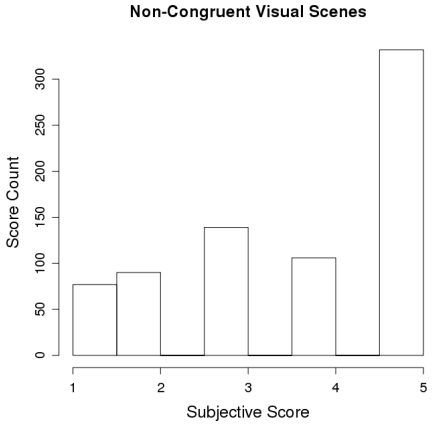


Figure 3: distribution of subjective scores for non-congruent stimuli.

In order to investigate the evaluations in more detail, scores 4 and 5 were collapsed under the label 'GOOD' and scores 1 and 2 under the label 'BAD', to overcome data sparseness.[9]

[9]The scores of 3 were labeled as 'NUTRL'

Fig. 4 presents GOOD and BAD judgments over all the congruent (C) and non-congruent (NC) stimuli. We observe that utterances in a congruent visual context were more often judged GOOD (66.26%) than BAD (22.53%). However, the distribution of judgement for the non-congruent visual context is fairly similar to that for the congruent context. About 58.87% of the stimuli in the non-congruent visual context were judged GOOD. That is, although the pitch accent placement was not licensed by the visual context of the scenes, the utterances were often judged GOOD.

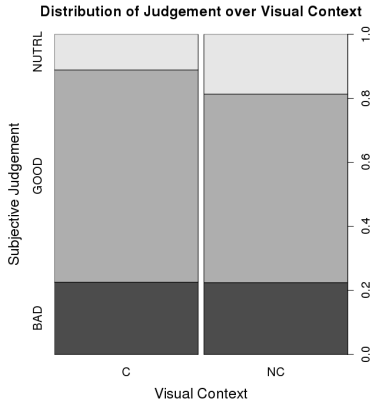


Figure 4: Distribution of subjective judgements over visual context.

Table 2 provides the distribution of the subjective judgements over tune A and tune B. It was expected that both tunes A and B would be more often accepted in congruent visual context than in non-congruent context.

Table 2: Distribution of subjective judgements over tunes.

Tune	GOOD	BAD	NUTRL
tune A	464	159	121
%	62.36%	21.37%	16.26%
tune B	467	176	101
%	62.76%	23.65%	13.57%

The distribution suggests that both tunes A and B were judged equally often GOOD, and nearly equally often BAD. Table 3 in turn shows that both tune A and B were more often judged GOOD in a congruent condition (C) than in a non-congruent condition (NC). Tune A has also been judged more often BAD in non-congruent condition than in congruent condition. However, tune B has been judged more often BAD in the congruent condition than the non-congruent condition.

Table 4 shows the distribution of the subjective judgments over all three conditions. The distribution shows that the judgments are always more often GOOD for the congruent condition than for the non-congruent condition.

The distribution of the subjective judgements over the robot's correct and incorrect hypothesis was also analyzed. This reveals that for a correct hypothesis, i.e., when the human response is 'Yes', both tune A and B are judged more often GOOD in the congruent condition than in the non-congruent condition and they are judged more often BAD in the non-congruent condition than in the congruent condition. However, for an incorrect hypothesis, i.e., when the human response is 'No', both tune A and B are more often judged BAD in the congruent condition than in the non-congruent condition.

Table 3: Distribution of GOOD and BAD over tunes–visual context.

Tune	GOOD	BAD	NUTRL
A-C	241	77	54
%	64.78%	20.69%	14.51%
A-NC	223	82	67
%	59.94%	22.04%	18.01%
B-C	252	91	29
%	67.74%	24.47%	7.79%
B-NC	215	85	72
%	57.79%	22.84%	19.35%

Table 4: Distribution of GOOD and BAD over tunes–visual-context–response.

Tune	GOOD	BAD	NUTRL
A-C-YES	156	13	17
%	83.87%	6.98%	9.13%
A-NC-YES	140	23	23
%	75.26%	12.36%	12.36%
A-C-NO	85	64	37
%	45.69%	34.40%	19.89%
A-NC-NO	83	59	44
%	44.62%	31.72%	23.65%
B-C-YES	167	11	8
%	89.78%	5.91%	4.3%
B-NC-YES	139	27	20
%	74.73%	14.51%	10.75%
B-C-NO	85	80	21
%	45.69%	43.01%	11.29%
B-NC-NO	76	58	52
%	40.86%	31.18%	27.65%

Fig. 5 presents the distribution of the subjective judgement over the human responses ('Yes' and 'No'), respectively. It can be inferred from the plot in Fig. 5 that the robot's clarification utterances with human response 'Yes' were more often judged GOOD than those with the response 'No'. This indicates that the subjects were more likely to judge the correctness of the robot's hypothesis than the appropriateness of the request in the context of the visual scene.

3. DISCUSSION AND CONCLUSIONS

Existing attempts to model the intonation of dialogue system output in practical systems include [2, 8, 5, 12]. These systems illustrate various approaches to model the role of linguistic context in realizing intonation.

For example, in [5] intonation assignment in system turns that are direct answers to questions is done based on *information structure partitioning* according to the preceding linguistic context, both in terms of what question is being answered and what alternatives are salient. Accent placement is determined using *semantic parallelism*: two basic

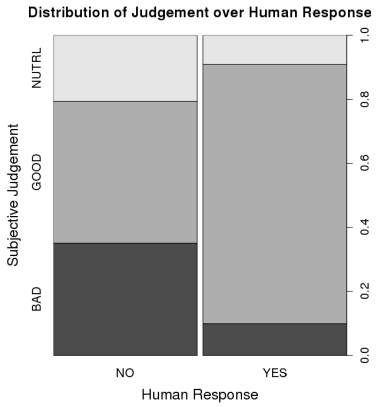


Figure 5: Subjective judgements vs. Human response.

terms are parallel when they are either identical or alternative (i.e., belonging to same *sort* but non-identical). A perception experiment comparing system generated responses with controlled intonation against defaults indicated that contextual appropriateness of system output improves when intonation is assigned based on information structure.

A method of synthesizing contextually appropriate intonation with limited domain unit selection voices is presented in [12]. In a pilot study, they built an APML-aware limited domain voice for use in flight information dialogues, which involve comparing and contrasting the most compelling attributes of the most relevant flights available, rather than simply listing the query results [13]. In a perception experiment comparing the APML voice to a default version built using the same recordings without the additional structure, the intonation produced by the APML voice was judged significantly more contextually appropriate than that of the default voice.

Situated human-robot dialogue differs from the type of dialogue in these applications in that the dialogue context is not the only source of contextual information: the *visual context* is also part of the discourse context, and should be used for determining the placement of nuclear accent in system utterances. Moreover, whereas the abovementioned systems address intonation assignment in statements answering user's questions, we concentrate on clarification requests pertaining to changes in the visual context. Such clarification requests may not be related to prior mentions in the dialogue; they may concern objects or properties that exist in the visual scene but have not been spoken about.

The analysis of our experiment data reveals that the acceptability of a clarification request is influenced by the visual context. The findings also support the claim that placement of nuclear accent is governed by the visual context. Assignment of nuclear accent to the type or color property of an object is preferred when the visual context licenses the placement.

In all combinations of experimental conditions we observe that utterances in which the nuclear accent placement is congruent with the visual context are perceived more often as good than those where the accent placement is not congruent with the visual context. The converse, that utterances in which the nuclear accent placement is not congruent with the visual context are more often perceived as bad, holds for those cases where the hypothesis expressed in the robot's clarification request is correct, i.e., the human response is 'Yes'. In these cases, the robot and the subject perceive the visual scene in the same way, and therefore have the same evaluation of whether the accent placement is congruent or not.

On the other hand, when the robot's hypothesis is incorrect, i.e., the human response is 'No', the robot's accent assignment and the subject's appropriateness judgment are based on a different perception of the visual scene. These scenes either contain two identical objects (congruent condition from the robot's viewpoint) or two objects which differ in both color and type (non-congruent condition from the robot's viewpoint). In the subject's view these scenes are thus ambiguous w.r.t. accent placement, i.e., they are able to license both tune A and B. Therefore, the comparison between judgements for the congruent and non-congruent condition is not informative in these cases.

We observe that the distribution of judgements is biased by the correctness vs. incorrectness of the robot's hypothesis, i.e., the human response 'Yes' vs. 'No', respectively. This clarifies to an extent why we do not find a difference between the subjective judgement for congruent and non-congruent visual contexts (cf. Fig. 5). If the subjects were focussing on the (in)correctness of the robot's hypothesis, they perhaps paid attention to only the object being introduced. The presence of another object in the visual context and the nuclear accent placement in the intonation perhaps did not influence their decisions as strongly as expected. This might be due to the fact that the subjects were not really involved in the interaction in this experimental setup.

In order to overcome the problems noted above and to investigate further the role of the visual scene and intonation in comprehension, we are preparing an eye tracker experiment for verifying if the subjects pay attention to the already present object when making a judgement. We modify the design of this experiment so that the subject is required to answer the robot's clarification request. In this manner the subjects will be directly involved in the interaction with the system. Moreover, since the subjects will be required to respond to the robot's utterances, the objective nature of the task will enable us to measure the influence of the visual scene and the intonation on their reaction. The hypothesis for this experiment is that with congruent intonation the subject will be looking more at the right object, and that they will react faster. At least for the cases where the hypothesis is correct. It is an interesting question whether there will be any differences between the intonation patterns when the robot's hypothesis is wrong.

4. ACKNOWLEDGMENTS

This work was supported by the EU Project CogX (FP7-ICT-215181).

5. REFERENCES

[1] G. Altmann and Y. Kamide. Discourse-mediation of the mapping between language and the visual world: Eye movements and mental representation. volume 111, pages 55–71. 2009.

[2] J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost, and M. Stone. Animated conversation: Rule-based generation of facial expression, gesture spoken intonation for multiple conversational agents. pages 413–420, 1994.

[3] M. W. Crocker, P. Knoeferle, and M. R. Mayberry. Situated sentence comprehension: The coordinated interplay account and a neurobehavioral model. *Brain and Language*, 112(3):189–201, 2010. MC.

[4] E. Garbe, B. S. Rosner, J. Garciá-Albea, and X. Zhou. Perception of english intonation by english, spanish, and chinese listeners. *Language and Speech*, 46(4):375–401, 2003.

[5] I. Kruijff-Korbayová, S. Ericsson, K. J. Rodríguez, and E. Karagjosova. Producing contextually appropriate intonation is an information-state based dialogue system. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 227–234. ACL, 2003.

[6] J. Pierrehumbert. *The Phonology and Phonetics of English Intonation.* PhD thesis, Massachusetts Institute of Technology, 1980.

[7] J. Pierrehumbert and J. Hirschberg. The meaning of intonation in the interpretation of discourse. In P. Cohen, J. Morgan, and M. Pollack, editors, *Intentions in Communication.* MIT Press, Cambridge MA, 1990.

[8] S. A. Prevost. *A Semantics of Contrast and Information Structure for Specifying Intonation in Spoken Language Generation.* Phd thesis, University of Pennsylvania, Institute for Research in Cognitive Science Technical Report, Pennsylvania, USA, 1996.

[9] M. Schröder and J. Trouvain. The german text-to-speech synthesis system mary: A tool for research, development and teaching. *International Journal of Speech Technology*, 6:365–377, 2003.

[10] M. Steedman. Information structure and the syntax-phonology interface. *Linguistic Inquiry*, 31:649–689, 2000.

[11] M. Tanenhaus, M. Spivey-Knowlton, K. Eberhard, and J. Sedivy. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268:1632–1634, 1995.

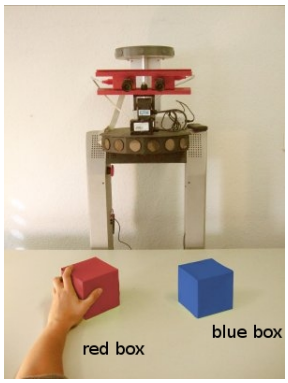
[12] M. White, R. Baker, and R. A. J. Clark. Synthetizing contextually appropriate intonation in limited domains. In *Proceedings of the 5th ISCA Speech Synthesis Workshop*, 2004.

[13] M. White, J. D. Moore, M. E. Foster, and O. Lemon. Generating tailored, comparative descriptions in spoken dialogue. In *FLAIRS Conference*, 2004.

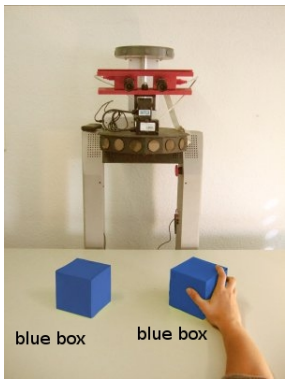
APPENDIX

A. SAMPLE VISUAL STIMULI

Fig. 6 exemplifies the patterns of the visual scenes used in the experiment. It shows the visual scenes presented with the utterance *"Is that a red box"* in the 8 different experimental conditions.



(a) C-A-YES

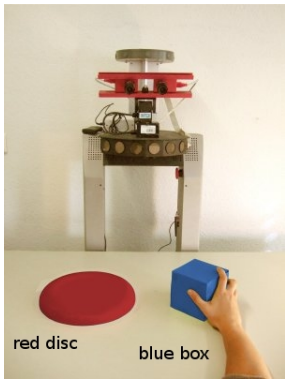


(b) C-A-NO

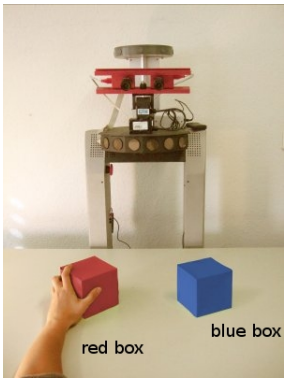
(c) C-B-YES

(d) C-B-NO

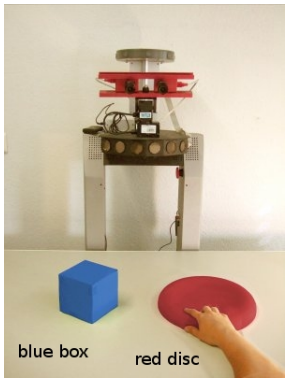
(e) NC-A-YES



(f) NC-A-NO



(g) NC-B-YES



(h) NC-B-NO

Figure 6: The 8 experimental conditions and the corresponding visual scenes for the utterance *"Is that a red box."*