The Reliability of Non-verbal Cues for Situated Reference Resolution and their Interplay with Language - Implications for Human Robot Interaction

Stephanie Gross
Austrian Research Institute for Artificial Intelligence
Freyung 6/6
Vienna 1010, Austria
stephanie.gross@ofai.at

Brigitte Krenn
Austrian Research Institute for Artificial Intelligence
Freyung 6/6
Vienna 1010, Austria
brigitte.krenn@ofai.at

Matthias Scheutz
Tufts University
200 Boston Avenue
Medford, MA 02155, USA
matthias.scheutz@tufts.edu

ABSTRACT

When uttering referring expressions in situated task descriptions, humans naturally use verbal and non-verbal channels to transmit information to their interlocutor. To develop mechanisms for robot architectures capable of resolving object references in such interaction contexts, we need to better understand the multi-modality of human situated task descriptions. In current computational models, mainly pointing gestures, eye gaze, and objects in the visual field are included as non-verbal cues, if any. We analyse reference resolution to objects in an object manipulation task and find that only up to 50% of all referring expressions to objects can be resolved including language, eye gaze and pointing gestures. Thus, we extract other non-verbal cues necessary for reference resolution to objects, investigate the reliability of the different verbal and non-verbal cues, and formulate lessons for the design of a robot's natural language understanding capabilities.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**;
• **Human-centered computing** → *Interaction paradigms*;
• **Computer systems organization** → Robotics;

KEYWORDS

multi-modal human-human and human-robot interaction, situated task descriptions

ACM Reference format:
Stephanie Gross, Brigitte Krenn, and Matthias Scheutz. 2017. The Reliability of Non-verbal Cues for Situated Reference Resolution and their Interplay with Language - Implications for Human Robot Interaction. In *Proceedings of ACM International Conference on Multimodal Interaction, Glasgow, Scotland, November 2017 (ICMI'17),* 9 pages.
https://doi.org/10.475/123_4

ICMI'17, November 2017, Glasgow, Scotland

ACM ISBN 123-4567-24-567/08/06...$15.00
https://doi.org/10.475/123_4

1 INTRODUCTION AND BACKGROUND

Imagine a robot that can analyse, interpret, and learn from task-oriented presentations where a human teacher shows a task to the robot learner and explains what she/he is doing by means of task-accompanying speech. For robots to be able to deal with the multi-modal complexity of human communication, we need to better understand general principles of human task-based descriptions within a shared environment in order to distil the critical interaction principles that have to be integrated into robotic control architectures (i.e., the software and hardware framework for controlling a robot).

Human instructors use not only speech, but various multimodal communication cues such as eye gaze and gestures, when showing and explaining a task to a learner, especially when the learner is physically co-present [see 7, 19, 22, 26]. While language could theoretically be used as the major, possibly even only information channel, it will often be underspecified and is heterogeneously used by individual speakers [see 3, 11, 12].

Incorporating visual information is thus a necessary prerequisite to deal with situated task descriptions. In order to investigate human task descriptions in more detail, an experimental setup was designed, and data was collected and analysed where a teacher explains and shows a task to a learner. By letting different people explain the same task, insights can be gained about how humans naturally structure and present information and the variation between and within task descriptions a robot would have to deal with if it were in the learner's position.

There exists converging psycholinguistic evidence that pointing, eye gaze, placing objects etc. play an important role during language understanding in humans [4, 6, 7]. Accordingly it is vital for computational models aimed at understanding human language to account for its multi-modal complexity.

However, most computational approaches for resolving and generating referring expressions focus only on language and leave non-verbal communicative cues aside (see [29] and [13] for an overview), e.g., the Centering Theory by [16] or the Incremental Algorithm [9] and some of their more recent adaptations (e.g., [14, 23]).

Some computational approaches, which take also non-verbal cues into account for the resolution of referring expressions, are for example [1, 5, 20, 21, 24, 25, 27, 30]. Gundel et al. [17] have formulated the Givenness Hierarchy (GH), assigning different pieces of information to different cognitive statuses (in focus (*it*) ⊂ activated (*that, this, this* N) ⊂ familiar (*that* N) ⊂ uniquely identifiable (*the* N) ⊂ referential (indefinite *this* N) ⊂ type identifiable (*a* N)). The hierarchy is nested, so any information that is in focus, is also activated, familiar etc., any information, that is activated, is also familiar, uniquely identifiable etc. but not in focus and so on. In their resultant coding protocol [18], also eye gaze and pointing gestures are taken into account, for example targets of eye gaze or gesture are automatically activated.

Several computational approaches have build upon and extended this coding protocol. Kehler [21] proposed an adapted version of the GH including four simple rules to resolve references: (i) if an object is gestured to, chose that object; (ii) if the currently selected object meets all semantic type constraints imposed by the referring expression, choose that object; (iii) otherwise, if there is a visible object that is semantically compatible, then choose that object; (iv) otherwise, a full noun phrase (NP) is used that uniquely identified the referent.

Chai et al. [5] applied a greedy algorithm for combining the GH with Conversational Implicature by Grice [15]. Their algorithm is able to handle ambiguities and multiple references in one utterance and their hierarchy looks the following: *gesture* ⊂ *focus* (subsuming Gundel's *in focus* and *activated* tiers) ⊂ *visible* (subsuming Gundel's *activated* and *uniquely identifiable* tiers) ⊂ *others* (subsuming Gundel's *referential* and *type identifiable* tiers).

Williams et al. [31] propose an implementation of the GH handling definite and indefinite noun phrases, and pronominal expressions. Thus, the algorithm is able to deal with a wider range of linguistic expressions than previous approaches. Their algorithm is also able to handle open world and uncertain contexts, though it has not yet been evaluated on a robot.

In addition to the GH, there are also other computational approaches dealing with multi-modal reference resolution. Prasov & Chai [27] for example developed a probabilistic framework to combine linguistic referential expressions and eye gaze for reference resolution, to decrease the need for a complex pre-defined domain model.

Kranstedt et al. [24] aim to model the focussed area of pointing gestures (the "pointing cone") in combination with verbal references, in order to investigate the usage of pointing gestures and linguistic referring expressions. Van der Sluis & Krahmer [30] on the other hand developed a graph based model, assigned costs to linguistic properties and pointing gestures in the generation of multi-modal referring expressions.

Lemaignan et al. [25] present an approach to extract, represent, and use knowledge from real-world perception as well as from human-robot verbal and non-verbal interaction. Strategies for disambiguating concepts include for example whether the previous interaction involved a specific action and whether the user is looking or pointing at a specific object. Currently, their implementation relies on a small, predefined set of action verbs that can be recognized from natural language.

In another approach, Huang & Mutlu [20] develop a dynamic Bayesian network (DBN) for modelling how humans coordinate speech, gaze, and gesture behaviour in narration. Model parameters are learned from annotated data, and the learned model is used to coordinate the modalities on a robot.

In general, non-verbal cues accounted for in current computational models of reference resolution include up to three different cues: objects in the visual field, eye gaze, and pointing gestures. An important exception is the work by Foster et al. [10], who include haptic ostensive references, i.e., reference which involves manipulating an object, in their model. However, their focus is on the generation of referring expressions not reference resolution.

None of the above mentioned models propose solutions for how to deal with verbal and non-verbal aspects of inherently multi-modal situated communication, i.e., which non-verbal cues need to be accounted for, as well as their reliability and interlinkage for automatic reference resolution. Hence, it is crucial to develop more comprehensive computational models of human reference resolution in task-based contexts where instructor and instructee are co-located.

The major goal of this paper is to extract non-verbal cues relevant for reference resolution in situated task descriptions as well as their interplay with the accordant linguistic form (Section 3). Based on these results, we formulate general principles for robot architectures on how to resolve multimodal object references (Section 4). In Section 2, the data collection procedure and data annotation is presented.

2 DATA COLLECTION

All in all, 22 people working or studying at Universities in Munich with German as their mother tongue participated in the data collection activity. They had an average age of 27 and explained four different tasks to either a human or a robot. In this paper, we focus on human-human interactions of one of the four tasks (n=16), because of its comparably large number of different objects and thus the large number of referring expressions to objects (207).

2.1 Procedure

Audio and video data (a frontal video of the teacher, a frontal video of the learner, a video of the setting) of the recordings are used for analysis and annotation.

The task was directed towards a human learner, who was told to carefully watch and listen to the explanations of the teacher to be able to pass the information on to a new learner. In the subsequent trial, the learner became the new instructor. A calibration trial was introduced at least after every fifth trial where the experimenter functioned as an instructor: a) to counteract that the task descriptions get altered over time;

b) to keep communicative variation to a minimum in order to assess which kinds and amount of variation still remains for a robot to deal with even under constrained, however, natural conditions. Additionally, before each task the teachers received a schematic "cheat sheet" depicting the course of action during the task to reduce their cognitive load.

In the task analysed in this paper, an instructor explains and shows to a learner how to connect two separate parts of a tube and then to mount the tube in a box with holdings. The learner stands in front of the table at the left side of the instructor (see Figure 1) and is only observing while the instructor is explaining and conducting the task.

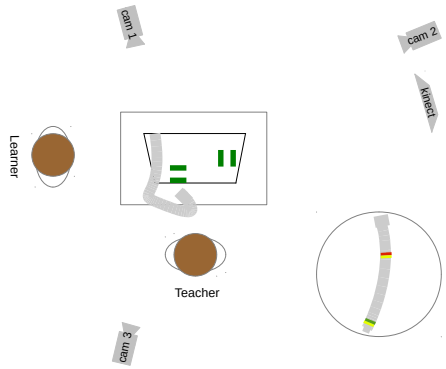

Figure 1: The task setup for mounting a tube.

2.2 Tools used for annotation

For the annotation and synchronisation of the data, the following tools were employed: (i) ELAN[1] for manual annotations and for synchronising audio, video and representation tiers; (ii) Praat[2] for transcribing the utterances. The Praat tiers were then imported in Elan for further analysis; (iii) TreeTagger[3] for part of speech tagging [28].

In addition, Python programs were written to automatically extract temporal sequences of object references and respective multi-modal cues. A mixture of quantitative and qualitative methods was used. For the quantitative analysis, only frequencies and percentages were used, as the data size is too small to conduct statistical tests.

2.3 Annotations

To represent the various kinds of information present in the multi-modal task descriptions, the following annotation tiers were defined and respective annotations were made by two independent annotators and then consolidated.

Transcription of instructor utterances. The sound files with the utterances were manually transcribed, using graphemic representation, being as close as possible to the spoken utterance, i.e., keeping disfluencies, dialectal utterances, concatenations of words or elisions.

Transliteration. In addition to the transcription, an extra tier is added where concatenations typical for spoken language are separated, elisions are recovered, etc. so that the utterances are as close to written text as possible.

POS. The transliterated utterances were used as input to the TreeTagger [28] and the thus resulting part-of-speech sequences were imported to Elan and manually corrected.

Gesture of the instructor. There exists a number of gesture coding schemes, some of which are rather extensive such as the MUMIN [2] and the BAP [8] coding schemes. In the present data the following gesture types were identified and manually annotated: pointing, iconic gestures (depicting aspects of objects, actions, etc.), beat gestures (spontaneous gestures when speaking), emblem gestures (symbolic gestures substituting words), exhibiting gestures (e.g., raising an object in order to direct the interlocutors attention on it) and poising gestures (e.g., poising with the hand above an object before grasping it). In addition to the gesture type, the following information was annotated: for (i) pointing gestures, object, location or person the gesture is directed at, for (ii) iconic gestures, the accordant action, for (iii) emblem gestures the kind of emblem that is used (e.g., "thumbs up" for "great"), (iv) for exhibiting and poising gestures, the object emphasised by the gesture.

Eye gaze of the instructor. At which object, location or person in the scenario the instructor is looking was manually annotated.

Relevant objects. On the "relevant objects"-tier the salient objects in the respective task description scene were manually annotated. The salience of an object is identified by a linguistic reference in the instructor's speech, by the instructor's gaze behaviour, pointing, exhibiting, or poising gestures, whether the instructor is holding or still holding an object, or whether an object is moving towards a target. Linguistic indicators are, for instance, full or elliptic noun phrases, e.g., "the tube" (*den Schlauch*), "tube" (*Schlauch*), pronouns, e.g., "it" (*er*), "the" (*der* for "the tube" (*der Schlauch*), determiners combined with spatial indexicals, e.g., "the one here" (*den hier*), spatial indexicals, e.g., "here", "there" (*hier, da*), adjectives, e.g., "red-yellow" (*rot-gelb*) for the red and yellow marker attached to the tube.

Holding object and still holding object. These cues are manually annotated on the same tier. If an instructor just grasped an object and is currently holding it, it is annotated (with the same tags as for relevant objects). The annotation starts as soon as the instructor's hand touches an object to hold it and ends when it is released, or when the instructor grasps a new object and still holds on to the old one (then both, the old and the new object are annotated).

[1] https://tla.mpi.nl/tools/tla-tools/elan/
[2] http://www.fon.hum.uva.nl/praat/
[3] http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/treetagger.en.html

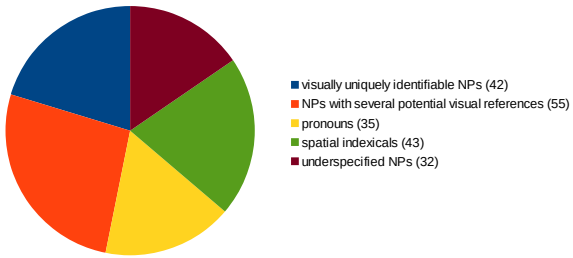

Figure 2: This figure shows all linguistic types of references in the analysed task.

Moving towards a target. As soon as an object is moving towards another object or location, the target of the moving object is annotated.

An important prerequisite for investigating the importance and interplay of different modalities is to determine which object is actually intended by the participant with a certain referring expression. For example both object references "the thing" and "the green and yellow marker" need to be resolved to a physical object. For all object references the annotator had to label the intended object (based on her competence) on the "relevant objects"-tier. Cohen's kappa was computed to measure inter-rater agreement for the relevant object, and with 0.918 the kappa coefficient agreement between annotators is high, showing that humans are rather consistent in interpreting multi-modal references to objects.

3 THE INTERPLAY OF LINGUISTIC FORMS AND NON-VERBAL MODALITIES IN OBJECT REFERENCES

In the literature, gaze and gesture are frequently mentioned as non-verbal cues directed at the referred object and are used for the resolution of references. A first analysis of the data has shown, that in 50% of all referring expressions, either lexically specified noun phrases, pronoun resolution, pointing gestures, or eye gaze can be used to resolve the references.

All in all, 207 object references were uttered referring to: the loose part of the tube, the mounted part of the tube, the two parts of the tube connected, the green and yellow marker, the red and yellow marker, the pair of green holdings on the right side of the instructor, the pair of green holdings on the left side of the instructor, the motor block, and the round table. Language only covers 48 references, i.e., 23.19% via uniquely identifiable noun phrases (42) and pronoun resolution (7) via a congruent, proximate, and uniquely identifiable antecedent. Now the question arises, how all references can be resolved, which additional (non-verbal) cues come into play, and what is the interplay between verbal and non-verbal cues.

3.1 Research questions and results

This section investigates the following research questions (RQ) and discusses the results of an in-depth analysis investigating all 207 object references.

RQ1 Which cues are needed to resolve all object references in situated task descriptions?

RQ2 Is there a correlation between linguistic form and non-verbal cues?

RQ3 How reliable is eye gaze for reference resolution in a task description where different objects are involved?

Uniquely identifiable noun phrases include examples such as "the red and yellow marker" (*die rot-gelbe Markierung*) but also "the red and yellow section" (*der rot-gelbe Abschnitt*) and even "the red one" (*das Rote*) if there is no other object with the attribute "red". The reliability of language, gestures, and eye gaze differs, see Table 1. In the data, uniquely identifiable noun phrases and pronoun resolution via a congruent, proximate and visually uniquely identifiable antecedent are 100% reliable. The other verbal references are insufficient for reference resolution.

Table 1: The reliability of language, gestures, and eye gaze for all 207 object references.

Reliability of	in %	occurrences
Language	100%	48/207
Gestures	76%	37/207
Eye gaze	40%	207/207

With regard to object references, the following types of gestures were of interest: pointing, poising, and exhibiting gestures. Clark (1996) argues that gestures are considered composite parts of references made with deictic expressions.

When explaining the task, the instructors often had both hands occupied when they connected or mounted objects and thus had no hand free to point. Thus, poising and exhibiting gestures are of equal importance as pointing gestures to direct the attention of the interlocutor at a certain object or location. In this respect, all three gestures are deictic acts. The reliability of gestures is 76% and it can be increased, if the following aspects are taken into account:

(i) If the temporal sequence of the gesture is so long, that it lasts during several object references (occurs 6 times), knowledge about verbs needs to be taken into account for reference resolution (see below for details).

(ii) If a person is gesturing with two hands in two different directions, in the data the moving hand was the one directed at the relevant object and the other one was for keeping the attention also at another object.

In general, before using non-verbal cues (such as gestures) to resolve object references, these aspects need to be investigated:

(1) **A summary of the task** before starting the task description in detail can be ignored for reference resolution

to objects as it might not always be possible to link the mentioned object (or constellation of objects) to the already existing objects in the scene. This aspect can be investigated by looking at the words at the beginning of the utterance and the beginning of the task description, e.g., sentences starting with "now it is about" (*hier geht es darum*) or "the task consists of three steps" (*die Aufgabe besteht aus drei Schritten*).

(2) **Utterances that describe the task and meta- descriptions referring to the performance** need to be identified for the same reason, e.g., "that is the task" (*das ist die Aufgabe*), "it is a bit difficult" (*es geht ein bisschen schwer*).

(3) **Semantic knowledge about verbs**: e.g., when two objects of the same type are assembled or connected, the separate objects are not referred to anymore (e.g., if there are two tubes, before their assemblance, they need to be disambiguated, but after their assemblance, there is only one tube although its parts are still visible). Another example is that for certain verbs a hand of the person conducting the action is involved, e.g., for "take" (*nehmen*).

These three aspects are important for the resolution of all references to objects. However, depending on the type of verbal reference, see Figure 2, different additional non-verbal cues are relevant and the reliability of these cues also depends on the type of verbal reference.

Eye gaze as a potential cue is the least reliable, as it is always directed somewhere, but it can not always be used for reference resolution. No pattern could be extracted in which contexts eye gaze can be used as a potential cue and in which it can not be used. Thus, it is too unreliable for automatic reference resolution in situated task descriptions.

3.1.1 Resolution of noun phrases. The majority of uttered references to objects are NPs: 62.93%. They can be grouped into (i) visually uniquely identifiable NPs, (ii) NPs with more than one potential visual referent (e.g., "the marker" if there are two markers), and (iii) underspecified NPs, such as general concepts (e.g., "the thing") and similar semantic concepts (e.g., "the channel" for the holdings).

Visually uniquely identifiable noun phrases

Verbal references of visually uniquely identifiable objects (20.49%) is the easiest case for object reference resolution. In case a NP is uttered including the noun as well as the attributes (if there are any necessary for disambiguation) of a visually uniquely observable object, the references can be resolved. Note: There was no occurrence of an utterance including a visually uniquely identifiable object and the instructor intended to refer to another object (based on the evaluation of the annotators). Thus, the reliability of language is 100% and there are no other cues needed.

NPs with several potential visual references

NPs with several potential visual references refer to more than one potential object in the scene and thus additional visual cues are needed for reference resolution, e.g., "the tube" (*der Schlauch*) or "the green part" (*das grüne Teil*). Out of these 55 references, 28 can be resolved, if **the object the instructor grasped last and is currently holding** is added as a cue, see Table 2. 2 out of these 28 utterances are additionally accompanied by a gesture. In addition to these 28 references, 5 can be resolved via the **gesture** of the instructor directed at the referred object. Gesture and holding an object are very reliable cues in this context (i.e., seldom misleading). Information about the object, the instructor grasped last and is currently holding is only misleading in three cases. These three cases are special and can still be correctly resolved, if aspects (1) – (3) are taken into account.

However, after adding gesture and information about the object, the instructor grasped last as an additional cue, there are still 22 references unresolved. Another important cue is **semantic knowledge about the verb** in combination with visual information about **where a certain object moves** or which other object or location it touches. Referring to NPs with several potential visual references, there is already a preselection of visually perceivable objects. For example, in the utterance "insert with the right hand in this pair of holdings" (*mit der rechten Hand in die Halterung einführen*), the verb transmits the information that during an "inserting"-action two objects touch. The right hand of the instructor is moving towards or touching a certain object and that is the object needed for object resolution. This cue is relevant for 15 object references out of these 55 references.

Table 2: Important non-verbal cues to resolve NPs with several potential visual referents.

	Important cues	occurrences
(4)	Gestures	7/55
(5)	Objects, the instructor is holding	28/55
(6)	Where a certain object moves/ moved	15/55

The missing seven reference can be dealt with by (i) identifying whether the utterance is a **summary of the task before starting the task description in detail**, (ii) the knowledge that **after assembling two objects of the same type, they are not referred to separately anymore**, e.g., if "tube" is uttered after assembling the two tube, it refers to the assembled one and not to the separate parts anymore, (iii) the knowledge that **two colours mentioned one after the other refer to one object**, if there is an object with this attribute, e.g., "the green and the yellow one" (*das Grüne und das Gelbe*), and (iv) the knowledge that **a person can do something with the right and the left hand in combination with knowledge about the verb**, e.g., that the utterance "and then you put with the left one" (*und dann tust du mit der Linken*) refers to the left hand.

Gaze partially overlaps with other non-verbal cues. However, it is very often not directed at the referred object and therefore no reliable cue on its own.

Underspecified NPs

32 underspecified NPs are uttered by the 16 instructors. This group of object references can be divided in NPs lexically underspecified for their conceptual content, such as "the thing" (*das Ding*) and NPs with a similar conceptual content.

19 verbal references were uttered containing a general concept, such as "the whole" (*das Ganze*), "this end" ((*dieses Ende*), "the other part" (*das Andere*).

13 verbal references were uttered containing a noun which does not fully match the referred object but a similar semantic concept, e.g., "pipe" (*Rohr*) instead of "tube" (*Schlauch*) or "channel" (*Kanal*) instead of "holdings" (*Halterung*).

Table 3: Important non-verbal cues to resolve underspecified NPs.

	Important cues	occurrences
(4)	Gestures	7/32
(5)	Objects, the instructor is holding	13/32
(6)	Where a certain object moves/ moved	8/32

Important information to allow reference resolution of these underspecified NPs is (i) to identify an utterance as a task summary four times, (ii) the gestures of the instructor seven times, (iii) knowledge about the verb in combination with information about the object, the instructor grasped last 13 times, as well as (iv) knowledge about the verb in combination with visual information about where a certain object moves or which other object or location it touches eight times, see Table 3.

3.1.2 Resolution of pronouns. Out of these 33 pronouns, seven can be resolved via discourse, via a proximate, congruent and specific antecedent, e.g., in "I take the green and yellow end of the tube and connect it [...]" (*ich nehme das gruen-gelbe Ende des Schlauches und verbinde es [...]*).

As pronoun resolution via discourse fails in the majority of cases, additional cues are needed. Also gestures are seldom used in combination with pronouns. As opposed to NPs, for pronouns the object the instructor grasped before he/she grasped the last object and is **still holding** is a more important cue than the object he/she grasped last, see Table 4. Only in six cases where there was no object he/she was "still holding", the object the instructor grasped last and is **currently holding** could be used to resolve the reference.

However, nine times the instructors were assembling and holding the two parts of the tube while they already referred to it as a whole. Thus, **knowledge about the verb** is needed in order to know that after combining two objects, it can be referred to as one. Only in one case, holding and still holding were misleading: "you would have to insert the tube with the right hand in the pipe, insert no that is somehow, and when it is then inserted [...]" (*du müsstest den Schlauch mit der rechten Hand reinstecken am Rohr reinstecken nein das ist irgendwie und wenn er dann drinnen steckt [...]*). This is a rare case of a summary of an already described process and needs to be identified via the verb. The object which is inserted first is the same object as the one which is in the next step already inserted. It also includes a sentence fragment of a meta-description when the instructor commented that it is not working the way he wanted it. In another case an instructor was holding no object and it was during a summary before starting the task description in detail. With including this knowledge about verbs and excluding meta-descriptions, still holding and holding are very reliable cues for resolving references.

With regards to knowledge about verbs, it is also important if there are, for example, two very similar objects such as the loose part of the tube and the fixed part of the tube and a verb for assembling the two objects, such as "connect" (*stecken, anstecken, hineinstecken, zusammenstecken*) or "combine" (*verbinden, kombinieren*), it takes two objects. It is also possible to utter "one connects that", then the instructor already refers to the assembled object. In the other cases, where the pronoun takes two objects or an object and a location, they are separate during the assembling action, but immediately after the assembling action, there is a new object: the assembled object. Additionally, knowledge is needed that e.g., objects move during "put" (*legen*) or "take" (*nehmen*) to/from a certain location.

Additionally, 11 pronouns could be resolved via pronoun resolution, but for these pronouns, the accordant antecedent had to be resolved via visual cues.

Table 4: Important non-verbal cues to resolve pronouns.

	Important cues	occurrences
(4)	Gestures	1/35
(5)	Pronoun resolution – via a proximate, congruent, and visually uniquely identifiable antecedent	7/35
(6)	Objects, the instructor is still holding	11/35
(7)	Objects, the instructor is holding	6/35
(8)	Pronoun resolution – via a visually resolvable antecedent	11/35

3.1.3 Resolution of spatial indexicals. For spatial indexicals, holding is often misleading and thus not very reliable as a cue. The most important cue for spatial indexicals is the combination of **knowledge about the verb, whether there is a pause before of after "here" (*hier, da*)**, and **towards which object the already mentioned argument of the verb moves / which it touches**. This occurred 41 times, two times accompanied by deictic gestures. In the majority of the cases, the already mentioned argument of the verb is still moving towards the object, but in some task descriptions, the verbal description is a bit slower than

conducting the action. In these cases, the last movement of two objects towards each other can be used as a cue. One instructor omitted the verb "and now this end through here" (*und jetzt dieses Ende noch hier durch*), still "here" can be identified as a location, "this end" can be resolved via visual cues and moves towards the left pair of green holdings, thus, the reference can be resolved. There are three spatial indexicals uttered by the instructors during **summaries before actually conducting the task**.

To resolve spatial indexicals in general, again knowledge about verbs is a necessary prerequisite: for example to resolve "you have to put this tube here" (*du musst den Schlauch hier reinstecken*), and "you have to take this tube here" (*du musst den Schlauch hier nehmen*). Due to the knowledge how many arguments a verb takes or allows (the argument structure of a verb), it can be determined that the first "here" refers to another object / location, the second to the same object as the preceding noun. Another example refers to objects which are parts of other objects, such as the red and yellow marker, which is part of the tube. The verb "insert" (*einführen*) for example takes two objects. If the instructor talks about the red and yellow marker and is holding the tube, there needs to be an additional object, where the instructor puts the tube, because the marker and the tube count as one object referring to the verb. Also in case there is a NP immediately preceding or following the spatial indexical, it is important whether there is a pause before or after "here". Based on knowledge about the verb, it is important where the already resolved arguments of the verb move or if they moved immediately before and are now touching an object. Gestures on the other hand are not reliable to resolve spatial indexicals. Out of five gestures, two are directed somewhere else. Also, holding is not very reliable for reference resolution of spatial indexicals, as locations are often not touched.

It can also be observed that spatial indexicals refer more often to fixed objects (e.g., the markers on the tube, the pair of green holdings or the fixed part of the tube) and very seldom to loose objects (e.g., the loose part of the tube, or the connected part of the tube). It is only used once to refer to the loose part of the tube, and in that case, the instructor refers successively at the two parts of the tube and it is not definitely clear which tube is meant by which referring expression. It is also not essential to understand the utterance: "Ok first we take this tube and this one here" (*also nehmen wir zuerst den Schlauch und den hier*).

4 LESSONS FOR AGENT DESIGN

The reliability of the different cues for reference resolution. With regard to reliability, **language** takes a special role: if there is a linguistic reference to an object in a situated task description, it is always intended to refer to an object in the scene, although its lexical content might often not be enough to uniquely identify an object.

Although **gestures** were rarely the only cue to resolve references, they are still very important for directing attention. However, some gestures are also misleading and do not refer to the intended object. These cases can be identified and avoided by including the following aspects: (i) is the temporal sequence of the gesture so long, that it lasts during several object references, and (ii) is a person gesturing with two hands in two different directions. The first aspect can be resolved via knowledge about the verb, e.g., its argument structure. Does the instructor use both hands for gesturing? If yes, is one of the two hands moving? If yes, the moving hand might be the one directed at the relevant object. A challenge for robot architectures is that gestures valuable for reference resolution to objects are not only pointing, but also exhibiting and poising gestures. Thus, a robust gesture recognition system is needed that also allows for the detection of exhibiting and poising gestures.

Knowledge about the verb such as its argument structure (e.g., in "You have to take this tube here" (*Du musst den Schlauch hier nehmen*) versus "You have to insert this tube here." (*Du musst den Schlauch hier einfügen*) is relevant for the resolution of all verbal referring expressions. This information is not sufficient to resolve referring expressions to objects on its own, but it is very valuable information to distinguish between two or more potential objects identified via other cues.

Information about **the object the instructor grasped last and is currently holding** is a major cue to resolve noun phrases with several potential visual referents. However, this cue is less important to resolve references of pronouns and spatial indexicals.

Information about **the object the instructor grasped before the last one and is still holding** is in particular relevant for the resolution of pronouns. Only in cases where this cue was not present in the task, the pronoun could be resolved via the object the instructor grasped last.

In case a NP or a pronoun was uttered preceding or following a spatial indexical, **a pause** can be used to distinguish, whether this spatial indexical refers to the same object as the pronoun or NP, or to another object or location.

Visual information about where a certain object moves or which object or location it touches is important information to resolve NPs and spatial indexicals. In tasks, often two objects are close to or touch each other. This information was also frequently needed and provides reliable information for reference resolution. In case an object moves towards another object (e.g., when a verb for "put" is uttered), first the biggest moving object should be selected (e.g., the tube moves towards the box) and when it is then clearer that the marker on the tube moves towards a pair of green holding in the box, the marker and the pair of green holdings can be selected.

The interlinkage of verbal referring expressions and additional cues for reference resolution. The results have shown a tight interlinkage between the non-verbal cues relevant for reference resolution and the linguistic form. Also, the multi-modal channels need to be checked in a certain order to successfully resolve reference. While the first four steps for the resolution

of referring expressions overlap, the subsequent – and most relevant cues for the accordant linguistic form – differ.

1 Is the utterance a summary of the task before starting the task description in detail? This kind of utterance can be identified via lexical markers and can be ignored for object reference resolution in shared environments. If not:
2 Is the utterance a meta-description, e.g., referring to the performance (e.g., "I am not very good in doing this" (*das kann ich nicht so gut*), "it is a bit difficult" (*es geht ein bisschen schwer*)? These utterances can also be identified via lexical markers and they can also be ignored for reference resolution to objects. If not:
3 Extract information about the verb including its argument structure, the spatial relation of the manipulated objects before and after the action, and whether a hand is involved to conduct the action. Is/are the other argument(s) of the verb already resolved? This information might be needed in an upcoming step.
4 Is a deictic gesture conducted by the instructor? If yes, check the plausibility (according to information extracted about the verb), whether the object gestured at could be the object referred to. If it is plausible, extract the object.

Noun phrases with several potential visual referents as well as underspecified noun phrases need the following multimodal cues for reference resolution in the following sequence:

5a Has the instructor grasped an object and is currently holding it? If yes, check the plausibility (according to information extracted about the verb), whether the object could be the one referred to. If it is plausible, extract the object.
6a Are already mentioned arguments of the verb moving towards a certain object? If yes, check the plausibility (according to information extracted about the verb), whether the object could be the one referred to. If it is plausible, extract the object.

For pronoun resolution, the object the instructor grasped before the last one and is still holding is the most relevant. In case this cue is not present, the object, he/she grasped last can be used for resolution purposes:

5b Can the pronoun be resolved via a proximate, congruent, and visually uniquely identifiable antecedent? If yes, extract that object.
6b Has the instructor grasped an object before the last object he/she grasped and is still holding it? If yes, check the plausibility (according to information extracted about the verb), whether this object could be the object referred to. If it is plausible, extract the object.
7b In case the instructor has not grasped an object before grasping the last one, has he/she grasped an object at all and is still holding it? If yes, check the plausibility (according to information extracted about the verb), whether this object could be the object referred to. If it is plausible, extract the object.
8b Can the pronoun be resolved via a proximate and congruent antecedent which was already visually resolved? If yes, extract that object.

Spatial indexicals refer more often to fixed than to loose objects. For the resolution of these verbal referring expressions, the most important cue is thus whether the already resolved argument(s) of the verb move(s) towards an object, or just moved towards and now touch(es) an object:

5c Is there a NP immediately preceding or following the spatial indexical? If yes, in case there is no pause in between, the spatial indexical is probably a reference to the same object as the NP, while it is probably referring to another object or location in case there is a pause in between.
6c Based on information about the verb, do the already resolved arguments of the verb move or did they immediately move before and are now touching an object? If yes, extract the object.

For robot architectures, the parallel processing of verbal and non-verbal cues is needed in order to extract and merge the information transmitted via different channels.

5 DISCUSSION AND CONCLUSION

An analysis of the data has shown that in situated task descriptions, language takes more of a secondary role for information transmission. From language, it can be determined whether there is an object reference or an action, e.g., via information about part-of-speech. Eye gaze of the instructor was directed at the object referred to in approx. 40% of the cases and elsewhere otherwise, and thus is not a very reliable cue for automatic references resolution. However, gestures and other non-verbal cues including the position of the hands of the instructor and the relation between objects need to be continuously tracked. Based on knowledge about the verb, this information then needs to be merged in a certain sequence. Also, the consulted cues depend on the linguistic form of the verbal referring expression.

As an upshot, the relevant linguistic and visual information will need to be incrementally incorporated in a robot architecture for the robot to be able to resolve referents. The resulting design suggestions have high potential to enhance human-robot situated task-based interaction.

Even though, the discussed data stem only from one task, due to the qualitative approach, the results show a minimum – though very high – variation with which a robot has to deal with. We argue that thus the results are transferable to other situated task descriptions. There might be some differences for example on the basis of whether (i) the instructor is conducting the task and the learner is only observing or whether it is a collaborative task, (ii) whether both hands of the instructor are frequently occupied, or whether at least one hand is free for gestures etc. However, although the lessons for agent design might thus not be extensive, it is an important starting point already covering a large spectrum of human multi-modal referring expressions.

ACKNOWLEDGMENTS

This research is supported by the Vienna Science and Technology Fund (WWTF, project ICT15-045) and the CHIST-ERA project ATLANTIS.

REFERENCES

[1] Henny Admoni, Christopher Datsikas, and Brian Scassellati. 2014. Speech and Gaze Conflicts in Collaborative Human-Robot Interactions. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society (CogSci 2014)*.

[2] Jens Allwood, Loredana Cerrato, Kristiina Jokinen, Costanza Navarretta, and Patrizia Paggio. 2007. The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources and Evaluation* 41, 3-4 (2007), 273–287.

[3] Susan E Brennan. 1996. Lexical entrainment in spontaneous dialog. *Proceedings of ISSD* (1996), 41–44.

[4] Susan E Brennan. 2000. Processes that shape conversation and their implications for computational linguistics. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 1–11.

[5] Joyce Yue Chai, Zahar Prasov, and Shaolin Qu. 2006. Cognitive Principles in Robust Multimodal Interpretation. *Journal of Artificial Intelligence Research (JAIR)* 27 (2006), 55–83.

[6] Herbert H Clark. 2003. Pointing and placing. *Pointing: Where language, culture, and cognition meet* (2003), 243–268.

[7] Herbert H Clark and Meredyth Krych. 2004. Speaking while monitoring addressees for understanding. *Journal of Memory and Language* 50, 1 (2004), 62–81.

[8] Nele Dael, Marcello Mortillaro, and Klaus R Scherer. 2012. The body action and posture coding system (BAP): Development and reliability. *Journal of Nonverbal Behavior* 36, 2 (2012), 97–121.

[9] Robert Dale and Ehud Reiter. 1995. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science* 19, 2 (1995), 233–263.

[10] Mary Ellen Foster, Ellen Gurman Bard, Markus Guhe, Robin L Hill, Jon Oberlander, and Alois Knoll. 2008. The roles of haptic-ostensive referring expressions in cooperative, task-based human-robot dialogue. In *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction*. ACM, 295–302.

[11] G Furnas, T Landauer, L Gomez, and S Dumais. 1984. *Statistical semantics: Analysis of the potential performance of keyword information systems*. 187–242.

[12] G Furnas, T Landauer, L Gomez, and S Dumais. 1987. The Vocabulary Problem in Human-system Communication. *Commun. ACM* 30, 11 (1987), 964–971.

[13] Albert Gatt, Emiel Krahmer, Kees van Deemter, and Roger PG van Gompel. 2014. Models and empirical data for the production of referring expressions. *Language, Cognition and Neuroscience* 29, 8 (2014), 899–911.

[14] Martijn Goudbeek and Emiel Krahmer. 2012. Alignment in interactive reference production: Content planning, modifier ordering, and referential overspecification. *Topics in cognitive science* 4, 2 (2012), 269–289.

[15] H Grice. 1975. *Logic and conversation*. New York, 41–58.

[16] Barbara J Grosz, Scott Weinstein, and Aravind K Joshi. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational linguistics* 21, 2 (1995), 203–225.

[17] Jeanette K Gundel, Nancy Hedberg, and Ron Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language* (1993), 274–307.

[18] Jeanette K Gundel, N. Hedberg, R. Zacharski, A. Mulkern, T. Custis, B. Swierzbin, A. Khalfoui, L. Humnik, B. Gordon, M. Bassene, and S. Watters. 2006. Coding protocol for statuses on the Givenness Hierarchy. (2006). unpublished manuscript.

[19] Joy E Hanna and Susan E Brennan. 2007. Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory and Language* 57, 4 (2007), 596–615.

[20] Chien-Ming Huang and Bilge Mutlu. 2014. Learning-based modeling of multimodal behaviors for humanlike robots. In *Proceedings of the 2014 International Conference on Human-Robot Interaction (HRI)*. ACM, 57–64.

[21] Andrew Kehler. 2000. Cognitive status and form of reference in multimodal human-computer interaction. In *Proceedings of AAAI/IAAI*. 685–690.

[22] Adam Kendon. 2004. *Gesture: Visible action as utterance*. Cambridge University Press.

[23] E Krahmer and M Theune. 2002. *Efficient context-sensitive generation of referring expressions*. Stanford.

[24] Alfred Kranstedt, Andy Lucking, Thies Pfeiffer, Hannes Rieser, and Ipke Wachsmuth. 2006. Deictic object reference in task-oriented dialogue. *Trends in Linguistic Studies and Monographs* 166 (2006), 155.

[25] Séverin Lemaignan, Raquel Ros, E Akin Sisbot, Rachid Alami, and Michael Beetz. 2012. Grounding the interaction: Anchoring situated discourse in everyday human-robot interaction. *International Journal of Social Robotics* 4, 2 (2012), 181–199.

[26] David McNeill. 1992. *Hand and mind: What gestures reveal about thought*. University of Chicago Press.

[27] Zahar Prasov and Joyce Y Chai. 2008. What's in a gaze?: the role of eye-gaze in reference resolution in multimodal conversational interfaces. In *Proceedings of the 13th International Conference on Intelligent User Interfaces*. ACM, 20–29.

[28] Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to German. In *Proceedings of the ACL SIGDAT-Workshop*. Citeseer.

[29] Kees Van Deemter, Albert Gatt, Roger PG van Gompel, and Emiel Krahmer. 2012. Toward a computational psycholinguistics of reference production. *Topics in Cognitive Science* 4, 2 (2012), 166–183.

[30] Ielka Van der Sluis and Emiel Krahmer. 2007. Generating multimodal references. *Discourse Processes* 44, 3 (2007), 145–174.

[31] Tom Williams, Stephanie Schreitter, Saurav Acharya, and Matthias Scheutz. 2015. Towards Situated Open World Reference Resolution. In *Proceedings of the 2015 AAAI Fall Symposium on AI and HRI*.