# PentoRef: A Corpus of Spoken References in Task-oriented Dialogues

**Sina Zarrieß[1], Julian Hough[1], Casey Kennington[1], Ramesh Manuvinakurike[2], David DeVault[2], Raquel Fernández[3], David Schlangen[1]**

[1]Dialogue Systems Group, CITEC, Department of Linguistics and Literature, Bielefeld University

[2]USC Institute for Creative Technologies, Playa Vista, CA

[3]Institute for Logic, Language and Computation, University of Amsterdam

**Abstract**

PentoRef is a corpus of task-oriented dialogues collected in systematically manipulated settings. The corpus is multilingual, with English and German sections, and overall comprises more than 20000 utterances. The dialogues are fully transcribed and annotated with referring expressions mapped to objects in corresponding visual scenes, which makes the corpus a rich resource for research on spoken referring expressions in generation and resolution. The corpus includes several sub-corpora that correspond to different dialogue situations where parameters related to interactivity, visual access, and verbal channel have been manipulated in systematic ways. The corpus thus lends itself to very targeted studies of reference in spontaneous dialogue.

**Keywords:** Reference, Task-oriented Dialogue, Situated Interaction

## 1. Introduction

We present PentoRef, a corpus of task-oriented spoken dialogues recorded in a puzzle-playing domain where players have to manipulate and communicate about *Pentomino* pieces.[1] PentoRef presents a rich resource for investigating human conversational strategies for referring to objects, on different levels of linguistic realization (including speech and timing/turn-taking) and in different yet consistently represented interactive and visual contexts. In particular, PentoRef is useful for developing automatic systems for, and studying the human mechanisms for, two concrete tasks, namely *reference resolution* (RR) and *referring expression generation* (REG).

The corpus is a meta-collection that bundles up a range of experimental data collected over recent years in the Dialogue Systems Group, first at Potsdam University and then Bielefeld University, and by collaborators. The individual sub-corpora have been used for empirical studies of conversational behaviour in spoken language interaction as well as work on building statistical reference resolution systems in situated environments, in German and English (Fernández et al., 2006; Schlangen and Fernández, 2007; Fernández et al., 2007; Schlangen et al., 2009; Heintze et al., 2010; Kennington et al., 2013; Kennington and Schlangen, 2015).

The common property of the experiments in this collection is that participants have to produce spoken referring expressions to puzzle pieces in a game, normally to instruct another player to carry out a certain move on the Pentomino game board. At the same time, some important parameters of the respective experimental settings were manipulated, such as the way communication was mediated (speech channel and/or visual channel), and the presentation of the scene (virtual or real-world). The original versions of the sub-corpora could not be directly exploited for systematic studies of referring expressions across these settings, due to inconsistent conventions used for segmenting, transcribing and annotating the audio recordings. Moreover, in each experiment, the visual scenes and visual attributes of pieces in a scene were represented in different ways (e.g. either as sets of logical properties or as low-level features from machine vision) such that additional annotation and standardization is needed to exploit the data as an actual corpus of spoken references.

This paper presents the upcoming inaugural release of PentoRef, a unification of these resources that contains high-quality transcriptions of spoken utterances, consistent representations of visual scenes, mark-up of referring expressions and mappings between referring expressions and pieces present in a visual scene. In addition to a consistently structured resource of the raw and derived data, we also provide a light-weight relational database that can be easily processed and queried across the different experimental settings in PentoRef.

[1]*Pentomino* is a puzzle game with pieces based on the 12 different shapes that can be constructed from arranging 5 squares next to each other.

## 2. Related Work

Compared to other resources used in dialogue research, PentoRef follows a tradition perhaps best exemplified by the HCRC Map Task Corpus (Anderson et al., 1991; MacMahon et al., 2006) in that it combines the naturalness of unscripted conversation with the advantages of task-oriented dialogue, such as careful control over aspects of the linguistic and extralinguistic context. Recent comparable data collection efforts are relatively rare, but see (Tokunaga et al., 2012; Gatt and Paggio, 2014).

Related studies in REG research showed that the linguistic phenomena found in the elicited referring expressions vary widely with the modality, task, and audience, cf. (Mitchell et al., 2010; Koolen and Krahmer, 2010; Clarke et al., 2013). Inspired by a recently increasing interest in image description and labelling tasks, data sets of real-world photographs (paired with references to specific entities in the image) have also been created for REG (Kazemzadeh et al., 2014; Gkatzia et al., 2015). Real-world images pose interesting challenges for REG, as the set of visual attributes and, consequently, the distractor objects (objects present in the scene which are not the target of a referring expression) cannot be directly controlled.

Although attempts have been made to systematically assess the effects of the different domains on the reference task (Gkatzia et al., 2015), the comparability of existing reference corpora is limited as they are based on very different types of visual stimuli.
PentoRef provides an unusually wide spectrum of experimental settings that have been investigated in a single domain, combining various levels of interactivity and mediation on the one hand, and variation between virtual and real-world scenes on the other.

## 3. PentoRef Overview

PentoRef consists of different manipulations on task-oriented puzzle-playing using the 12 Pentomino pieces, individuated by their shape. When more than one set of Pentominoes is used, the object type may also be individuated by colour. An important difference to standard reference resources is that control over the set of distractors was not a major consideration during experiment design. Different settings vary widely with respect to number of pieces in a scene, and the properties that a target piece shares with distractor objects. For instance, in some settings, all pieces had the same color. In other settings, each piece had a unique color. Taken together as a corpus, the experiments thus provide an interesting test-bed for REG and RR systems that need to adapt to different types of visual contexts within a common domain.

### 3.1. General Task

In the puzzle games, a player can have one of the following roles: (i) the Instruction Giver (IG), the player who has complete knowledge about the game's goal (e.g. a picture of a shape constructed out of Pentomino pieces), but who cannot manipulate the pieces herself, or (ii) the Instruction Follower (IF) who can manipulate pieces, but does not have knowledge about the game's goal. In order to achieve the goal, the IG has to formulate verbal instructions which the IF has to execute in terms of actions on the game board (i.e. selecting, moving, rotating, or placing pieces).
In this task-oriented setting, it is possible to directly assess the communicative success (effectiveness) of an utterance or a referring expression in that if the IF could quickly identify the intended Pentomino piece in the scene, the referring expression formulated by the IG was immediately effective. In some of the interactions, only the piece selection is required of the IF rather than the construction of the entire puzzle, however reference identification is common to all domains.
The corpus contains two main types of task-oriented interactions:

**Human-wizard interaction:** A human IG has the task to instruct what they believe to be a machine to select or move certain pieces on a game board or desk. Depending on the setting the IG can use speech, and sometimes, gesture. Behind the scenes, a human wizard performs the game actions as the IF. The IG receives signals of the wizard's game actions (e.g. via highlighted pieces on the screen, or audio signals). In some cases, the IG can react to these signals.

**Human-human dialogues:** The IF is a human player that communicates with the IG via speech. Both players collaboratively perform the task (i.e. building a shape out of Pentomino pieces). The IG has the desired solution to the puzzle, but cannot manipulate pieces, whereas the IG can manipulate pieces but does not have the solution.

### 3.2. Experimental Settings

Table 1 shows an overview of the data that we have bundled up for PentoRef, and introduces the sub-corpora with their labels, as they were used in previous research. Experimental settings have been manipulated along the following dimensions.

**Scene:** In virtual settings, Pentomino pieces are shown as graphical objects on a computer screen. In the real-world settings, participants had to interact with real pieces on a physical game board. There is also an intermediate level of "images" in the RDG-Pento experiment, a version of the RDG-Image game described in (Manuvinakurike et al., 2015), using the same web-based data-collection methods using photographs of real Pentomino pieces.

**Pre-solved game:** When the game plan is pre-solved, the IG cannot decide on the pieces that the IF has to select and actions that the IF has to perform, but has to follow some plan given to them as a stimulus. When the game is not pre-solved, the IG can freely decide on the order of game actions, and potentially, the types of pieces the IF has to select.

**Vision:** When vision is available, IGs can observe what the IF is doing, e.g. via a camera feed of the IF's game board and their hands, or the IF's mouse movements on a screen. Otherwise, participants only communicate via speech.

### 3.3. Scenes and Distractors

In each experimental setting, players had to interact with Pentomino pieces. Beyond that common property, the different settings vary widely with respect to number of pieces in a scene, and the properties that a target piece would share with distractor objects. This is illustrated in Figure 1, showing four example scenes from Take, Take-CV, Visual Pento, and WOz-Pento. For instance, in Visual Pento, all pieces initially have the same color (blue) and their shape uniquely distinguishes them from all other pieces.
For the Take experiment, the scenes were randomly generated and contained a large number of pieces in various colors such that there were always pieces that had the same color and/or shape. As another example, the scenes in Take-CV were composed of real Pentomino pieces taken from 3 sets and randomly distributed on a desk. In this case, some colors only occur with a particular shape (e.g. red crosses). Moreover, there were wooden pieces or pieces with different shades of the same color.
Another difference between the virtual and the real scenes concerns the orientation of the pieces. In the virtual scenes, the pieces were arranged on a regular rectangular grid. The

**Human-wizard Interactions**

| Experiment | Scene | Pre-solved Game | WOz Task | Language |
|---|---|---|---|---|
| WOz Pento | Virtual | Yes | Select, move | German |
| Take | Virtual | No | Select | German |
| Take-CV | Real-world | Yes | Select | German |

**Human-human Interactions**

| Experiment | Scene | Pre-solved Game | Vision | Language |
|---|---|---|---|---|
| Push-to-talk | Virtual | No | No | German |
| Noise/No-Noise | Virtual | Yes | No | English |
| Visual Pento | Virtual | Yes | Yes | German |
| Pento-CV | Real-world | No | Yes | German |
| RDG-Pento | Images | Yes | No | English, German |

Table 1: Overview of experimental settings in the PentoRef corpus

real scenes were more cluttered, and pieces can have various orientations.

## 4. Experimental Settings

### 4.1. WOz Pento

**Task** In this Wizard-of-Oz study, users gave instructions to the system (the wizard) in order to manipulate (select, rotate, mirror, delete) puzzle pieces on an upper board and to put them onto a lower board, reaching a pre-specified goal state. Each participant took part in several rounds in which the distinguishing characteristics for puzzle pieces (color, shape, pro- posed name, position on the board) varied widely.

### 4.2. Take

**Task** In this Wizard-of-Oz study, the participant was confronted with a game board containing 15 randomly selected Pentomino puzzle pieces (out of a repertoire of 12 shapes, and 6 colors). The positions of the pieces were randomly determined, but in such a way that the pieces grouped in the four corners of the screen. They were instructed to (silently) choose a Pentomino tile on the screen and then instruct the computer system to select this piece by describing and pointing to it. When a piece was selected (by the wizard), the participant had to utter a confirmation (or give negative feedback) and a new board was generated and the process repeated.

**Procedure** The participants were seated at a table in front of the screen. Their gaze was then calibrated with an eye tracker (*Seeingmachines FaceLab*) placed above the screen and their arm movements (captured by a Microsoft Kinect, also above the screen) were also calibrated. The utterances, board states, arm movements, and gaze information were recorded in a similar fashion as described in (**?**). The wizard was instructed to elicit pointing gestures by waiting to select the participant-referred piece by several seconds, unless a pointing action by the participant had already occurred. When the wizard misunderstood, or a technical problem arose, the wizard had an option to flag the episode.

### 4.3. Take-CV

**Task** In this Wizard-of-Oz setting, participants were seated in front of a table with 36 Pentomino puzzle pieces that were randomly placed with some space between them. The task of the participant was to refer to that object using only speech, as if identifying it for a friend sitting next to the participant.

**Procedure** Above the table was a camera that recorded a video feed of the objects, processed using OpenCV to segment the objects; of those, one (or one pair) was chosen randomly by the experiment software. The video image was presented to the participant on a display placed behind the table, but with the randomly selected piece (or pair of pieces) indicated by an overlay. The wizard had an identical screen depicting the scene but not the selected object. The wizard listened to the participants RE and clicked on the object she thought was being referred on her screen. If it was the target object, a tone sounded and a new object was randomly chosen. If a wrong object was clicked, a different tone sounded, the episode was flagged, and a new episode began. At varied intervals, the participant was instructed to "shuffle" the board between episodes by moving around the pieces.

**Phases** The first half of the allotted time constituted Phase 1. After Phase 1 was complete, instructions for Phase 2 were explained: the screen showed the target and also a landmark object, outlined in blue, near the target. The participant was instructed to refer to the target using the landmark. (In the instructions, the concepts of landmark and target were explained in general terms.) All other instructions remained the same as Phase 1. The targets identifier, which was always known be- forehand, was always recorded. For Phase 2, the landmarks identifier was also recorded.

### 4.4. Noise/No-noise

**Task** The IG instructs the IF on how to build a Pentomino puzzle—an elephant shape built out of tiles that are composed out of five squares (see Figure 1). The IG has the solution of the puzzle, while the IF is only given the outline and a set of 12 loose pieces. The Pentomino pieces available to the IF, while distinct in shape, are all the same colour and do not have an identifying label.

**Conditions** In Noise/No-Noise, there were two conditions: a *Noise* condition (experimental group) where the channel from the IG to the IF was manipulated by replacing, in real time and at random points, all signal with noise (brown noise, matched to loudness level of channel); and

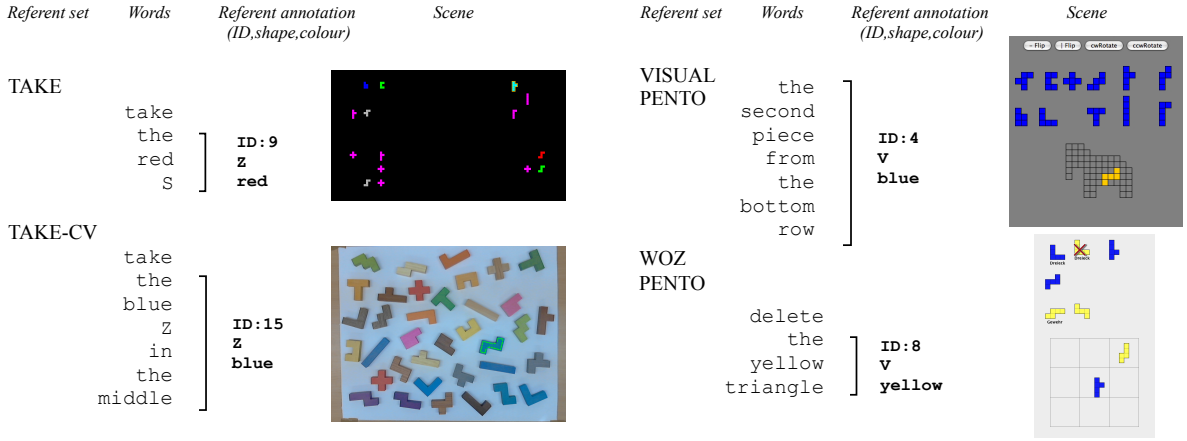


Figure 1: A common reference mark-up across the PentoRef settings (the letters V and Z serve as shape identifiers)

a *No-noise* condition (control group) where there were no manipulations.

**Procedure** Subjects were jointly greeted by the experimenter, who briefly explained the tasks to be carried out and allowed them to choose their roles as either IG or IF. They were then placed in different sound-proof rooms and were given written instructions for the Pentomino task. The IF was allowed a few minutes to get used to the Pentomino program. After subjects had read the instructions, the experimenter asked each of them whether they had any questions. Before leaving the IF room, the experimenter said to the IF something to the effect of: "There might be some problems with the audio, which we can't fix at the moment, so please just go ahead". This was done in order to prevent subjects in the noise condition from coming out of the room to complain about the quality of the audio. Finally the experimenter left the rooms and the first phase of the run began.

### 4.5. Visual Pento

**Task** Same as Noise/No-noise.

**Procedure** The setting in this experiment was very much like the one described for the Pentomino task in the *Noise* experiment, except that there was a visual channel between IG and IF that allowes IG to see the actions performed by IF on the board. This was realised technically through a Virtual Network Computing (VNC) connection between the IF computer and a computer in IG's room, which replayed the GUI of the Pentomino program on which the IF was executing the instructions. Recording was done as described for the *No-noise* condition.

### 4.6. Pento-CV

**Task** In this human-human set-up, two participants worked together to construct objects out of 12 pentomino tiles, one person could see the goal shape (the IG), the other could manipulate the objects (the IF). Each game was further subdivided into an initial selection phase and the actual game. In the selection phase, the IF picked some objects and presented them to the IG. The IG had to find a shape in a database with those objects. After that, the IG directed the IF in creating that shape.

**Procedure** Subjects were jointly greeted by the experimenter, who briefly explained the tasks to be carried out and allowed them to choose their initial roles as either IG or IF. They were then placed on different tables in the room. Above the table of the IF was a camera that recorded a video feed of the objects and his hands. The video image was presented to the IG on his screen. For each pair of participants, several games were recorded. After the first half of the allotted recording time, participants were asked to switch roles.

### 4.7. RDG-Pento

**Task** This is a Pentomino version of the Rapid Dialogue Game (RDG) described in (Manuvinakurike et al., 2015), a human-human set-up where participants have audio access to each other through microphones and headsets. The participants had mutual visual access to a set of images, which are changed for each new round in the game. The participant playing the IG role would have one of the images on their screen highlighted as a target. They would describe the target to the participant in the IF role, who would try to identify it as fast as possible and click on the image they guessed to be the target. Participants were motivated by time pressure with the incentive to score as many points as possible in each fixed-duration round.

**Procedure** Participants were recruited and their technical set-ups tested via the web in the way described by (Manuvinakurike et al., 2015). Participants would follow on-screen instructions then begin their first round in one of the roles (IG or IF). In each round, the pairs were presented 8 images of Pentomino pieces at a time on their own screens. The participant roles were switched every round. There were several rounds per difficulty level, starting with the easiest task with images of single Pento pieces, then progressing to sets of 2-6 pieces in each image. See Figure 2 for an example of the level with 2 pieces per image.

## 5. Referring in Spoken Dialogue: Examples

PentoRef consists of recordings of spontaneous speech. Most REG corpora have been collected in written, non-interactive domains. However, it is well-known that when humans use referring expressions in more natural, interac-

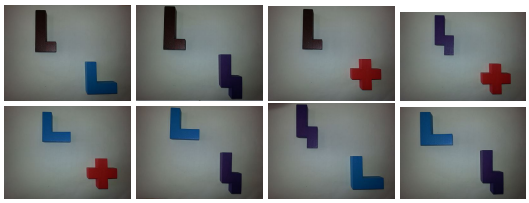

Figure 2: Game board in the RDG setting

tive and situated contexts, conversational strategies are entirely different (Clark and Krych, 2004). Importantly, in a situated dialogue, conversation partners typically collaborate to identify a particular target object, often coordinating on a referring strategy. While the IG utters the RE, the listener (the IF) can give feedback signals (verbal or action-based), or ask for clarifications and engage in repair sequences. A frequent phenomenon is 'reference in installments' where speakers split the reference across several utterances to incrementally build common ground with the listener. On the other hand, in spoken interactions, speakers (instruction givers) do not have unlimited time to ponder an optimal RE to refer to a particular object in a potentially complex scene. As a result, spoken referring expressions (as spoken language in general) typically contain disfluent material, including interruptions, pauses, hesitations, repetitions and self-repairs. To illustrate that PentoRef captures these types of referring, we present a few examples.

Example (1) taken from Visual Pento (cf. Figure 1) illustrates typical phenomena in spoken referring expressions, such as repair, interruption and hesitation.

(1) a. IG: Jetzt kommt der Rüssel.
IG: now comes the trunk.
b. IG: hm, das ...
IG: ehm, the ...
c. IF: *selects distractor*
d. IG: nee ja nich, nicht das ganz lange Teil,
IG: no yes not, not the really long piece,
e. IG: das in der zweiten Reihe links das zweite Teil...
IG: that in the second row on the left the second piece
f. IG: Und das muss einmal.. .
IG: and this has to be once ...
g. IF: Ok ...
IF: okay ...
h. IF: *selects distractor*
i. IG: n weiter links
IG: ehm more to the left
j. IF: *selects target*

In Example (1), the IG first uses an analogical expression to refer to a piece. This is misunderstood by the IF who does not select the intended referent. The IG immediately produces utterances that correct the IF's action and provides more information about the target.

In Example (2), again taken from Visual Pento, the IG is not certain how to name the properties of the target piece in an optimal way (i.e. shape or location) so he uses the location of the mouse pointer as a landmark, and produces a hesitation, and a hypernym. The IF interrupts him and asks for feedback about his current piece selection.

(2) a. IG: Genau da wo der Pfeil jetzt ist ... das Ding...
IG: exactly there where the arrow now is ... the thing

...
b. IF: Das da ?
IF: that one ?
c. IG: Ja .
IG: Yes .

The following example illustrates a human-wizard interaction from the Take data. In this setting, the IG does not have visual access to the wizard's actions (i.e. what he thinks are the machine's actions). Misunderstanding is signaled by silence/inactivity of the wizard. In order to achieve her goal, the IG has to reformulate the initial expression (and infer possible causes of misunderstanding, namely missing information, acoustics problems etc.).

(3) a. IG: ähm das grüne Objekt oben links in der Ecke
IG: em the green object top left in the corner
b. IF: *waits*
c. IG: das grüne Objekt das aussieht wie ein T oben
IG: the green object that looks like a T top left
links in der Ecke.
in the corner.

In the RDG Pento data, participants had to refer to sets instead of individual Pentomino pieces. The following Example illustrates a referring expression from that sub-corpus (produced for the second set in the bottom row in Figure 2).

(4) blue L on the top and the harry potter sign on the right

Finally, we want to point out that our corpus also contains references to locations and a restricted set of actions. In the following example, taken from Pento-CV, the IG tries to explain to the IF how to position and rotate the object on the game board. As this example illustrates, this data is rich in disfluencies which are marked up according to the transcription and segmentation guidelines developed by (Hough et al., 2015).

(5) a. IG: dann kommt das W
IG: then comes the W
b. IG: das steckst du . {also} . mehr so .. (warte ...
IG: that stick you . {well} . more like .. (wait ...
wieviel Grad <v=ist denn>is' n< /v> das dann) ..
how many degree is that then) ...
neunzig Grad nach rechts ... drehen
ninety degree to right ... turn

## 6. Data Representation for Dialogue, Scenes and References

Here we briefly describe the representations we provide in the corpus. The available annotations and overall corpus statistics including word types and tokens in each experimental setting are summarized in Table 2.

### 6.1. Transcription and Segmentation

We provide high quality utterance segmentation and transcription according to the manual in (Hough et al., 2015), all of which was quality checked by the first two au-

| Experiment | # tokens | # types | # utts | # games | # participants | Annotations |
|---|---|---|---|---|---|---|
| **Human-wizard Interactions** | | | | | | |
| WOz Pento | 9149 | 237 | 1686 | 284 | 12 | scene-logical, target |
| Take | 13863 | 383 | 1045 | 1214 | 8 | scene-logical, target, dialogue act tags, disfluencies |
| Take-CV | 15053 | 736 | 870 | 870 | 9 | scene-perceptual, target, landmark, relation |
| **Human-human Dialogues** | | | | | | |
| Noise/No-Noise | 29057 | 1482 | 6073 | 11 | 22 | scene-logical, target, disfluencies |
| Visual Pento | 4610 | 907 | 1158 | 6 | 12 | scene-logical, target, dialogue act tags |
| Pento-CV | 89373 | 1828 | 6108 | 32 | 16 | scene-perceptual, target, dialogue act tags, disfluencies |
| RDG-Pento (En) | 55238 | 1371 | 8030 | 24 | 48 | scene-perceptual, target, dialogue act tags, disfluencies |

Table 2: Corpus statistics and available annotations for PentoRef

thors. For a subset of our corpora, *disfluency* and *laughter* annotation is also included in-line in the way described therein, making it suitable for training and testing disfluency detection. For a subset of the corpora the segments are given dialogue act type tags such as `Instruction`, `Confirmation` and `ClarificationRequest`.

## 6.2. Referent and Scene Representation

Across all datasets we provide a common mark-up for objects, whereby each puzzle piece in a game has a unique ID. Also common across every setting are the two high-level attributes of piece shape[2] and colour from a closed set which is sufficient to identify all piece types across all settings. All referring expressions to pieces are marked with this identifying information over word spans. See Figure 1 which shows the commonality of this mark-up between the virtual and real-world settings. The reference annotation links the transcribed utterances to unique identifiers of pieces in the corresponding scene. In Take-CV, at the time of writing is the only corpus with landmark referents and relations such as 'next to' to be annotated in addition to the target referring expression.

**Visual Information from Scenes** For RR and REG automatic tasks, one wishes to identify a referent in a scene given a representation of the scene and the words, so we make available both logical features and, for the real-world scenes, automatically derived real-valued machine vision captured features of each object in the scene. For example in Figure 1, while the Take dataset provides logical features for a piece such as `colour=red`, in Take-CV, the features provided are from machine vision and will provide features such as `RGB value`, `hue` and `saturation`.

[2]Each object shape name is the letter that corresponds most closely to its shape in its normal orientation.

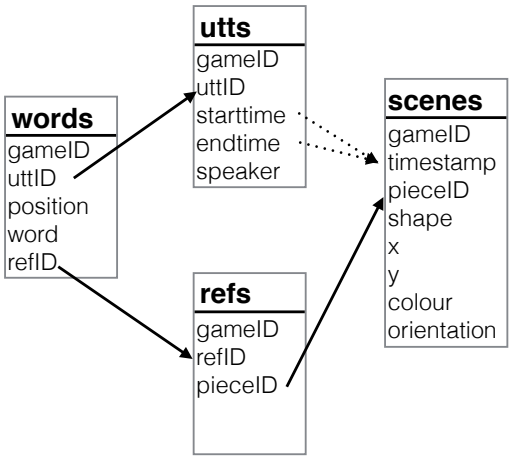


Figure 3: Database design for representing the mapping between dynamic visual context, words and references

**Lightweight database** Our data therefore represents the following layers of information: (i) transcribed words, (ii) segmentation of sequences of words into utterances, (iii) annotation of referring expression on word spans, (iv) representations of visual scenes. We use a light-weight relational database format to represent the data in PentoRef, shown in Figure 3. Information on words, utterances and scenes are kept in tables that can be linked via the identifiers for pieces and referring expressions. Therefore, it is straightforward to query the database for all expressions referring to pieces with a particular shape across the different sub-corpora. In the general case, the scenes in our experiments are dynamic. This means that the location of pieces and their orientation on the game board changes over time. We include timestamps as unique identifiers for scenes.

## 7. Release

PentoRef transcriptions and annotations are made available under a public PDDL license (doi:10.4119/unibi/2901444). Please contact the authors for obtaining audio data.

## 8. Conclusion

We have presented PentoRef, a spoken dialogue corpus consisting of several sub-corpora collected in systematically manipulated settings. The corpus includes a variety of dialogue situations that differ systematically with respect to interactivity, verbal channel, and visual access, which allows for interesting comparisons between experimental settings. The corpus is fully transcribed and enriched with different representations of visual scenes and annotations of referring expressions, providing a rich resource for reference in spontaneous spoken language.

## 9. Acknowledgements

This work was supported by the German Research Foundation (DFG) through the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) at Bielefeld University and the DUEL project (grant SCHL 845/5-1).

## 10. References

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., et al. (1991). The HCRC Map Task corpus. *Language and Speech*, 34:351–366.

Clark, H. H. and Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, 50(1):62–81.

Clarke, A. D., Elsner, M., and Rohde, H. (2013). Where's wally: the influence of visual salience on referring expression generation. *Frontiers in psychology*, 4.

Fernández, R., Lucht, T., Rodríguez, K., and Schlangen, D. (2006). Interaction in task-oriented human–human dialogue: The effects of different turn-taking policies. Proceedings of the First International IEEE/ACL Workshop on Spoken Language Technology.

Fernández, R., Schlangen, D., and Lucht, T. (2007). Push-to-talk ain't always bad! comparing different interactivity settings in task-oriented dialogue. Proceeding of DECALOG, the 11th International Workshop on the Semantics and Pragmatics of Dialogue (SemDial07).

Gatt, A. and Paggio, P. (2014). Learning when to point: A data-driven approach. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2007–2017, Dublin, Ireland, aug. Dublin City University and Association for Computational Linguistics.

Gkatzia, D., Rieser, V., Bartie, P., and Mackaness, W. (2015). From the virtual to the real world: Referring to objects in real-world spatial scenes. In *Proceedings of EMNLP 2015*. Association for Computational Linguistics.

Heintze, S., Baumann, T., and Schlangen, D. (2010). Comparing local and sequential models for statistical incremental natural language understanding. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 9–16. Association for Computational Linguistics.

Hough, J., de Ruiter, L., Betz, S., and Schlangen, D. (2015). Disfluency and laughter annotation in a light-weight dialogue mark-up protocol. In *The 6th Workshop on Disfluency in Spontaneous Speech (DiSS)*.

Kazemzadeh, S., Ordonez, V., Matten, M., and Berg, T. L. (2014). ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pages 787–798, Doha, Qatar.

Kennington, C. and Schlangen, D. (2015). Simple learning and compositional application of perceptually grounded word meanings for incremental reference resolution. Proceedings of the Conference for the Association for Computational Linguistics (ACL), pages 292–301. Association for Computational Linguistics.

Kennington, C., Kousidis, S., and Schlangen, D. (2013). Interpreting situated dialogue utterances: an update model that uses speech, gaze, and gesture information. Proceedings of SIGdial 2013.

Koolen, R. and Krahmer, E. (2010). The d-tuna corpus: A dutch dataset for the evaluation of referring expression generation algorithms. In *LREC*.

Kousidis, S., Pfeiffer, T., Malisz, Z., Wagner, P., and Schlangen, D. (2012). Evaluating a minimally invasive laboratory architecture for recording multimodal conversational data. In *Proceedings of the Interdisciplinary Workshop on Feedback Behaviors in Dialog, INTERSPEECH2012 Satellite Workshop*.

MacMahon, M., Stankiewicz, B., and Kuipers, B. (2006). Walk the talk: Connecting language, knowledge, and action in route instructions. *Def*, pages 1475–1482.

Manuvinakurike, R., Paetzel, M., and DeVault, D. (2015). Reducing the cost of dialogue system training and evaluation with online, crowd-sourced dialogue data collection. In *Proceedings of SEMDIAL 2015 goDIAL*.

Mitchell, M., van Deemter, K., and Reiter, E. (2010). Natural reference to objects in a visual domain. In *Proceedings of the 6th international natural language generation conference*, pages 95–104. Association for Computational Linguistics.

Schlangen, D. and Fernández, R. (2007). Speaking through a noisy channel - experiments on inducing clarification behaviour in human-human dialogue. Proceedings of Interspeech 2007.

Schlangen, D., Baumann, T., and Atterer, M. (2009). Incremental reference resolution: The task, metrics for evaluation, and a bayesian filtering model that is sensitive to disfluencies. Proceedings of SIGdial 2009, the 10th Annual SIGDIAL Meeting on Discourse and Dialogue.

Tokunaga, T., Iida, R., Terai, A., and Kuriyama, N. (2012). The REX corpora : A collection of multimodal corpora of referring expressions in collaborative problem solving dialogues. In *Proceedings of the Eigth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 422–429.