Towards a Multimodal Taxonomy of Dialogue Moves for Word-Guessing Games

Eli Pincus, David Traum

Institute for Creative Technologies, University of Southern California
12015 Waterfront Drive, Playa Vista, CA 90094
pincus@ict.usc.edu, traum@ict.usc.edu

Abstract

We develop a taxonomy for guesser and clue-giver dialogue moves in word guessing games. The taxonomy is designed to aid in the construction of a computational agent capable of participating in these games. We annotate the word guessing game of the multimodal Rapid Dialogue Game (RDG) corpus, RDG-Phrase, with this scheme. The scheme classifies clues, guesses, and other verbal actions as well as non-verbal actions such as gestures into different types. Cohen kappa inter-annotator agreement statistics for clue/non-clue and guess/non-guess are both approximately 76%, and the kappas for clue type and guess type are 59% and 75%, respectively. We discuss phenomena and challenges we encounter during annotation of the videos such as co-speech gestures, gesture disambiguation, and gesture discretization.

Keywords: rapid dialog, RDG-Phrase, clues, guesses, gesture disambiguation, gesture discretization

1. Introduction

In this work we develop a taxonomy of dialogue moves for team word guessing games in which one or more team members (called clue receivers) try to guess a target word or phrase known to the other partner (clue giver). The clue giver can use verbal or non-verbal means to elicit the target from the receiver. Generally, there are also restrictions on what the giver can say or do, which includes not saying (parts of) the target, but also might include other forbidden words or expressions. Variations of this game are popular as parlor games, card games, electronic games, and television game shows.

The taxonomy, presented in Section 2., seeks to capture strategies and typical behavior of both givers and receivers. This is done as a first step towards construction of computational agents capable of simulating human players of word-guessing games. To this end, we define categories for different types of clues, different delivery methods of clues, different types of guesses, as well as more generic actions such as hesitations. We also define several attributes that these actions can possess.

This taxonomy was used to annotate parts of the multimodal Rapid Dialogue Game (RDG) corpus (Paetzel et al., 2014). One of the games in this corpus, called RDG-Phrase, is a word-guessing game. This game has a single clue receiver, who is face to face contact with the clue giver. The clue giver has an opportunity to view, order, and prioritze the set of target words before each round. There is also a strict time limit, encouraging rapid interaction. Each pair in the corpus alternates rounds as clue giver and clue receiver. An interested reader should refer to Table 2 for a sample dialogue (with annotation) or for a longer sample dialogue (Paetzel et al., 2014).

2. Annotation Scheme

We divide actions that occur during word guessing game play into two categories according to role: clue giver or clue receiver. Both giver and receiver actions come in verbal and non-verbal form. Giver verbal actions are classified as either clues or non-clues. Receiver verbal actions are classified as either guesses or non-guesses. In order to address the multi-functionality nature inherent in utterances as discussed in (Bunt, 2010) , we use the "code high" approach (Condon and Cech, 1995) and specifiy a hierachy of tag types, so that lower priority tags are used only if no higher priority tags are used. Clues and guesses are higher priority than non-clue and non-guess. We further subdivide all of these categories by type. Clues are also associated with a delivery method attribute according to the structure of the sentence(s) utilized by the giver to deliver the clue to the receiver. Besides delivery method; we have defined several other attributes for verbal actions that we will define below. Non-verbal actions are broken into 7 categories: turn-taking, metaphoric, iconic, deictic, positive symbolic, negative symbolic, and other.

2.1. Giver Verbal Clues

There are 16 clue types defined below. Example instances can be found in Table 1. Each clue type is given a priority [A,B, or C], shown to in parentheses to the right of the type name, below. The "code high" principle is used to code clue types only from the highest category, in the case that more than one applies.

Analogy ***(A)*** clues set up a relationship between two entities and then attempt to elicit the receiver to recognize the same relationship between the target and another entity.
AssocAction ***(B)*** clues are utterances that describe what the target word does, what it is used for, or what uses it.
CitePast ***(B)*** clues reference previous turns or segments.
Contrast ***(A)*** clues supply a contrasting word or concept.
DescriptionDef ***(C)*** clues either describe or define the target word.
Disabuse ***(B)*** clues are meant to convey to the receiver that his guesses are off track.
Hypo ***(A)*** and **Hyper** ***(A)*** clues occur when the giver provides hyponyms or hypernyms of the target, respectively.
GeneralContext ***(A)*** clues cite knowledge that depends on aspects of the conversation situation (time, visible objects, etc.) concepts such as the current time, the current specific location, or the objects that are present in the room.
PartialPhrase ***(A)*** clues refer to instances where the giver states words that are commonly used with the target word or describe words that are commonly used with the target.
SemanticClass ***(A)*** clues are giver utterances that contain words with the same hypernym as the target word or utterances that request the receiver to say words with the same hypernym as the receiver's previous guess.
Synonyms ***(A)*** provide a synonym of the target word or a

Clue Type	Delivery Method	Target	Instance	Next Guess
Analogy	Complete	Night	"light versus dark but daytime"	Incorrect
Descr/Def	Fill-in-Blank	Alley	"The pathway behind a building is called a"	Correct
Contrast	Fragment	Video	"Not audio"	Correct
CitePast	Complete	Today	"you mentioned it before"	Incorrect
AssocAction	Complete	Doll House	"A place little girls play in"	Correct
Hyper	Fragment	Bus	"Public Transportation"	Correct
Hypo	Fragment	Gas Guzzler	"Cadillac"	Incorrect
SemanticClass	Fragment	Hour	"Um minute"	Incorrect
Partial-Phrase	Fill-in-Blank	Cabin	"Abraham Lincoln lives in a log"	Correct
Synonym	LeadingQue.	Main Street	"Whats another word for major"	Incorrect
Disabuse	Fragment	Electric	"nope" (after prius)	Incorrect
GeneralContext	Fragment	Today	"friday" (said on a Friday)	Incorrect
RequestSynonym	Complete	Hair Care	"another word for that" (after guess of "nurture")	Incorrect
Widen	Fragment	Hair Care	"more in general" (after guess of "washing hair")	Incorrect

Table 1: Example Clue Types

close approximation to a synonym of the target word.
RequestSynonyms ***(A)*** and **RequestAntonyms** ***(A)*** are clues where the giver directs the receiver to provide synonyms or antonyms, respectively, of words recently said.
Rhyming ***(A)*** clues have words that rhyme with the target.
Widen ***(A)*** clues ask the receiver to generalize what he is saying while **Narrow** ***(A)*** clues ask the receiver to state something more specifically.
Each clue has a **Delivery Method** that specifies the manner in which it is said. **Fill-in-Blank** clues are given as a sentence containing a missing word that is intended to be the target. Clues given in the form of a **LeadingQuestion** are expressed in the form of a question whose answer is supposed to be the target. A clue stated as a full sentence that does not fall into the other categories is considered **Complete** while a clue that is not a fully formed sentence and is not a Fill-in-Blank is a **Fragment**. If the delivery method of the clue is not clear, the clue's delivery method is said to be **None**. Refer to Table 1 for some example clues and their associated delivery methods.

2.2. Receiver Verbal Guesses

Receiver guess types are assigned to one of 6 categories. **Correct** guesses state the target word. **PartialCorrect** guesses contain the target within a larger word or phrase while **AbbreviatedCorrect** guesses state an abbreviated version of the target. **Partial** guesses are ones that state a part of the target but not the whole target. A guess is considered **Incorrect** if it contains no part of the target. Finally, guesses that are incomplete and therefore can not be unambiguously classified into one of the other categories are labeled as **None**. If a receiver utterance contained multiple guesses annotators marked the guess in the following order of priority: Correct, Partial Correct, Abbreviated Correct, Partial, Incorrect, None.

2.3. Non-Clues & Non-Guesses

Giver and receiver non-clue and non-guess actions have several categories in common. Both players can state an **Acknowledgement** indicating understanding of what the other player has said or a **Clarification** indicating that the player requires additional information about what was just said. Alternatively, either player can state a **Delay**, a filler utterance said while a player is thinking about his next action. The former three non-clue/non-guess types are instances of core dialogue dimensions discussed in (Bunt, 2010) as none of the types qualify as a RDG-Phrase dependent dialogue act. Acknowledgement and Clarification lie in the *Auto-Feedback* dimension and Delay has the communicative function *Stalling* in the *Time-Management* dimension. In addition, either player can utter an **Encouragement** in an attempt to boost the other player's morale or request to **Skip** to the next target. Either player can also **Evaluate** their performance by expressing thoughts on current game-play or emit **Laughter**. Evaluate, Skip, and Encouragement lie in the *Task* core dimension defined in (Bunt, 2010).
The giver can state a **Confirmation** in order to convey to the receiver that he has made a correct guess or partially correct guess. On the other side, the receiver may **Reject** by communicating his lack of knowledge of the target based on current information or **RequestRepeat** by asking the giver to repeat his last clue. Confirmation and RequestRepeat can be viewed as lying in Bunt's *Auto-Feedback* dimension while Confirmation can be viewed as lying in Bunt's *Task* dimension. Note that we only consider Laughter and Delay tags if none of the other tags seem appropriate.

2.4. Additional Verbal Attributes

We have also defined a number of attributes for clues and guesses. **Repeat** clues or guesses have already been used for the current target, **Incomplete** ones have been cut short, while clues or guesses assigned **ProsodyCompletion** are identified by their extended prosody. A **Multiple** guess is a receiver utterance composed of multiple guesses. Any clue labeled as **MultiWord** is a clue intended to elicit only part of the whole target from the receiver. **Recast** clues are clues that have adopted content words used by the receiver to guess the current target. Clues labeled with the **Clarification** attribute are ones that could not be understood without knowledge of previous clues. If the annotator feels that one clue spans either sequential giver utterances or giver utterances that are separated by Delay utterances or Laughter utterances only; then the blocks that span the clue are labeled **Partial** to indicate the multiple-block span nature of the clue. The delivery method attribute is then assigned to each of these blocks by considering all of the blocks as a single entity rather than assigning a delivery method attribute to each individual block. Table 2 shows a partial

Speaker	Utterance	Type	Attributes
Giver	"Not a large car but a"	Contrast	DM:Fill-in-Blank
Receiver	"Small car sedan"	Incorrect	Multiple
Giver	"Small"	Synonym	DM:Fragment;Recast
Receiver	"Small car"	Incorrect	Repeat
Giver	"Small car"	Hyper	DM:Fragment;Recast
Receiver	"Suburban [laughter] oh suburban"	Incorrect	-
Giver	"Sub"	PartialPhrase	DM:Fragment
Receiver	"Oh subcompact"	PartialCorrect	-
Receiver	"Right got you"	Acknowledgment	-

Table 2: Sample RDG-Phrase Dialogue with Target: Compact

transcription of a RDG-Phrase game, with annotations.

2.5. Non-Verbal

Initially, we divided non-verbal actions into 7 categories, loosely based on the categories of (McNeill, 1995), with a few specialized to timed guessing games: turn-taking, metaphoric, iconic, deictic, positive symbolic, negative symbolic, and other.

3. Annotation Method & Evaluation

3.1. Method

We utilize the multi-modal annotation tool Anvil (Kipp, 2012) to perform our annotation. Speech was segmented in the transcriptions of the RDG-Phrase videos if it was separated by 300 milliseconds of silence or more. We automatically convert these segmented utterances to instantiate utterance block elements in Anvil. Each speaker's utterance blocks are assigned their own "track" in Anvil. Each utterance block is labeled with its type in corresponding blocks in either the giver track or the receiver track and appropriate attributes selected.

3.2. Challenges

Several conversational phenomena arose during the course of our annotation. Co-speech gestures occurred frequently during game-play. We came across many verbal utterances whose semantic content was only clear when one considered the gesture the speech co-occurred with. For instance, in an attempt to elicit the target *playing cards* one giver pantomimed dealing cards while saying "I'm just gonna do this."

As pointed out by Susan Duncan[1], gestures are often multifunctional and segmentation can be particularly challenging as gestures repeat and blend into each other. For example, we frequently came across instances where the giver would utter an uninterrupted stream of clues of the same type synchronously with a rhythmic forward-backward hand extension. These gestures were unequivocally beat gestures but also appeared to serve a turn-taking cue function each time the giver's hand extended forward toward the receiver; seemingly to provide a chance for the receiver to interject with a guess. After initial attempts, we deferred non-verbal coding until we can suitably refine the annotation scheme to focus on those elements that are most crucial for game play.

3.3. Scheme Evaluation

We perform a small inter-annotator agreement study on four sequential seventy-second RDG-phrase rounds played by one pair (team), this includes 90 giver and 57 receiver utterances. Table 3 contains Cohen's Kappa statistics and absolute agreement statistics for each of the major verbal categories in our annotation scheme.

Category	Cohen's Kappa	Absolute
Clue/Non-Clue	76.18%	88.89%
Guess/Non-Guess	75.63%	89.47%
Giver Type	59.00%	64.44%
Receiver Type	74.96%	80.70%
Clue Delivery Method	53.00%	64.71%

Table 3: Inter-Annotator Agreement Statistics

The tags causing the most disagreement for utterances both annotators label as clue are DescriptionDef and AssocAction. This type of disagreement accounts for 3 out of the 10 or 30% of clue type disagreements. One example of this disagreement occurs with the giver utterance "yeah and then this one is on the ocean" where the target had been beach house and the receiver had just correctly guessed country house. This clue seems to fit in both categories as it describes the target like a DescriptionDef but in some sense it also answers the question: what is it used for? like an AssocAction. Instances such as this might lead us to further refine the definitions of these two categories for future annotation efforts.

The most common disagreement for the clue delivery method attribute occurs when one annotator feels the delivery method is not clear and therefore chooses the None value. This scenario accounts for 7 of the 18 tags that did not match; close to 40%. None of the other delivery method disagreements account for more than 3 of the delivery method tags that do not match.

4. Preliminary Annotation Results

The first author annotates all of the speech in 18 70-second RDG-phrase rounds played by three different pairs of people. The speech was segmented into 762 utterances according to our 300 milliseconds of silence criterion. 439 (58%) of the total utterances were said by the giver while 323 (42%) utterances were said by the receiver. See Table 4 for a further breakdown of these utterances.

4.1. Clues & Guesses

Figure 1 shows the relative frequency of Clue types. We find no instances of RequestAntonym or Rhyming clues in the annotated rounds and therefore these two types do not appear in Figure 1. The two most common clue types are AssocAction clues (28%) and DescriptionDef clues

[1] http://mcneilllab.uchicago.edu/pdfs/susan_duncan/Annotative_practice_REV-08.pdf

Giver Utt. Categ.	**# of Utt.** (% Giver Utt.)
Clues	247 (60%)
Non-Clues	162 (40%)
Rec. Utt. Categ.	**# of Utt.** (% Rec. Utt.)
Guesses	224 (69%)
Non-Guesses	99 (31%)

Table 4: Giver & Receiver Utterance Breakdown

(16%). One possibility is that this indicates that the definition of AssocAction captures important properties of the most common conceptual model for a noun or noun-phrase (all targets fall into one of these two syntactic categories). These statistics also imply that givers find word-relations (a category most of the other clue-types fall under) either more difficult to construct or consider them a less effective way of eliciting the target. We calculate a little less than

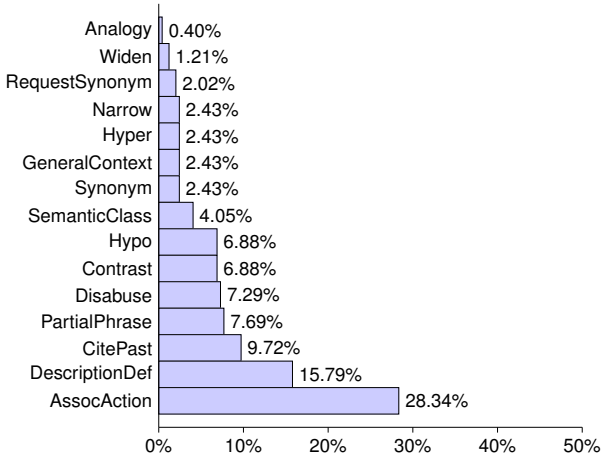

Figure 1: Clue Type Relative-Frequency

a quarter of the total guesses are correct (24%) and 55% contain at least part of the target or an abbreviated version of part of the target. More specifically, the breakdown of guesses are as follows: AbbreviatedCorrect (2.23%), PartialCorrect (2.68%), Correct (23.66%), Partial (26.79%), Incorrect (44.64%).

Clue Delivery Method Table 5 shows clue delivery statistics. The Fragment (39%) and Complete (28%) delivery methods were the most common for clues. This indicates that human givers find non-complete sentences the most efficient manner to deliver a clue and frequently consider structuring a grammatically correct sentence a task that does not contribute a significant amount of value. This also implies human givers use Fill-In-Blank and Leading Question delivery methods less often; possibly due to the time needed to construct clues in these forms.

Delivery Method	**# of Clues (%)**
LeadingQuestion	16 (7%)
None	21 (9%)
Fill-In-Blank	38 (17%)
Complete	64 (28%)
Fragment	89 (39%)

Table 5: Clue Delivery Method Statistics

4.2. Non-Clues and Non-Guesses

We tag 133 (17% of all utterances, 51% of Other Verbal utterances) utterances of either the giver or the receiver as Delay. 74 of these delays were said by the giver and 59 by the receiver. One third of all non-clues said by the giver were Confirmations. 18% of receiver's non-guesses were Acknowledgements. The other non-clue categories and the other non-guess categories each comprised a small relative percentage of all non-clue and non-guess utterances; 21% and 22% respectively. Further annotation and deeper investigation into these statistics should provide us data relevant to constructing a computational agent player that is able to perform behaviors such as backchannels, filled pauses, and turn-taking in a natural manner.

5. Conclusions

We present a taxonomy of dialogue moves for word-guessing games as a first step towards implementing a computational agent that can simulate a human player. Evaluation of our scheme yields reasonable inter-annotator reliability.

In future work, we intend to further refine our annotation scheme including providing guidelines for non-verbal annotation that minimize issues such as gesture disambiguation and gesture discretization. We will also continue our study of word-guessing game strategy by examining the relationship between prior clues and a current guess if the current guess is viewed as the current target. This investigation should also help determine how receivers interpret clues. We also have plans to implement a computational giver that is able to generate clue types such as Synonym, Contrast, Hyper, Hypo and DescriptionDef. We will accomplish this task by linking the giver to a database of word relations such as WordNet (Miller, 1995).

6. Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. IIS-1219253. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

7. References

Sherri L. Condon and Claude G. Cech. 1995. Problms for reliable discourse coding systems. In *AAAI Technical Report SS-95-06 Working Notes AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation.*, pages 27–33, March.

Harry Bunt et. al. 2010. Towards an iso standard for dialogue act annotation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA), May.

Michael Kipp. 2012. Anvil: A universal video research tool. In J. Durand, U. Gut, and G. Kristofferson, editors, *Handbook of Corpus Phonology*. Oxford University Press.

David McNeill. 1995. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press.

George A. Miller. 1995. Wordnet: A lexical database for english. *Communications of the ACM*, 38:39–41.

Maike Paetzel, David Nicolas Racca, and David DeVault. 2014. A multimodal corpus of rapid dialogue games. In *Language Resources and Evaluation Conference (LREC)*, May.