Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 5919–5927
Marseille, 11–16 May 2020

A Formal Analysis of Multimodal Referring Strategies Under Common Ground

Nikhil Krishnaswamy and James Pustejovsky
Brandeis University Department of Computer Science
Walthan, MA, USA
{nkrishna,jamesp}@brandeis.edu

Abstract

In this paper, we present an analysis of computationally generated mixed-modality definite referring expressions using combinations of gesture and linguistic descriptions. In doing so, we expose some striking formal semantic properties of the interactions between gesture and language, conditioned on the introduction of content into the *common ground* between the (computational) speaker and (human) viewer, and demonstrate how these formal features can contribute to training better models to predict viewer judgment of referring expressions, and potentially to the generation of more natural and informative referring expressions.

Keywords: multimodality, interfaces, referring expressions, semantics, common ground

1. Introduction

Multimodality has been a topic of study in computational linguistics and natural language processing since at least the mid-1990s (Johnston et al., 1997), but has seen increased interest from the CL/NLP communities in recent years. This has been due to a number of factors, including the increase in processing power; the availability of large datasets of text, images, and video; the rise of depth sensors (e.g., Microsoft Kinect); and the availability of GPUs for deep model training. This has resulted in a number of new datasets and approaches to cross-modal linking (Yatskar et al., 2016; Goyal et al., 2017), shared tasks (Barrault et al., 2018), and grounding tasks (Beinborn et al., 2018; Zhou et al., 2018).

The most common modalities under study in the CL/NLP communities are text, audio/speech, and images/video, but "modality" can in principle refer to any channel of information. Therefore, multi-channel transmission of information can be separated by channel into the particular information transmitted by each modality (i.e., objects depicted in images with their descriptions in text, or spoken demonstratives with aligned deixis via a gesture). Such disjunct mechanisms allow us to package, quantify, measure, and order our experiences, creating rich conceptual reifications and semantic differentiations. By examining the nature of these differentiations, we can study the conceptual expressiveness of these systems (Pustejovsky, 2018).

Demonstrating such knowledge is needed to ensure a shared understanding between interlocutors, and when one such interlocutor is a computer whose multichannel expressions are quantitatively defined, this allows us to measure certain aspects of the *computational common ground* created by the computer's representation of information it has shared with its interlocutors, including humans.

When two agents are co-situated and attending to the same situation (*co-attending*), it is the introduction of such information into the discourse that creates the "shared situated reference" (Pustejovsky et al., 2017) between them, and the introduction of particular information into the common ground may be more or less informative depending not only on the prior contents of the common ground but also the *modality through which the new information is introduced*. The task is then to assess this, either quantitatively or formally.

In this paper, we present an analysis of the common ground structures presented in a dataset of *Embodied Multimodal Referring Expressions* (EMRE) (Krishnaswamy and Pustejovsky, 2019a). These are references to definite objects performed by an avatar in a simulated world using gesture, language, or both. The appropriateness of each referring technique depicted was then evaluated by annotators on Amazon Mechanical Turk. The virtual environment allows saving a number of quantitative and qualitative parameter values for each depicted referring technique, allowing further analysis, including for our purposes here, of the introduction of elements between the avatar and the annotators (as proxy for the human interlocutors), and the subsequent update to the common ground caused by each new element. We analyze both the formal parameters of the common ground updates, and their quantitative effects on annotator preferences for referring techniques within the data.

2. Related Work

Referring expressions of course pervade natural language dialogues and are a prominent subject of study in natural language processing (Krahmer and Van Deemter, 2012). Dale and Reiter (1995) identify a successful referring expression as one that identifies the intended target to the hearer without introducing false implicatures a la Grice (1975). Paraboni et al. (2007) discuss generating referring expressions in hierarchically structured domains, and explore the hypothesis that reducing search for the identifying referent with a referring expression can be improved by including logically redundant information, such as denoting the same content using different methods. Thus a successful referring expression a la Dale and Reiter may not necessarily be quantitatively optimal as long as it is sufficiently Gricean.

Current approaches to referring expressions include neural approaches with high-dimensional word embeddings (Ferreira et al., 2018) and spatial expression generation in human-robot interaction (Wallbridge et al., 2019)—including grounding referring expressions in an environment using visual features and attributives (Shridhar and Hsu, 2018; Cohen et al., 2019; Magassouba et al., 2019).

Studies in the interaction between language and gesture also have a long history in computational linguistics (Claassen, 1992; Bortfeld and Brennan, 1997; Van Der Sluis and Krahmer, 2001; Krahmer and van der Sluis, 2003; Funakoshi et al., 2004; Viethen and Dale, 2008). Despite this, there has been comparatively little research from the community into the ways that multiple modalities interact during *real-time* communication and how to replicate such structures computationally. Most work in this area originated in the psychology and cognitive science communities, and has been explored in related communities such as robotics (Petit et al., 2012; Matuszek et al., 2014; Whitney et al., 2016; Kasenberg et al., 2019), but has direct relevance to computational language understanding and generation.

McNeill (2000) argues that thought is multimodal, and that the combinatorics of gesture do not correspond to the syntagmatic values that emerge from the combinatorics of speech. Quek et al. (2002), holds that speech and gesture are coexpressive and processed partially independently, and therefore complement each other. Thus, if interlocutors agree that the meaning of a gesture in a description and the meaning of accompanying speech share the same referent, this must be tested to see if 1) the gesture and speech align, and 2) they share the same denotative content. Thus rather than by abstract combinatoric analysis, the appropriateness of the referencing operation must be tested within a shared common ground. This is where we feel that both formal and statistical analysis can be used together to establish computational principles for combining multimodal streams while maintaining maximum interpretability. First we will describe the dataset we examined, then outline the formal principles of computational common ground, and finally present the methodology and results of our analysis.

3. Embodied Multimodal Referring Expressions

Previously, we gathered a dataset of what we called *Embodied Multimodal Referring Expressions* (EMRE): that is, visualizations of an agent (here a virtual avatar in a simulated environment) referring to definite objects in her world using various means, including gesture, language, or a multimodal mixture ("ensemble"). The dataset consists of videos of the avatar using various techniques to refer to a given object in a given configuration in her virtual world, along with associated parameters used in the generation of each video. The details of the dataset generation process are given in Krishnaswamy and Pustejovsky (2019a).[1]. In brief, annotators (eight per each video) were presented with a scene depicting objects on a table, given a target object, and then asked to rank each of the depicted methods with which the avatar in the scene then referred to the target object (see Fig. 1). The avatar used one of the three available modal options (gesture, language, or ensemble), with variants in the language used, to distinguish the target object with regard to its distinct properties or relations to other objects in the scene. Annotators were asked to rank, on a scale of 1 (least) to 5 (most), the "naturalness" of the referring techniques presented relative to the indicated target object, as the goal was to gather data with which to generate multimodal referring expressions in real-time that are appropriate, salient, and natural in context. Annotation results, links to video, and parameters of each scene depicted are stored in a SQL database. Stored parameters include some specific to the target object, such as its identity or the distance from it to the simulated agent; some specific to the referring expression, such as modality, utterance used, and relational descriptors in the utterance; and some global to the scene, such as object raw coordinates or total relation set present in the simulation.

[1]The dataset itself is available at https://github.com/VoxML/public-data/tree/master/EMRE

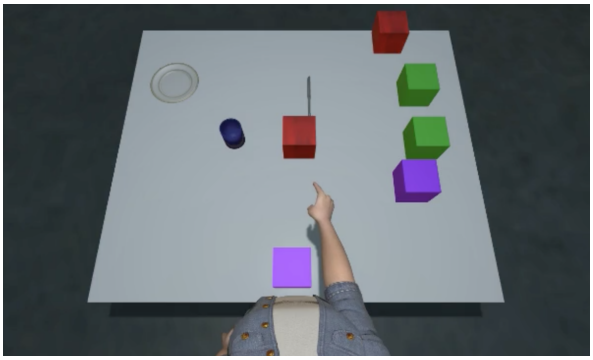
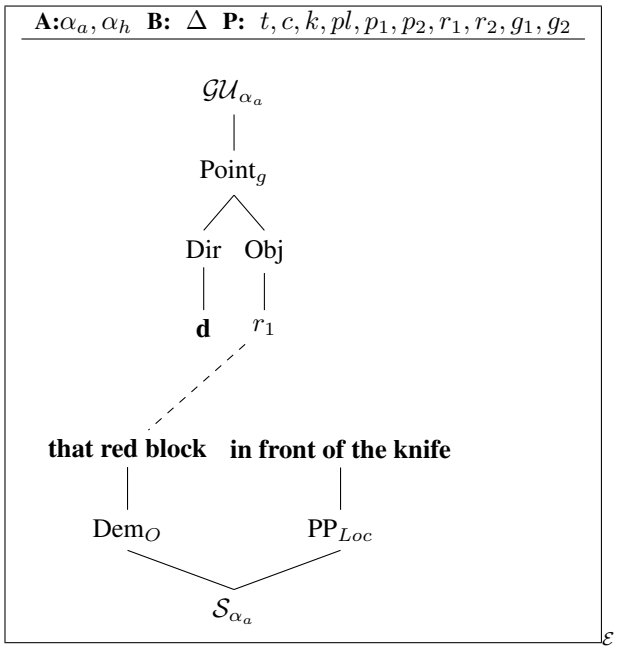


Figure 1: Frame from sample EMRE video, with accompanying utterance "that red block in front of the knife."

4. Computational Common Ground

The theory of common ground has a rich and diverse literature concerning what is shared or presupposed in human communication (Clark and Brennan, 1991; Stalnaker, 2002; Asher, 1998; Tomasello and Carpenter, 2007). Adopting and extending the model in Pustejovsky (2018), given a context of a co-situated interaction, the common ground is a state monad, the components of which are: **A**, the agents engaged in communication; **B**, the shared belief space; **P**, the objects and relations jointly perceived in the environment; and $\mathcal{E}$, the embedding space occupied by the agents. In the scenario under analysis here, we can specify **A** as $\{\alpha_a$ (the avatar), α_h (the human observer)$\}$, **B** as $\subseteq \{$*beliefs about the existence, affordances, and relative placement of objects, and the interlocutor's knowledge thereof*$\}$ (the set is denoted as Δ); and **P** as $\subseteq$ $\{$TABLE, CUP, Knife, PLATE, PURPLEBLOCK1, PURPLEBLOCK2, REDBLOCK1, REDBLOCK2, GREENBLOCK1, GREENBLOCK2, *locations within* $\mathcal{E}\}$. Each element may be introduced into the common ground at any time, such as at t_0 or subsequently based on an action taken by one of the agents. For instance, an agent might introduce a new object into the scene, making common the knowledge of its existence. Or (as happens in the EMRE dataset), one agent may use certain terms in a definite description, making public their knowledge of the meaning of those terms.

Given the common ground, a communicative act C_α, performed by agent, α, is a tuple of expressions from the modalities available to α, involved in conveying information to another agent. Here, we restrict this to the modali-



$\lambda k_s \otimes k_g(\textbf{that}(x)[\text{block}(x) \wedge \text{red}(x) \wedge \text{in\_front}(x, k, v)] \wedge k_s \otimes k_g(x)]$, where $v = \alpha_a$

Figure 2: Common-ground structure for "that red block in front of the knife" (cf. Fig. 1). The semantics of the RE includes a *continuation* (in the abstract representation sense in computer science, cf. Van Eijck and Unger (2010)) for each modality, k_s and k_g, which will apply over the object in subsequent moves in the dialogue.

ties of a linguistic utterance, $\mathcal{S}$, and a gesture, $\mathcal{G}$. Thus there are three possible configurations in performing C:

1. $C_\alpha = (\mathcal{G})$
2. $C_\alpha = (\mathcal{S})$
3. $C_\alpha = (\mathcal{S}, \mathcal{G})$

In the case of co-gestural speech, $(\mathcal{S}, \mathcal{G})$, we assume an aligned language-gesture syntactic structure, for which we provide a continuized semantic interpretation. Both of these are contained in the common ground state monad (see Fig. 2).

In co-gestural speech, the modal channels can be *aligned* or *unaligned*. Each input updates the common ground and each update to the common ground may change the probability of a subsequent communicative act being more or less salient, based on the content that *it* introduces into the common ground. Thus we propose that the formal characteristics of common ground updates serve as predictors of the naturalness of a referring expression, based on the saliency of the content communicated through the update.

Common ground updates execute modal operations over the belief space **B** such that each element of Δ is introduced via a *public announcement logic* (PAL) formula or an analogous formula denoting what the agents see or perceive (Plaza, 2007; Van Ditmarsch et al., 2007; Van Benthem, 2011). To avoid confusion between the two, we use the standard syntax of Plaza's public announcement logic with the following exceptions: we will use $\mathcal{K}_\alpha\varphi$ to denote "α *knows* φ", $\mathcal{L}_\alpha\varphi$ to denote "α *believes* φ", and $\mathcal{P}_\alpha\varphi$ to denote "α *perceives* φ". These are employed in place of a generic doxastic/epistemic update $[\alpha]\varphi$ ("agent α knows/believes ϕ"), so that an utterance like "You see it," as in Fig. 3, serves to express the update $[\mathcal{K}_{\alpha_h}\mathcal{P}_{\alpha_a}b!]\mathcal{K}_{\alpha_h}\mathcal{P}_{\alpha_a}b$, glossed as "$\alpha_h$ publicly announces (indicated by the bang, !) that α_h knows α_a perceives b."

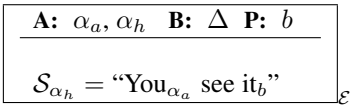


Figure 3: Common-ground structure for "You see it."

The types of update that we will examine as pertaining to referring expressions (REs) in the EMRE dataset are:

1. At t_0, the beginning of each video, the scene is presented. All objects displayed populate **P**, the elements of the jointly-perceived environment. $\forall b\ (b \in \mathbf{P} \rightarrow \mathcal{K}_{\alpha_h}\mathcal{P}_{\alpha_a}b \wedge \mathcal{K}_{\alpha_a}\mathcal{P}_{\alpha_h})$. This is shown in Fig. 4 (L). This is derived by performing transitive closure of perception over the agents who are co-situated in the perceptual environment: $[(\mathcal{P}_{\alpha_h} \cup \mathcal{P}_{\alpha_a})^*]\phi$.
2. At t_1, a circle is drawn around one particular object (b), raising it to the status of target object. The human observer α_h now knows that b is the target (but does not necessarily know that the avatar α_a knows this as well). $\mathcal{K}_{\alpha_h} target(b) \wedge \neg\Box\mathcal{K}_{\alpha_h}\mathcal{K}_{\alpha_a} target(b)$. This is shown in Fig. 4 (R).
3. At t_2 (shown above in Fig. 1):
   (a) The avatar points to b. This demonstrates the avatar can point, and knows that b is the target. $[C_{\alpha_a} = Point_g \rightarrow Dir\ b!]\mathcal{K}_{\alpha_h}\mathcal{K}_{\alpha_a}(Point_g \wedge target(b))$.
   (b) The avatar describes b using some combination of b's attributes (here, color), and relations to other objects. This demonstrates that α_a knows the meaning of the terms she uses ($[\![u]\!]$ being the *interpretation* of some utterance u) under a model $\mathcal{M}$ and a common ground cg, and also situates some of those terms (e.g., spatial relations) relative to her frame of reference. $[C_{\alpha_a} = \mathcal{S}!]\forall u(u \in \mathcal{S} \rightarrow \mathcal{K}_{\alpha_h}\mathcal{K}_{\alpha_a}[\![u]\!]_{\mathcal{M},cg})$.

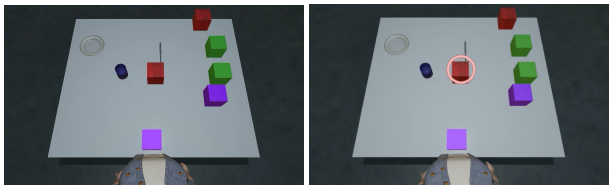


Figure 4: Additional frames accompanying common ground updates.

Time intervals in the video data are all constant, allowing us to maintain consistent timesteps in analysis: the initial presentation (t_0) is shown for 1.5 seconds, the target circle is drawn and held for 1.5 seconds (t_1), and .5 seconds later, at t_2, the agent indicates the target object, through gesture, language, or both.

5. Analysis Methodology

The EMRE dataset distribution contains an analysis script to evaluate the probability of a given annotator judgment over arbitrary sets of parameters in the scene. Parameters are presented in the form of SQL conditions to filter the results over, allowing the extraction of conditional probabilities over any parameters recoverable from the data using standard SQL syntax. Krishnaswamy and Pustejovsky (2019a) presents a basic statistical analysis of annotator judgments over parameters directly stored in the data. A full description of all parameters examined is contained therein. These data showed a clear preference for referring expressions using the gesture+language "ensemble" modality, and preference for longer descriptive strings, providing coarse-grained parameters over which to train a deployable multimodal referring expression generation model. However, we believe that examining the formal properties of the referring expressions shown in the data provides further discriminative features for better generation, as indicated by optimized saliency, naturalness, and informativity of the generated expression to human interlocutors. We extract formal and propositional values as features from the data based on the information each feature introduces into the common ground. If it is inferable from the content denoted by either $\mathcal{G}$ or $\mathcal{S}$ modality in the referring expression that an agent α either knows or perceives some propositional content p relative to the belief space **B** or the jointly perceived entities in **P**, this prompts an update to the common ground, and therefore new features for possible examination. Some examples include the following:

- Through use of a spatial term T, α introduces that interpretation of spatial term T into the common ground: $C_\alpha = (\mathcal{S} \mid T_{sp} \in \mathcal{S}) \rightarrow \mathcal{K}_\alpha[\![T_{sp}]\!]_\mathcal{M}$
- Through use of an attributive term T, α introduces that interpretation of attributive term T into the common ground: $C_\alpha = (\mathcal{S} \mid T_{att} \in \mathcal{S}) \rightarrow \mathcal{K}_\alpha[\![T_{att}]\!]_\mathcal{M}$
- By referencing b in a description string, α introduces that she perceives b into the common ground (this is different from b itself, which is already in **P**, the set of jointly-perceived objects; this explicitly encodes the *knowledge* that $\mathcal{P}_\alpha b$): $C_\alpha = (\mathcal{S} \mid b_s \in \mathcal{S}) \rightarrow \mathcal{P}_\alpha b$
- By differentiating two similarly colored objects (e.g., by use of "other"), α introduces that she knows the same attribute predicates over b_1 and b_2 but that b_1 and b_2 are distinct: $C_\alpha = (\mathcal{S} \mid [\text{“}other\text{”}, b_{1s}, b_{2s}] \in \mathcal{S} \wedge b_{1s} = b_{2s}) \rightarrow \mathcal{K}_\alpha[\![Att(b_1 \wedge b_2)]\!]_\mathcal{M} \wedge \mathcal{K}_\alpha b_1 \neq b_2$
- By distinguishing demonstratives in an ensemble RE, α introduces that she is meaningfully distinguishing between "near" and "far" regions of sfc, the table surface: $C_\alpha = (\mathcal{S}, \mathcal{G} \mid \mathcal{G} = Point_g \wedge \text{“}this\text{”} \in \mathcal{S}) \rightarrow \mathcal{K}_\alpha[\![near(sfc)]\!] \neq [\![far(sfc)]\!]_\mathcal{M}$

The visualizations in the EMRE dataset are produced using the VoxSim event/agent simulation platform (Krishnaswamy and Pustejovsky, 2016a; Krishnaswamy and Pustejovsky, 2016b), which employs the VoxML modeling language, enabling object and event visualization semantics (Pustejovsky and Krishnaswamy, 2016). Because a simulator is an extension of a model checker (Pustejovsky and Krishnaswamy, 2014), a simulation can be evaluated formally. Because a simulator requires numerical parameter values to run (Davis and Marcus, 2016), it can be evaluated quantitatively. Values extracted from the simulator and collated in the dataset may be either real numbers or vector values (e.g., distance values or coordinates) or symbolic (e.g., object labels or qualitative attributes). Thus, we can conduct ablation tests on the effects of formal, symbolic, and quantitative features on the predictive model trained over data extracted from a simulation.

Each of these features and others can be extracted from a common ground structure of the kind shown in Figs. 2 or 3. In addition, they can be linked with each other and other features by virtue of the linkages established in the common ground structure (e.g., the link between $Dem_O \rightarrow$ **that red block** and $Point_g \rightarrow Obj \rightarrow r_1$ in Fig. 2), effectively allowing us to search the data for sets of related parameters that predict given annotator ratings of a referring technique, formally reanalyze them in terms of computational common ground, and use segments of common ground structures as input features into a prediction algorithm.

5.1. Model Architecture

Here, the extracted data and quantitative features are those described in Section 3. The formal common ground features are those as described above. What we are trying to predict, then, is the likelihood of an annotator evaluating a referring expression at a given naturalness (1-5), with the expectation that the best or most natural REs will come with a consistent set of features that predict a high score.

We feed all features into a multilayer perceptron (MLP) written in Keras with the TensorFlow backend. Our reasons for choosing this type of architecture is its relative simplicity, and therefore training speed, but also ability to distinguish dependencies between points in linearly-inseparable regions of data (Cybenko, 1989). Our architecture consists of three fully-connected hidden layers of 32, 128, and 64, respectively, prior to a $softmax$ output layer. The layers use $tanh$, ELU, and $tanh$ activation, respectively. The model uses categorical cross-entropy loss and Adam optimization, and is trained for 1000 epochs with a batch size of 50. Due to the relatively small size of the sample data, we validate all results using 7-fold cross-validation in order to achieve a more balanced sample across all classes of annotator judgments. $k = 7$ is chosen here to approximate a leave-one-out cross-validation approach over the 8 annotator judgments on each visualized referring expression. Because in the EMRE dataset, 8 separate annotators evaluated each RE, the "most likely" annotator judgment is in fact a probability distribution. Therefore, we regard a "correct" prediction by the classifier not as one that returns the exact integer value representing the argmax of all annotator judgment counts, but one that falls within the correct quintile of the distribution over all annotator judgments of that visualized referring expression.

6. Results and Evaluation

6.1. Baseline

The EMRE dataset already contains quantitative and some qualitative features about scenes, referring expressions generated within them, and annotator judgments thereof. As a baseline, we used the raw features used in the generation of videos in the EMRE dataset to try and predict the most likely annotator rating of that video. The baseline features include: 1) the target object; 2) the referring modality: one of gesture, language, or ensemble; 3) the distance from the object to the agent; 4) whether the linguistic description uses a near/far distance distinction; 5) whether that distinction is relative to embedding space of similar objects or the entire world (*n/a* if no distance distinction is used. We also add in 6) the linguistic description used and 7) the individual relational descriptors used, which are represented as 200-dimensional sentence vectors trained using a Skip-Gram model over the entire vocabulary that occurs in the dataset. In gesture-only referring expressions, where all the avatar does is point to the target object, these are vectors of all 0s.

The top half of Table 1 (see Section 6.2.) shows baseline results. The raw features extracted from the EMRE dataset are successful in predicting the correct quintile of annotator judgment of the associated multimodal referring strategy approximately ⅔ of the time. Interestingly, the addition of sentence embeddings caused the average accuracy to drop about 3.28%, suggesting that sentence embeddings caused some confusion in the classifier. Discussion of these results follows in Section 7.

6.2. Formal Features

To keep track of formal features, such as those described in Section 5., we maintain two lists of the propositional content within the common ground structure available to each agent (i.e., what each agent—here the avatar and the annotator—knows and perceives about the scene and about each other). Each element in these lists is correlated with features extracted from the EMRE dataset using the provided analysis script (refer to Section 5.) and the value is inserted into a data structure representing a common ground structure of the form shown in Fig. 2: consisting of a gesture, the speech string, and links between the constituents of each.

The encoding of formal features is done by creating one-hot vectors representing the state of the belief space **B** (Δ) as it pertains to the agents **A** and jointly perceived content **P**. That is, propositional content that is formally denoted as C_α = ($\mathcal{S}$, $\mathcal{G}$ | $\mathcal{G} = Point_g \wedge$ "$this$" $\in \mathcal{S}$) $\rightarrow$ $\mathcal{K}_\alpha[\![near(sfc)]\!] \neq [\![far(sfc)]\!]_\mathcal{M}$, as above, is treated as a one-hot vector for α's knowledge of the distance distinction of $near(sfc)$ and $far(sfc)$ in Δ, whereas content formally denoted in the form C_α = ($\mathcal{S}$ | ["$other$", b_{1s}, b_{2s}] $\in$ $\mathcal{S} \wedge b_{1s} = b_{2s}$) $\rightarrow$ $\mathcal{K}_\alpha[\![Att(b_1 \wedge b_2)]\!]_\mathcal{M} \wedge \mathcal{K}_\alpha b_1 \neq b_2$ is treated as *three* one-hot vectors, one for $\mathcal{K}_\alpha[\![Att(b_1)]\!]_\mathcal{M}$, another for $\mathcal{K}_\alpha[\![Att(b_2)]\!]_\mathcal{M}$, and a third for $\mathcal{K}_\alpha b_1 \neq b_2$.

MLP prediction results of annotator judgment using formal features are shown below. However, if performance increases when the formal features are added, it could be due to the fact that since they are (under our hypothesis) dependent features, and they reinforce each other, giving stronger prediction results. Therefore, to demonstrate the effect of formal features, we present ablative results using raw features with formally-derived features, raw features with formally-derived features including sentence embeddings, and formally-derived features only.

We present, as before, the mean and standard deviation of classification accuracy over a 7-fold cross-validated sample.

	Raw features	**Raw feat. + SE**
μ **Acc. (1K)**	0.6757	0.6429
σ **Acc. (1K)**	0.0230	0.0111

	Raw + form.	**Raw + form. + SE**	**Formal only**
μ **Acc. (1K)**	0.7214	0.6671	0.7471
σ **Acc. (1K)**	0.0398	0.0243	0.0269

Table 1: Classification accuracy after 1000 epochs using formal features (mean and standard deviation over 7-fold cross-validated sample)

Features that correlate formally with elements of the common ground structure equivalent to the referring strategy depicted do between 7-11% better at predicting the category label (annotator judgment) on the referring strategy than raw features alone, or raw features augmented with sentence embeddings.

The above data shows that formal features derived from the common ground structure provide a modest but appreciable improvement in the quality of predicting *how well* a referring strategy is likely to be perceived as natural, based on the content it encodes, but tells us little about *what* propositional content is likely to produce a natural, salient multimodal referring expression, and *how* it should be assembled, which is an important question for generating multimodal REs.

Lascarides and Stone's formal semantics of gesture (Lascarides and Stone, 2009) separates gestural and speech assignment functions in order to distinguish entities that can satisfy interpretations of referents in speech from entities used to ground references in gesture. It should be pointed out that we are focusing on *co-gestural speech* ensembles rather than *co-speech gesture* (Schlenker, 2018). Further, since here we focus only on *deictic* gesture rather than depicting gestures, we do not evaluate gesture that conflicts semantically with the speech, but we can draw some inferences analogically regarding the information provided by each.

1. If information provided by gesture is constant between referring expressions for the same object, then the "best" ensemble RE should be that which maximizes the score of its linguistic component if taken alone.

2. If information provided by gesture is *not* constant between referring expressions for the same object, the "best" ensemble RE should be that which maximizes the information gain provided by each of the individual modalities.

If (1) is true, then we should expect that features dependent only on the language, including sentence embeddings but not including things like the distance from the agent to the target object, should predict the quality of linguistic-only referring expressions better than the full set of features, including parameters dependent on the gesture and embodiment of the agent, predict the quality of ensemble referring expressions. If (2) is true, this should not be the case. Given that the gestural component of a well-formed ensemble referring expression is always deixis, we hypothesize that the gestural information is constant across referring expressions. To test this, we run different subsets of the total raw and formal feature set (depending on which modality each feature explicitly depends on) through the classifier, over either only the linguistic-only referring expressions from the EMRE dataset, or over only the ensemble referring expressions. Classifier results are given below.

	Raw features	**Raw feat. + SE**
μ **Acc. (1K)**	0.7471	0.6329
σ **Acc. (1K)**	0.0468	0.0577

	Raw + form.	**Raw + form. + SE**	**Formal only**
μ **Acc. (1K)**	0.7471	0.6443	0.7985
σ **Acc. (1K)**	0.0213	0.0469	0.0405

Table 2: Classification accuracy after 1000 epochs using formal features and linguistically-dependent features only, over purely linguistic EMRE referring expressions (mean and standard deviation over 7-fold cross-validated sample)

	Raw features	**Raw feat. + SE**
μ **Acc. (1K)**	0.6014	0.5842
σ **Acc. (1K)**	0.0537	0.0281

	Raw + form.	**Raw + form. + SE**	**Formal only**
μ **Acc. (1K)**	0.6014	0.5842	0.6171
σ **Acc. (1K)**	0.0302	0.0840	0.0550

Table 3: Classification accuracy after 1000 epochs using formal features, over only ensemble (multimodal) EMRE referring expressions (mean and standard deviation over 7-fold cross-validated sample)

Tables 2 and 3 demonstrate that not only do linguistically-dependent features predict the quality of language-only referring expressions better than *all* features predict ensemble referring expressions, meaning that the level of information provided by solely deictic gesture is likely to be of roughly constant relevance across the dataset (i.e., directly grounding to a location and object(s) in that location), but that the addition of formal features provide a larger net increase in classifier accuracy over the raw feature baseline for the language only REs than they do for the ensemble REs.

For linguistic REs, adding formal features to raw features plus sentence embeddings improved accuracy by only about 1%, but *only* using formally-derived features improved accuracy by approximately 5-16%, depending on if the baseline compared includes sentence embeddings or not. For ensemble REs, the addition of formal features made no difference in the average classification accuracy compared to simply raw features (with or without sentence embeddings) and formal features alone resulted in a small ($\sim$1%) average improvement over the baseline. From these results we can see that since the referring modality is already a strong predictor of referring expression naturalness and salience according to the dataset (see Fig. 5, taken from (Krishnaswamy and Pustejovsky, 2019a)), most of the improvement in the multimodal referring expressions compared to low-rated gesture-only referring expressions, where the avatar just wordlessly points to the target object, comes from the information gain associated with an appropriately informative linguistic utterance accompanying the co-speech gesture, which is an important consideration to take into account when generating quality referring expressions, particularly multimodally.

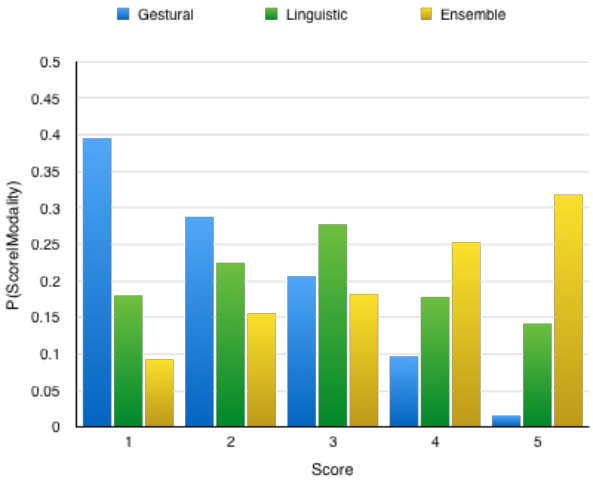


Figure 5: Probability of annotator judgment score given referring expression modality (taken from Krishnaswamy and Pustejovksy (2019a))

7. Discussion and Conclusions

Our results demonstrate an appreciable increase in the ability of a formal feature set derived from a common ground structure to predict the naturalness and salient quality of a referring expression associated with that common ground structure. We hypothesize that this is because formal features make the model explainable on a finer-grained level, and that the propositional content extractable from the linguistic utterances used correlates more closely with the quality of the overall referring expressions than less-symbolically defined features like sentence embeddings.

Basic features provide a solid baseline upon which to improve RE classification accuracy (Zhang et al., 2016). Here, however, using sentence embeddings actually seemed to hurt the accuracy. In the data, the purely linguistic "the red block in front of the knife" is more likely to be rated as "average" while "that red block in front of the knife" is multimodal (accompanied by deictic gesture) and more likely to receive a high rating (see Fig. 5). However, the sentence embeddings for these sentences are very similar, due to an alternation of two words ("the"/"that") that already tend to be similar in distributional semantic space.

"The" vs. "that" captures little of the distinction introduced by the ensemble in the common ground. For this data and task, therefore, this suggests that either simple sentence embeddings are not a very useful feature or should should be trained in a different way, other than a Skip-Gram model.
Meanwhile the formally-defined features extracted from common-ground structures, are much more adept at distinguishing the types of salient information introduced by the referring agent into the common ground and by encoding what type of knowledge is introduced or publicly perceived in the common ground, we are able to quite effectively predict how our annotators would judge the depicted referring expression.
Using the formal features alone usually performs best at this task, likely since the common ground structure is designed specifically to capture the type of information we seek to disambiguate in a multimodal referring expression classification task, compared to raw features that describe either the physical environment or vague contours of the priors that go into the referring expression generation procedure in the EMRE dataset. Thus we propose that formal common ground structures would be an effective medium through which to interpret and generate multimodal referring expressions and other types of multimodal communicative acts in a co-situated interaction.

8. Future Work

The composition of gesture and speech plays an important role in multimodal communication. The two modalities display complementary strengths at communicating different types of information—it is hard to communicate certain types of spatial configurations solely through language, and deictic gesture may prove more economical; conversely, attributives like color are much more aptly communicated through language. It is through the combination of the two that successful referring expressions can be generated in co-situated space, and by digging into the data we previously gathered, we have found evidence that while the addition of gesture provides a boost in naturalness and salient quality of a referring expression in co-situated space, the best and most natural REs are those that maximize the salience and naturalness of their linguistic components, even if the linguistic information overlaps with the gestural information (cf. Paraboni et al. (2007)).
As such, given that we have trained a prediction model to expose these considerations, the next step is to train a generation model that can be deployed "live" in a multimodal interaction where the situation encountered at any given time may not cleanly map to a situation from the EMRE dataset. The process of maximizing the contextually-salient information content provided by the linguistic component of the multimodal referring expression, and by extension by all modalities including iconic gesture and action, could be handled by a composing and constructing expressions with a probabilistic grammar.
Existing work in multimodal grammars, particularly on gesture and speech (cf. Alahverdzhieva et al. (2012), Alahverdzhieva et al. (2017)) often focuses on timing and aligning the gesture and speech components using edge-based constraints to generated a syntax tree of both speech and gesture. To this we would propose the addition of a continuation-based semantics (Krishnaswamy and Pustejovsky, 2019b) to capture additional content from common ground structures, such as the formally-derived features that we have shown here can be stronger predictors of gesture-speech ensemble quality, particularly in the domain of referring expressions.
Given demonstrated success in the prediction of multimodal referring expression quality, for which formally-derived features are an asset, we propose to use similar formal analysis methods using common ground structures as the medium within which to both recognize and generate multimodal referring expressions by maximizing the information content provided by each applicable modality.
As common ground structures provide a formal and explainable way of segmenting multimodal content and the information specified by each modal channel, we are also exploring other tasks in which common ground structures may be useful representations. Some examples include:

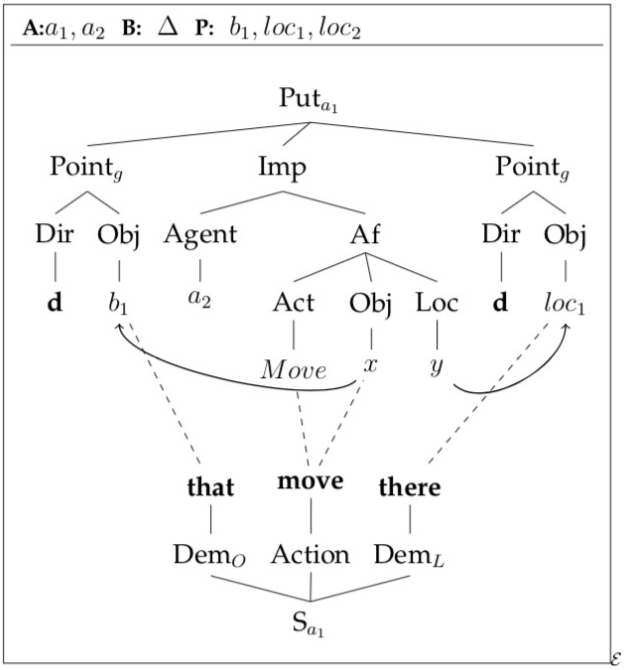


Figure 6: Action command using gesture-language ensemble

- *Multimodal dialogue parsing*. Given a situation where both gestures and natural language can indicate both objects and actions or events, common ground structures should be helpful in extracting both object and action information separately from each modality and in disambiguating the information provided by one modality with information from the other (see Fig. 6).

- *Scene classification*. By exploiting the relation sets between objects that populate the belief space, common ground structures can cluster and classify novel scenes and configurations with known examples, providing a way to transfer dialogue or referring strategies from a known situation to a novel one (see Fig. 7).

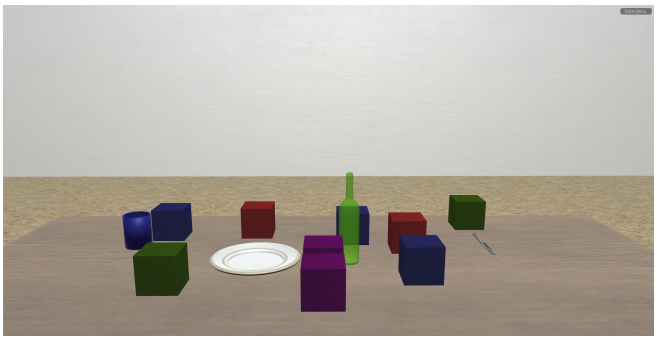


Figure 7: Sample novel situation

- *Intelligent modality switching*. There may be cases when an agent cannot use one modality or another—e.g., hands are full, prohibiting gesture, or the environment is loud, prohibiting language (cf. Kim et al. (2016), Drijvers et al. (2018))—in this case common ground structures can be deployed to maximize the information content in the remaining available modalities for optimal communication in sub-optimal circumstances.

Acknowledgments

We would like to thank the reviewers for their helpful comments. This work was supported by the US Defense Advanced Research Projects Agency (DARPA) and the Army Research Office (ARO) under contract W911NF-15-C-0238 at Brandeis University. The points of view expressed herein are solely those of the authors and do not represent the views of the Department of Defense or the United States Government. Any errors or omissions are, of course, the responsibility of the authors.

9. Bibliographical References

Alahverdzhieva, K., Flickinger, D., and Lascarides, A. (2012). Multimodal grammar implementation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 582–586, Montréal, Canada, June. Association for Computational Linguistics.

Alahverdzhieva, K., Lascarides, A., and Flickinger, D. (2017). Aligning speech and co-speech gesture in a constraint-based grammar. *Journal of Language Modelling*, 5.

Asher, N. (1998). Common ground, corrections and coordination. *Journal of Semantics*.

Barrault, L., Bougares, F., Specia, L., Lala, C., Elliott, D., and Frank, S. (2018). Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323, Belgium, Brussels, October. Association for Computational Linguistics.

Beinborn, L., Botschen, T., and Gurevych, I. (2018). Multimodal grounding for language processing. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2325–2339, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Bortfeld, H. and Brennan, S. E. (1997). Use and acquisition of idiomatic expressions in referring by native and non-native speakers. *Discourse Processes*, 23(2):119–147.

Claassen, W. (1992). Generating referring expressions in a multimodal environment. In *Aspects of automated natural language generation*, pages 247–262. Springer.

Clark, H. H. and Brennan, S. E. (1991). Grounding in communication. In Lauren Resnick, et al., editors, *Perspectives on Socially Shared Cognition*, pages 13–1991. American Psychological Association.

Cohen, V., Burchfiel, B., Nguyen, T., Gopalan, N., Tellex, S., and Konidaris, G. (2019). Grounding language attributes to objects using bayesian eigenobjects. *arXiv preprint arXiv:1905.13153*.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314.

Dale, R. and Reiter, E. (1995). Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive science*, 19(2):233–263.

Davis, E. and Marcus, G. (2016). The scope and limits of simulation in automated reasoning. *Artificial Intelligence*, 233:60–72.

Drijvers, L., Özyürek, A., and Jensen, O. (2018). Hearing and seeing meaning in noise: Alpha, beta, and gamma oscillations predict gestural enhancement of degraded speech comprehension. *Human brain mapping*, 39(5):2075–2087.

Ferreira, T. C., Moussallem, D., Kádár, Á., Wubben, S., and Krahmer, E. (2018). Neuralreg: An end-to-end approach to referring expression generation. *arXiv preprint arXiv:1805.08093*.

Funakoshi, K., Watanabe, S., Kuriyama, N., and Tokunaga, T. (2004). Generating referring expressions using perceptual groups. In *International Conference on Natural Language Generation*, pages 51–60. Springer.

Goyal, R., Kahou, S. E., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al. (2017). The" something something" video database for learning and evaluating visual common sense. In *ICCV*, volume 1, page 3.

Grice, H. P., Cole, P., Morgan, J., et al. (1975). Logic and conversation. *1975*, pages 41–58.

Johnston, M., Cohen, P. R., McGee, D., Oviatt, S. L., Pittman, J. A., and Smith, I. (1997). Unification-based multimodal integration. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 281–288. Association for Computational Linguistics.

Kasenberg, D., Roque, A., Thielstrom, R., Chita-Tegmark, M., and Scheutz, M. (2019). Generating justifications for norm-related agent decisions. In *12th International Conference on Natural Language Generation*.

Kim, J.-H., Jo, S., and Lattimer, B. Y. (2016). Feature selection for intelligent firefighting robot classification

of fire, smoke, and thermal reflections using thermal infrared images. *Journal of Sensors*, 2016.

Krahmer, E. and Van Deemter, K. (2012). Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

Krahmer, E. and van der Sluis, I. (2003). A new model for generating multimodal referring expressions. In *Proceedings of the 9th European Workshop on Natural Language Generation (ENLG-2003) at EACL 2003*.

Krishnaswamy, N. and Pustejovsky, J. (2016a). Multimodal semantic simulations of linguistically underspecified motion events. In *Spatial Cognition X: International Conference on Spatial Cognition*. Springer.

Krishnaswamy, N. and Pustejovsky, J. (2016b). VoxSim: A visual platform for modeling motion language. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. ACL.

Krishnaswamy, N. and Pustejovsky, J. (2019a). Generating a novel dataset of multimodal referring expressions. In *Proceedings of the 13th International Conference on Computational Semantics-Short Papers*, pages 44–51.

Krishnaswamy, N. and Pustejovsky, J. (2019b). Multimodal continuation-style architectures for human-robot interaction. *arXiv preprint arXiv:1909.08161*.

Lascarides, A. and Stone, M. (2009). A formal semantic analysis of gesture. *Journal of Semantics*, page ffp004.

Magassouba, A., Sugiura, K., and Kawai, H. (2019). Multimodal attention branch network for perspective-free sentence generation. *arXiv preprint arXiv:1909.05664*.

Matuszek, C., Bo, L., Zettlemoyer, L., and Fox, D. (2014). Learning from unscripted deictic gesture and language for human-robot interactions. In *AAAI*, pages 2556–2563.

McNeill, D. (2000). *Language and gesture*, volume 2. Cambridge University Press.

Paraboni, I., Van Deemter, K., and Masthoff, J. (2007). Generating referring expressions: Making referents easy to identify. *Computational linguistics*, 33(2):229–254.

Petit, M., Lallée, S., Boucher, J.-D., Pointeau, G., Cheminade, P., Ognibene, D., Chinellato, E., Pattacini, U., Gori, I., Martinez-Hernandez, U., et al. (2012). The coordinating role of language in real-time multimodal learning of cooperative tasks. *IEEE Transactions on Autonomous Mental Development*, 5(1):3–17.

Plaza, J. (2007). Logics of public communications. *Synthese*, 158(2):165–179.

Pustejovsky, J. and Krishnaswamy, N. (2014). Generating simulations of motion events from verbal descriptions. *Lexical and Computational Semantics (* SEM 2014)*, page 99.

Pustejovsky, J. and Krishnaswamy, N. (2016). VoxML: A visualization modeling language. *Proceedings of LREC*.

Pustejovsky, J., Krishnaswamy, N., Draper, B., Narayana, P., and Bangar, R. (2017). Creating common ground through multimodal simulations. In *Proceedings of the IWCS workshop on Foundations of Situated and Multimodal Communication*.

Pustejovsky, J. (2018). From actions to events. *Interaction Studies*, 19(1-2):289–317.

Quek, F., McNeill, D., Bryll, R., Duncan, S., Ma, X.-F., Kirbas, C., McCullough, K. E., and Ansari, R. (2002). Multimodal human discourse: gesture and speech. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 9(3):171–193.

Schlenker, P. (2018). Gesture projection and cosuppositions. *Linguistics and Philosophy*, 41(3):295–365.

Shridhar, M. and Hsu, D. (2018). Interactive visual grounding of referring expressions for human-robot interaction. *arXiv preprint arXiv:1806.03831*.

Stalnaker, R. (2002). Common ground. *Linguistics and philosophy*, 25(5-6):701–721.

Tomasello, M. and Carpenter, M. (2007). Shared intentionality. *Developmental science*, 10(1):121–125.

Van Benthem, J. (2011). *Logical dynamics of information and interaction*. Cambridge University Press.

Van Der Sluis, I. and Krahmer, E. (2001). Generating referring expressions in a multimodal context: An empirically oriented approach. In *Computational Linguistics in the Netherlands 2000*, pages 158–176. Brill Rodopi.

Van Ditmarsch, H., van Der Hoek, W., and Kooi, B. (2007). *Dynamic epistemic logic*, volume 337. Springer Science & Business Media.

Van Eijck, J. and Unger, C. (2010). *Computational semantics with functional programming*. Cambridge University Press.

Viethen, J. and Dale, R. (2008). The use of spatial relations in referring expression generation. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 59–67. Association for Computational Linguistics.

Wallbridge, C. D., Lemaignan, S., Senft, E., and Belpaeme, T. (2019). Generating spatial referring expressions in a social robot: Dynamic vs non-ambiguous. *Frontiers in Robotics and AI*, 6:67.

Whitney, D., Eldon, M., Oberlin, J., and Tellex, S. (2016). Interpreting multimodal referring expressions in real time. In *Robotics and Automation (ICRA), 2016 IEEE International Conference on*, pages 3331–3338. IEEE.

Yatskar, M., Zettlemoyer, L., and Farhadi, A. (2016). Situation recognition: Visual semantic role labeling for image understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5534–5542.

Zhang, X., Pacheco, M. L., Li, C., and Goldwasser, D. (2016). Introducing drail–a step towards declarative deep relational learning. In *Proceedings of the Workshop on Structured Prediction for NLP*, pages 54–62.

Zhou, M., Cheng, R., Lee, Y. J., and Yu, Z. (2018). A visual attention grounding neural model for multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3643–3653, Brussels, Belgium, October-November. Association for Computational Linguistics.