Annotation-Based Semantics for Dialogues in the Vox World

Kiyong Lee
Korea University, Seoul
ikiyong@gmail.com

Abstract

This paper aims at enriching Annotation-Based Semantics (ABS) with the notion of small visual worlds, called the *Vox worlds*, to interpret dialogues in natural language. It attempts to implement classical set-theoretic models with these Vox worlds that serve as interpretation models. These worlds describe dialogue situations while providing background for the visualization of those situations in which these described dialogues take place interactively among dialogue participants, often triggering actions and emotions. The enriched ABS is linked to VoxML, a modeling language for visual object conceptual structures (vocs or vox) that constitute the conceptual basis of visual worlds. Each Vox world is characterized by a set of visualized situation types, possibly depicted by static pictures or dynamic videos, to interpret dialogues. This paper focuses on annotating and interpreting a few illustrative dialogues for such a small visual world.

Keywords: annotation-based semantics (ABS), partial information, situation types, small visual world, visual object concept structure (vocs, or vox), Vox world,

1. Introduction

1.1. Aim and Overview

This paper aims to enrich Annotation-Based Semantics (ABS), proposed by Lee (2020, 2023), with the notion of *small visual worlds* to annotate and interpret dialogues in natural language. Small visual worlds form the Vox world, consisting of visual conceptual object structures (vocs or vox) in the modeling language VoxML (Pustejovsky and Krishnaswamy, 2014). These small visual worlds may be forming *scenes* or "*visually perceived situations*" (Barwise, 1989) with formal constructions involving human perceptions of the surroundings in interactive communications or dialogues.

ABS makes two but related uses of a set of small visual worlds. One use is to describe a dialogue situation in which the dialogue participants interact with each other linguistically through verbal exchanges. The other use is to form a background situation à la Barwise and Perry (1983) or bring in the linguistic or world knowledge for interpreting communicative exchanges and the things involved in them. In annotating dialogues for their act types and content, ABS refers to these two *situation types*, one for describing situations and another for providing background for interpreting them.

For example, part of a dialogue transcript "Husband to Wife: Take this." describes a situation in which the husband says to his wife: "Take this." Suppose the wife responded with a smile to her husband by saying "Thanks. Delicious." Then, this response provides a contextual background for inferring that the deictic expression "this" must refer to something edible or potable for tasting while showing the wife's satisfaction with gratitude. Furthermore, a picture or scene showing how such a dialogue was enacted provides a background situation for interpreting more vividly what is meant by the husband's utterance "Take this." Such a picture depicts a small visual world or part of it.

1.2. Scope, Focus, and Motivation

The scope of the paper is very much restricted in its form for presentation and data for analysis. This is not a formal paper that formulates the key notions rigidly in logico-mathematical terms. It illustrates how a few short dialogues are annotated and interpreted for such a visually perceptible world, the Vox world or part of it. The data for analysis is also very restricted to the extent that no statistical justification is presented for the claims made in the paper.

The paper focuses on the complementary roles of dialogue scripts and related images or pictures that I claim depict a small world providing background for the interpretation. It treats very simple dialogues, having only a few words in the utterances, for illustrations while avoiding the treatment of various dialogue act types and dimensions (e.g., task-oriented vs. expressive (of emotions)) (Bunt, 2022).

Dialogues are chosen as specific data for analysis in this paper because they present the most challenging task for natural language processing in at least three respects. First, annotation may work while syntax fails to process because dialogues have many deictic expressions (e.g., "this" as in "Husband to Wife: Try this.") or ellipses (e.g., "Wife to Husband: Thanks. Delicious.") with syntactic variations and aberrations from regular grammar, unlike written text. Second, dialogue acts often trigger the actions of dialogue participants as agents or objects with some other semantic roles and emo-

ISA-20 Proceedings @LREC-COLING-2024, pages 133–143

tions (e.g.: emotive and evaluative as in "Wife to Husband: Thanks. Delicious.") lie involved in the content of the dialogue conveyed. Annotation can easily mark up such actions enacted and emotions expressed by dialogue participants. Third, the interpretation of dialogue contents requires background information, especially in the applicational context of Human-Computer Interactions (HCI) or Human-Object Interactions (HOI) (e.g.: "Husband to Wife: Try this." requires a variety of actions as responses, depending on its context of use). For these reasons, the treatment of dialogue acts and content is well-motivated, challenging, and most interesting as a linguistic task, especially for computational applications. Computers or robots may participate in a dialogue as artificial agents in a computational application.

1.3. Claim, Proposal, and Basic Assumptions

This paper claims that the set-theoretic model structures for interpreting natural language or its logical forms, as in Montague Semantics (Montague, 1974b; Dowty et al., 1981) should be re-envisioned and re-designed. This must be implemented with visualized small worlds or situation types delimited by the visual object conceptual structures that are well-defined, for instance, by the modeling language VoxML.

In VoxML, in contrast, each object, action, or relation is a first-class citizen in a small world, as proposed in Situation Semantics (Barwise, 1989), that forms a visual object conceptual structure. These structures are then represented by a complex attribute-value matrix (AVM) structure with embedded AVM's that carry a variety of relevant information. Likewise, various types of relations in an interpretation model are defined similarly.

Figure 1 shows how the Annotation-Based Semantics (ABS) is linked to VoxML, [i] linguistically supporting it. ABS annotates communicative language segments including dialogues, [ii] generating annotation structures **a** while referring to the voxicon V of VoxML. It then translates annotation structures to logical forms σ(**a**) in typed first-order logic [iii]. These logical forms are then interpreted with respect to the minimal models constrained by the habitats, affordances, and embodiments of denotative elements. For such processes, VoxML as a modeling language introduces Voxicon to list the voxemes that augment the Generative Lexicon (Pustejovsky and Batiukova, 2019) with the notions of Habitat theory (Pustejovsky, 2013) and Gibsonian affordance structures (Gibson, 1977, 1979). These voxemes are represented in complex feature structures, in which some features (attributes) have feature structures as values, as illustrated in

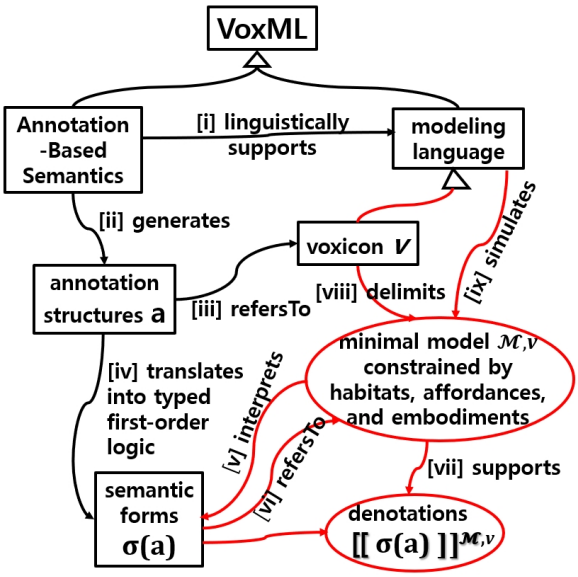

Figure 1: VoxML-linked ABS

Section 5.

This paper claims that the ABS thus designed is linked to VoxML, which constitutes the structural basis of visual worlds. Its sub-module, the *Voxicon*, which comprises *voxemes*, characterizes various visual object concept structures in specific forms. Each Vox world, composed of these structures, is represented by a set of visualized situation types, possibly accompanied by static images or dynamic videos (motion pictures) to interpret dialogues.

As in the Situation Semantics of Barwise and Perry (1983) and Barwise (1989), the *partiality* of information is a foundational notion for ABS. Annotation targets the particular points of information and focuses on them. The basic assumption of ABS is that rational agents with a limited perception act on partial information and concentrate on a task thereby. The information provided by static or dynamic pictures and enriched linguistic and world knowledge with voxemes is too much for these agents to act properly. The annotation focuses on particular viewpoints on objects or aspects of information conveyed by language. Pictures carry too much information, while the annotated language, for instance in dialogues, focuses on a *small part* of it with perspectives. With this annotated partial information with particular views, the agents focus on their task and act intelligently. This paper claims such a focused interaction between the small restricted environment and the task is a fact.

2. Background study

Here are two views of dialogue. I use these views as a background study when analyzing and interpreting dialogues.

2.1. Classical Common Views of Dialogues

Dialogues are interactive linguistic exchanges among at least two participants, conveying or receiving information for actions or emotive reactions. The participants are message senders, recipients, and others directly or indirectly involved with specific intentions or forced responses, differentiating the various types of *dialogue acts* (Bunt, 2019; ISO, 2020). These participants can be either human agents or artificial rational agents like robots.

Question-answering is a typical type of dialogue. One party raises a question, while the other responds if a dialogue succeeds. Negotiations constitute another type of dialogue: one party proposes by requesting, while the other party accepts, modifies, or rejects the proposal by taking linguistic or non-linguistic actions. There may be mediators.

Dialogues are heavily grounded in various types of participant attitudes, background situations, and affordances. They thus license qualifications, restrictions, redundancies, or utterance omissions, much depending on their described situations, as spelled out by Barwise and Perry (1983) and with their later work on situation theory and semantics.

2.2. Dialogues in the Vox World

Pustejovsky and Krishnaswamy (2014, 2016) introduces a modeling language VoxML for visual object conceptual structures in language actions. As stated in Section 1, one of the key notions in VoxML is the *Vox World*. Pustejovsky and Krishnaswamy define this notion more formally with rich implications as a multimodal simulation framework for modeling embodied human-computer interactions and communication between agents engaged in a shared goal or task.

In the Vox World, dialogues are modeled as part of HOI (human-object interactions) or HCI (human-computer interaction) through language. Dialogue participants can be humans (H) or computers (C), all as rational agents that may include artificial agents like computers, while some other objects also participate or get involved indirectly in dialogues. Task-oriented dialogues are embodied interactions between agents, where language, gesture, gaze, and actions are situated within a common ground shared by all agents in the communication. Situated semantic grounding assumes a shared perception of agents with co-attention over objects in a situated context, with co-intention towards a common goal. Dialogues are thus viewed as complex linguistic phenomena in the Vox World.

3. Issues in Interpreting Dialogues

In this section, some dialogues are presented to focus on the issues of interpreting them with illustrations.

3.1. Interpreting Dialogue 1

Dialogue 1 illustrates the complexity of actions even in a short dialogue. It shows how a husband and his wife interact with each other in a shared task of making a cocktail punch and tasting it. The spoken part of the dialogue itself is simple, consisting of three words: "try," "it," and "delicious."

(1) Dialogue 1
Husband: Try it.
Wife: Delicious!

The script alone cannot be understood unless a situation, depicted visually like Figure 2, is given as a background. The script, as it is, only tells that it is a dialogue between two participants (a husband and his wife) and that the pronoun "it" refers to something edible or potable with a taste. The verb "try" means "try to eat or sip and see how it tastes." It is a task-oriented dialogue through which the couple tries to work together on some common goal, namely to make a good cocktail punch.

Figure 2: Dialogue Situation Visualized

Figure 2 supports the situation in which the dialogue has developed.[1] Two dialogue participants, the husband and his wife, are holding a glass together. The glass looks like containing a cocktail punch. The couple may have been preparing a good punch, possibly for a party. The husband

[1]This picture is provided by Ghang Lee (2023), who worked on it through $Dalle3^{+}ChatGpt$.

Figure 3: Empty Glass to be Washed

made a cocktail punch in a punch bowl, poured part into a small container from which one can drink, and handed over the glass to his wife to taste by saying, "Try it." See the difference between a punch bowl and a punch glass: you won't lift the bowl and try the punch from it. Here, the pronoun "it" refers to the punch the husband prepared, but it could have referred to anything edible or that can be tasted. What was presented to the wife was the glass containing the punch. So the wife took the glass of cocktail punch in her hand, raised it to her mouth, sipped the punch, and said "Delicious!". The utterance of the single word "Delicious" followed a series of actions with satisfaction on her facial expression. The wife tasted the punch the husband prepared, and approved the husband for it, making him feel good.

All these actions are not shown in Figure 2. The dialogue implies them only when the picture is looked at. Both the dialogue and the picture are *interpreted* coherently. An adequate interpretation model should be constructed to interpret the cooperative roles of the two dialogue participants, the wife and the husband, who made a punch and tasted it, and the two objects, the glass and the punch contained in it. The punch bowl and other material not shown in the picture may have been somewhere in the kitchen.

3.2. Interpreting Dialogue 2

Dialogue 2 is even shorter than Dialogue 1. It is a short script with two words, supposedly for a dialogue between a couple, Husband and Wife.

(2) Dialogue 2
Husband: Take this.
Wife: [says nothing.]

Dialogue 2 records the husband uttering the two words "Take this.", asking the wife to take something that is referred to by the demonstrative pronoun "this" and should be located near the speaker himself, but the wife says nothing. There were *two dialogue participants*, and the husband's act was *task-oriented*, telling or ordering his wife to take something near him. This is all that a dialogue act annotation can capture.

A visualized situation, depicted with Figure 3,[2] for the dialogue provides detailed information on the interactions between the husband and the wife. The wife didn't say a word, but one should see her face in the picture, Figure 3. It says a lot. A husband, sitting on a couch in the living room, told his wife, standing by the dishwasher, to take the glass in his hand, expecting her to put it in the dishwasher. The wife was angry at her husband, who played the king. A situation like this may be considered disgusting in some cultures.

3.3. Dialogue 3 in Contrast to Dialogue 2

A dialogue almost the same as Dialogue 2 has a totally different interpretation. In Dialogue 3, the wife expresses her appreciation.

[2] Ghang Lee also provided this image.

(3) Dialogue 3 with Appreciation
Husband: Take this!
Wife: Thanks. Looks delicious.

The husband mixed a cocktail punch and *offered* it to his wife. The wife says "Thanks" in an appreciative way by saying a little more, "Looks delicious." The following picture[3] depicts a delightful scene that says more than words.

Figure 4: Punch offered to the Wife

I have presented the two pictures that visualized dialogue situations. They show how much visual information contributes to the rich interpretation of dialogues or interactive communications. The same imperative "Take this!" is interpreted differently, one as an *order* and the other as an *offer*.

3.4. Dialogue 2 Extended

Dialogue 4 illustrates with Script 4 how Dialogue 2 is extended with another round of exchanging the turns.

(4) Dialogue 4 Extending Dialogue 2
Husband: Take this.
Wife: [Got angry, saying nothing.]
Husband: Sorry. I'll do it.
Wife: [Facial expression changed to exasperation. She is still silent.]

Looking at his wife's angry face, the husband realized he had mistakenly asked her to take the glass to the dishwasher. He thus apologized and took the glass himself to the washer.

The dialogue has four turns, although the wife does not respond verbally. Such a situation can easily be imagined and turned into a short video. However, the current technology has not fully developed to convert text to videos.[4]

As one of the reviewers pointed out, it must be emphasized that it is not so much the picture itself providing the background context of the dialogue but rather the situation type we construct based on the picture. We imagine or visualize appropriate situations or create such scenes to interpret dialogues. Dialogues, on the other hand, help interpret visually perceptible scenes by helping us focus on some specific parts of them.

4. Annotating Dialogues

4.1. Basic Annotation Structure of Dialogues

The annotation of dialogues follows Bunt (2019) and ISO (2020). The basic structure of a dialogue consists of two parts, the dialogue act and the semantic content. In the simplest case, the dialogue structure is a quadruple $<<s, A, f_d>, c>$, where the triple represents the simplest dialogue act structure consisting of a sender s, addresses A, and a dimension-specific function f_d while the last component c represents the dialogue content. For general purposes, this list can be extended to the most complex case with a 7-tuple (ISO, 2020) plus the content c, where the three bracketed components need not be specified:

(5) $<<s$, A, [h], f, d, [Q], [E]$>$, C$>$ of attributes,
where s is a sender (speaker),
A addresses,
H other participants,
f a general-purpose communicative function,
d a dimension,
Q qualifiers,
E dialogue units that the act depends on, and
c the semantic content of the dialogue.

The first seven components specify the act type of dialogues while the last component C refers to the dialogue content. The content c directly *plugs in* the semantic content, which carries the information of a dialogue associated with a dialogue act. The 7-tuple plus the content c forms a complex feature structure such that the value of c is directly linked to another annotation scheme. No link like contentLink needs to be introduced, although it is a preference recommended in Bunt (2019) and proposed in ISO (2020).

4.2. VoxML-linked Annotation

The VoxML-linked annotation (Lee et al., 2023) refers to the Voxicon, a component of VoxML, consisting of complex feature structures, called *vox-*

[3] Ghang Lee also provides this image.

[4] The Open AI just announced Sora for such a task.

emes. These voxemes represent the visual object conceptual structures of VoxML basic categories such as **object**, **program** (event, motion, or action), and **relation** that includes property and function). Each voxeme is associated with a linguistic expression (e.g., "glass"), its morpho-syntactic or lexical information, and semantico-pragmatic or physical information associated with it such as information about its habitat, affordance structures, and embodied interactions (Pustejovsky, 1995; Gibson, 1977; Pustejovsky and Krishnaswamy, 2016, 2021). The reference to these structures is expected to free the VoxML-oriented ABS from its reliance on syntactic or pragmatic analysis.

For illustration, consider the annotation of Dialogue 3. The annotation takes two steps: Step 1 focuses on the dialect act, while Step 2 on its content.

4.2.1. Step 1: Annotating Dialogue Act

The first part of the whole script, which includes the information about the speaker and the addressee, is annotated as in (6).

(6) Annotating the Dialogue Act of Dialogue 3
   a. Segmented Dialogue Script (id="d3S"):
      Husband$_{w1}$ to Wife$_{w3}$: Take$_{w4}$ this$_{w5}$.
   b. Dialogue Act Annotation:
      <dialogue id="#d3", target="#d3S">
       <dAct id="d3A", sender="#w1",
        addressee="#w3", dimension="task",
        cFunction="offer", content ="#d3C"/>
      </dialogue>

The dialogue act annotation marks up not just what has been uttered by the speaker, but the whole dialogue script that describes all the components that constitute the act of a given dialogue.

4.2.2. Step 2: Annotating the Content

The proposed VoxML-linked ABS annotates the content c of a dialogue by referring to the dialogue utterance and the background situation, possibly depicted by an associated picture. The content of Dialogue 3 is annotated as in (7):

(7) Annotating the Content of Dialogue 3
```
<dialogue id="#d3", target="#d3S">
 <dContent id="d3C", linkedTo="#d3A">
   <object id="o1", target="#w1"
     type="human", pred="husband",
     relatedTo="#w3"/>
   <object id="o2", target="#w3"
     type="human", pred="wife",
     relatedTo="#w1"/>
   <action id="a3", target="#w4"
     type="transition",
     pred="take:consume[5]",
     agent="#o2", theme="#o6:punch"/>
   <object id="o3", target=" ",
     type="physicalObj:artifact",
     pred="glass", definite="yes",
     grabbedBy="#o4:hand",
     comment="See Figure 4"/>
   <object id="o4", target="",
     type="physicalObj", pred="hand",
     definite="yes", partOf="#o1:husband",
     comment="See Figure 4"/>
   <object id="o5", target="",
     type="physicalObj:liquid:beverage",
     pred="punch", definite="yes",
     containedIn="#o3:glass",
     comment="See Figure 4" />
 </dContent>
</dialogue>
```

With the comment "See Figure 4", the demonstrative pronoun "this" is annotated as referring to the punch in the glass held by the husband in his hand. It does not refer to the glass, for it is already in the wife's hand. The verb "take" is thus understood as meaning *to consume the punch*, instead of meaning *to grab the glass with a hand*. The dialogue does not mention "glass," "punch," or "hand" but the annotation introduces them all as *non-consuming tags*. Figure 4 shows that the glass is in the husband's hand and also in the wife's hand.

4.3. Abstract Syntax and the Metamodel

The annotation of the content structure in the Vox World as presented in (7) requires the specification of an annotation scheme. Such a specification is done partially with the formulation of an abstract syntax. For this, the abstract syntax, named $\mathcal{ASyn}_{vox}$, is minimally formulated for the annotation in the Vox World, as in (8):

(8) $\mathcal{ASyn}_{vox}$ is defined as a tuple $<M, B, @>$, such that, given a language L,
   a. M is a nonempty subset, called *markables*, of L, delimited by B;
   b. B is a set $\{o, a, r\}$ of *base categories*:
      o stands for category **object**;
      a, category **action**, a subcategory of **eventuality**;
      r, category **relation** that includes the subcategories **function** and **property**;
   c. $@$ is a set of *assignment functions* from features (attributes) to values associated with each category in B.

Note that this syntax has no links. Instead, some attributes are plugged into other annotation structures. See Annotation Structure (9).

[5] See WordNet-3 for the sense of "take."

(9) Semantic Roles:
<action id="a3" type="transition", pred="take", agent="#o2", theme="#o3"/>

The semantic roles for the action *take* are directly annotated into its base annotation structure by referring to the semantic role frames in a lexicon. There is no repeated application of a link like `srLink` for semantic role labeling.

The minimal abstract syntax $\mathcal{ASyn}_{vox}$ specified in (8) conforms to the metamodel for the Vox World as a markup language (Lee et al., 2023).

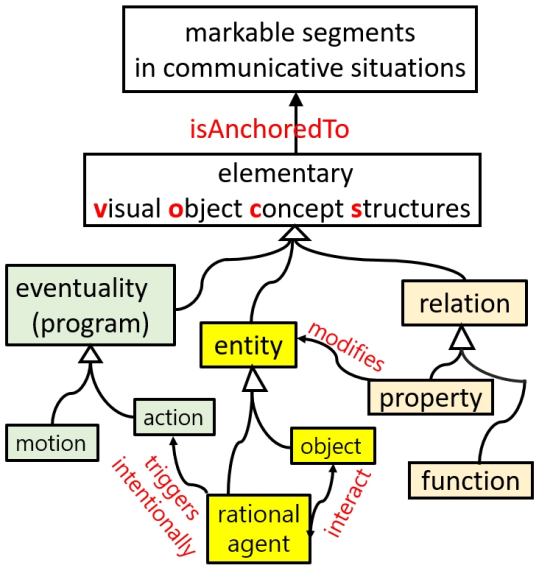

Figure 5: Metamodel of the Abstract Syntax

Here, the Vox World consists of anything perceptible in communicative situations including dialogues. Visual object concept structures (vocs or vox), either elementary or their relational compositions, are then anchored to the Vox World. The vox are categorized into three major categories with subcategories **eventuality (program):action**, **object:rational agent**, and **relation:property, function**. Actions are triggered by rational agents intentionally, while rational agents, either humans or robots, interact with one another or other objects. Properties (attributes) modify objects, while functions and relations operate or range over visual object conceptual structures.

5. Interpretation

5.1. Overview

ABS (Lee, 2023) interprets annotation structures for a model constrained by relevant parts of the Vox World. Implementing classical set-theoretic models as in Montague semantics (Montague, 1974b), or the Discourse Representation Theory (DRT) (Kamp and Reyle, 1993; Parson, 1990), these parts of the Vox world supplement those models $<D, R, [[\,]]>$ of denotational semantics, especially by formally delimiting the domain D of a model, which normally consists of individual entities, and the set of n-ary relations R over D or its Cartesian products with a small world in which some relevant visual object concept structures reside.

In the Vox World, everything in its small world is a first-class citizen, including properties and relations, as in Situation Semantics (Barwise, 1989), or else the notion of *functional types* is introduced to allow such objects as *event descriptors* of type $e \rightarrow t$ (Kracht, 2002; Pustejovsky et al., 2019) or as in Davidsonian Semantics (Davidson, 1967, 2001; Parson, 1990). ABS then interprets annotation structures in two steps. First, annotation structures **a** are translated to semantic forms σ(**a**) in typed first-order logic. Second, these logical forms are interpreted for a well-defined model M constrained by the Vox World v: $[[\sigma(\mathbf{a})]]^{M,v}$.

5.2. Translating Annotation Structures to Logical Forms

To interpret annotation structures, ABS translates them into semantic forms directly. ABS does not require syntactic analysis to derive semantic forms because the annotation already contains the necessary information for adequate translation. In contrast, Montague Semantics (Montague, 1974b) uses Categorial Grammar for analyzing input data to trees, for instance, to capture scope ambiguity, before translating the analyzed trees to semantic forms in Higher-order Intensional Logic.

Translation (10) shows how the annotation structures of category **object** are translated.

(10) a. <object id="o1", target="#w1"
type="human", pred="husband",
relatedTo="#w3"/>
$\sigma(o1) := [human(x_1), husband(x_1, x_2)]$
b. <object id="o2", target="#w3"
type="human", pred="wife",
relatedTo="#w1"/>
$\sigma(o2) := [human(x_2), wife(x_2, x_1)]$

The attribute @relatedTo in the annotation structures treats the predicates *husband* and *wife* as binary relations in the semantic forms.

The transitive verb "take" denotes an action of type *transition* with two required arguments. Translating the annotation structure that marks up its semantic content is straightforward. The two semantic roles associated with the two verb arguments are marked up.

(11) <action id="a3", target="#w4",
type="transition", pred="take",
agent="#o2", theme="#o3"/>
$\sigma(a3)$:=

$[transition(e_3), take(e_3), agent(e_3, x_2),$
$theme(e_3, x_3)]$

The semantic form $\sigma(a3)$ here does not add new information to the annotation. The predicate *take* is a transition, thus involving a series of sub-actions: the wife, who was told to take something, referred to with the demonstrative pronoun "it", must reach a reachable position to grab the object and take it out, intending to move it to somewhere for some purpose. The annotation does not capture such information but must be captured at the interpretation stage, given an appropriate background.

5.3. Direct Interpretation vs. Enriched Logical Forms

Intuitively speaking, annotation structures should be interpretable without being translated into logical forms, as in Montague (1974a)'s English as a Formal Language. Translation carries no additional meaning except that it shows that the translated logical forms are expressed in lower-order logic. However, it is possible to generate enriched annotation structures by referring to the Voxicon.

VoxML contains the Voxicon that lists *voxemes* enriching annotation structures of those categories, **object**, **event: action**, and **relation: property, function** in the metamodel. For illustration, consider the annotation structure of category **object**: <object id="o3", target="#o4 (glass)"/> to enrich it with the voxeme of *glass* listed in the Voxicon.

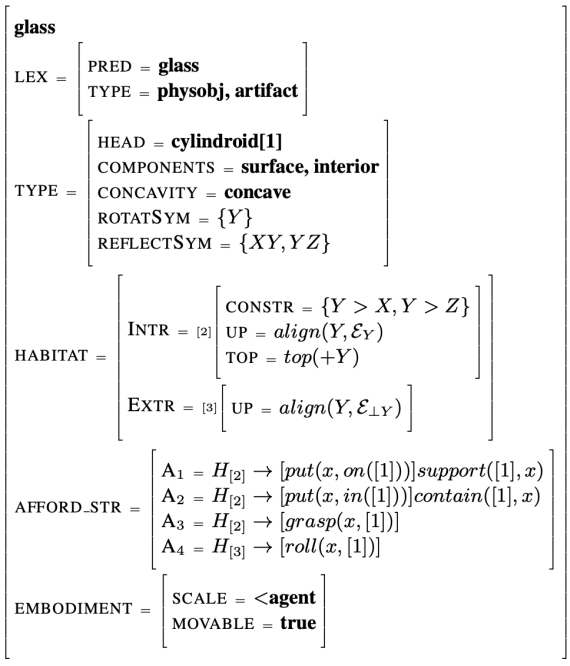

Figure 6: Voxeme of a Glass

The voxeme of *glass* in Figure 6 represents five sorts of information: [i] LEX, [ii] TYPE, [iii] HABITAT, [iv] AFFORD_STR, [v] EMBODIMENT.[6] Annotation (12b) represents part of the lexical information (LEX) and the affordance structure (AFFORD_STR) information about its being a container (A_2). The EMBODIMENT says that the object's size is smaller than the agent who carries or grabs it and can be moved by the agent. For illustration, consider annotating the noun "glass" in Dialogue 3. Its annotation structure can be enriched with contextual information by referring to the voxeme as in Figure 6.

(12) a. Basic Annotation, copied from 7:
<object id="o3", target=" ",
type="physicalObj:artifact",
pred="glass", definite="yes",
grabbedBy="{#o1,#o4:hand}",
refersTo="Figure 4"/>

b. Annotation Enriched with Voxeme 6:
<object id="o3", target=" ",
type="physObj:artifact",
pred="glass", definite="yes",
grabbedBy="{#o1,#o4:hand}",
form="cylindroid", shape="concave",
use="container", contains="o5:punch",
smallerThan="{#o1,#o4}",
refersTo="Figure 4, Figure 6"/>

c. Logical Form $\sigma(x3)$:=
$[physobj(x_3), artifact(x_3),$
$glass(x_3), definite(x_3),$
$grab(e_1), agent(e_1, x_1), theme(e_1, x_3),$
$instrument(e_1, x_4 : husband'sHand),$
$cylindroid(x_3), concave(x_3),$
$container(x_3), contains(e_2),$
$theme(e_2, x_6 : punch)]$

Annotation (12b) shows the enrichment of Annotation with some pieces of information obtained from the voxeme of "glass" presented in Figure 6. The logical form based on the enriched annotation states that the glass, which is small enough to be grabbed by the husband, contains punch.

5.4. Interpretation

The Vox World provides visual information for interpreting actions and interactive communications. Specifically, it controls the three processes of annotating, translating, and interpreting dialogue acts and contents interchanged among the participants. The utterance "Take this.", which is made by the husband in the two different dialogues, for instance, is annotated differently: in Dialogue 2, it is annotated as an *order*, whereas it is annotated as an *offers* in Dialogue 3. In Dialogue 2, the demonstrative pronoun "this" refers to the empty glass. In

[6] For the detailed explanation of the voxeme of *glass*, see Lee et al. (2023).

Dialogue 3, in contrast, the same pronoun refers to either the glass with a punch in it or the punch in the glass, for the wife says, "Looks delicious," referring to the punch, not the glass.

Annotation and the Vox World complement each other. Voxemes enrich annotation structures. Annotation can capture all these differences and refer to the appropriate figures for appropriate information, but the voxemes alone cannot.

Annotation structures and semantic forms are inadequate to capture finer-grained information associated with all aspects of dialogues. This especially concerns the interpretation of actions, for actions of type transition particularly involve a dynamic sequence of sub-events or sub-actions. The husband's order in Dialogue 2 is not a simple act, but a complex sequence of sub-situations and sub-actions.

(13) Sub-situations and sub-actions in Figure 2:
  a. The wife was standing near the washing machine.
  b. The husband was sitting on a sofa not far from the kitchen.
  c. The husband asked the wife to take the glass,
  d. and expecting
  e. her to come to him easily
  f. to pick it up from his hand and
  g. put it in the dishwasher.
  h. Her emotional reaction, displayed on her face with silence,
  i. indicated that his expectation was wrong.
  j. She rather expected
  k. him to come and
  l. put the glass in the dishwasher himself.

All this information cannot be captured in the annotation or represented in simple logical forms. It can only be *abduced*[7] by learning relevant perspectives on the informational content and the intention of dialogue or discourse participants, as mentioned by Hobbs (1996). In addition, such an abduction becomes possible by constructing appropriate background scenes with visual object conceptual structures (vox). The construction of such scenes is systematically constrained in the Vox World that characterizes not only the lexical features of the language used in human communications, but also the habitat, affordance structures, and embodiment of objects and actions, and their interactions mentioned in that language with perceptual (visual) conditions.

[7] I have intentionally used the term *abduce* to focus on the experiential and perceptual aspects of Peirce (1931–1958); Hobbs et al. (1993); Hobbs (1996, 2006) for understanding language and logic.

6. Concluding Remarks

The partiality of information is a basic motivation for annotation, for annotation marks up only some parts of a language. This paper has shown how this notion of partiality works in annotating and interpreting dialogues. Annotation also explicitly uses language such as dialogues by annotating the type of dialogue acts and content and interpreting them against a small visual world called the Vox World.

The paper treated the tripartite understanding of dialogues: annotation, visualization, and interpretation. Annotation focuses on some basic linguistic elements in described situations in which dialogue participants interact with relevant objects or each other. At the same time, visualization provides details of fine-grained perspectives with background information. Interpretation with logical forms validates such details of information with consistency.

The paper proposes using visual information in general and the Vox World in particular, to annotate and interpret dialogues or other interactive communications among rational agents or relevant objects. It even suggested that a set-theoretic semantics should be redesigned by restructuring its basic model structure $<D, R, [[\,]]>$. For instance, the domain D and the set R of n-ary relations can be modified with a small set of visual object concept structures. Or else, such a model is minimally implemented but constrained by something like the Vox World. However, the formal specification of such a task is left for the future.

The paper intentionally focused on simple dialogues to highlight the complementary roles of dialogue scripts and related images and on the role of VoxML-linked annotation that links them for coherent interpretation. Complex dialogues, such as those involving misunderstandings and subsequent repair strategies, require complex images, such as motion pictures, for their interpretation.

Pictures are extensively used to show how dialogues are annotated and interpreted. For this reason, the proposed VoxML-linked ABS may be understood mistakenly as a picture-based semantics that requires the generation of static or dynamic pictures as an essential process. It is a total misunderstanding. Pictures help visualize the situations in which dialogues are possibly enacted. Humans can easily visualize such situations through the power of imagination. It is, however, a different question of how artificial agents learn to visualize dialogue situations and interpret them or even to participate in a dialogue by understanding the flow of dialogues. Such a question is left for future work.

7. Acknowledgments

I owe many thanks to the three anonymous reviewers who provided detailed constructive comments to improve the paper extensively and to Jae-Woong Choe, Minhaeng Lee, and Chong-won Park, who helped write the preliminary version of the paper. I also would like to thank Byonrae Ryu, who reset the figures, and Ghang Lee with the production of dialogue-related images for their time-consuming work.

8. Bibliographical References

Jon Barwise. 1989. *The Situation in Logic*. CSLI (Center for the Study of Language and Information, Stanford, CA.

Jon Barwise and John Perry. 1983. *Situations and Attitudes*. The MIT Press, Cambridge, MA.

Harry Bunt. 2019. Plug-ins for content annotation of dialogue acts annotation. In *Proceedings of the 15th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-15)*, pages 33–45, Workshop at the 11th International Conference on Computational Semantics (IWCS 2019), Gothenburg, Sweden, Mary 23, 2019.

Harry Bunt. 2022. Intuitive and formal transparency in semantic annotation schemes. In *Proceedings of the 18th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-18)*, pages 102–109, Workshop at LREC2022, Marseilles, France.

Donald Davidson. 1967. The logical form of action sentences. In N. Rescher, editor, *The Logic of Action and Decision*, pages 81–120. University of Pittsburgh Press, Pittsburgh, PA.

Donald Davidson. 2001. *Essays on Actions and Events*. Oxford University Press, Oxford.

David Dowty, Stanley Peters, and Robert Wall. 1981. *Introduction to Montague Semantics*. Reidel, Dordrecht.

James Jerome Gibson. 1977. The theory of affordances. *Perceiving, Acting, and Knowing: Toward an Ecological Psychology*, pages 67–82. Reprinted as chapter 8 of Gibson (1979).

James Jerome Gibson. 1979. *Ecological Approach to Visual Perception*. Psychology Press, New York.

Jerry R. Hobbs. 1996. On the relation between the informational and intentional perspectives on discourse. In E. Hovy and D. Scott, editors, *Computational and Conversational Discourse: Burning Issues– An Interdisciplinary Account*, pages 247–260. Springer, Berlin, Germany.

Jerry R. Hobbs. 2006. Abduction in natural language understanding. In Laurence R. Horn and Gregory Ward, editors, *The Handbook of Pragmatics*, chapter 32, pages 724–741. Blackward Publishing, London. Https://doi.org/10.1002/9780470756959.ch32.

Jerry R. Hobbs, E. Stickel, Mark, Douglass Appelt, and Paul Martin. 1993. Intepretation as abduction. *Artificial Intelligence*, 63(1-2):69–142.

ISO. 2020. *ISO 24617-2 Language resource management – Semantic annotation framework – Part 2: Dialogue acts*. International Organization for Standardization, Geneva. 2nd edition.

Hans Kamp and Uwe Reyle. 1993. *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language Formal Logic and Discourse Representation Theory*. Kluwer Academic Publishers, Dordrecht.

Marcus Kracht. 2002. On the semantics of locatives. *Linguistics and Philosophy*, 25:157–232.

Kiyong Lee. 2020. Annotation-based semantics. In *Proceedings of the 16th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-16)*, pages 37–49.

Kiyong Lee. 2023. *Annotation-Based Semantics for Space and Time in Language*. Cambridge University Press, Cambridge, UK.

Kiyong Lee, James Pustejovsky, and Nikhil Krishnaswamy. 2023. An abstract specification of VoxML as an annotation language. In *Proceedings of the 19th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-19)*, pages 66–74, June 20, 2023, Workshop at IWCS 2023, Nancy, France. ACL anthology L16-1730.

Richard Montague. 1974a. English as a formal language. In *Formal Philosophy: Selected Papers of Richard Montague*, New Haven and London. Yale University Press.

Richard Montague. 1974b. The proper treatment of quantification in ordinary english. In *Formal Philosophy: Selected Papers of Richard Montague*, New Haven and London. Yale University Press.

Terence Parson. 1990. *Events in the Semantics of English: A Study in Subatomic Semantics*. The MIT Press, Cambridge, MA.

Charles S. Peirce. 1931–1958. *Collected Papers of Charles Sanders Peirce*. Harvard University Press, Cambridge, MA. Edited by Hartshorne, C. and Weiss, P. and Burks, A.

James Pustejovsky. 1995. *The Generative Lexicon*. The MIT Press, Cambridge, MA.

James Pustejovsky. 2013. Dynamic event structure and habitat theory. In *Proceedings of the 6th International Conference on Generative Approaches to the Lexicon (GL2013)*, pages 1–10. Association for Computational Linguistics, Pisa, Italy.

James Pustejovsky and Olga Batiukova. 2019. *The Lexicon*. Cambridge University Press.

James Pustejovsky and Nikhil Krishnaswamy. 2014. Generating simulations of motion events from verbal descriptions. In *Proceeding of the 3rd Joint Conference on Lexical and Computational Semantics. (*SEM 2014)*, pages 99–109.

James Pustejovsky and Nikhil Krishnaswamy. 2016. VoxML: A visualization modeling language. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4606–4613, Portorož, Slovenia. ELRA. ACL anthology L16-1730.

James Pustejovsky and Nikhil Krishnaswamy. 2021. Embodied human-computer interaction. *KI-Künstliche Intelligenz*, 35(3-4):307–327.

James Pustejovsky, Kiyong Lee, and Harry Bunt. 2019. The semantics of ISO-Space. In *Proceedings of the 15th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-15)*, pages 46–53. May 23, 2019, at IWCS2019, Gothenburg, Sweden.

9. Copyrights