# Production In A Multimodal Corpus: How Speakers Communicate Complex Actions

**Carlos Gómez Gallo[†], T. Florian Jaeger[◇], James Allen[†], Mary Swift[†]**

Department of Computer Science, University of Rochester, Rochester, NY, USA[†]
Department of Brain and Cognitive Science, University of Rochester, Rochester, NY, USA[◇]
cgomez@cs.rochester.edu

**Abstract**

We describe a new multimodal corpus currently under development. The corpus consists of videos of task-oriented dialogues that are annotated for speaker's verbal requests and domain action executions. This resource provides data for new research on language production and comprehension. The corpus can be used to study speakers' decisions as to how to structure their utterances given the complexity of the message they are trying to convey.

## 1. The Corpus

The Fruit Carts corpus is a collection of multimodal dialogues collected at the University of Rochester (Aist et al., 2006). The Fruit Carts domain was designed to elicit requested manipulations of both simple and complex referring expressions in unrestricted natural language.

A speaker is given a map showing a specific configuration of fruits and geometric shapes in different regions (see map on Figure 1). The speaker's task is to instruct a listener or actor to reorganize the objects so the final state of the world matches the map first given. The speaker can request to MOVE, ROTATE and PAINT objects on the screen and the actor performs requested actions as soon as he recognizes them. This multimodal dialogue corpus is particularly interesting since it interleaves the speech signal of one dialogue partner with the action execution of the second partner.

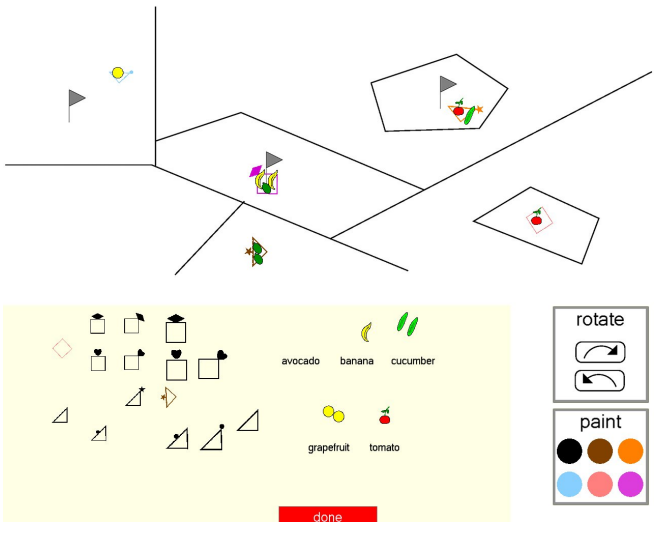


Figure 1: Fruit Carts Map.

The corpus consists of 104 digital videos of 13 participants, recruited from the university community. The dialogues range from 4 to 8 minutes in duration. The number of utterances per dialogue ranges from 20 to a little over 100, resulting in a total of approximately 4,000 utterances in the corpus. The average length of utterances is 11 words.

The corpus is being annotated by six University of Rochester undergraduate research assistants with the annotation tool Anvil (Kipp, 2001). The result will be a rich data set that captures continuous understanding at the word level with XML readable format for referring expressions, spatial relations, domain actions, semantic roles and speech acts. See (Gómez Gallo et al., 2007) for annotation details.

The Fruit Carts corpus was originally motivated by research on language comprehension (Tanenhaus et al., 1995, e.g.) and has since then been successfully employed to aid the development of dialogue agents within an incremental understanding framework (Stoness et al., 2004; Aist et al., 2006). The corpus has also been used to evaluate dialogue agents by measuring user satisfaction when using either incremental or non-incremental dialogue agents (Aist et al., 2007). Here we demonstrate that the Fruit Carts corpus is also suited for the investigation of language *production*.

The domain consists of a variety of objects and regions that these objects are located in (see figure 2). Some objects have known labels (fruit types), others are geometrical figures differing in features such as shape, size, decoration type, and decoration location. Therefore a referring expression may be as complex as *"The small triangle with a heart on the hypothenuse"* or as simple as *"a tomato"*. Region names were more uniform in complexity (see figure 2), but the complexity of the descriptions used to describe the goal locations of MOVE actions also differed in complexity, because speakers often elaborated in great detail where precisely within a region an object had to be placed (see below for examples).

It is this variety in description complexity, combined with the annotation of the conveyed message (which is reflected in the actions performed by the actor), and the relatively naturalness of the task that make the Fruit Carts corpus ideally suited for the study of language production. We illustrate this point using a case study on the relation between message complexity and speakers' planning of request acts at the clausal level.

## 2. Speaker's Planning of Request Acts

In ongoing work (Gómez Gallo et al., 2008), we investigate what determines how much information speakers convey in

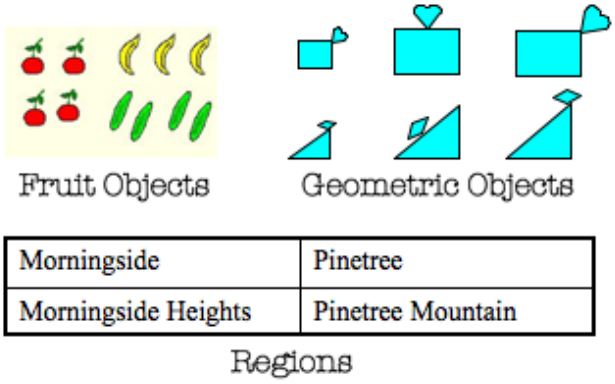


Figure 2: Objects in the Fruit Carts Domain

a single clause. In particular, we hypothesize that speakers prefer to keep the overall complexity of clauses relatively uniform.

Consider the two scenarios in (1a,b) vs. (2). Both (1a,b) and (2) request that an object be selected and moved to a specific location, as evidenced by the actions performed by the dialogue partner. The structural realization, however, differs between the two requests.

In (1a,b) the speaker chooses to esxplicitly introduce the theme (‘‘the square with the heart”) into the discourse using a separate utterance (1a) and only then describes the requested MOVE action (1b), using a pronoun to refer to the theme. We refer to this realization of SELECT+MOVE action as a bi-clausal realization.

This contrasts with (2), where the speaker conveys both parts of the MOVE request in one single utterance. The SELECT action is implicit. Only the MOVE action is explicitly mentioned. We refer to this as a mono-clausal realization.

(1a) S: Take [$_{theme}$the square with the heart]
A : (*actor grabs the theme*)

(1b) S: And move [$_{theme}$it] [$_{loc}$ into Forest Hills]
A : (*actor moves square in the region*)

(2) S: Then put [$_{theme}$an apple] [$_{loc}$ inside the triangle]
A: (*actor grabs *and* moves theme to location*)

Note that the location descriptions are similar in complexity in the two scenarios (*into Forest Hills* and *inside the triangle* in (1b) and (2)). The two theme descriptions, however, differ greatly in length (and hence complexity). In (2), with the less complex theme, the speaker chose a mono-clausal request, while in (1) with a more complex theme, a bi-clausal request was used.

Next, we show that this apparent link between theme complexity and speakers’ choice between mono- and bi-clausal request realizations seems to be systematic. Below we focus on the effect of theme complexity, then on theme giveness and location complexity. We refer to (Gómez Gallo et al., 2008) for more detail on other factors.

## 3. Message Complexity and Structural Realization

We hypothesize that description length of referring expressions are correlated with message structure of a request act. Specifically, we hypothesize that **speakers prefer a bi-clausal structure, if the theme becomes too complex**.

To test this hypothesis, we annotated 21 session from 8 speakers of the Fruit Cart corpus. We annotated the theme of all 534 utterances with MOVE actions in those sessions. From this annotation, we extracted the length of theme description in number of words without counting disfluencies or repeated words. We perform a binary logistic regression model with theme description length as the only predictor. The modeled outcome variable was whether speakers used a mono or bi-clausal structure (MOVE only vs. SELECT-MOVE realization).

We found that theme description length is positive correlated with speakers’ decision to use bi-clausal message ($\beta$=1.89; SE($\beta$)=0.23; p$<$ 0.0001). Speakers are more likely to produce two clauses rather than one, the longer theme description is. Figure 3 illustrates the result.

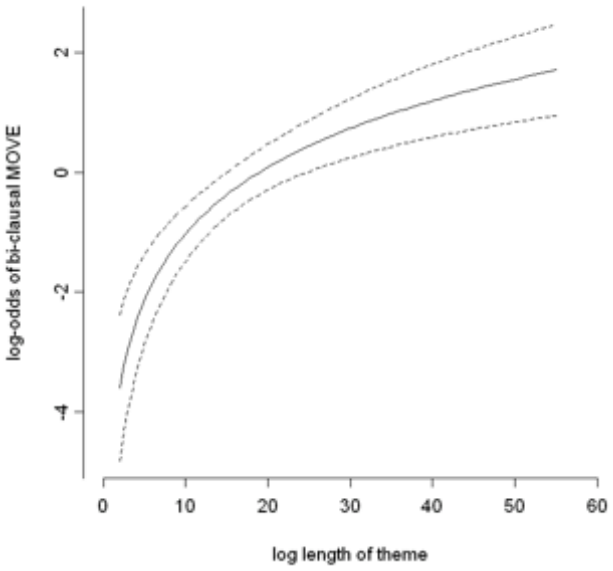


Figure 3: Fitted Effect of Theme Description Length on Speakers’ Decision to use a Bi-clausal Structure

This is evidence that theme description length is correlated with speaker’s planning of utterances. However, this correlation could be an artifact of information structural constraints. It is well-known that repeated reference to the same referent correlate with shorter and shorter referential expressions for that referent (**?**, e.g.)]Ariel01. It could thus be that the shorter theme descriptions in our sample are descriptions of themes that have been mentioned before in the discourse (i.e. given themes), while the longer descriptions may mostly refer to first-time mentions (i.e. new themes). The observed effect may then be entirely due to a preference of speakers to introduce new themes via a SELECT request, thereby directing their interlocutor’s attention to the

relevant theme before more detailed requests are uttered.
In examining the corpus, one can find sentences whose theme description is inversely correlated to the structure of the message according to our initial hypothesis. Looking at these sentences directly may give us some insight in how speakers are planning their utterances. We have two situations to consider. In the first case the theme has a short description and yet the message structure is realized in a bi-clausal way (utterance 3). In the second case longer theme descriptions occur in a mono-clausal realization (utterance 4).

(3) S: Take [$_{theme}$ one tomato]
S: Put [$_{theme}$ it] [$_{loc}$ in the center of that triangle]

(4) S: Add [$_{theme}$ two bananas and a tomato] [$_{loc}$ inside of it]

Both cases suggest that there are other factors besides the theme description length that should be taken into consideration. For instance, the type of phrases instantiating other semantic roles. Utterance 3 shows a simple theme in a bi-clausal structure. However, the location is much longer in comparison with the theme. Conversely, utterance (4) shows a simple location with an imbedded pronoun. This suggests that the location is salient or repeated and its reference simpler. Notice that the verb used *"add"* almost allows the omission of the location. Since the complexity of the location is lower, the theme could actually take a fuller form within the same sentence to realize a mono-clausal request.
These examples suggest that we should account for the overall complexity of both theme and location. Thus we can refine our hypothesis to say that **the description length of *all* verb arguments affect the production choice between mono or bi clausal structure**. To test this new hypothesis and to address the issue related with theme's previous mentions, we annotated for other features to be included as predictors in the regression model.
These are main verb used, theme and location, theme givenness, location elaboration, and speech disfluencies and repetitions. We coded four levels of givenness: new, given, implied and set. Implied themes referred to objects which were not directly present in the discourse, but that could be inferred using world or domain knowledge (**?**). Set themes referred to objects present in the discourse individually and are now being referred as a group. In this analysis we excluded location elaborations (e.g. *"a little more to the right"*) since by definition, these elaborations do not require a preceding SELECT action.
This left 280 MOVE actions. A logistic regression model including these new predictors find that speakers preferred a bi-clausal message both for complex theme and for complex location descriptions ($\beta$=1.64; SE($\beta$)= .27; p$<$0.0001 and $\beta$=0.64; SE($\beta$)= .26; p$<$0.01), as well as for new themes ($\beta$=3.48; SE($\beta$)= 1.04; p$<$0.001).

## 4. Summary and Conclusions

The Fruit Carts corpus is a novel resource for the study of language production, providing researchers with control over the conveyed message while maintaining economic validity. Here we have illustrated that data from the Fruit Carts corpus evidence that speakers prefer to convey complex messages by distribution the information across several clauses. This suggests some sort of limited mental resource at the level of clausal planning. Crucially, the specific result presented here, the effect of theme complexity, goes beyond earlier results and is unexpected given standard theories of sentence production (Levelt and Maassen, 1981; Dell and Brown, 1991). For further discussion and a proposal that accounts for the observed effect, we refer to (Gómez Gallo et al., 2008).

## 5. References


G. Aist, J. Allen, E. Campana, L. Galescu, C. Gómez Gallo, S. Stoness, M. Swift, and M. Tanenhaus. 2006. Software architectures for incremental understanding of human speech. In *Interspeech*.

G. Aist, J. Allen, E. Campana, C. Gómez Gallo, S. Stoness, M. Swift, and M. Tanenhaus. 2007. Incremental dialogue system faster than and preferred to its nonincremental counterpart. In *CogSci*.

M. Ariel. 2001. Accessibility theory: an overview. In T. Sanders, J. Schilperoord, and W. Spooren, editors, *Text representation: Linguistic and psycholinguistic aspects*, pages 29–87. John Benjamins, Amsterdam.

G. Dell and P. Brown. 1991. Mechanisms for listener-adaptation in language production: Limiting the role of the 'model of the listener'. *Bridges between psychology and linguistics*.

C. Gómez Gallo, G. Aist, J. Allen, W. de Beaumont, S. Coria, W. Gegg-Harrison, J. Pardal, and M. Swift. 2007. Annotating continuous understanding in a multimodal dialogue corpus. In *DECALOG*, Trento, Italy.

C. Gómez Gallo, T. F. Jaeger, and R. Smyth. 2008. Incremental syntactic planning across clauses. *Submitted to CogPsy*, 40:296–340.

M. Kipp. 2001. Anvil – a generic annotation tool for multimodal dialogue. In *Eurospeech*.

WJM Levelt and B Maassen. 1981. Lexical search and order of mention in sentence production. *Reidel*.

S. Stoness, J. Tetreault, and J. Allen. 2004. Incremental parsing with reference interaction. In *ACL Workshop on Incremental Parsing*.

M.K. Tanenhaus, M.J. Spivey-Knowlton, K.M. Eberhard, and J.E. Sedivy. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268:1632–1634.