

Production in a Multimodal Corpus: How Speakers Communicate Complex Actions

LREC 2008

Carlos Gómez Gallo

T. Florian Jaeger

James Allen

Mary Swift



Rochester Corpus: Incremental understanding data built in the TRIPS dialog system architecture

TRAINS (logistics) – constructing a plan to use boxcars to move freight between cities on an onscreen map

Monroe (emergency) – build plan for an emergency situation

Chester (medicine) – consult with patient on drug interactions

CALO (personal assistant) – purchasing computer equipment

PLOW (procedure learning) – computer learns from show & tell

Fruit Carts (continuous understanding / eye-tracking testbed)
– describing out loud how to place, rotate, colour, and fill shapes on a computer-displayed map



Talking about and executing commands

Fruit Carts testbed

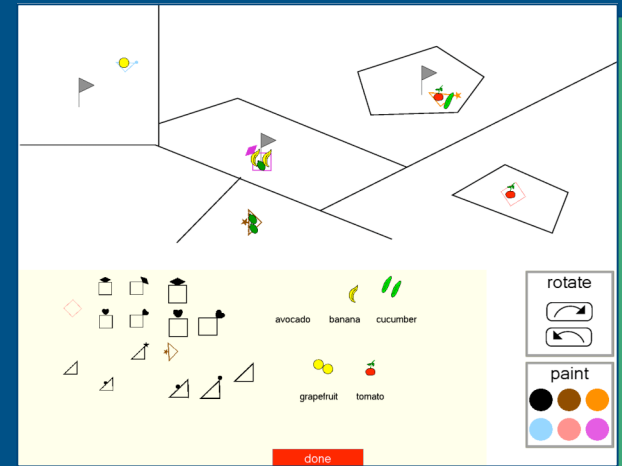
Subject (Speaker, User, Human) is given a map, and says how to manipulate objects on the screen.

Confederate (Actor, Listener, Computer) listens and acts accordingly

13 undergraduate participants.

104 sessions (digital video)

4,000 utterances (mean of 11 words per utterance).

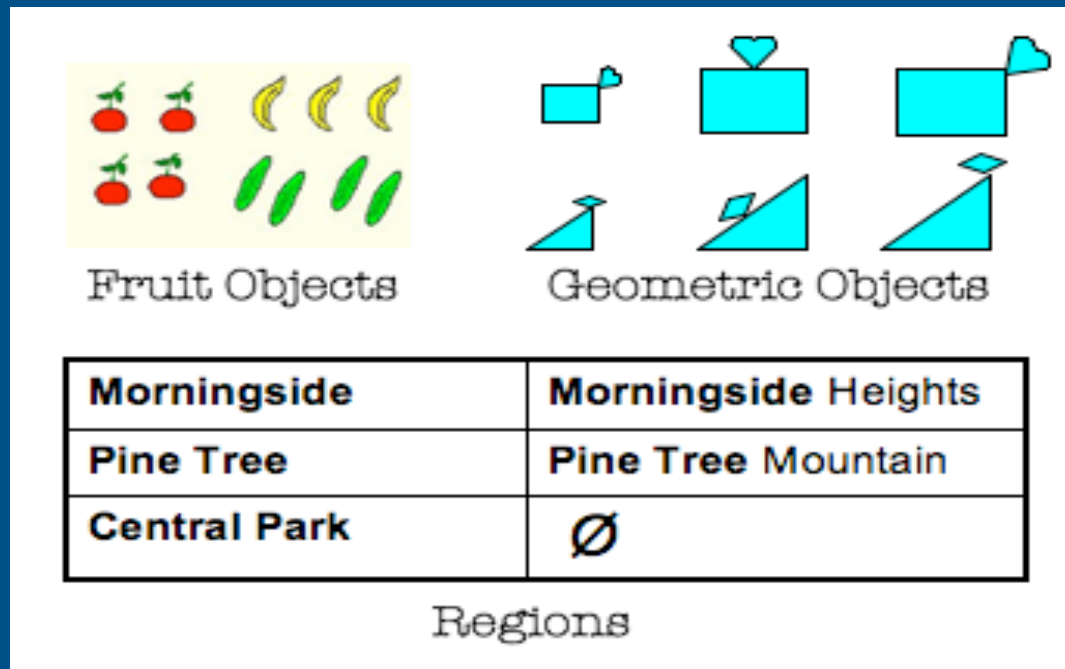


Corpus combines speech and visual modalities in a Speaker-Actor dialog and allows investigation of incremental production and understanding → Multi-modal Dialog



Fruit Carts Domain

- **Variety in actions:** MOVE, ROTATE, or PAINT objects
- **Variety in object:** contrasting features of size, color, decoration, geometrical shape and type.
- **Variety in regions:** contain landmarks and share similar names for ambiguity



Fruit Objects

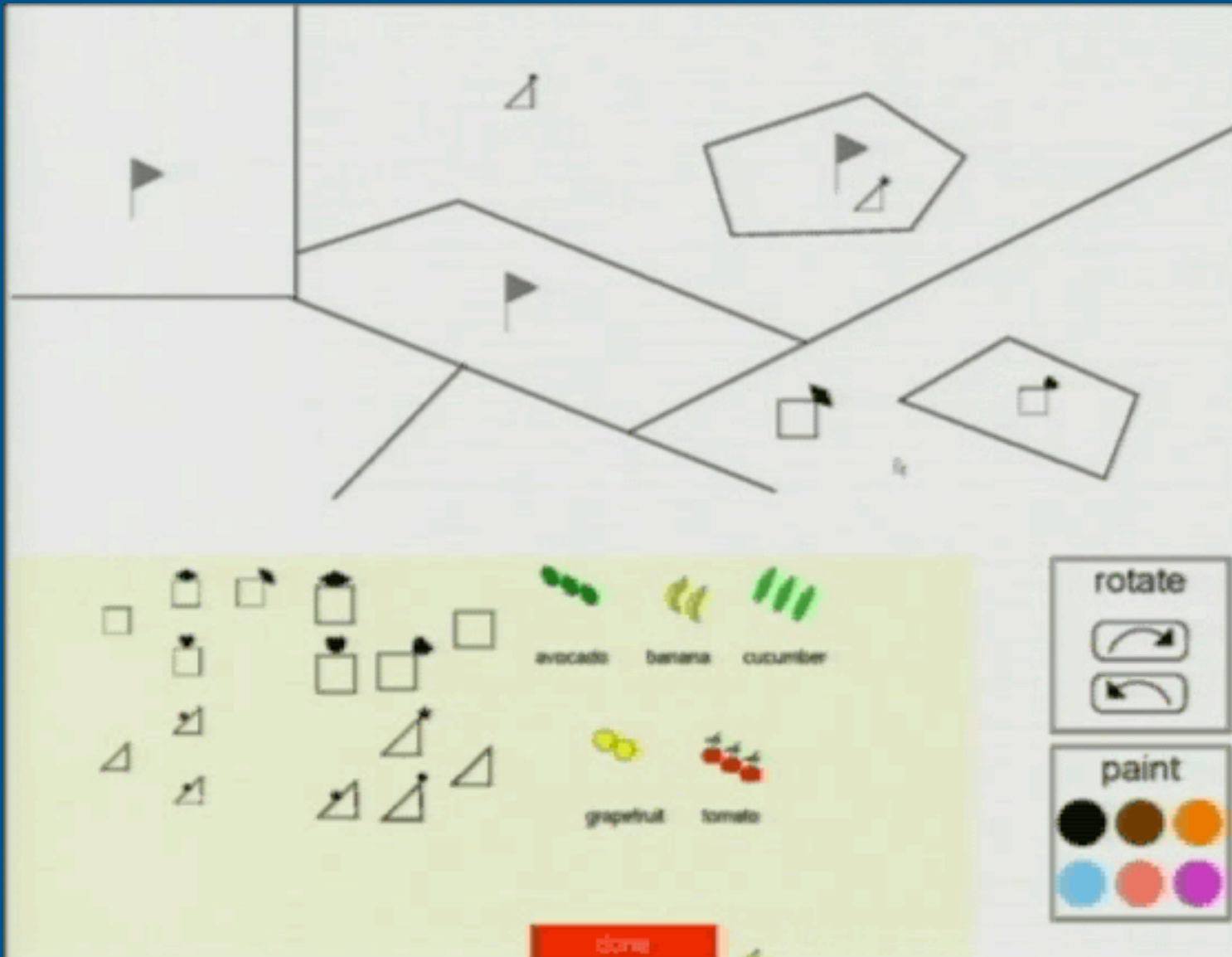
Geometric Objects

Morningside	Morningside Heights
Pine Tree	Pine Tree Mountain
Central Park	\emptyset

Regions



Fruit Carts Video



Dialog Example

SPEAKER

[ACTOR]

take the triangle with the diamond on the corner

[actor grabs object]

move it over into morning side heights

[actor moves it to region]

to the bottom of the flag

[actor adjusts location]

right there

(speaker confirms new location)

a little to the right..

[actor adjusts location]

and now a banana..

(speaker request new action)

[actor grabs object]

in ocean view..

[actor places object in location]

- Incremental production
- Non-sentential utterances
- Dynamic interpretation



Questions

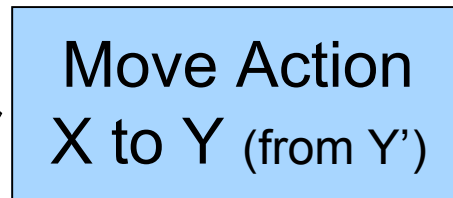
- Why do speakers decide to distribute information in multiple clauses?
- When are those ‘decisions’ made? What is the time course of such clausal planning?
- Is this behavior guided by a speaker centered model or listener center model?



Why/How speakers distribute an action across clauses

Precond's

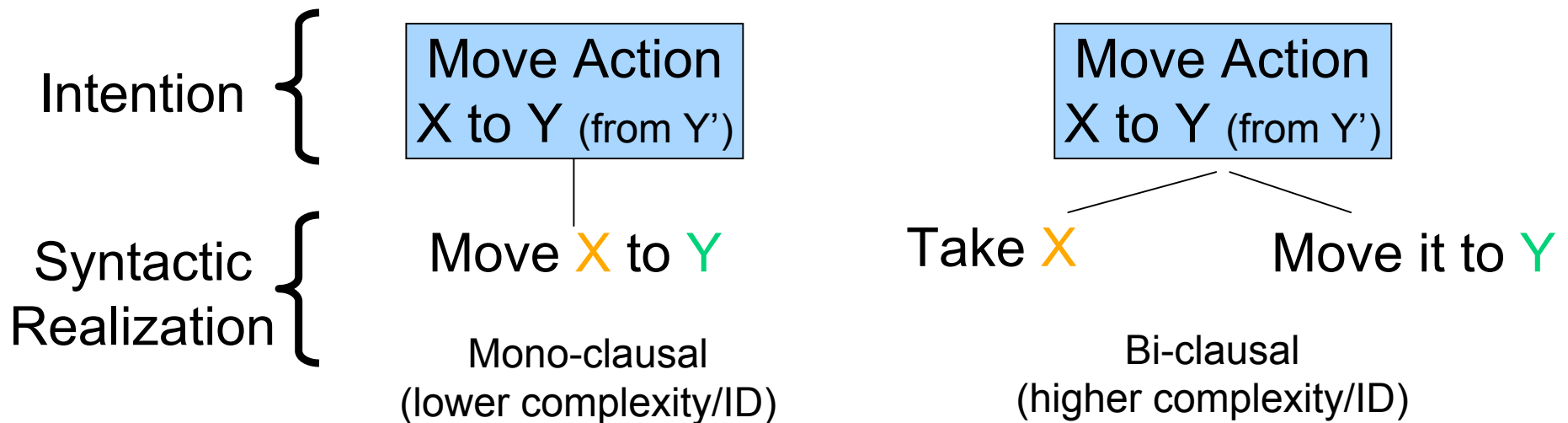
- select X
- Y is not Y'
- etc



Effects

- X is in Y (not Y')
- X is still X..
- etc

HYPOTHESIS: when a precondition has a high degree of complexity/information density(ID), speaker will produce a separate clause for it. Otherwise, speaker will tend to chunk the action in a single unit

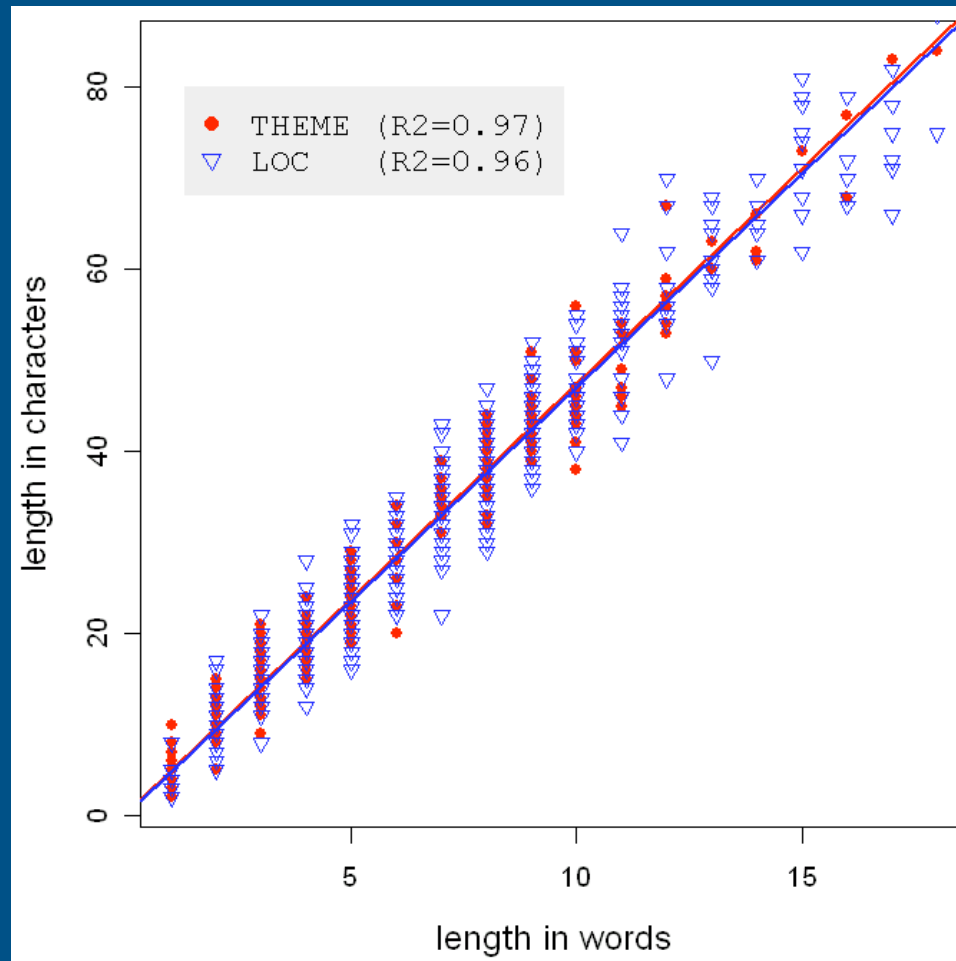


How to measure complexity?

- Semantic roles of MOVE: **theme** and **location**
- Givenness
 - New/given
- Description length:
 - Number of syntactic nodes, words, characters, syllables, moras, etc
- Presence of disfluencies and pauses:
 - “take **the [ban-] banana**”



High Correlation between word and character counts



- Number of characters, words, and syntactic nodes are highly correlated in English (Wasow, 1997; Smrecsanyi, 2004).
- Smrecsanyi (2004): word counts are a "nearly-perfect proxy" for measuring complexity.



Information Density

- Upper bound on *information or complexity* (number of words/syntactic nodes) during clause planning?
- **Uniform Information Density:** Speakers prefer a uniform amount of information per unit/time (Genzel&Charniak'02; Jaeger'06; Levy&Jaeger'06)
- We can measure information density in MOVE actions as well:
 - Event is the sequence of words that realizes a role ($w_1 \dots w_n$)
 - Information Content = $-\log P(w_1 \dots w_n)$
 - Information Density = IC / description length
 - $P(w_1 \dots w_n)$ estimated by $P(w_i | w_{i-2} w_{i-1})$ a smoothed backoff trigram model built from semantic roles extracted from Fruit Carts



How is this relevant?

- We can gain insight into how language is produced
- We can learn about the order of necessary steps in order to linearize a thought (lexical retrieval, syntactic frame selection)
- How does limited resources work such as working memory affect language production
- Only a handful of psycholinguistic studies on choice above the phrasal level (Levelt&Maassen'81; Brown&Dell'87):
What determines how speakers package and structure their message into clauses?



Gap in studies beyond the clause level

(but see Levelt&Massen'81, Dell&Brown'91)

- Most studies address issues at the phonological, lexical and intra-clausal level (Bock&Warren'85, FoxTree&Clark'97, Ferreira&Dell'00, Arnold et al'03, Jaeger'06, Bresnan et al'07, and others)
- Availability Accounts: successfully applied to choice above the phrasal level
 - NP vs. Clause conjunction (Levelt&Maassen'81)
 - “the triangle and circle went up”
 - “the triangle ... went up and the coin went up”
- Explain low lexical/conceptual accessibility of location → postpone production of location → bi-clausal realization
 - “Put an apple into Forest Hills” (Mono-clausal)
 - “Take an apple. And put it into Forest Hills” (Bi-clausal)
- Note the first conjunct is predicted not to matter (same position)
- Dell&Brown'91 discuss explicit mention of optional instruments in scene description. Their model does not make predictions on our data.



Annotation

We designed a multi-layer annotation to capture the incremental nature of this multimodal dialog (Gómez Gallo et al'07) with the annotation tool ANVIL (Kipp'04)

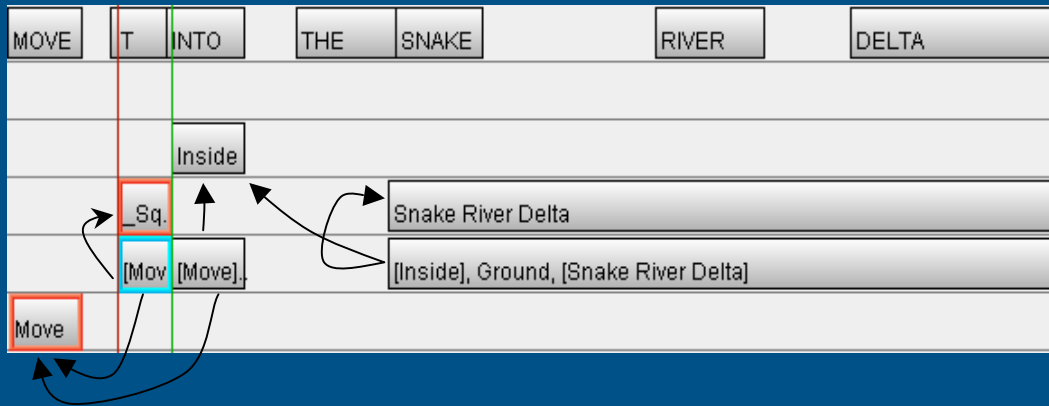
Annotation Layers: Speaker, Actor and Transaction Layers.

- The Speaker layer includes:
 - ***Object, Location, Atomic, Domain Action and Speech Acts.***
- Actor Actions include mouse movement, pointing objects, dragging objects.
- Transaction layer summarizes commitments between Speaker and Actor.

Speaker	Transcript	{text}
	Object Layer	{Anchor types}
	Location Layer	{Vertical, Horizontal, Modifiers}
	Atomic Layer	{Color, Size, Object_Ids}
	Id-Role	{Anchor, Role Type, Role Value}
	Domain Actions	{Actions}
	Speech Act	{Speech Act, Speech Act Content}
Actor		{Actor Actions}
Transactions		{Transaction Summary}

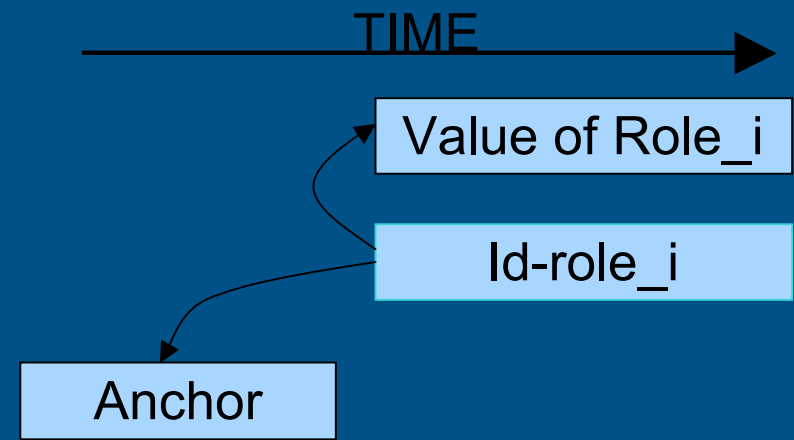


Annotating Incremental Understanding



Annotation Principles

1. Annotation is done at the word level
2. Annotation is done in minimal semantic increments
3. Semantic content is marked at the point it is disambiguated without looking ahead
4. Reference is annotated according to speaker's intention



Id-role: a speech act that identifies a particular relationship (the role) between an object (the anchor) and an attribute (the value).

This construct is used for incrementally defining the content of referring expressions, spatial relations and action descriptions.



Data

- So far: 1,100 MOVE and SELECT actions and their labeled semantic roles (**theme**, **location**)
- Of these, ~600 utterances are elaborations on a prior MOVE (e.g. “a little bit to the left”)
- Excluding elaborations, ~300 mono/bi-clausal MOVE actions



Data Analysis

- Mixed logit model predicting choice between mono-/bi-clausal realization based on:
 - **Theme**
 - Information Density
 - Givenness (*explicit vs. implicit mention vs. set vs. new*)
 - Log length (in words)
 - Pauses
 - Disfluencies: editing, aborted words
 - **Location**
 - Information Density
 - Log length (in words)
 - Pauses
 - Disfluencies: editing, aborted words



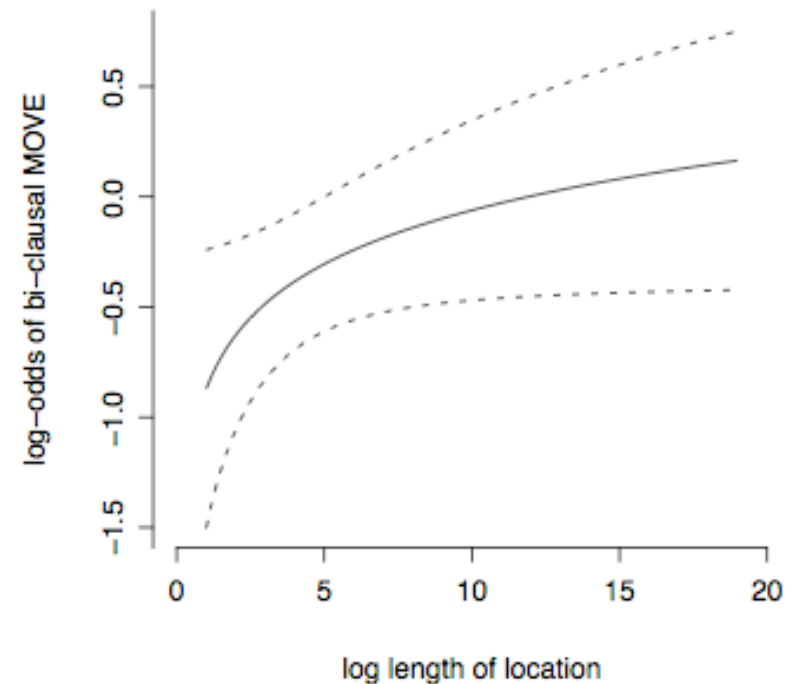
Results: Location

Speakers preferred a bi-clausal with:

- **disfluent locations** ($\beta=0.64$; $p<0.007$)
Significant Effect
- **location length** only marginal effect
when ID not included in the model
- **No other location effects** reached
significance

→ “Take **an apple**, .. and..
Move .. **it** .. into **Forest Hills**”

**This effect is explained by Availability-
based Theories**



Results: Theme

Speakers preferred:

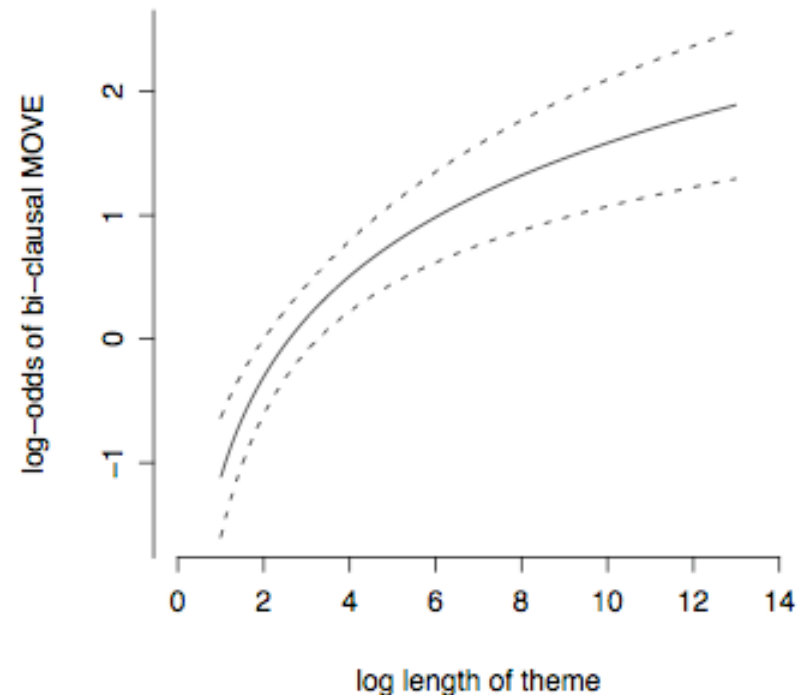
- bi-clausal with:
 - Longer themes ($\beta=2.01$; $p<0.0001$)
 - Higher ID themes ($\beta=1.58$; $p<0.003$)
 - New themes ($\beta=1.8$; $p<0.0002$)
- mono-clausal with:
 - Disfluent themes ($\beta= -0.79$; $p<0.007$)

No other theme effects reached significance

Unexpected for Availability-Based accounts:
Mono/Bi clausal plan has the same theme position

→ Bi: “Take an apple,”

→ Mono: “Move an apple there”



Most **theme** measures correlate with bi-clausal plan ...

- Except for.. The presence of disfluencies in object descriptions are positively correlated with single chunk actions.
- Unexpected.. But this may have something to say about the cognitive load in incorporating multiple semantic roles in one single chunk...
 - Single-chunk: move [a [ban--] banana] to Y
 - Two-chunk: take a banana, move it to Y
- Gibson'91 shows how people minimize long distance dependencies favoring certain parses during comprehension



Discussion: When do speakers decide on a production plan?

- When is the choice for a mono/bi-clausal structure made?
- Most cases in our database begin with the verb
- Hence there are two facts:
 - 1) **Theme** complexity and ID
 - 2) Verb distribution asymmetry

1st Verb	Mono-clausal	Bi-clausal
<i>take</i>	0%	73%
<i>move</i>	28%	0%
<i>put</i>	27%	1%
<i>be</i>	43%	7%
others	2%	19%



Discussion – Time course

- If speakers have access to complexity estimate early, before lexical selection, and before thematic assignment, both facts (1+2) are accounted for
- Otherwise, the complexity~clausal choice correlation does not follow from verb distribution asymmetry alone



Conclusions

- Fruit Carts Resource:
 - Fruit Carts is a multi-modal dialog corpus with rich features in domain objects, regions, and actions
 - Eye tracking and Semantic annotation at the word level
 - New resource to study language understanding and production
- Language Production Results
 - Speakers are sensitive to the complexity and information density of a clause
 - Speakers have *early access* (**prior to lexical selection and prior to functional encoding**) to *some* measure of complexity of overall clausal complexity



Future work

- More data:
 - Continue data annotation
 - Additional data, gathered by Susan Wagner-Cook et al, may provide further evidence for the presented effect.
 - “Simon has *a red striped bag*. He gave *it* to *the woman on the left*”
 - “Simon gave *the bag* to *the woman*”
- Further questions:
 - Is it information or complexity speakers keep uniform?
 - Analysis of disfluency effect due to unexpected direction
 - Is this effect due to speaker centered model or listener?
 - ONLY MOVE ACTIONS..! How about rotate
- Analyze eye tracking data to see how it can explain mono- versus bi-clausal production



Questions?

- Thanks for listening..



Studies at different levels of representation

- When speakers translate an intended message into an utterance, there are many choice points at many levels of production
 - **Phonological and phonetic level**
“thee” vs. “the” (Arnold,Fagnano&Tanenhaus’03; Fox-Tree&Clark’97)
 - **Word level**
“How big is the family (that) you cook for?” (Ferreira&Dell’00; Jaeger’06)
 - **Phrasal level**
“She gave {him the key/the key to him}” (Bresnan et al.’07; Givon’84)
Active vs. Passive (Bock&Warren’85; Prat-Sala&Branigan’00)
“She stabbed him (with a knife)” (Dell&Brown’91)



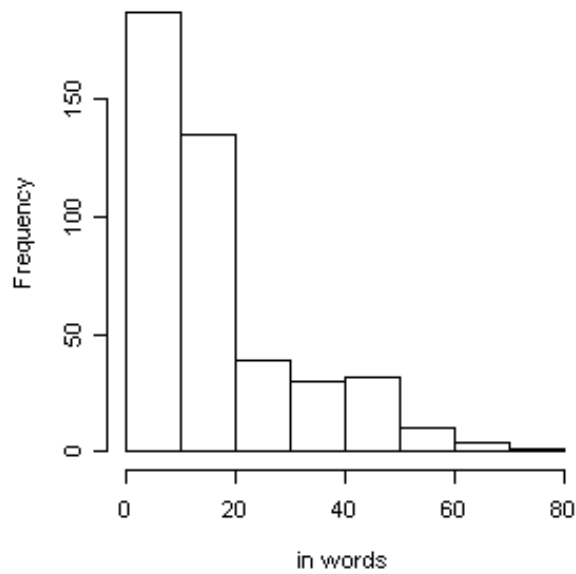
DEPTH FIRST LINEARIZATION STRATEGY

- When Speech Act preconditions are complex, then bi-clausal realization chosen.
 - GIVE-ACTION: Precondition: HAVE
 - MOVE-ACTION: Precondition: SELECT
 - Humans have a Depth-first problem-solving strategy (Newell&Simon'72)
- Computational Model for mono/bi-clausal choices

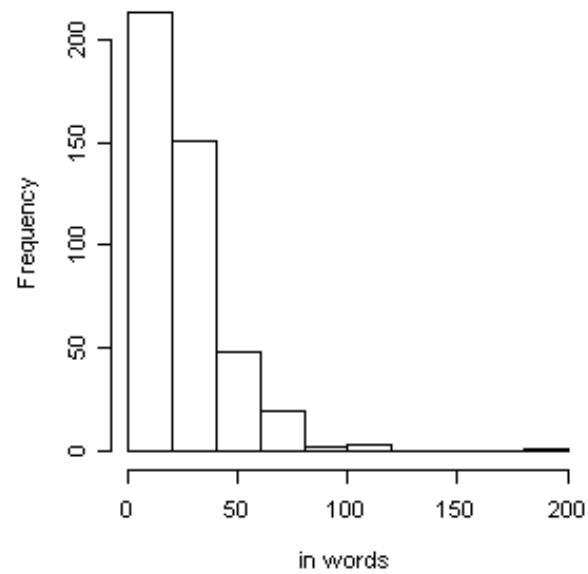


Overall distributions (histograms)

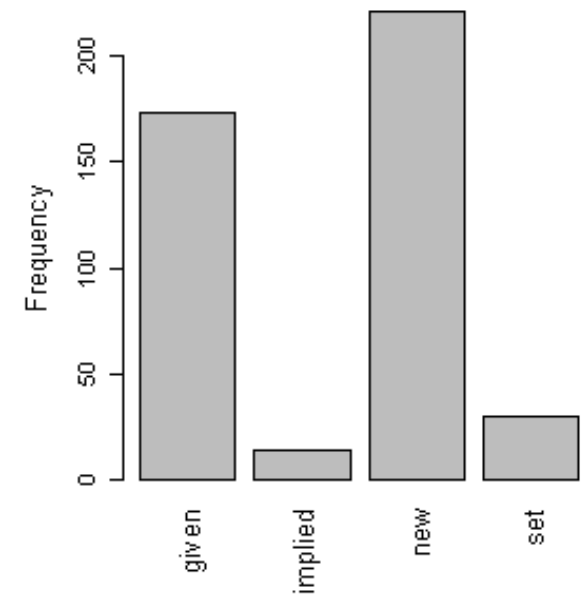
Length of theme



Length of location



Givenness of theme



Elaborations

- “Move it to central park”
 - “a little more to the right”
 - “a little more”
 - “sorry.. not that far”



Disfluencies annotated in both theme and location

- Words repetition: “*take the [the] square ...*”
- Aborted words: “*move the [ban-] tomato...*”
- Pauses: “*the square <pause> with a heart...*”
- Ignored restarted acts: “[*move ah.. I mean..*] take the..”



Additive effect between other semantic roles

- P values for model with both goal and theme added
- Intra Clause Message with Long Theme Description
 - Add *two bananas and a tomato inside of it*
- Inter Clause Message with Short Theme Description
 - Take *one tomato*
 - Put *it in the center of that triangle*
- These results suggest that the complexity of the semantic roles add up to the complexity of the message



Limited Resource Account

- We hypothesize that not only availability, but the overall complexity and information of a clause determines clausal structure.
- **Hypothesis:** Speakers prefer to keep the amount of information per clause uniform

