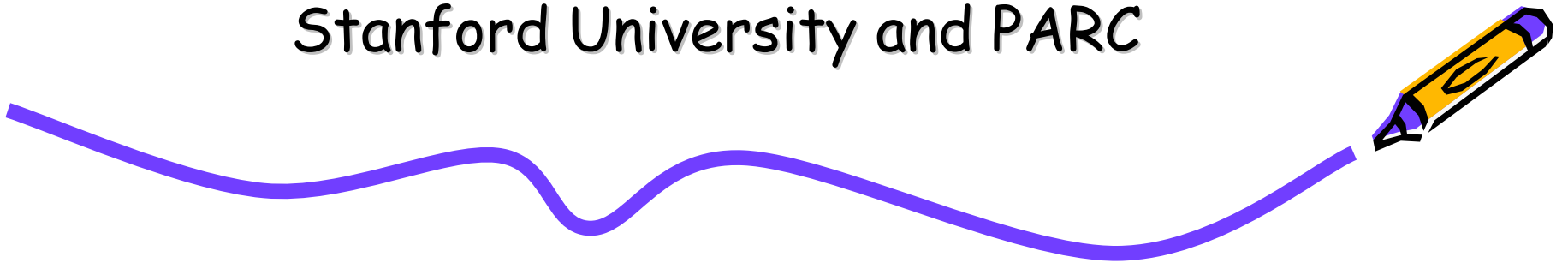




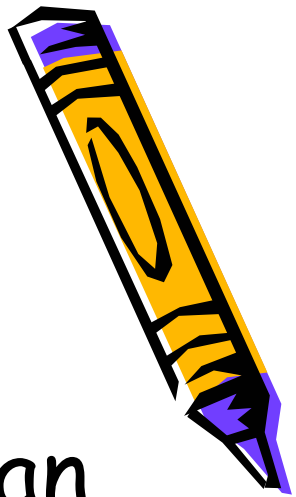
# Boundaries to the Influence of Animates

Anna Cueni, Neal Snider, Annie Zaenen  
Stanford University and PARC



# Question

- What determines how we express an idea
  - More specifically: what determines the choice between syntactic paraphrases
    - More specifically: what is the role of animacy in the syntactic realization of an NP



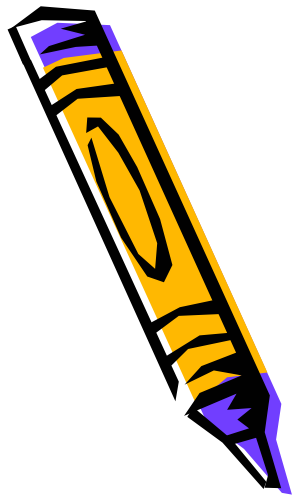
# Outline



- Existing hypotheses
- Our studies
  - 'Topicalization' and left dislocation
  - Subjects and adjuncts
- Hypotheses compatible with the English data
- Summary of some German and Spanish data
- Combined picture
- Consequences for generation models
- General conclusion



# Outline



- Existing hypotheses
- Our studies
  - 'Topicalization' and left dislocation
  - Subjects and adjuncts
- Hypotheses compatible with the English data
- Summary of some German and Spanish data
- Combined picture
- Consequences for generation models
- General conclusion



# Argument Salience and Syntactic Realization



## *Salience influences*

- the grammatical function realization of a semantic argument following the "Accessibility Hierarchy": SUBJ > OBJ > OBJ<sub>θ</sub> > OBL (Hypothesis 1)
- the linearization of arguments directly (Hypothesis 2)

*Salience* is an amalgam of functional factors such as animacy, information status. There is only one notion of salience (the various factors that compose salience are combined in fixed proportions for all phenomena that depend on it).



# Data compatible with both hypotheses



- The choice between passive and active voice in English.
  - E.g Prat-Sala, 1997: descriptions of pictured scenes

	Animate patient	Inanimate patient
passives	78%	43,5%

- See also Bock et al. 1992, McDonald et al. 1993
- Animates in the dative constructions (Cueni & Bresnan, 2005)



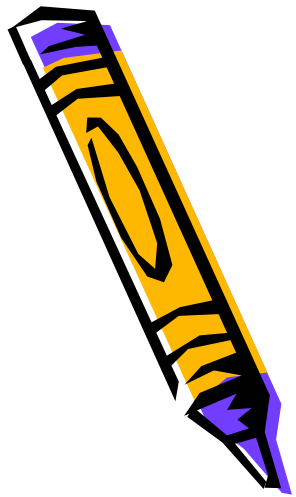
# Ways of distinguishing between the two hypotheses



- Look at other languages where word order precedence does not correlate with the GF hierarchy
- Look at cases in English where elements come earlier in a sentence without being higher on the GF hierarchy.



# Outline



- Existing hypotheses
- Our studies
  - 'Topicalization' and left dislocation
  - Subjects and adjuncts
- Hypotheses compatible with the English data
- Summary of some German and Spanish data
- Combined picture
- Consequences for generation models
- General conclusion



# Two different kinds of English data

- Comparison of Left Dislocations and 'Topicalizations' with in situ arguments
- Linearization of adjuncts and arguments (subjects)
- Both studies are done on part of the Switchboard, annotated for animacy and information status. Most of the annotations were done in Edinburgh-Stanford Paraphrase-LINK project (Nissim et al, 2004, Zaenen et al, 2004)



# Left Dislocation and 'Topicalization'

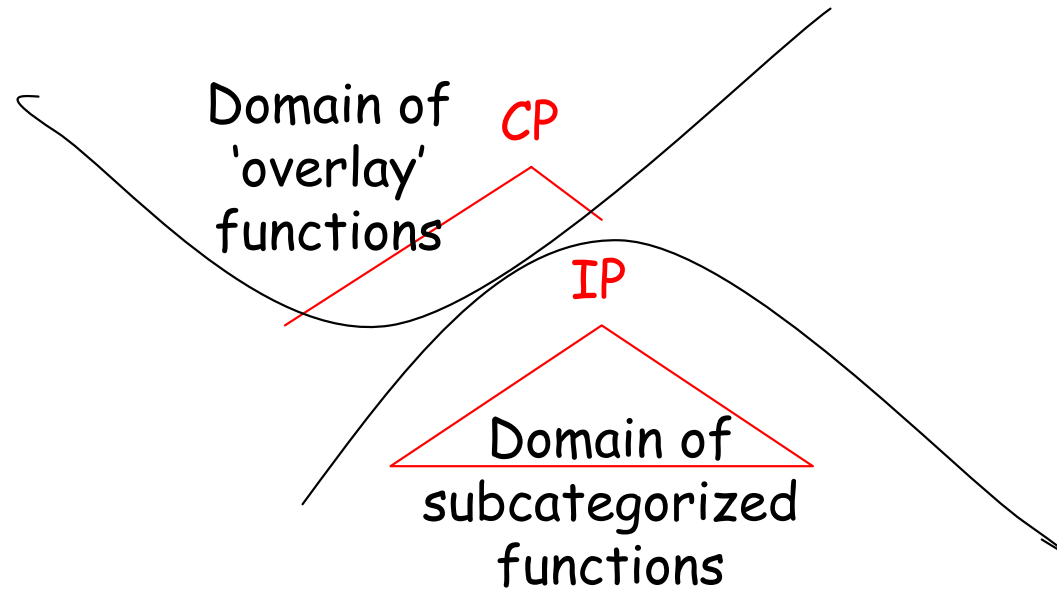


Annotated SWB:

'topicalizations' n=30 Beans, I like.

left dislocations n=123 That guy, I saw him last week  
in the store.

Structural assumptions (English)



# Left Dislocation and 'Topicalization' (1)



- For each LD, TPZ, and 5800 control (argument and adjunct) NPs:
  - Extracted animacy (human, organization, animal, place, time, concrete entity, nonconcrete entity)
  - Extracted information status:
    - Old - identity, relative, generic, generic coreferential, general, event
    - Inferrable - bound, general, event, aggregation, function value, set, possessive, part, situation
    - New



# Left Dislocation and 'Topicalization' (2)



- Grammatical function - subject and non-subject
- Grammatical weight - number of words
- Speaker ID



# Left Dislocation and 'Topicalization' (3)



- A mixed-model linear-logistic regression was performed for each construction
- Used the aforementioned factors (speaker id was taken as a random factor)
- Predicted the presence or absence of the construction (LD or TPZ)



# Left Dislocation and 'Topicalization' (4)



- All factors *except animacy* were significant for 'Topicalization' and Left Dislocation. ( $p > 0.5$ )



# Arguments and Adjuncts (1)



- Hypothesis: If animacy has a direct effect on linearization, we expect adverbs to be postponed more frequently to allow an animate subject to occur sentence-initially.
- Study: Subset of instances in which two variants are truth-conditionally equivalent: temporal adverbial phrases without scoping

Example: I met a girl from Boston one time  
One time I met a girl from Boston



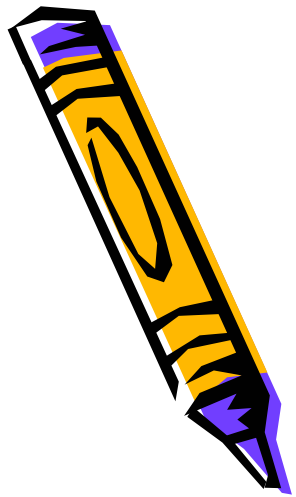
# Arguments and Adjuncts (2) : Results



- When the subject is animate, the adverb is postponed 69.5% of the time. When the subject is inanimate, the adverb is postponed in 68.9 % of the time. (not significant by Fisher's Exact Test)
- Looking only at lexical subjects (no pronouns): When the subject is animate, the adverb is postponed 76.7 % of the time (n=124). When the subject is inanimate the adverb is postponed 88.2 % of the time (n=170).  $p = .01$  by Fisher's Exact Test.
- No evidence for an animate-first effect. In fact there is a small anti animacy effect.



# Outline



- Existing hypotheses
- Our studies
  - 'Topicalization' and left dislocation
  - Subjects and adjuncts
- **Hypotheses compatible with the English data**
- Summary of some German and Spanish data
- Combined picture
- Consequences for generation models
- General conclusion



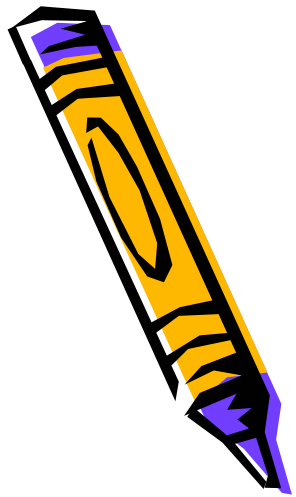
# Hypotheses compatible with the new English data



- The influence of animacy on word order precedence only counts in the IP domain.
- Only the hierarchy of subcategorized GFs (SUBJ<OBJ<OBJ<sub>θ</sub>) is relevant for animacy effects.



# Outline



- Existing hypotheses
- Our studies
  - 'Topicalization' and left dislocation
  - Subjects and adjuncts
- Hypotheses compatible with the English data
- **Summary of some German and Spanish data**
- Combined picture
- Consequences for generation models
- General conclusion



# Some German data



- Kempen-Harbusch study of Negra corpus:
  - Corpus: around 20,000 newspaper syntactically annotated sentences
  - 1168 pairs of argument orderings (distinguished by function, SUBJ, OBJ, IOBJ, and full or pronominal NP status) in *embedded clauses*
  - Results:

Linear order:	SBF<DOP	DOP<SBF	Linear order:	IOF<SBF	SBF<IOF
SBFI	11	64	IOI	3	39
SBFA	52	56	IOA	17	20

- The domain-based hypothesis is not contradicted by the German data (*reordering in embedded clauses is A-movement*) but the simple GF hierarchy one is.



# Some Spanish and Catalan data



- Prat-Sala (1997) Elicitation study; data for Spanish

- A un hombre le golpea una pelota
- To a man him hits a ball

(Prat-Sala gives a slightly different version, in fact a relative clause)

Animate	patient	inanimate	patient
passive	dislocation	passive	dislocation
27,5 %	6,5 %	17,5 %	1,5 %

- Prat-Sala also looked at cases where the subjects described the situation with a non-passivizable predicate; in that case too, there were more dislocations with the animates than with the inanimates



# CLLD and LD are different



- Pragmatic difference: shown by the data itself (no dislocations were produced for English), see also Escobar (1997), who describes the following as a felicitous context for CLLD

- Has visto a María últimamente?  
Did you see Mary recently?
- A María, no la he visto desde hace tiempo  
%Mary, I haven't seen her for a long time.  
?Mary I haven't seen for a long time.

A constituent in CLLD position seems to be old information in Spanish, in English that is not the function of LD or 'topicalization' (see e.g. Prince, 1995) but often the function of the subject.



# CLLD and LD : Syntactic Differences



- No connectivity phenomena ~ Connectivity phenomena
  - Case/pied-piping
    - A Juan lo conozco (Escobar, 1997)  
John, I know him  
\*To John, I talked to him
  - Reflexives
    - Nadie baila con su propia hermana.  
Nobody dances with his own sister.
    - Con su propia hermana nadie baila.  
With his own sister nobody dances.
    - María no va a testificar contra si misma.  
Mary will not testify against herself.
    - Contra si misma María no va a testificar.
- Root ~ non-root phenomenon
  - Piensa que a Juan lo llamé María  
\* He thinks that John, Mary called



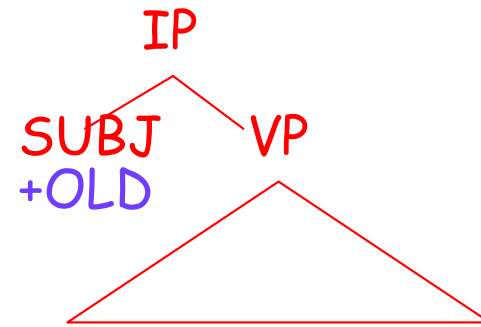
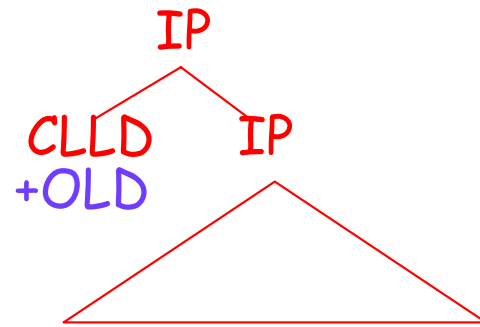
# CLLD and LD: Structural differences



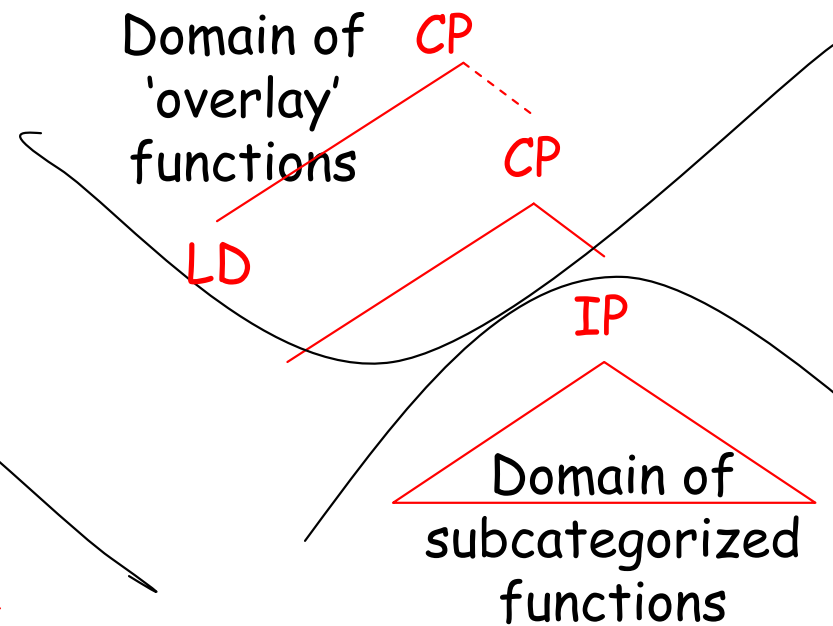
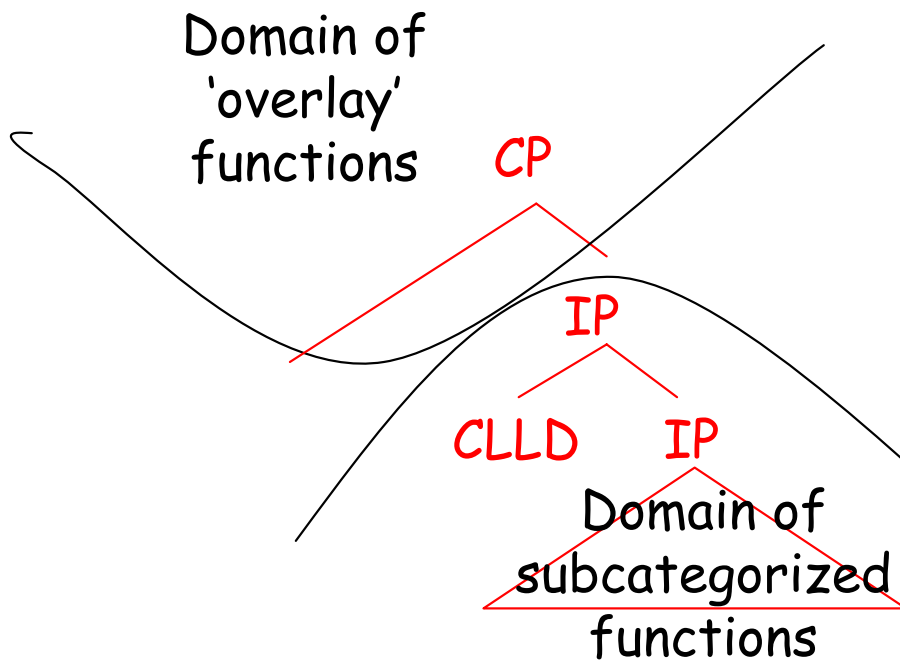
- Structural difference: IP instead of CP, as argued by e.g. Anagnostopoulou, 1997, for Greek and Italian
- Functionally, the clitic is an agreement marker and it is the CLLD element that is the real argument.



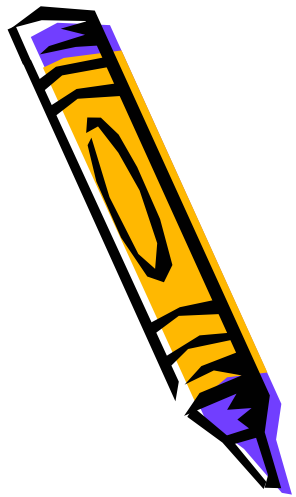
# Discourse mappings for Spanish and English: an hypothesis



# Structure of CLLD and LD



# Outline



- Existing hypotheses
- Our studies
  - 'Topicalization' and left dislocation
  - Subjects and adjuncts
- Hypotheses compatible with the English data
- Summary of some German and Spanish data
- **Combined picture**
- Consequences for generation models
- General conclusion



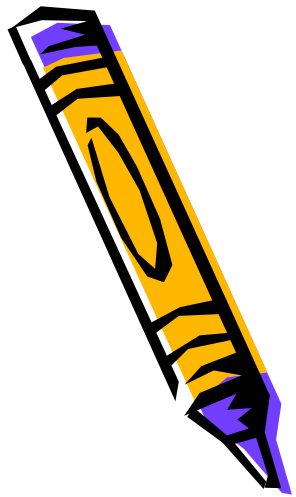
# Combined Picture



- Animacy influences linearization of arguments in the IP domain but the linearization doesn't follow the GF hierarchy (Conclusion based on German and Romance data).
  - Complication: Cross linguistically, animacy plays a role in GF mapping independent from linearization (morphological marking, grammaticality constraints, etc.)
- Outside of the IP domain, animacy might not play a direct role (Based on our English data)
  - Complication: these non-IP elements precede the IP in surface realization



# Outline

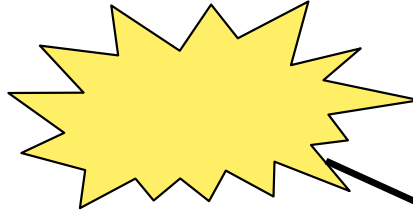


- Existing hypotheses
- Our studies
  - 'Topicalization' and left dislocation
  - Subjects and adjuncts
- Hypotheses compatible with the English data
- Summary of some German and Spanish data
- Combined picture
- **Consequences for generation models**
- General conclusion



# Generation model

Concepts

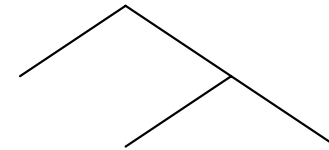


Animacy information

Functional Structure

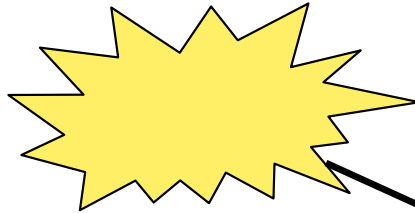
PRED  
SUBJ  
OBJ

Constituent Structure  
linearization



# Generation model

Concepts



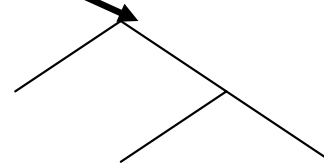
Animacy information

Functional Structure

PRED  
SUBJ  
OBJ

Animacy information

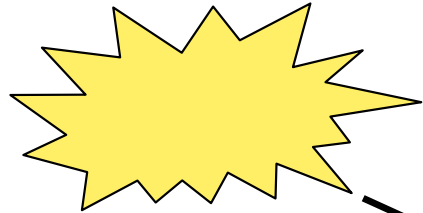
Constituent Structure



# Generation models



Concepts

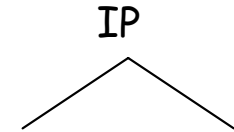


Animacy information

Thematic roles: CS

Subcategorisation realization  
All prominence information

PRED  
SUBJ  
OBJ

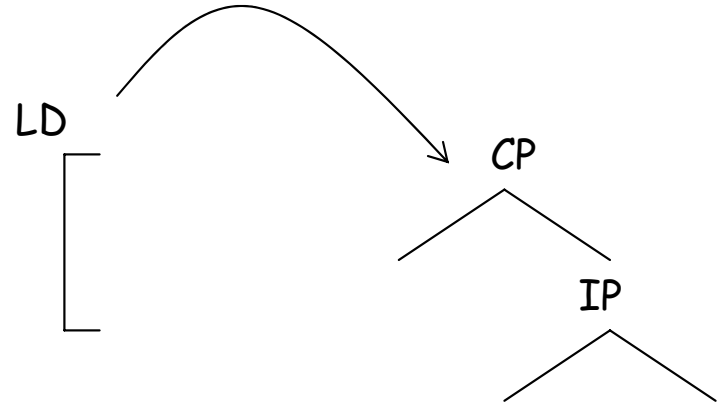


Discourse realization  
discourse prominence information

LD

CP

IP



# What is at issue



- How much processing gets done before linearization/vocalization starts?
- The new model is not compatible with early vocalization



# Caveats



- Corpora statistics extrapolate from a very small amount of data
- Different tasks might lead to different results
- Different statistical analysis methods



# Outline



- Existing hypotheses
- Our studies
  - 'Topicalization' and left dislocation
  - Subjects and adjuncts
- Hypotheses compatible with the English data
- Summary of some German and Spanish data
- Combined picture
- Consequences for generation models
- **General conclusions**



# Conclusions



- Linearization preferences are sensitive to syntactic domains
- Saliency is not a unitary notion: different forms of saliency can play a role in different domains or at least the proportion in which the various factors that make up 'saliency' play a role seems to be different in different domains.



# Further work



- Do a uniform study of the importance of animacy over the CP and the IP domain for English
- Similarly for languages like German and Spanish.
- See how material that is extraposed to the right behaves.

