

A Corpus Study of Left-Dislocation and Topicalization

Neal Snider

April 8, 2005

Introduction

- How to choose among truth-conditionally equivalent constructions?
- Left-dislocation construction (LDC):
Burlington's crime_{*i*}, **it**_{*i*} doesn't involve children.
- Topicalization construction (TPZ):
And **that**_{*j*} I couldn't watch []_{*j*}

Factors

- Which factors are significant in differentiating these constructions?
- Hypotheses from the literature:
 - General
 - Specific to these constructions

General Factors

- Animacy (e.g. Dative alternation Bresnan and Nikitina 2003)
- Grammatical Weight (e.g. Heavy NP Shift, Wasow 1997)

Hypotheses specific to LDC and TPZ

- Information status factors (Prince 1992,1995,1998)
- 3 disjoint functions for LDC:
 - Simplifying LDC
 - P(artially) O(rdered) Set LDC
 - Island Topicalization LDCs

Prince's functions of LDC: Simplifying LDC

- removes fronted NPs that refer to discourse-new entities from a syntactic position (subject) that disfavors them
- 'Two of my sisters were living together on 18th Street. They had gone to bed, and this man, their girlfriend's husband came in. He started fussing with my sister and she started to scream. **The landlady_i, she_i went up**, and he laid her out. So my sister went to get a wash cloth to put on her, he stabbed her in the back...' Terkel, *Welcomat*, 12/2/81, p.15
- predicts that there should be more discourse-new left dislocations from subject

Prince's functions of LDC: Poset LDC

- triggers the hearer to infer that the entity to which the fronted NP refers is in a salient 'partially-ordered set' relation to some previous entity in the discourse
- A: I would like to be a little more into investigating some of the other countries in the world and their educational problems. And to come up with something a little better than what we've got.
B: Uh-huh. Yeah, it's tough to, to say what, uh, you know, what, uh, as far as this, that good or bad or what.
A: But , uh, I was just talking to somebody else, and **all those European countries_i, they_i** pay all the way through college and stuff like that.
- predicts more set relations between left-dislocated NPs and previous discourse

Prince's functions of LDC: Island Topicalization LDCs

- topicalization from extraction islands surfaces as LDC
- 'My first book_i, I paid half of each trick to the person who gave it_i to me.'
(Terkel)
- none in Switchboard corpus

Prince's functions of TPZ

- Poset inference
- Focus-Focus Frame

Prince's functions of TPZ: Focus-Focus Frame

- the focus is the prosodically prominent constituent within the clause that follows the fronted NP
- focus frame is the rest of the clause, with the focussed constituent replaced by a variable
- 'She had an idea for a project. She's going to use three groups of mice. One, she'll feed them mouse chow, just the regular stuff they make for mice. Another, she'll feed them veggies. And **the third_i, she'll feed e_i junk food.**'
- predicts a strong correlation between the accented constituent following the fronted NP and discourse newness

Hypotheses specific to LDC and TPZ

- Referential Distance (Givón 1983)
 - LDC ‘re-introduces’ a topic
 - There once lived a **gracious king** in an enchanted forest. He was married to a beautiful queen, and she wasn’t only beautiful but also smart, so she soon became the real power in the realm. In a forest clearing near the palace there lived a poor prince, and the queen used to visit him and have lunch. **Now the king, he** didn’t like that one bit...
 - predicts referential distance should peak at higher number of utterances back, like Givón found for a smaller corpus (11-20 utterances back)

The Corpus

- Treebank Switchboard - syntactically annotated
- annotated for information status by Paraphrase-Link project (Nissim et al, 2004)
 - **old**: identity, event, general, generic, identity-generic
 - **mediated**: general, bound, possession, part, situation, event, set, func-value, aggregation
 - **new**

- also annotated for animacy (Zaenen et al, 2004)
human, organization, animal, machine, vehicle, place, time, concrete, nonconcrete

Data Extraction

- used “tgrep2”
 - 121 LDC in annotated part of corpus (406 total)
 - 29 TPZ in annotated part (104 total)
- Factors:
 - those mentioned above (referential distance was hand coded on a smaller sample)
 - grammatical function (of gap or resumptive pronoun)
 - speaker
- also extracted control NPs, which were in S’s that were not in questions, topic-constructions (5750)

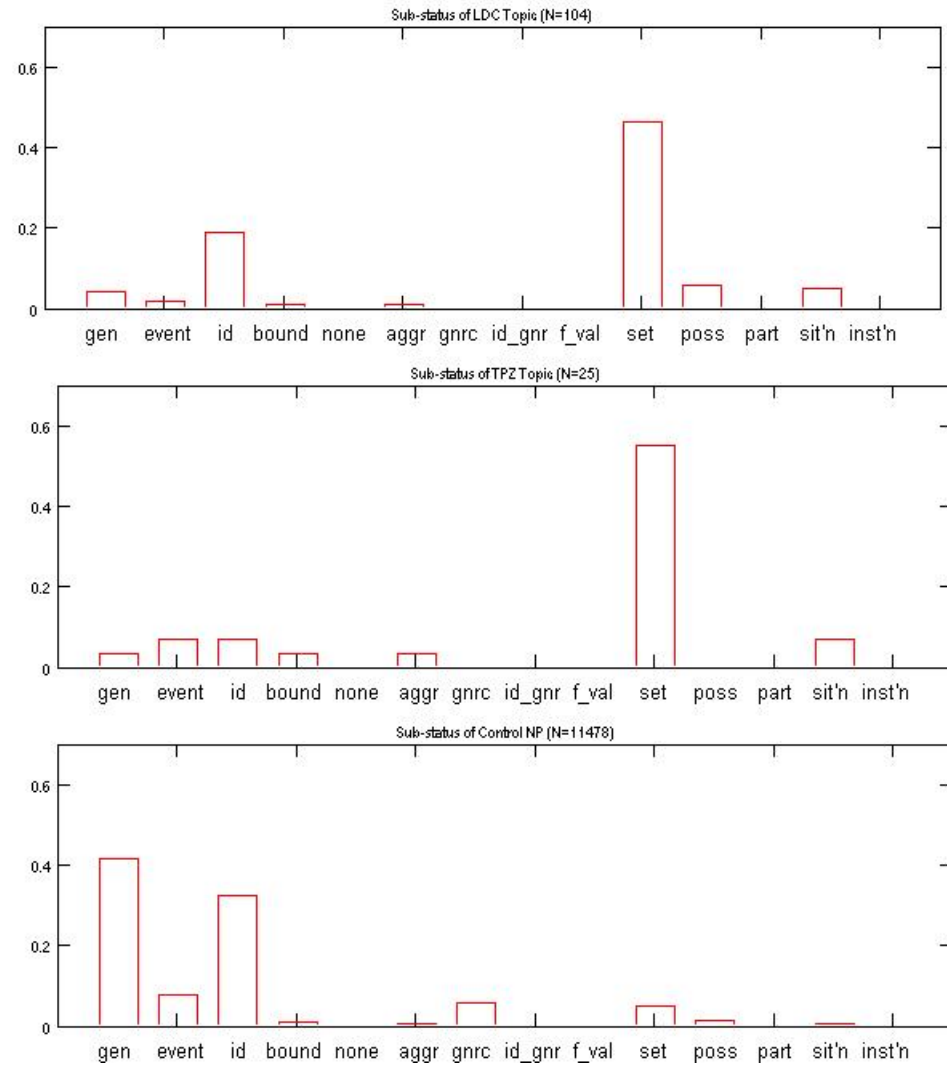
Data Analysis

- mixed-model Logistic Regression with speaker as random factor
- significance tests by ANOVA
 - LDC: animacy significant at $p < 0.05$, all other factors significant at $p < 0.001$
 - TPZ: animacy, information status, weight significant at $p < 0.01$, GF at $p < 0.001$

Results: Information status categories

- set relation is by far the most common for LDCs (46.3%) and TPZs (55.2%), as compared to the controls (5.2%)
- logistic regression: LDC is 80% more likely to be in a set relation than others, verifying Prince's POSet function
- TPZ is 1.6 times more likely to be in a set relation than others

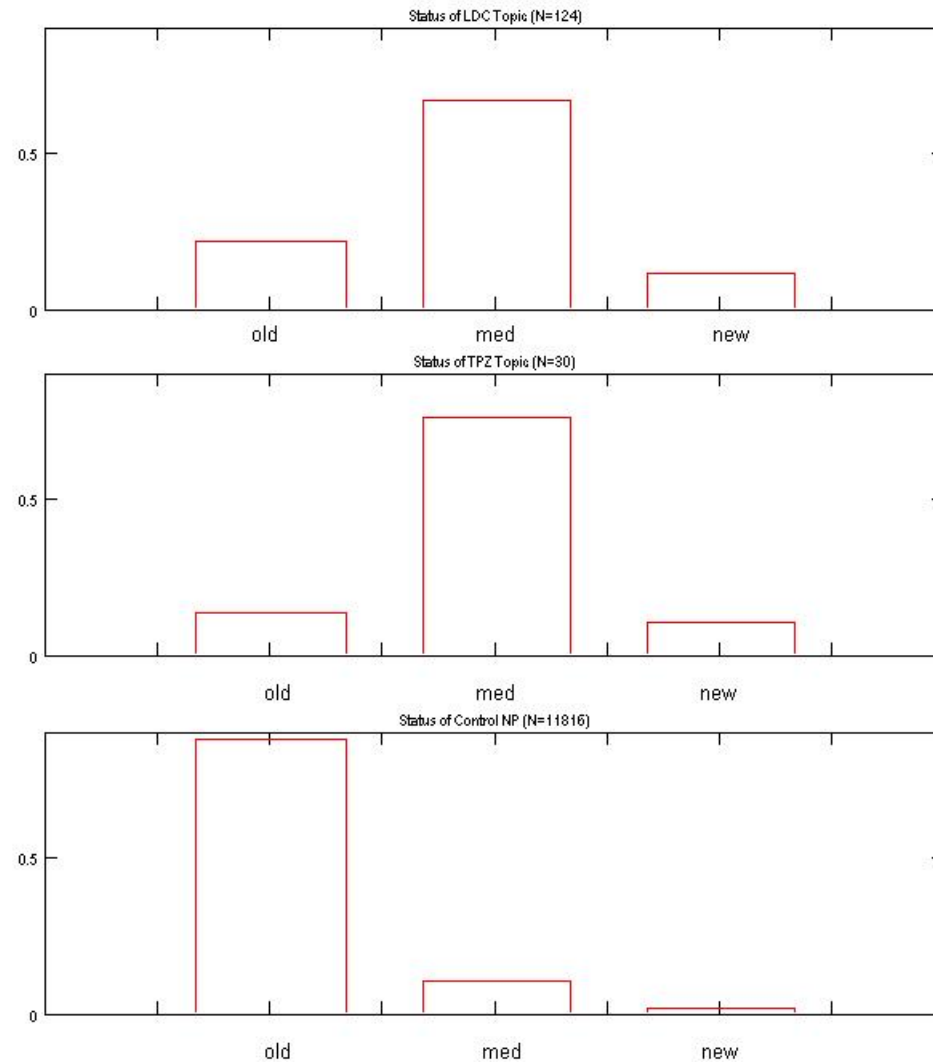
Results: Fine information status categories



Results: Information status categories

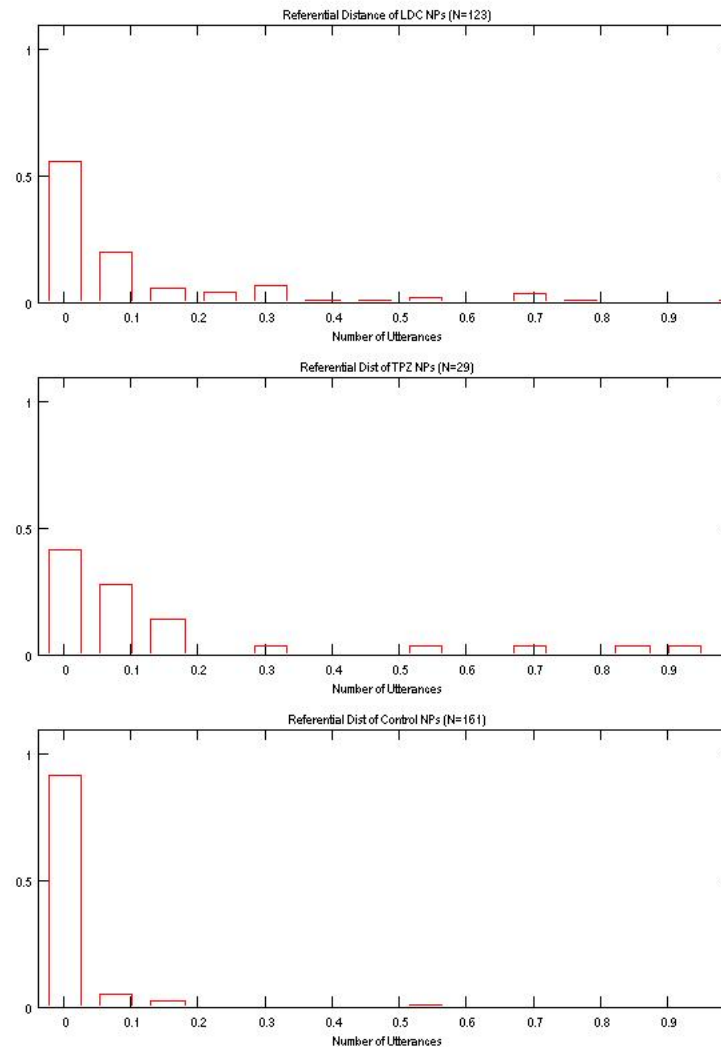
- Prince's 'Simplifying' function predicts that there should be more discourse-new left dislocations from subject
- also verified: χ^2 , Fisher's exact tests $p < .0001$
- but vastly fewer new LDCs than old or mediated, so function is quite small

Results: Coarse information status categories



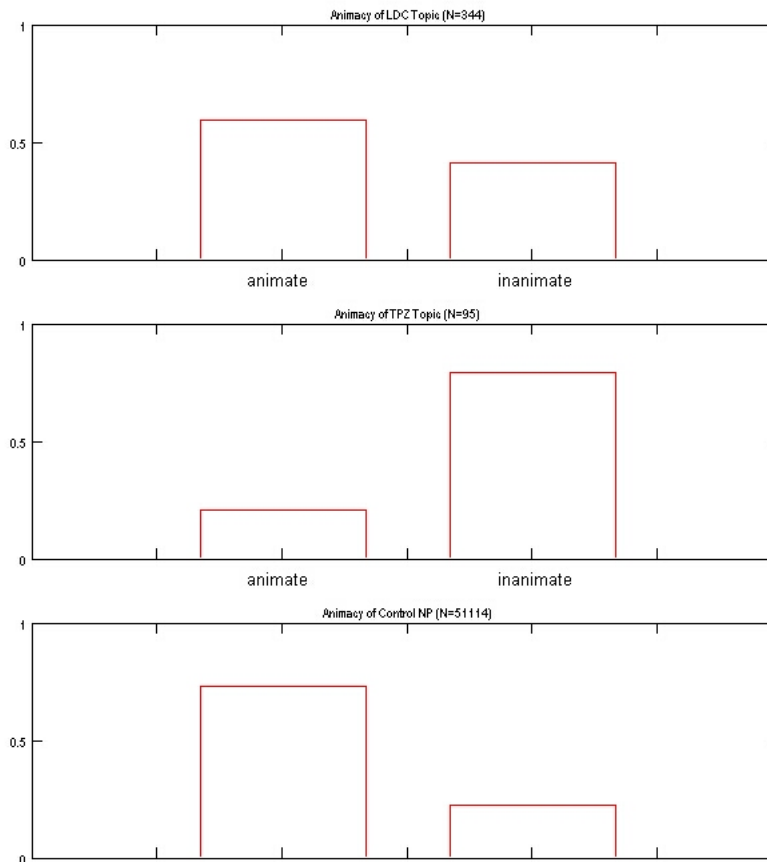
Results: Referential distance

LDCs and TPZs tend to have been last referred to further back in the discourse than control NPs, but not the qualitative difference Givón predicts.



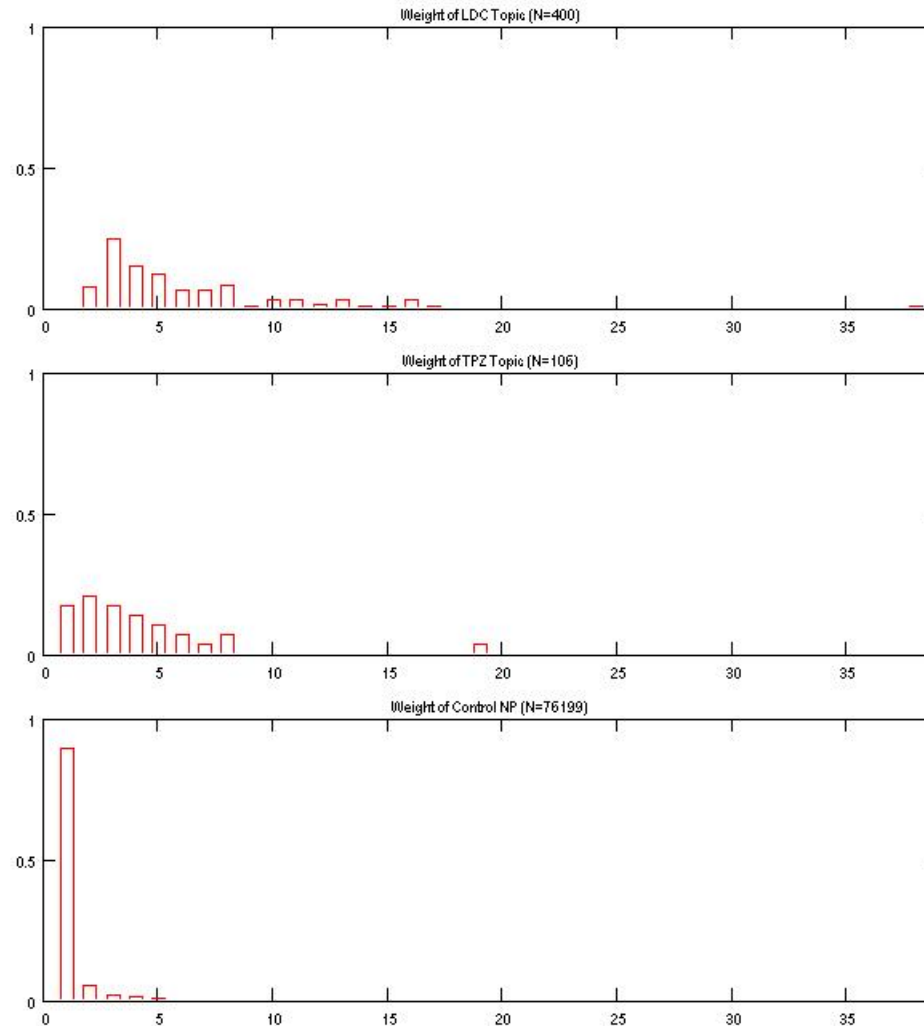
Results: Animacy

- animates are 1.5 times more likely than inanimates to be in a LDC
- inanimates are 7.6 times more likely than animates to be in a TPZ



Results: Grammatical weight

LDCs and TPZs tend to be heavier than control NPs



Conclusion

- logistic regression of corpus data can determine the factors that contribute to the choice of construction, even for infrequent constructions such as TPZ and LDC
- tested hypotheses from the literature
 - verified Prince's predictions that were testable with this corpus
 - Givón's predictions not supported
- new results: 'anti-animacy' effect for TPZ problematic for theories that predict animacy has a direct effect on the linearization of arguments (Cueni, Snider, & Zaenen 2005)

Acknowledgments

- valuable direction: Annie Zaenen and Tom Wasow
- also the rest of my committee: Dan Jurafsky and Ivan Sag
- help with the analysis: Anna Cueni and Joan Bresnan
- data extraction: Doug Rohde and Jean Carletta

LDC Logistic Regression

	Coefficient	F-value	p-value
(Intercept)	-1.212015	360.0704	< .0001
animacy		3.8718	0.0492
inanimate	-0.420939		
status (coarse)		27.3778	< .0001
old	-1.961611		
new	-0.270017		
status (fine)		0.7501	0.3865
set	0.235180		
gf		157.8188	< .0001
subj	-3.209837		
weight	0.401827	144.0611	< .0001

TPZ Logistic Regression

	Coefficient	F-value	p-value
(Intercept)	-2.032766	57.61879	< .0001
animacy		9.36251	0.0022
inanimate	1.477782		
status (coarse)		4.48857	0.0113
old	-1.908018		
new	-0.199473		
status (fine)		0.63553	0.4254
set	0.463401		
gf		89.80565	< .0001
subj	-5.846039		
weight	0.219027	6.23979	0.0125