A Corpus Study of Left Dislocation and Topicalization
Neal Snider

This presentation describes a corpus study of the left dislocation construction (LDC) and the topicalization construction (TPZ), and the logistic regression models that were run in order to compare the important factors that determine the choice of construction.

The constructions are exemplified as follows:

(1)a. LDC: **Burlington's crime**$_i$, **it**$_i$ doesn't involve children

b. TPZ: And **that**$_j$ I couldn't watch [ ]$_j$

LDC is defined by a fronted NP, followed by a clause containing a resumptive NP that is coreferential with it. TPZ is a long distance dependency where the fronted filler is coreferential with a gap in the following clause. This study only examined NP topicalizations.

Several hypotheses from the pragmatics literature helped guide the choice of predicting factors for these constructions. Prince (1997) has proposed several explicit functions for TPZ and LDC involving the information status of the fronted NP. Thus, a corpus was chosen for this study that made available fine-grained distinctions in the information status of NPs: the Switchboard corpus annotated by the Edinburgh-Stanford LINK project (Nissim *et al.* 2004; Zaenen *et al.* 2004) .

The first function of LDC's in Prince's classification is 'Simplifying' Left-Dislocation. The purpose of these LDC's is to 'simplify' discourse processing by removing fronted NPs that refer to discourse-new entities from a syntactic position (subject) that disfavors them. This hypothesis predicts that there should be more discourse-new left dislocations from subject LDC's than there are discourse new subjects in control NPs (those eligible but not topicalized or left-dislocated). This prediction is supported by the corpus data. However, this function must be a smaller one than the second LDC function Prince describes because there are very few left-dislocated subjects compared to non-new LDC subjects.

Prince's second function of LDC's is 'Poset' left dislocation. These LDC's trigger 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. The logistic regression showed that the left-dislocated NP being in a set relationship is a significant predictor of the occurrence of an LDC and that it is also 80% more likely to be in a set relation than all the other relations.

The third, and final, function Prince proposes for LDCs is to amnesty 'island' violations in constructions that would have otherwise been topicalizations. There were none of these LDCs in the corpus.

Prince proposes two overlapping functions for topicalization. The first function is to trigger a *Poset* inference, which is the same function as *Poset* LDCs. The evidence from this corpus supports Prince's theory about the first function

of TPZs. The second function of topicalization is to "trigger an inference on the part of the hearer that the proposition is to be structured into a focus and focus-frame." This study did not test this function of topicalization.

Givón (1983a) has proposed a relationship between left dislocation and the re-introduction of a discourse entity. This predicts that the referential distance distribution for LDCs will have a peak at a number utterances further back than control NPs. The LDC referential distance distribution peaks at 3 utterances, which, given that the utterance count includes switching turns, is far too low to be consistent with the re-introduction hypothesis.

There have been no previous studies that examined the effects of animacy on left dislocation and topicalization. This study found that there is a small tendency for left dislocated NPs to be animate, and a much larger tendency for topicalized NPs to be inanimate. The logistic regression shows that this inanimacy effect for TPZ is not merely due to the fact that they are extracted from a non-subject position, where inanimates are preferred, because this factor was included in the regression model. It is clear that animacy, along with grammatical function, are the major factors that differentiate LDC and TPZ. This tendency for inanimates to topicalize is problematic for theories of production that predict the accessibility of referents to directly influence linearization of NPs in the clause (Kempen & Harbusch 2004).

In addition to the information status and animacy factors mentioned above, weight and grammatical function were also found to be significant predictors of LDC and TPZ.

This study is relevant to several areas of linguistics. It is of interest to formal linguistics because it provides empirical evidence that tests the linguistic theories about these specific constructions. However, its larger goal of examining the factors that motivate construction choice is also relevant to both computational linguistics and psycholinguistics. The statistically significant factors in the model presented here could be used as features in a machine learning algorithm that decides when to use a LDC, TPZ, or canonical declarative sentence. Also, this study presents data about the relative importance of factors such as animacy, information status, and weight, all of which have been suggested as factors in production and comprehension models.

References

Givón, Talmy. 1983a. Topic continuity in discourse: An introduction. In *Topic Continuity in Discourse: A Quantitative Cross-Language Study*, ed. by Talmy Givón, p. 142. Amsterdam: John Benjamins.

Kempen, Gerard, & Karin Harbusch. 2004. How flexible is constituent order in the midfield of german subordinate clauses? a corpus study revealing unexpected rigidity. In *Linguistic Evidence*, Tbingen, Germany.

Nissim, Malvina, Shipra Dingare, Jean Carletta, & Mark Steedman. 2004. An annotation scheme for information status in dialogue. In *4th Conference on Language Resources and Evaluation (LREC2004)*.

Prince, Ellen. 1997. On the functions of left-dislocation in english discourse. In *Directions in functional linguistics*, ed. by A. Kamio, 117–44. Philadelphia/Amsterdam: John Benjamins.

Zaenen, Annie, Jean Carletta, Gregory Garretson, Joan Bresnan, Andrew Koontz-Garboden, Tatiana Nikitina, M. Catherine O'Connor, & Tom Wasow. 2004. Animacy encoding in english: why and how. In *ACL Workshop on Discourse Annotation*, ed. by D. Byron & B. Webber, Barcelona.