

Qualifying Paper 1: A Corpus Study of Left Dislocation and Topicalization

Neal Snider

March 28, 2005

Contents

1	Introduction	2
2	The Data	3
2.1	The structure of the corpus	3
2.1.1	Information status annotation	4
2.1.2	Animacy annotation	6
2.1.3	Annotation of corpus sample	8
2.2	Data extraction	8
2.2.1	Animacy and Information status	11
2.2.2	Grammatical function	11
2.2.3	Grammatical weight	12
2.2.4	Speaker	13
2.2.5	Referential Distance	13
3	Results	14
3.1	Patterns in the Data	14
3.2	The Models	19
3.2.1	Left Dislocation	22
3.2.2	Topicalization	24
3.2.3	Predicting LDC vs. TPZ	25
4	Implications for Previous Explanations	25
4.1	Information status factors	25
4.1.1	Prince	25
4.1.2	Gregory and Michaelis	30
4.2	Animacy factors	30
4.3	Referential Distance	31
5	Conclusion	32
A	Appendix	35

1 Introduction

Language allows many different ways of expressing the same truth-conditionally equivalent content. A primary question that motivates studies of linguistic production is how speakers choose among the many options offered by their language.

Pragmatic motivations are often cited as important factors in deciding among constructions, and the choice of two constructions in particular has been discussed as being driven purely by pragmatic factors: the left dislocation construction (LDC) and the topicalization construction (TPZ). 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. I considered only NP topicalizations in this study.

This paper describes a corpus study of LDC and TPZ, and the logistic regression models that were run in order to compare the important factors that determine the choice of construction. A corpus study is a good medium in which to study these constructions because the conditions that produce them are rather subtle, and they would be difficult to elicit in a controlled experiment. Logistic regression was chosen as the analytical tool because of the infrequency of the constructions and the strong correlations between the relevant linguistic factors. This is a relatively new tool to formal linguistic analysis, but is proving useful at determining the relative strengths of highly correlated factors in, for example, recent studies on the dative alternation (Bresnan & Cueni 2005; Bresnan *et al.* 2005).

Several hypotheses from the pragmatics literature helped guide the choice of predicting factors for these constructions. Prince (1992, 1995, 1997) has proposed several explicit functions for TPZ and LDC involving the information status of the fronted NP and, in the case of LDC, the grammatical function of the resumptive pronoun and other NPs in the rest of the clause. Her theories about the information status of the left-dislocated NP are far more precise than can be captured by the traditional distinction between old and new information.

Thus, a corpus was chosen for this study that made available fine-grained distinctions in the information status of NPs. Further, Givón (1983a) has proposed a relationship between left dislocation and the re-introduction of a discourse entity. This motivates the need for the factor of referential distance that was used in this study. These hypotheses will be described in Section 4, where this study is shown to provide crucial tests for them.

This study is clearly relevant 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.

In natural language generation, choosing among truth-conditionally equivalent paraphrases is a key problem. The statistical model presented here shows the relative importance of various factors to the choice of these two constructions. The important factors could be used as features in a machine learning algorithm that decides when to use a LDC, TPZ, or canonical declarative sentence.

Construction choice is equally important in psychological models of sentence processing. 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. Animacy and information status are important because they are two measures of conceptual accessibility. Psycholinguistic studies have shown that the syntactic prominence of nominals is related to how easily their referents can be mentally accessed (Bock *et al.* 1992; Prat-Sala & Branigan 2000). This *conceptual accessibility* has been hypothesized to derive from two different sources: the inherent accessibility of the referent, and the accessibility derived from being mentioned in the discourse. Animacy is one possible contributor to the inherent accessibility of a referent, and information status is an indicator of derived accessibility.

2 The Data

2.1 The structure of the corpus

The Switchboard corpus (Godfrey *et al.* 1992) is a corpus of spoken English that was compiled from telephone conversations. It is particularly useful for the current study because it is a large collection of spoken English. This corpus has also been annotated over the years to make it more useful for syntactic studies.

The Treebank Project (Marcus *et al.* 1994) of the Linguistic Data Consortium released a version of Switchboard annotated for part of speech and hierarchical syntactic structure. This annotation was essential for the extraction of the LDCs and TPZs from the corpus.

The Switchboard corpus has also been annotated for various semantic and pragmatic features associated with nominals. A central focus of the Edinburgh-Stanford Paraphrase LINK annotation project (Nissim *et al.* 2004; Zaenen *et al.* 2004) was this annotation. The project annotated a subsection of the corpus for animacy and information status. The Edinburgh half of the project annotated for information status, while Stanford did the animacy annotation. For this paper, I also annotated an even smaller sample of the corpus for referential distance in order to test predictions from the literature of the behavior of LDCs and TPZs based on this factor.

In the following sections, I will briefly describe how the nominals in the corpus were annotated for various pragmatic and semantic factors.

2.1.1 Information status annotation

The simplest conception of information status is the well-known distinction between old and new information. The old/new distinction is very coarse-grained, and models with finer grained distinctions have been proposed over the years. The annotation scheme used on this version of the Switchboard is a novel attempt at capturing the advantages of both coarse-grained and fine-grained models.

The annotation scheme distinguishes among many fine-grained relationships between entities in the dialog. The fine-grained distinctions are arranged hierarchically into three coarse-grained information status relationships: *old*, *mediated*, and *new*. If the entity has already been mentioned in the conversation, and is known to the hearer, that is, it is speaker-old and hearer-old, it is coded as **old**. If an entity is unknown to the hearer, and it is new to the discourse, it is defined as **new**. If it is known to the hearer, but new to the discourse, it is coded as **mediated**. Mediated entities are those that can be inferred by the hearer. Examples include entities that are generally known, such as “the moon” and “the president”.

In this hierarchical model of information status, there are six subtypes of **old**:

- The **identity** relation is defined as that represented by an anaphoric chain.

For example, if “it” corefers to a previous instance of “the book” in the discourse. There are 1859 tokens of this relation out of the 5880 NPs in the subset of the corpus I used for my analysis.

- If the antecedent is a verb phrase, the nominal is coded with subtype **event**. There are 444/5880 tokens of this relation.
- A nominal is coded **general** if it refers to the dialog participants, i.e., “I” or “you”. There are 2343/5880 tokens of this relation.
- The subtype **generic** is used for pronouns in their generic use, like “you” or “they”. There are 347/5880 tokens of this relation.
- **identity-generic** is used when there is a coreference chain of generic pronouns. There are no tokens of this relation in the subset of the corpus, due to its infrequency.

There are nine subtypes of **mediated**:

- **general** is the code given to generally known entities like “France” and “the pope”. There are 61/5880 tokens of this relation.
- **bound** is the code given to pronouns that refer to an indefinite or generic NP in the same clause. There are 74/5880 tokens of this relation.
- **poss** marks intra-clausal possessive relations. There are 94/5880 tokens of this relation.
- **part** is used to mark part-whole relations that occur within or across clauses. There are 11/5880 tokens of this relation.
- **situation** is used if an entity is part of a situation set up by a previously introduced entity. There are 50/5880 tokens of this relation.
- **event** is used when an entity could be inferred from a previously introduced VP. There are 10/5880 tokens of this relation.
- **set** is used when an entity is a subset, superset, or member of a previously introduced set. There are 353/5880 tokens of this relation.
- **func-value** is used when an entity is the value of a previously introduced function, like temperature and time. There are no tokens of this relation in the subset of the corpus, due to its infrequency.

- **aggregation** is applied to coordinations whose entities have not all been previously introduced into the discourse. There are 48/5880 tokens of this relation.

There is no finer subtyping of **new** entities. There are 106/5880 tokens of this relation.

Nissim *et al.* conducted a reliability study of their annotation scheme and found it to be reliable. They used Kappa statistics to test for reliability (Carletta 1996). A Kappa score of 0.8 indicates good reliability on discourse annotation. They found their scheme yielded $K = 0.845$ for the coarse-grained categories and $K = 0.788$ on the subtypes, with some subtypes found to be more reliable than others. The annotators expected more difficulty with the *mediated* and *new* categories, because it is often a subjective decision between what may be inferred and what is new to the discourse. They did indeed find that the *old* annotation was more reliable ($K = 0.897$). However, the *mediated* and *new* categories were still quite reliable ($K = 0.794$ and $K = 0.791$ respectively).

2.1.2 Animacy annotation

The animacy hierarchy has been hypothesized to influence the syntactic prominence of nominals. Binary animacy distinctions have occasionally been proposed, such as human/non-human and animate/inanimate. One major problem with devising an appropriate animacy hierarchy, even with the apparently simple ones above, is that the notion of animacy that is linguistically relevant does not directly correspond to biologically-based distinctions.

The following nine-valued animacy scale was used to annotate this part of the corpus. This scale was aggregated into animates and inanimates for the analysis in this study. The animates are subdivided as follows:

- **human** refers to one or more humans, including imaginary entities that are presumed to act like humans. There are 4020 tokens of this relation out of the 5880 NPs in the subset of the corpus I used for my analysis.
- **org** is the tag for organizations of humans that display some degree of group identity. The difference between *human* and *org* is having a “collective voice”, so a group with “collective voice and purpose” is an *org*, while a group without such, like a mob, is not. There are 283/5880 tokens of this code.

- **animal** refers to non-human animates, including viruses and bacteria. There are 114/5880 tokens of this code.
- **mac** refers to intelligent machines and robots. There are 3/5880 tokens of this code.
- **veh** is the tag for vehicles, which are often treated as living things in many contexts. There are 49/5880 tokens of this code.

The inanimates are:

- **place** is used to tag nominals that ‘refer to a place as a place’. The coding scheme assumes that only potential locations for humans are places. There are 47/5880 tokens of this code.
- **time** refers to expressions of periods of time. There are 19/5880 tokens of this code.
- **concrete** refers to concrete substances. For example, body parts are concrete, but air, voice, and other intangibles are not. There are 232/5880 tokens of this code.
- **nonconc** refers to non-concrete inanimate entities. This is the default category and is used for anything that is not prototypically concrete and clearly inanimate. There are 1113/5880 tokens of this code.

Zaenen *et al.* also evaluated their animacy annotation of the corpus for reliability. Again using a Kappa statistic as a reliability benchmark, they found their annotation reliable with $K = 0.92$. They discuss that this high reliability masks some problems with the annotation. Part of the reason the reliability is so high is the fact that the most common categories are *human* and *nonconc*, which are easily distinguished. The cross-coder reliability for these is high, but it is less so for the other categories. This led the authors to conclude that these other categories were not defined well enough to allow reliable coding. For example, *time* and *place* appeared not to be defined in a way that corresponded to the coders intuitive understanding of them. There was also a lot of disagreement about the *human* and *org* codes, depending on how coders interpreted the referents of expressions.

2.1.3 Annotation of corpus sample

In order to test some of the predictions from the literature (Givón 1983a; Givón 1983b), I personally annotated a much smaller sample of the corpus for referential distance. I annotated all the fronted NPs of the LDC and TPZ constructions as well as a randomly chosen sample of the control NPs from the data I used in my analysis. The method by which control NPs were chosen will be described below. Thus, I annotated 122 fronted NPs from the LDCs, 29 fronted NPs from the TPZs, and 163 control NPs.

Referential distance is a measure of how far back in the discourse the entity referred to by a NP was last mentioned. I measured referential distance by number of utterances, which is different from the clause measure used by Givón (1983a). This difference arises because my measure includes changes in turn, as well as many affirmations and fillers, so the referential distances calculated by the utterance technique will be higher than those calculated by counting clauses. I determined the last-referred entity by finding the last NP in the previous discourse that had a relationship with the current NP that could be described by the relations under **old** and **mediated** in the information status annotation scheme described above.

2.2 Data extraction

The product of the annotation projects mentioned above was a corpus in XML format, with separate interlinked files for each dimension of annotation. For example, the hierarchical syntactic information is stored in one file (actually a set of files given the complex nature of such information), and the information status and animacy coding were stored in one other file each. The corpus is searched using the NITE query language (Carletta *et al.* 2004) in a Java application written for such queries. However, in order to facilitate compatibility with pre-existing linguistic tools for corpus analysis, the corpus was “back-translated” by Jean Carletta (at Edinburgh) into the original Treebank single-file format with the information status and animacy annotations appearing as extra labels on the name of each node. An example of this formatting is as follows:

```
(2) <NP-TPC_MARKABLE_human_med_set <ANTEC 35>  
      <NP_MARKABLE_human_old_generic <ANTEC 34>  
        <DT those>>  
      <SBAR <WHNP_MARKABLE_hunan_ol_d_relative <N 402501>  
        <ANAPH 34>  
        <UP who>>  
      <S <NP-SBJ_MARKABLE <-NONE- <N 402501>>>  
        <UP <UBP find>  
          <PRT <RP out>>  
          <NP_MARKABLE_nonconc_med_set <JJ such>  
            <NN information>>>>>>>
```

The annotations following the NP above signify the following:

- Its part of speech is noun phrase topic (NP-TPC), according to Treebank notation.
- It is a “MARKABLE”, meaning that it is a nominal eligible for information status or animacy annotation according to the Paraphrase-LINK annotation project’s criteria.
- Its animacy is “human”.
- Its information status is “mediated”, with the subtype “set”.

There are several other features of the above tree that differ from the Treebank standard. The Paraphrase-LINK project also included annotation for coreference, which is represented by the “ANAPH” and “ANTEC” codes above. The coreference scheme works by assigning anaphors the same ANAPH value as their preceding coreferential antecedents’ ANTEC value. In the original back-translation, these codes were included in the node names, like the animacy and information status codes. However, to facilitate searches that refer to the coreference codes, I wrote a script to make them into preterminal nodes that are sisters to the head of their NP. The reasons for this will be explained below. A similar transformation was done for the filler-gap dependency codes (marked by “N” in the example above). These hexadecimal codes were in the original Treebank annotation and were used to mark the dependency between a filler and a gap: the gaps (coded “*T*” or “-NONE-”) have the same N value as their filler NP.

I extracted the LDCs and TPZs from the corpus with “tgrep2”, by Doug Rohde. The back-translation of the XML corpus was done to enable searches with this program, which requires Treebank-formatted corpora. TPZs can be extracted rather easily because the fronted NP is in a filler-gap relationship to some missing NP in the rest of the sentence. Therefore one need only require identity between the N-value of the fronted NP (coded NP-TPC), and some gap later in the sentence (coded -NONE-). Such a search could not be done in the version of tgrep that existed before my project, however. This version allowed one to refer to a node twice in a search pattern, but it required token-identity, that is the node was assumed the same node each time it was referred to. However, a TPZ search needs the ability to require that only the name of the two nodes be identical. I corresponded with Rohde, and he added this type-identity

function. With all these tools assembled, I was able to use the search pattern in the appendix (28) to extract 106 TPZs. There is an additional restriction that the NP-TPC not dominate the word “more” was put in to remove a few instances where a fronted NP (not in a topicalization) was involved in a comparative, which were also coded as filler-gap dependencies in Treebank.

LDCs can also be extracted using the node type-identity function, except to get a LDC, one requires that the ANTEC value of the fronted NP-TPC have the same value as the ANAPH node of a NP later in the sentence (usually a pronoun). This reflects the fact that LDCs involve a relationship of coreference between the fronted NP and another NP, while in TPZs, the relationship is a filler-gap dependency. The pattern in (29) extracts LDCs. Unfortunately, this pattern will only extract 115 LDCs from coreference-annotated part of the corpus (which is the same as the information status annotated part). It also misses 8 LDCs due to incorrect annotation. Therefore, I extracted the LDCs manually. I did this by first extracting all the sentences containing NP-TPCs, because all fronted NPs in LDCs would be coded this way, and then manually retrieving all the sentences that were LDCs. The other sentences that had NP-TPCs were topicalizations, dangling topic constructions, and right dislocations (where the NP-TPC is not fronted). As I extracted each LDC, I tagged the resumptive NP with a tag RES, because there is no way to automatically extract it without coreference coding, and I needed to be able to access information about the resumptive NP. I extracted a total of 401 LDCs with this technique.

I also extracted many “control” NPs in order to construct a model that predicts the presence or absence of the constructions in question. I constructed the control set of sentences by excluding those sentence types which could not participate in LDC or TPZ constructions, and obviously excluding sentences with LDCs and TPZs. I did this by excluding all sentences that were, or contained, questions, embedded questions, and gerundive phrases, as well as sentences which contained a fronted NP. The gerundive phrases were removed because I did not think they could participate in LDC and TPZ. They include examples like the following:

- (3) I think that great strides are being made nowadays in, in **caring for the elderly**, you know, in several, in a, in a whole lot of areas.

This left control 69350 clauses of the 70341 in the sampled corpus. The extraction of the control sentences was accomplished with the pattern in (30):

I also extracted the unique node identifier supplied by `tgrep2` for each LDC, TPZ, and control NP. This enabled me associate each NP with its features as I extracted them in successive searches.

2.2.1 Animacy and Information status

These annotations were very easy to extract. I simply had `tgrep2` output only the NP (as opposed to the sentence containing the NP). In this format, the node label, which contains the annotations, is at the beginning of the line containing the NP. I then wrote a perl script that used regular expressions to take out the animacy and information status codes, which always occurred in the same order.

2.2.2 Grammatical function

Extracting grammatical function information was somewhat more difficult. As part of the manual extraction of LDCs and TPZs, I created separate corpora such that all the sentences in the corpus were the construction in question. I then came up with several patterns which would extract NPs of the grammatical functions that I thought were able to participate in LDC or TPZ constructions. I developed patterns for first objects, second objects, prepositional objects, and prepositional adjuncts. I assumed that other types of adjuncts, like temporal NPs, would not participate in LDC and TPZ.

These patterns actually cut several corners to make the search pattern easier to construct. I justify this with the fact that I will sample the control NPs anyway, since there are too many for the statistical package to analyze, and also aggregate the GFs to subject and non-subject. I assume that the data that I throw out, for example any progressive tenses (as shown below) would not correlate with the constructions in question, given that no previous research has suggested that verbal tense or aspect are predictive factors in LDC or TPZ. Therefore, my shortcuts will not bias my data.

For control NPs, extracting the grammatical function was just a matter of combining the control NP pattern (30) with the relevant grammatical function pattern. For the LDC and TPZ corpora, I just added a condition that the NP whose GF was being constrained also contain a RES code (for LDC) or a NONE code (for TPZ).

To extract transitive first objects and second objects of ditransitives, I had to make two patterns, as the objects of raising and control verbs are coded as

subjects in Treebank. First, to get transitive objects and ditransitive second objects, the pattern requires that a S immediately dominate a VP, which did not contain a form of the word “be” (in order to avoid predicate nominals, which do not participate in LDC or TPZ). Excluding VPs that contain forms of “be” also has the effect of removing all progressive tenses. As I said above, I assume that this will not bias my data. The VP has a last daughter which is the NP in question. The last NP daughter will be a transitive object or a ditransitive second object. The pattern further requires that the NP not contain a trace (NONE tag) or be part of a parenthetical (PRN). The search pattern is (31).

To get the objects of raising and control verbs, the pattern finds VPs that immediately dominate an S (as opposed to the SBAR of sentential complements). The NP-SUBJ of this S is the object of a control or raising verb. This pattern is (32).

I extracted first objects of ditransitives by requiring that the VP immediately dominate two NPs, printing the first (33). The prepositional object of prepositionally ditransitive verbs can be extracted using the PP-DTV annotation of treebank. I used a pattern (34) that requires a VP to immediately dominate a NP and an PP-DTV, and prints the object NP of the PP-DTV. I extracted objects of prepositional arguments and adjuncts by a similar pattern that requires a VP that immediately dominates a PP, which is not a PP-DTV, then printing the object NP of this PP (35).

As I mentioned above, I aggregated these grammatical functions into subject and non-subject, and then matched them to the NPs whose annotation I had already extracted by their unique node codes.

2.2.3 Grammatical weight

Extracting grammatical weight in words is quite simple in general. I extracted the NPs in my now growing data table by a `tgrep2` function that takes a list of codes instead of a pattern, then prints the corresponding nodes. I printed the terminal symbols of the NPs, and wrote a script to count them. Here my previous alterations to the corpus to facilitate coreference matching become a problem. The coreference codes become terminal nodes in my scheme, so I must filter these out in the word count. I did this by filtering out all numbers. Some numbers in Switchboard are transcribed with numerals, but there are very few of these in the Switchboard conversations, so I assume this will not bias my data.

2.2.4 Speaker

The original Switchboard project collected a lot of personal information about each speaker. Factors such as gender, age, and region are available. Sociolinguistic factors are not being examined in this study, but it is necessary to track which speaker uttered each NP in the analysis to control for the possibility that use of TPZ or LDC could vary with speaker independent of the factors studied here.

In the individual Switchboard dialogs, the two speakers are denoted A and B. The Switchboard release includes a file that lists the dialog number and the 4-digit speaker identity code that corresponds to each speaker. Unfortunately, the tgrep2 formatted corpus is a concatenation of all the individual dialog files, so it is not easy to determine which dialog an NP appears in. However, as I mentioned before, the tgrep2 corpus has a unique numeric label for each node. By extracting the label of the node that contains the dialog number, I made a file that listed the id of each speaker and label of the node that begins the dialog that contains that speaker. Since the tgrep2 node labels are consecutive, I wrote a script that compared the node label of each NP with the list to retrieve the speaker id.

2.2.5 Referential Distance

Referential distance may be extracted automatically from this corpus using the coreference annotations. However, these annotations have serious limitations, so I ended up coding a part of the corpus manually, with criteria that will be described below.

As mentioned above, NPs that refer to entities that have been previously mentioned in the discourse have the same ANAPH value as the ANTEC value of the previous NP. To automatically calculate referential distance, one need only extract the top-level node number from the antecedent NP and its anaphor, and then subtract the two. The top-level node number is the first number in the colon-separated tgrep2 node code. This top-level number corresponds to the utterance number. Thus subtracting the top-level numbers of the antecedent and anaphor will yield the number of utterances that separate them. I actually implemented this with a script that first extracted the ANAPH number of each NP in the table (if it had one) and then calling tgrep2 again to get all NPs in the corpus whose ANTEC code matched it.

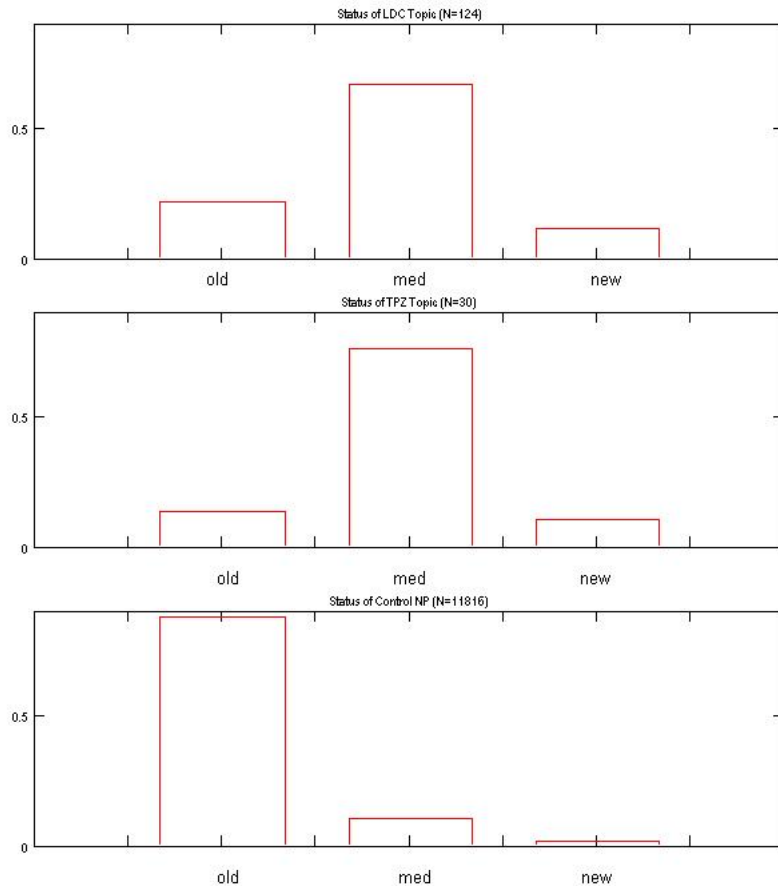
The problem with this automatic extraction is that these anaphoric relations

were only coded for one particular type of coreference relation, the *identity* relation defined above. Therefore, the automatically extracted data was too sparse. To get around this, I coded by hand the sample of the corpus mentioned previously. I did this by extracting the utterance containing each NP in the sample, as well as the previous 80 utterances. I then looked for a previous NP that stood in one of the above information status relationships with the sample NP. I then recorded how many utterances back was the *first* previous mention of the entity referred to by the sample NP.

3 Results

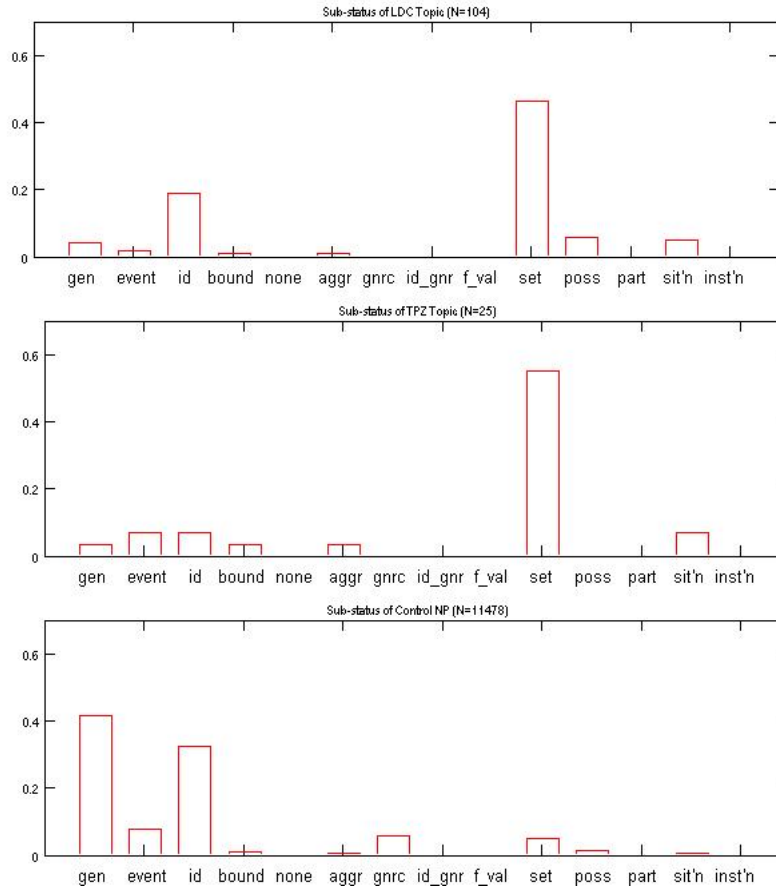
3.1 Patterns in the Data

Before running the models, several patterns were evident in the data. With respect to information status, the fronted NPs in LDCs (66.9%) and TPZs (75.9%) is much more likely to be mediated than control NPs (10.7%):



(4)

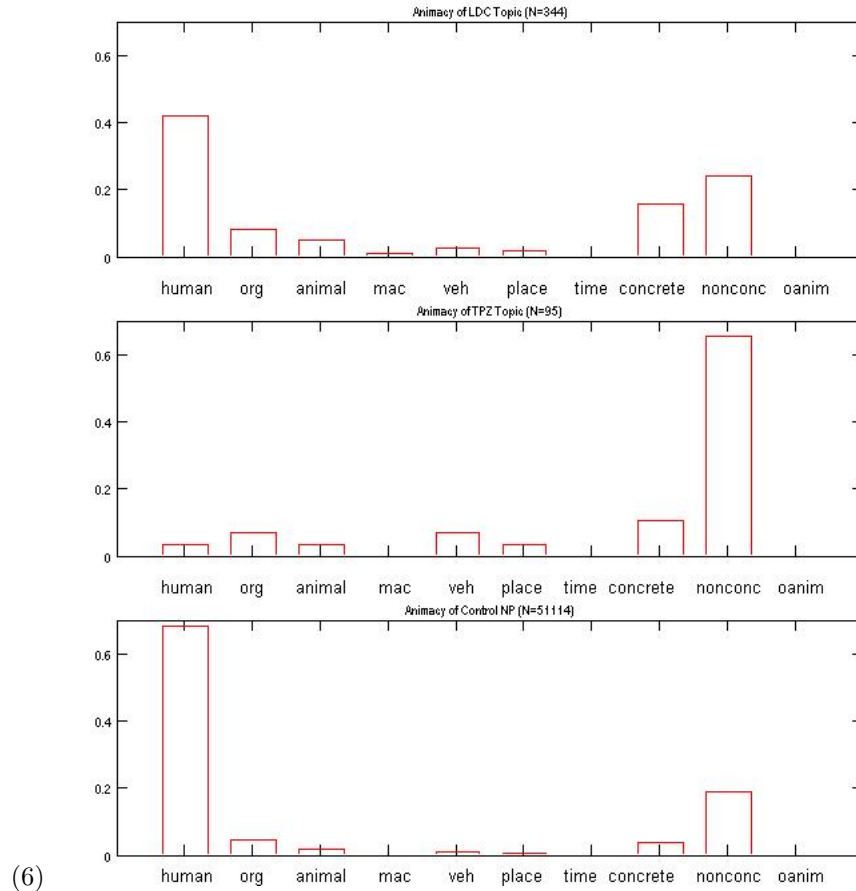
Also, in the fine-grained information status categories, the set relation is by far the most common for LDCs (46.3%) and TPZs (55.2%), as compared to the controls (5.2%):



(5)

These patterns are very suggestive, but many of the factors being considered in this study are correlated, and there are relatively fewer tokens of the LDCs and TPZs than the controls, so the logistic regression will determine whether the patterns are statistically significant.

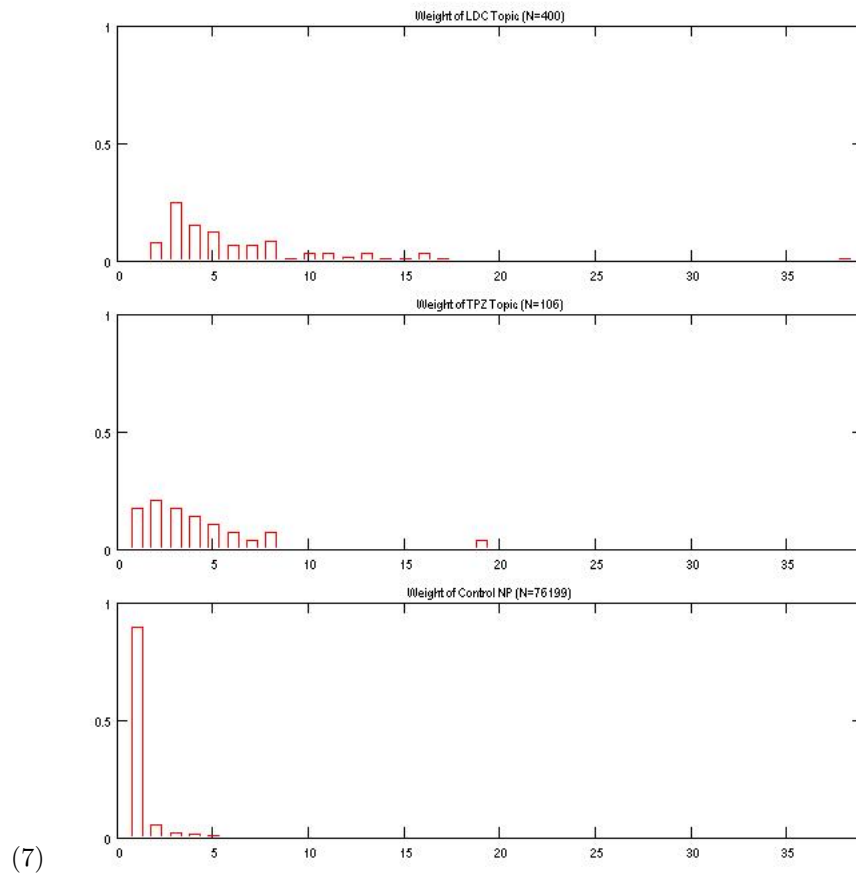
The animacy data is somewhat less clear, although there does appear to be a tendency for TPZs to be inanimate (65.5%) compared to LDCs (24%) and controls (18.7%). LDCs are less likely to be animate than controls, but not to the extent of the TPZs. Again, however, only the logistic regression will determine whether the patterns are statistically significant. The animacy data graphed below comes from the full animacy-annotated part of the corpus, so there are more LDCs and TPZs than above:



With respect to grammatical function, most LDCs are subjects (69.4%). Topicalization is generally regarded to be impossible from subject position. In any case, a subject topicalization would have the same structure as a declarative sentence. 11 sentences were coded as topicalizations from subject (a fronted NP coded as NP-TPC, with a gap in the NP-SBJ of the following clause), but these are all parentheticals. The LDCs and TPZs may be compared to controls by assuming a left dislocated NP that is coreferential with a NP with a given grammatical function is comparable to a control NP with that grammatical function. Thus, for the discussion of the models below, LDCs with resumptives in subject position are compared to control NPs that are in subject position. The histogram for the grammatical function data may be found in the appendix.

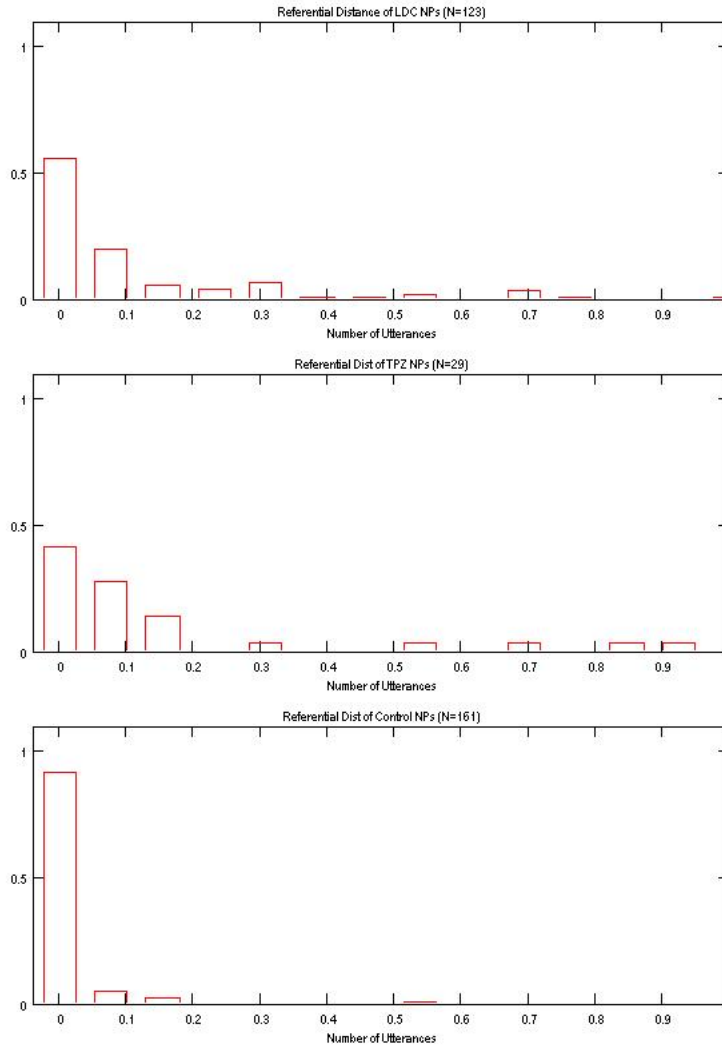
The grammatical weight data indicates that LDCs and TPZs tend to be heavier than control NPs, with weight peaking at 3 words (24.8%) for LDC and

2 words (20.7%) for TPZ. Control NPs are overwhelmingly likely to be 1 word long (89.4%), confirming the well-known prevalence of pronouns in spoken corpora. Personal pronouns are quite rare in LDC and TPZ, so it is not clear from the data alone that weight has an independent effect. A logistic regression, with its ability to control for multiple factors, will be essential in determining the independent effect of weight.



(7)

The data for referential distance is much more sparse, given it comes from my hand annotated data set. It is clear, however, that LDCs and TPZs tend to have been last referred to further back in the discourse than control NPs. Qualitatively, the pattern is similar to that of control NPs: the histogram peaks at 1 utterance back and steadily decreases. There is not a second peak in the histogram, a fact that will be important in comparing these results to the hypothesis of Givón 1983a.



(8)

3.2 The Models

I modelled the choice of the two constructions with multi-variable logistic regression models: one predicting presence or absence of LDC (LDC vs. “canonical” declarative sentence) and another modelling choice of TPZ vs. canonical declarative sentence. I used the open-source statistical package *R* to perform the regressions.

A logistic regression model is particularly well suited to modelling categorical responses, such as the choice of construction being investigated here. Logistic

regression assumes that the contribution of the predicting factors is linear with respect to the *logit*, or log odds, of the probability of the response. When a logistic regression is performed, the parameters β are calculated in the logistic regression equation:

$$(9) \log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots$$

P is the probability of the response and the X_i 's are the predicting factors. The logistic regression equation may also be written in terms of the product of the odds ratios of the factors:

$$(10) \frac{P}{1-P} = \frac{p_1}{1-p_1} \frac{p_2}{1-p_2} \frac{p_3}{1-p_3} \dots$$

Thus, for a continuous factor, determining the coefficients β_i gives the odds ratio for increasing the factor by 1:

$$(11) X_i + 1 : X_i = e^{\beta_i}$$

An example of a continuous factor is grammatical weight as used in this analysis. Grammatical weight can vary continuously over integers greater than zero. Logistic regression assumes a linear increase in odds with respect to weight. Applying the equation above, the log odds ratio of a weight of 15, say, to a weight of 14 is just the coefficient for the weight factor, β_W , so the odds ratio equation is as follows:

$$(12) \frac{o(W=15)}{o(W=14)} = e^{\beta_W}$$

For binary categorical factors, R recodes the data into a boolean value, with one level being arbitrarily chosen as 0. Therefore, when only one factor is being modelled, the probability of the 0 level is the intercept term β_0 . However, in a multivariate regression such as this study, the intercept term will contain the reference values for all the factors, so one cannot calculate the probability of a level directly, but the odds ratio is still indicated by the coefficient of the weight factor. The equation for this odds ratio is as follows:

$$(13) (X_i = 1) : (X_i = 0) = e^{\beta_i}$$

The animacy scale mentioned above is aggregated into a binary *animate/inanimate* distinction in the model, and R chose *animate* as the zero level. Thus, the log odds ratio of *inanimate* to *animate* is the coefficient of the animacy factor, and the odds ratio equation is as follows:

$$(14) \frac{o(\text{inanimate})}{o(\text{animate})} = e^{\beta_A}$$

For multi-valued categorical factors, R arbitrarily chooses one as the reference value and then recodes the other levels as binary distinctions with respect to the reference. Thus, an n -level distinction will become $n-1$ factors, and their regression coefficients will represent the odds ratio of each level with respect to the reference. The coarse-grained information status scale is an example of a multi-valued factor. R chose *mediated* as the reference value, so the effect of information status in the model is captured by two odd ratio equations: one for *old*, and one for *new*. The log odds ratio of *old* to *mediated* is the coefficient for $status = old$, and the log odds ratio of *new* to *mediated* is the coefficient for $status = new$. The odds ratio equations that describe the effect on the model of the coarse information status factor are as follows:

$$(15)\text{a. } \frac{o(old)}{o(med)} = e^{\beta_{IS=old}}$$

$$\text{b. } \frac{o(new)}{o(med)} = e^{\beta_{IS=new}}$$

Like other statistical methods, logistic regression is used to test hypotheses. The null hypothesis in a logistic regression is that $\beta_i = 0$ for one or more factors, and therefore the odds ratio is 1 , so the prediction of the dependent factor is no more or less likely with respect to factor i . R can return two tests of significance for a multivariate logistic regression. It gives a t -test for each factor, which, in a logistic regression with multi-valued categorical factors, is a test for each level of the factor, excluding the reference. The t -test for a factor tests the null hypothesis that the coefficient for that factor is significantly different from zero. However, for a study such as this, where one is interested in the significance of each factor, not necessarily all of its levels individually, the results of an analysis of variance (ANOVA) are more important. An ANOVA in R yields an F -test for each factor, instead of each level in the factor, which tests the null hypothesis that the coefficients for each factor are significantly different from zero. This is the test that is needed for this analysis.

Further, an ANOVA tests the significance of the model as each factor is added in turn. To determine if a factor X changes the likelihood of the model independent of other factors, one makes a model with all factors (including X), and one with all factors except X , and then test if the likelihood of the two models is significantly different. If the addition of a single factor is significant, then it has an independent effect. The ANOVA allows this test if one constructs several models with each factor in turn added last. The significance values for each factor reported below are the significance of the change in the likelihood of the model when that factor is added last.

3.2.1 Left Dislocation

In constructing logistic regression models of the data, I actually used a mixed model which included *speaker* as a random effect. As a factor, *speaker* has too many levels to be treated as a fixed effect: it would add too many degrees of freedom to the model. In this situation, when the factor has more levels than the underlying phenomenon presumably has (assuming that *speaker* is a proxy for dialect), it can be included as a random factor. As opposed to the other factors, a linearity assumption is not made about the contribution of the random factor, merely that it is normally distributed. It was important to include *speaker* in the model because some dialects use left dislocation more than others.

The first models I ran included all the levels of the animacy and information status in the regression. Including all these factors did not capture enough of the variation, and many of them were not significant predictors of the construction. Thus, I aggregated the factors and found that reducing animacy to a binary *animate/inanimate* distinction caused the model to account for the most variation in the data, compared to other aggregations of the animacy scale. The fine-grained information status was aggregated to a binary scale based on whether the NP was in a *set* relation or not. The coarse scale information status remained three-valued. The coefficients for a linear logistic regression with predicting factors animacy, coarse and fine information status, grammatical function, and weight are as follows:

	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

The significance tests for these factors bear some explanation. The test for fine-grained information status shows that this factor is not independently significant because a model made with this factor does not have significantly improved likelihood over one without. Note also that coarse information-status is significant at $p < .0001$ when added alone. However, in the model where coarse-grained

information status is added last, the fine-grained information status factor is significant at $p < .0001$ when it is added (it happened to be first factor added in this case, but the order does not matter as long as it is added before the coarse-grained factor). This demonstrates that there is an interaction between coarse- and fine-grained information status, so there is no way to determine from these data which of these factors is primary. This is no surprise given that an NP coded as *fine = set*, will always be coded *coarse = mediated* as well. Thus, it is clear that information status is a significant predictor of left dislocation, with *mediated, set*-coded entities most likely to left-dislocate, and the coarse- and fine-grained factors are highly correlated.

The significance of the animacy and grammatical function factors is also important because it is well established that, in many linguistic phenomena, animacy correlates strongly with grammatical function, with animates being attracted toward higher levels on the grammatical function hierarchy. However, The tests above show that, for this construction, each of these factors have an independent effect, because they significantly increase the likelihood of the model when added individually. Grammatical function is significant at the $p < .0001$ level, and animacy is significant at the $p < 0.05$ level.

These coefficients may be interpreted as odds ratios as described above. Thus, animates are 1.5 times more likely than inanimates to be in a LDC. The fronted NP in a LDC is 1.3 times more likely to be *mediated* than *new*, and 7.1 times more likely to be *mediated* than *old*. It is also 80% more likely to be in a set relation than all the other relations.

Weight is also a significant independent predictor of LDC. This is interesting because the control NPs are overwhelmingly one word long, due to the preponderance of pronouns. LDCs occur less often with pronouns, but the logistic regression controls for this possible confound with the information status factor. Pronouns are almost always old information, and the fact that weight is still a significant factor even when it is added after information status indicates that the effect of weight is independent, despite possible confound of pronouns in the control set. The logistic regression coefficient shows that the likelihood for a NP to be in a LDC increases 1.5 times with each 1 word increase in length. This is not the right prediction given the evidence above that the weight distribution peaks at 3 words and then decreases. Weight clearly violates the linearity with respect to log odds that is assumed in logistic regression.

The odds ratio for grammatical function bears some discussion because the model shows that subjects are disfavored, while the histogram above shows

that there are more subject LDCs than object LDCs. This is not the relevant comparison, however. The odds ratios from the model give the odds that a subject will be in a left dislocation, and in the data sample used to construct the model 33.9% of objects are left-dislocated, while only 1.5% of subjects. Thus, the odds ratio of the model is intuitively correct. The high percentage of left-dislocated objects is due to the sampling of control NPs without sampling LDCs (which would throw away too much data). This inflates the overall likelihood of LDC somewhat.

3.2.2 Topicalization

The coefficients for the topicalization model are as follows:

	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

All of these factors are independently significant predictors of topicalization, except the fine-grained information status factor, which interacts with the coarse-grained factor like in the LDC model. Thus, information status is a significant predictor of topicalization, it is just not determinable from these data whether coarse- or fine-grained measurement is primary.

The coefficients may also be interpreted as odds ratios. Similar to LDC, TPZs are more likely to be *mediated*, and in a *set* relation. Inanimates are 7.6 times more likely than animates to be in a TPZ. This is a surprising result, but the tests above show that animacy has an independent effect at $p = 0.0022$, so it cannot be merely due to near-categorical tendency for TPZs to be an extraction from a non-subject gap, where non-subjects tend to be inanimate. The relationship of TPZ to the weight factor is similar to that for LDC.

3.2.3 Predicting LDC vs. TPZ

A model was also created to determine the differences in function between LDC and TPZ, with LDC being the positive response. This model was made from considerably less data, only the 123 LDCs and 29 TPZs. The coefficients were as follows:

	Coefficient	F-value	p-value
(Intercept)	0.447162	6.48232	0.0129
animacy		16.69858	0.0001
inanimate	-3.540786		
status (coarse)		9.20495	0.0003
old	4.288793		
new	1.696242		
status (fine)		0.02521	0.8743
set	0.121352		
gf		37.65436	< .0001
subj	3.753414		
weight	0.289549	7.22701	0.0088

This model shows that *set* relation is not independently significant in predicting the realization of LDC or TPZ. Also, when coarse-grained information status is added to the model as the last factor, it does have a significant effect, but the fine-grained factor does not ($p = 0.5406$). This is different from the independent LDC and TPZ models, where the fine-grained information status factor became significant when the coarse-grained one was removed. Because the *set* relation is never significant, this model indicates that information status is a significant predictor of LDC vs. TPZ, but it is not due to the *set* relation factor. Coarse-grained information status is predictive, but for different reasons that are not clear in these data. As for the other factors, animates are 35 times more likely in LDCs than TPZs, as suspected given the aforementioned ‘anti-animacy’ effect in TPZ. LDCs are also slightly more likely to be heavy.

4 Implications for Previous Explanations

4.1 Information status factors

4.1.1 Prince

Ellen Prince (1992, 1995, 1997) has written extensively on the pragmatic function of left dislocation and topicalization. She proposes that LDC’s have three distinct functions, and TPZ’s two, with two functions common between the

constructions. This study provides some important tests of her hypotheses. It provides strong evidence for the significance of one function, but some of the others are inconsistent with the evidence from this corpus.

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. The 'simplification' occurs because, in her theory, the LDC creates a new 'processing unit' for the left-dislocated NP so it is no longer subject to the constraint that discourse-given entities be more syntactically prominent than discourse-new ones. Once the fronted NP is processed, its referring entity is now discourse old, and can be a subject, usually taking the form of a pronoun. As a supporting example, Prince presents the following discourse, where the left-dislocated NP refers to a discourse-new entity:

- (19) 'My sister got stabbed. She died. 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...' *Welcomat*, 12/2/81, p.15

The highlighted portion above is a left dislocation where the fronted NP refers the discourse-new *landlady*. This entity is the agent of the following sentence, but is disfavored for the subject position. Thus, it is left-dislocated. She demonstrates that discourse-newness is not a sufficient condition with the following hypothetical change to the discourse:

- (20) 'He started fussing with my sister and she started to scream. The landlady went up, and he laid her out. #So a **wash cloth_i**, my sister went to get **it_i/one_i** to put on her, he stabbed her in the back...'

The *wash cloth* is discourse new, but it is not in a syntactically prominent position in the following sentence, so a left dislocation is pragmatically anomalous.

This hypothesis can be operationalized for this study in that it predicts that there should be more discourse-new left dislocations from subject LDC's than there are discourse new subjects in the control data. This prediction is supported by the corpus data. A contingency table with the factors newness and subjecthood shows that there are about 8 times fewer new subject LDCs than mediated and old, but there are about 63 times fewer new subjects in the controls:

	LDC Subj	Control Subj
(21) new	9	181
med and old	76	11413

A χ^2 test of these data shows that the interaction between LDC subject and newness is significant at the $p < .0001$ level. The low numbers of LDC subjects may violate the normalcy assumptions of the χ^2 distribution, so I also performed a Fisher’s exact test, and still found significance at $p < .0001$. Also, this result persists if you remove the possible confound of first and second person pronouns, which occur rarely in LDC, but very frequently in control subjects, and are always old. Thus, there are significantly more discourse new left dislocations from subject, as Prince’s “Simplifying LDC” hypothesis predicts. 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 old and mediated 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. A partially-ordered set defined by a partial ordering on a set of entities. The ordering is reflexive, transitive, and antisymmetric, or irreflexive, intransitive, and symmetric. Examples of the Poset relation include *is-a-supertype-of*, *is-a-subtype-of*, *is-a-part-of*, among others. An example from Switchboard of the *is-a-subtype-of* relationship is as follows:

- (22) 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.

Speaker A has already introduced a a set, the other countries in the world, and then uses a left dislocation to signal that “all those European countries” is a subtype of that set. An example of the *is-a-supertype-of* is as follows:

- (23) A: I was told by somebody that works for J C Penney ’s that, uh, Ross, Ross is just one of those places that sell, sells seconds.
 B: The defect .
 A: Yeah, yeah, um, they don’t really buy the first quality, they buy the second. And, uh, **places like JC Penney’s_i**, that **they_i**’ll reject the seconds

Speaker A introduces a set of clothing stores, JC Penney and Ross, and then refers to a superset of stores like JC Penney by a left dislocation.

The fact that the set relation was one of the fine-grained information status categories available in the LINK-annotated Switchboard corpus makes this an obvious hypothesis to test. The histogram shown in figure (5) shows that there are a lot more left-dislocated NPs in set relations than NPs in control sentences. Further, as was discussed in section 3.2.1, the logistic regression shows that the fronted NP being in a set relationship is a significant predictor of the occurrence of an LDC.

The third, and final, function Prince proposes for LDCs is to amnesty ‘island’ violations in constructions that would have otherwise been topicalizations. She terms these ‘Resumptive Pronoun Topicalization’ LDCs. The intuition behind this theory is that some LDCs are actually “topicalizations in disguise” (Prince 1995). A resumptive pronoun occurs in the sentence following the fronted NP because the syntactic relationship between the two is not one that allows a filler-gap dependency. In other words, the speaker begins to produce a topicalization, but finds the gap would be inside an extraction island, so they insert a resumptive pronoun, and the construction becomes a LDC instead of a TPZ. As an example, Prince gives the following sentence (from Studs Terkel); the resumptive pronoun **it** is in a relative clause island:

- (24) **My first book**_{*i*}, I paid half of each trick to the person who gave **it**_{*i*} to me.

Prince indicates that such examples are quite rare. There are no clear examples of such LDCs in the Switchboard corpus, which is consistent with this intuition.

Prince proposes two functions for topicalization. These functions are of a different character than those hypothesized for LDCs. The first two LDC functions form two disjoint sets: a given LDC cannot have both. Prince’s third function of LDCs is basically that of topicalization. The functions of topicalization, on the other hand, overlap. Prince’s hypothesis is that all TPZs have these functions. The first function is to trigger a *Poset* inference, which is the same function as *Poset* LDCs. The data above suggests that this function is very common in topicalized NPs, although it was not a statistically significant factor in the logistic regression. Also, as with LDCs, *mediated* NPs were significantly more likely than *old* and *new*. Therefore, 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.” (Prince 1995) In this focus/focus-frame construction, as opposed to

clefts, the focus is not the fronted constituent. In topicalizations, the focus is the prosodically prominent constituent within the clause that follows the fronted NP. The focus frame is the rest of the clause, with the focussed constituent replaced by a variable. The focus frame represents the information that is “saliently and appropriately on the hearer’s mind”, and the prosodically stressed constituent is the instantiation of the variable in the focus-frame and the new information in the discourse. Prince gives an example, which contains left dislocations and topicalizations, that nicely demonstrates this singular function of topicalization and the differences between the two constructions:

- (25)a. “She had an idea for a project. She’s going to use three groups of mice.
 b. One, she’ll feed them mouse chow, just the regular stuff they make for mice.
 c. Another, she’ll feed them veggies.
 d. And **the third_i, she’ll feed e_i junk food.**”

Sentences (25b) and (c) are left dislocations because they signal that their fronted NPs, *one* and *another*, are members of the set of the three groups of mice. In (d), its fronted NP is part of the salient set as well, fulfilling the first function of topicalization. The object NP of the following clause *junk food* was tonically stressed. The open proposition of this clause, with the object NP replaced with a variable, is:

- (26) She’ll feed the third (\in {the three groups of mice}) X.

This proposition is clearly salient in the mind of the hearer, as it has been repeated in the previous two sentences. The stressed NP is the instantiation of the variable. Thus, the last sentence is instantiated as a topicalization because it serves both of the functions of this construction.

This hypothesis could be operationalized in that it predicts a strong correlation between the stressed constituent following the fronted NP and discourse newness. This could be tested by combining data from this corpus and the original Switchboard audio, which was beyond the scope of this study. One strong caveat with such an additional study is that there are only 29 tokens of topicalization in the information status-coded portion, so statistical significance would be unlikely. One would have to broaden the analysis to all 106 topicalizations and hand code the stressed NPs for newness, being careful to use the same annotation criteria as in the LINK-annotated corpus.

4.1.2 Gregory and Michaelis

Gregory & Michaelis also conduct a study using the Switchboard corpus to study LDC and TPZ. Their primary argument is that LDC has its own singular function, instead of the 3 functions that Prince argues for, and its function is specialized like TPZ. Their study compares the LDCs and TPZs they found by the criteria of givenness, anaphoricity, and persistence. A clear difference between our two studies is that they extracted LDCs and TPZs with *tgrep* syntax, and only found 177 LDCs and 44 TPZs. Clearly, any search syntax is too limited to find all of the LDCs, and so one needs to extract them by hand to some extent, as was done in this study. Although, as mentioned above, TPZs can be extracted automatically, and there are 106 of them in Switchboard.

Their givenness scale they used was just that of Gundel *et al.*. Their anaphoricity scale was a mix of the coarse- and fine-grained information status levels used in this study. Their three-valued anaphoricity scale had levels that directly corresponded to *old/mediated*, *set*, and *new* in the levels of this study. Their final scale was persistence, another measure from Givón, which is the number utterances a referent persists in the discourse after it is mentioned. Their predictions based on these scales present another problem with their methodology. Scales like givenness and anaphoricity, are highly correlated. For example, pronouns tend to be both *old/mediated* in the anaphoricity scale and *activated* in the givenness scale. To determine the individual effects of these scales, a logistic regression is needed that controls for both simultaneously.

Their final conclusion that that LDC is as distinct in pragmatic function as TPZ could be supported by the current study. It has been shown above that LDC and TPZ can be differentiated based on pragmatic factors such as information status, but this study does not elucidate exactly how. Other possible differentiating factors are Prince’s focus/focus-frame function for TPZ, and animacy, which will be discussed further below.

4.2 Animacy factors

There have been no previous studies that examined the effects of animacy on left dislocation and topicalization. The above results show 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.

The 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). Such theories would predict that a construction that caused a referent to occur first in the clause would choose the most accessible referent. Animate NPs have more inherent accessibility than inanimates, so these theories would predict animates should topicalize more. The data in this study show that this is not the case. There is other evidence that a simple ‘animate-first’ theory is inadequate. Data from the ordering of temporal adjunct PPs shows that the animacy of the subject does not affect the realization of of adjunct PPs before the subject or at the end of the clause. (Cueni *et al.* 2005)

4.3 Referential Distance

Givón (1983a, 1983b) hypothesizes that the main purpose of LDC is for ‘re-introduction’ of a topic. When a referent has been previously introduced into the discourse, but has not been mentioned recently, a left dislocation is used to ‘re-introduce’ it. He exemplifies this by the following example:

- (27) 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...

Here, *the king* is introduced at the beginning of the story but not mentioned again for three sentences, so it is re-introduced with a LDC. The re-introduction hypothesis can be operationalized by examining its prediction for the referential distance distribution of LDCs. If the fronted NP in a LDC is a re-introduction, then the referential distance distribution will have a peak at a number utterances further back than control NPs. Givón tested this hypothesis in his 1983b work, and found that LDCs have a referential distance distribution that peaks at 11-20 utterances back in the discourse. This result was found using a small corpus,; a transcript from a spoken monologue in an interview. As shown above, the data from the my annotated sample of the Switchboard corpus does not confirm Givón’s hypothesis. 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.

5 Conclusion

This study has shown that the factors that contribute to the choice of construction can be determined by corpus study and logistic regression, even for very infrequent constructions such as topicalization and left dislocation. Hypotheses from the literature about these constructions were confirmed, such as Prince's about signalling an inference of set relation. Other hypotheses were not supported by the data presented here. This study also suggests that animacy may be an important factor differentiating LDC and TPZ, a relationship that has not been studied before.

There is much future work to be done on these constructions investigating both information status and animacy factors. The correlation between prosody and information status in the focus-frame of TPZ that was suggested by Prince could be verified with information from this version of Switchboard, combined with acoustic prominence information from the audio portion. Also, the 'anti-animacy' effect seen for TPZ could be explained, perhaps by examining the features of intervening NPs in the clause.

References

- BOCK, J.K., H. LOEBELL, & R. MOREY. 1992. From conceptual roles to structural relations: Bridging the syntactic cleft. *Psychological Review* 99.150–171.
- BRESNAN, JOAN, & ANNA CUENI, 2005. Explaining the dative alternation with corpus data. paper presented at LSA Annual Meeting Jan 6-9, Oakland, CA.
- , —, TATIANA NIKITINA, & HARALD BAAYEN, 2005. Explaining the dative alternation with corpus data. Unpublished manuscript. Stanford University.
- CARLETTA, JEAN. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics* 22(2).249254.
- , SHIPRA DINGARE, MALVINA NISSIM, & TATIANA NIKITINA. 2004. Using the nite xml toolkit on the switchboard corpus to study syntactic choice: a case study. In *4th International Conference on Language Resources and Evaluation (LREC2004)*, Lisbon.
- CUENI, ANNA, NEAL SNIDER, & ANNIE ZAENEN, 2005. Boundaries to the influence of animates. paper presented at LSA Annual Meeting Jan 6-9, Oakland, CA.
- 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.
- . 1983b. Topic continuity in discourse: The functional domain of switch reference. In *Topic Continuity in Discourse: A Quantitative Cross-Language Study*, ed. by Talmy Givón, p. 5182. Amsterdam: John Benjamins.
- GODFREY, J.J., E.C. HOLLIMAN, & J. MCDANIEL. 1992. Switchboard: Telephone speech corpus for research and development. In *IEEE ICASSP*, 517–520.
- GREGORY, MICHELLE L., & LAURA A. MICHAELIS. 2001. Topicalization and left-dislocation: A functional opposition revisited. *Journal of Pragmatics* 33.1665–1706.

- GUNDEL, JEANETTE, NANCY HEDBERG, & RON ZACHARSKI. 1993. Cognitive status and the form of referring expressions in discourse. *Language* 69.274–307.
- 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.
- MARCUS, MITCHELL P., BEATRICE SANTORINI, & MARY ANN MARCINKIEWICZ. 1994. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics* 19.313–330.
- 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)*.
- PRAT-SALA, MERCE, & HOLLY P. BRANIGAN. 2000. Discourse constraints on syntactic processing in language production: A cross-linguistic study of english and spanish. *Journal of Memory and Language* 42.168–182.
- PRINCE, ELLEN. 1992. The zpg letter: subjects, definiteness, and information-status. In *Discourse description: diverse analyses of a fund raising text*, ed. by S. Thompson & W. Mann, 295–325. Philadelphia/Amsterdam: John Benjamins B.V.
- . 1995. On the limits of syntax, with reference to left-dislocation and topicalization. In *Syntax and semantics Vol. 29. The limits of syntax*, ed. by P. Culicover & L. McNally, 281–302. NY: Academic Press.
- . 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.

A Appendix

TPZ pattern:

- (28) TOP << (/NP-TPC/ ! << /more/ < (N < (/ [0-9A-Z]+/=antec .. (/NP/ < (/NONE/ < (N < (/ [0-9A-Z]+/ ~ =antec))))))

LDC pattern:

- (29) TOP << ((/NP-TPC/ < (ANTEC < (/ [0-9]+/=antec)) .. (ANAPH < (/ [0-9]+/ ~ =antec)))

Control clause pattern:

- (30) S ! << /NP-TPC/ ! << /SBARQ/ ! << /S-NOM/ ! << /SQ/

Transitive objects and ditransitive second objects:

- (31) S < (VP ! << /BE/|was|is|were|are|am|/:re\$/|/:m|being|been < ('/NP/ ! < /-N/ ! >> PRN))

Objects of raising and control verbs:

- (32) S < (VP ! << /BE/|was|is|were|are|am|/:re\$/|/:m|being|been < (S < (/NP-SBJ/ ! < /-N/ ! >> PRN)))

First objects of ditransitives:

- (33) S < (VP ! << /BE/|was|is|were|are|am|/:re\$/|/:m|being|been < ('/NP-/ ! < /-N/ ! >> PRN \$.. (/NP-/ ! < /-N/ ! >> PRN)))

Prepositional objects of prepositional ditransitives:

- (34) S < (VP ! << /BE/|was|is|were|are|am|/:re\$/|/:m|being|been < ('/NP-/ ! < /-N/ ! >> PRN \$.. (/PP-DTV/ < ('/NP/ ! < /-N/ ! >> PRN))))

Objects of prepositional arguments and adjuncts:

- (35) S < (VP ! << /BE/|was|is|were|are|am|/:re\$/|/:m|being|been < (/PP/ < ('/NP/ ! < /-N/ ! >> PRN)) ! < /PP-DTV/)