

Lexical Subjects and the Conflation Strategy

Laura A. Michaelis and Hartwell S. Francis

University of Colorado at Boulder

1 Introduction¹

People tend to focus on what is unusual, and theorists who study the relationship between discourse function and syntactic form in English are no exception to this principle. The analytic tradition that they have developed is founded on the study of MARKED sentence types—like *there*-constructions, topicalization, left dislocation and locative inversion—which allow for the noncanonical placement of a lexical argument expression. Particular attention has been paid to a class of sentence types that are hallmarks of conversational English: those constructions that permit a lexical NP which would otherwise be a SUBJECT to appear somewhere other than subject position. One such a construction is the nonstandard presentational amalgam exemplified in (1). Another is found in (2)—an instance of left dislocation involving a preclausal NP that corefers with a pronominal subject:

- (1) There was **a ball of fire** shot up through the seats in front of me.
(Lambrecht 1988: 319)
- (2) **The guy that's taken over for Gorbachev**, he's supposed to on
our side, isn't he?²

¹ This research was supported by a National Science Foundation grant (NSF-POWRE 9805829), awarded to Laura A. Michaelis. A preliminary report of the findings discussed here appears as Francis, Gregory and Michaelis 1999. We are grateful to Michelle Gregory for her participation in the initial phase of this research. We also owe thanks to Knud Lambrecht, Nancy Hedberg, Dan Jurafsky, Doug Roland, Gregory Ward and an anonymous reviewer for their insights and help.

² All examples are taken from the Switchboard corpus of English telephone conversations (Godfrey et al. 1992), unless otherwise noted. We will shortly describe the properties of the Switchboard corpus.

Why are sentences like (1-2) of such great interest to functional syntacticians? Since Kuno (1972), these sentence patterns have been used to show that the syntax of conversational English is shaped by level-mapping constraints of the kind described by Prince (1981:247). Prince argues that sentence patterns of the type illustrated in (1-2) belong to “a conspiracy of syntactic constructions resulting in the nonoccurrence of NPs low on the [familiarity] scale in subject position” (1981:247).

The current state of research in functional syntax appears to promote a paradox. On the one hand, we are required to assume that there is an unmarked, canonical, default, or pragmatically neutral alternative to the marked syntax displayed in (1-2). This pragmatic default, as described by Prince (1996) and Birner and Ward (1998) is the SV(O) pattern, exemplified in (3):

- (3) **The news coverage** showed all the, you know, the guys who didn't get hurt coming home.

On the other hand, we must acknowledge that the use of this putatively neutral option is severely restricted in spoken language. Research on conversational English has shown that sentences like (3), in which the subject role is expressed by a lexical NP, are rare, while those in which the object and oblique grammatical functions are so expressed are frequent (Prince 1992, Givón 1983b). Lambrecht (1987a) and Dubois (1987) make similar observations for, respectively, conversational French and spoken narratives in Mayan languages.³ These studies, coupled with typological findings concerning the asymmetrical behavior of subjects and objects in long-distance dependencies (Keenan and Comrie 1977,

³ An anonymous reviewer suggests that there is no paradox here if we acknowledge that, as pointed out by Gundel, Saunders and Houlihan (1983), a pattern may be unmarked with regard to its morphosyntax while being marked distributionally, i.e., rare. Certainly, a form may be marked according to one markedness diagnostic and unmarked according to another. However, it is important to notice that the decision to treat SV(0) as a default pattern for English has rarely, if ever, been based merely on a morphosyntactic simplicity metric; rather, neutral morphology has always been seen as entailing nonrestrictive use conditions, and, in particular, the lack of discourse-pragmatic constraints on argument positions. Thus, neutral syntax and neutral pragmatics have been taken as the same thing. It is this assumed equivalence that creates the paradox to which we refer here.

Raymond and Homer 1996), agreement marking (Bresnan and Mchombo 1987) and case marking (Aissen 2003), provide evidence that subjects and objects are distinct from the perspective of functional encoding.

How can the SV(O) sentence pattern be a discourse-pragmatic default if at the same time it strongly constrains the type of NP that can occupy the subject role? A coherent answer to this question will require reference to MARKEDNESS REVERSAL. As Battistella (1990:6) observes, “a feature will be unmarked in part because it is the prototypical choice in a context; but which element is prototypical in lexical and grammatical oppositions will depend to some degree on the background of social, cultural and discourse expectations”. If background can encompass genre, and genre can encompass rhetorical intent, it is clear that there are purposes for which the SV(O) pattern, and a lexical subject, are unmarked choices. One such purpose is that of grammatical description. As Lambrecht (1987a:217) observes, grammarians have traditionally chosen atypical patterns as basic objects of grammatical analysis, as when Sapir (1921) exemplifies basic English syntax by means of the “typical English sentence” *The farmer kills the duckling*. Certainly, there is nothing typical about this sentence; construction of a discourse context that would welcome it is an imaginative exercise: if the definite NPs *the farmer* and *the duckling* are to be construed as referential, the sentence is a report, but if it is a report, as opposed to, say, a gnomic statement, its inflection should be present progressive rather than simple present. Considered in another light, however, Sapir’s choice makes sense. Fully expanded sentences are useful precisely because they do not index any context. As a consequence, the reader need not perform referent recovery in a vacuum and can discern the basic argument order of the language under study, which might otherwise be obscured by the weakening of syntactic and prosodic integrity that is characteristic of indexical and anaphoric reference.

From the foregoing examples one could infer that markedness reversal, and accordingly lexical subject-encoding, is exclusively a function of genre. Certainly, the importance of genre to argument-encoding constraints should not be minimized, as we will see in the next section. However, our focus in this paper will not be inter-genre variation but intra-speaker variation within a single genre, conversational speech. In particular, we

will ask why a given speaker, at a given juncture, should choose one form of subject encoding over another. Accordingly, we will view markedness, and markedness reversal, as the real-time interplay between countervailing constraints on production, as in the recent Optimality-theoretic accounts of optional rule application (so-called stochastic OT; Boersma 1998). Our findings will be based on two different, although overlapping, data sources. One data source is the frequency and morphosyntactic distribution of lexical subjects, based on automatic sorting of the declarative sentences in an online corpus of conversational English. Wherever appropriate, this distribution is compared to that of lexical objects. The other data source is a balanced sample of the lexical subjects in the corpus, in which the authors hand-coded the discourse-pragmatic linkages of lexical-subject denotata to prior and subsequent context.

On the basis of these analyses, we will argue that lexical subjects are symptomatic of a CONFLATION STRATEGY, in which the speaker compresses into a single clause two pragmatic functions that are ordinarily performed in a sequence of two clauses: establishing a new topic and commenting about that topic. The markedness of the conflation strategy accounts for the relative rarity of productions like (3) in the corpus, but the mere fact that such productions do occur shows that speakers do not uniformly avoid the conflation strategy. Why not? We will argue that fully expanded instances of the SV(O) pattern are discourse-motivated syntactic amalgams in the sense of Lambrecht 1988. These amalgams allow speakers to simultaneously satisfy two competing optimization principles, namely, the two halves of the Gricean quantity maxim, as described by Horn (1984). These constraints have enjoyed a recent revival in optimality-theoretic accounts of case, in particular that of Aissen (2003), where they are referred to, respectively, as the ICONICITY CONSTRAINT (marked mappings receive morphological marks) and the ECONOMY CONDITION (specification of morphological marks is penalized). There is, however, a critical difference between Aissen's implementation and the current one. For Aissen the tension between the two constraints is resolved through constraint ranking, whereas in the production scenario to be described here the speaker strikes a morphosyntactic compromise between the two competing constraints, through the use of

the marked SV(O) amalgam and the forms of GIVENNESS encoding (Gundel et al. 1993) that are generally characteristic of highly accessible referents.

We begin in Section 2 with a review of models of the function of subjects in spoken English and an overview of the distribution of lexical versus pronominal NPs in the corpus under study. In Section 3 we discuss the morphosyntactic and textual properties of lexical subjects in the corpus. In Section 4 we consider Lambrecht's (1994) Principle of Separation of Reference and Role as a constraint on subject encoding in English, and describe an opposing principle that is simultaneously in play during functional encoding. In Section 5 we argue that patterns in the morphosyntactic coding of lexical subjects provide evidence for the claim that speakers' productions involve attempts to mediate between hearer- and speaker-based constraints.

2 The Discourse Function of Subjects

2.1 Topics and subjects

Researchers in functional syntax have converged on the finding that the grammatical role of subject is, at least canonically, the syntactic expression of the discourse role of TOPIC (Givón 1984, Lambrecht 1994, Van Valin and LaPolla 1997). For example, Mithun (1991:160) asserts that "[t]he function of subjects is clear: They are essentially grammaticized clause topics." The statistical connection between the roles of topic and subject also shows up in the statement of typological constraints on quantifier scope (Ioup 1975, Kuno 1991). These constraints are stated in the form of aligned scope rankings, in which topical NPs outrank nontopical NPs and subject NPs outrank nonsubject NPs. The need for separate statements concerning subjects and topics arises because subjects and topics are in principle distinct: there are, e.g., topical object denotata just as there are focal subject denotata. These scope rankings collude to induce a strong preference for wide scope of the subject quantifier in sentences whose argument structures are strongly biased in favor of topical subjects. One such argument structure is the 'transform' pattern exemplified in (4a). In this pattern, the 'raw material' argument maps to a direct grammatical function and the 'product' argument to an oblique

grammatical function (Basilico 1998). The discourse-pragmatic mapping constraints on this construction conspire to create a strong dispreference for both a topical ‘product’ argument, as shown in (4b), and a focal ‘raw material’ argument, as shown in (4c):

- (4) a. That tiny acorn grew into a beautiful oak.
 b. ??That tiny acorn grew into it.
 c. ??A tiny ACORN grew into that oak.

Because the ‘raw material’ argument must generally be topical, it must also have wide scope relative to the focal ‘product’ argument, as predicted by Ioup’s scope-assignment hierarchy. As a result, sentences like (5) are semantically anomalous: the ‘transform’ argument structure requires the subject NP to denote a topic, and thereby a specific individual. The result is a reading that conflicts with world knowledge, in which a single acorn grows into multiple oaks:

- (5) *An acorn grew into every oak.

It is clear that the scope constraint is not a function of thematic role or grammatical function, since in the pattern exemplified in (5’) the theme argument (in this case, the ‘product’ argument) can have narrow scope with respect to the ‘raw material’ argument:

- (5’) An oak grew out of every acorn.

The theme argument need not have wide scope because it need not be a topic. This is shown by the fact that this argument can be focal, as in the sentence *An OAK grew out of it*. In other words, the scope constraint follows from the linkage of a given thematic role to a given pragmatic role, as specified by a particular linking pattern.

The conception of topic that figures in the statement of discourse-pragmatic mapping constraints like those exemplified in (4-5) is that of *CLAUSE TOPIC*. Lambrecht (1994: Ch. 4) distinguishes clause topics from discourse topics by relating the former concept to the articulation of a proposition into asserted and presupposed portions.

While a clause topic is not presupposed in the way that a proposition is, a clause topic is ‘in the presupposition’ in the sense that its availability for commentary is taken for granted. The view of a clause topic as the predictable argument in a predication is captured by Gundel’s 1988 definition of topic:

Topic. An entity E is the topic of a sentence, S, iff in using S the speaker intends to increase the addressee’s knowledge about, request information about, or otherwise get the addressee to act with respect to E. (Gundel 1988a:210)

This definition of clause topic makes clear that the topic role is in principle distinct from the discourse (GIVENNESS or FAMILIARITY) status of a referent. As Lambrecht and Michaelis (1998) argue, EVOKED status does not entail topic status (pronouns, both deictic and anaphoric, may be foci). By the same token, as Gundel (1988a,b) has observed, topic status does not entail evoked status (a referent may be established in the topic role in the very act of commenting about it). Therefore, topic status and evoked status are not the same thing. However, as the “peg on which the message is hung” (Halliday 1970:161), a topic should be a referent whose appearance as an argument in the predication at hand is predictable.. The speaker and hearer, as intrinsically relevant participants, are ipso facto predictable participants; third parties, by contrast, typically become predictable participants only via prior mention. As a result, third-person subjects tend to have discourse antecedents. This idea is captured by the markedness hierarchy of shift types described in centering theory (Walker and Prince 1996): third-person subjects tend to form anaphoric chains, as in B’s response in (6):

- (6) A: But does your sister live in a big community?
 B: **She** lives, it’s a, it’s a fairly large community. **She** got real lucky, though. **She** had a boss who, uh, moved into a larger office.

Third-person subjects tend to be pronouns, but this fact is only indirectly attributable to the anaphoric function exemplified in (6).. Third-person subjects are pronouns for the same reason that deictic pronouns are: they denote entities that are already highly salient

to the conversants (Gundel et al. 1993). Third-person direct objects, by contrast, tend to be lexically expressed. This trend can be attributed to the strong tendency in conversational corpora for new referents to be introduced in postverbal position and resumed as pronominal subjects in subsequent predications:

- (7) We used to see **a husband and wife in there together** and **they** were in the same room which not all husband and wives were.

However, as mentioned, the two functions, topic-establishment and predication, may be conflated into one clause rather than distributed over two. One such example is given in (8):

- (8) **My sister** has a, she just had a baby. He's about five months old, and she was worrying about going back to work and what she was going to do with him.

In (8), both subject- and object-denotata are discourse new, and both persist as topics. While the referent of *a baby* is in the grammatical position canonically dedicated to discourse-new entities, the referent of *my sister* is not. It is this type of example, in which subject mapping represents a 'short-circuited' form of referent introduction, that will interest us here.

2.2 Distribution of subjects and objects in the corpus

The data for the current study were gathered from a syntactically parsed version of the Switchboard corpus of American English telephone conversations (Godfrey et al. 1992, Marcus et al. 1993). The Switchboard corpus is composed of approximately 2,400 diadic telephone conversations between previously unacquainted adults. The participants in the conversations vary in age and represent all major dialect groups. The parsed version consists of 400 of these conversations. Using `tgrep`, Unix search strings that allow hierarchical syntactic structures to be represented as regular expressions, we collected a total of 31,021 subjects of declarative sentences from the parsed version of the corpus. Of these, we found 91% to be pronouns, and only 9% to be lexical NPs. This distribution is shown in Table 1.

	Number	Percentage
Lexical Subjects	2,858	9%
Pronominal Subjects	28,163	91%

Table 1. Subject-type distribution for 31,021 declarative sentences

Of course, we must establish that this preference for pronominal encoding is not a fact about argument encoding as such. In order to do this, we have compared subject-type distribution to object-type distribution in the data. As shown by Table 2, the distribution of lexical objects versus pronominal objects is also highly asymmetric. However, this asymmetry is the reverse of that found among subject expressions .

	Number	Percentage
Lexical Objects	4,921	66%
Pronominal Objects	2,568	34%

Table 2. Object type distribution for 7,489 transitive sentences

Comparison of Tables 1 and 2 shows that lexical coding is strongly dispreferred for subject denotata in the corpus, while it is the preferred coding strategy for object denotata. The tendencies summarized in Tables 1 and 2 appear to be specific to the conversational genre (see Roland and Jurafsky 2002, for a discussion of genre-related biases in speech corpora). While Givón (1984) finds, in general accord with our results, that 25.6% of the subjects in a corpus of spoken English narratives are lexical, written genres yield very different results. For example, about 80% of subjects of declarative sentences in the Wall Street Journal corpus are lexical NPs (D. Roland, p.c.). Similarly, Prince (1992) finds that about 60% of the subjects in the ZPG fund-raising letter are lexical NPs. Of course, it makes sense that the subject-coding preferences should differ in spoken and written genres, since the processing of spoken language is temporally constrained in ways that visual processing is not. And yet a model of subject encoding

based solely on the recipient-design features of spoken language will have nothing to say about those cases in which spoken and written genres converge. How can we characterize the small class of lexical subjects in our conversational data? In the following section, we will pose three questions, the answers to which will determine the applicability of topic-encoding constraints, and in particular Lambrecht's (1994) Principle of Separation of Reference and Role, to our data. As we will discuss in greater detail below, Lambrecht's constraint states that the first mention of a referent cannot also be a predication about that referent. The questions are as follows:

- Do the lexical subjects in our data denote topical (as opposed to focal) entities?
- Do the lexical subjects in our data in fact denote discourse-new entities?
- Do the referents of lexical subjects persist in the discourse following introduction?

The first question pertains to the existence of an ABOUTNESS relation between the subject-referent and the proposition, as invoked by Lambrecht's constraint. The second and third questions pertain to the TOPIC INTRODUCTION function targeted by this constraint.

3 The Nature of Lexical Subjects

Given the relatively minute number of lexical NPs in subject position, one must consider whether the general discourse-pragmatic properties of subjects (topic status and evoked status) extend to this small and potentially highly anomalous class of subjects. Through an examination of sentences with lexical subjects, we find that this class is both anomalous *and* regular: like most subjects, lexical subjects denote topical arguments; unlike most subjects, however, they do not denote previously evoked referents.

3.1 Lexical subjects have topical denotata

Many researchers have pointed out that there is not a one-to-one mapping between the grammatical function of subject and the role of topic (Givón 1983a, Gundel 1988b, Lambrecht 1994). Subject denotata may instead be FOCAL. There are two focus constructions in which focal subjects can in principle occur: ARGUMENT FOCUS predications and THETIC or, equivalently, SENTENCE-FOCUS predications (Kuroda 1972, Lambrecht 1994). Upon examination of the lexical subjects in our data, we found that the semantico-pragmatic hallmarks of these two focus constructions were missing. In order to see what the relevant properties are, let us examine each of the two focus constructions in turn. Argument-focus sentences express pragmatically presupposed open propositions (Jackendoff 1972; Ch. 6), as in (9):

- (9) I was the only one who did not catch a single fish. **My daughter** caught fish, **his daughter** caught fish, he caught fish.

In the series of clauses following the first sentence, the subject NPs identify the variable in a presupposed open proposition ‘Someone caught fish’ ($x = \textit{my daughter, his daughter, etc.}$). Although argument-focus examples like (9) are found in the data, they are rare, accounting for fewer than 3 percent of the sentences with lexical subjects. In accordance with Prince (1992), who makes a similar observation, we find that argument focus is not a significant source of lexical coding in subject position. What then of sentence focus? Rather than identifying a variable in an open proposition, sentence-focus sentences present entities and/or report states of affairs. As Lambrecht argues (1994:320-322) the prosodically marked sentence-focus pattern in English is pragmatically equivalent to the syntactically marked inversion patterns of Italian and Spanish. An example of the latter is given in (10):

- (10) Sali-ó el médico
 exit:3SG:PAST the doctor
 ‘The DOCTOR came out.’
 (Ocampo 1993:356)

While the thetic pattern in Spanish reverses the canonical subject-predicate order, the English version of this construction reverses the canonical accentuation pattern (of English), in which primary prominence falls within the VP. English, as a PLASTIC-ACCENT language in terms of Vallduví (1991) and Ladd (1996), can express the thetic type by preserving canonical word order and shifting accent to the subject, as in the following example, from a conversation about the nature of the Russian military:

(11) They get real nasty, **the hyundee HELICOPTERS** come out.

If sentence-focus constructions were a significant source of lexically expressed subjects in our data, we would expect the lexical subjects in the corpus to prefer intransitive predicates, as sentence-focus sentences tend strongly to contain location and change-of-state verbs, which are predominantly intransitive (see Lambrecht 1994: Ch 5). In fact, the lexical subjects in our corpus data are no more highly correlated with intransitivity than are pronominal subjects: intransitive verbs account for about 65% of both the pronominal-subject and lexical-subject predications. Overall, predications in the Switchboard corpus are highly intransitive and highly stative, as is typical of spoken English (Thompson and Hopper 2001): copular predications (both predicational and identificational) account for about 45% of the declarative sentences in the corpus, and the stative verb *have* accounts for the majority of transitive predications.

Similarly lacking in the corpus are symptoms of sentence focus that relate to THEMATIC STRUCTURE. Because unaccusative and stative verbs select for undergoer-type subjects, sentence-focus sentences tend strongly to have patient subjects (Lambrecht 1994:310). However, the lexical subjects in our data appear to be no less agentive overall than the pronominal subjects. Using a sort based on verbal Aktionsart, we determined that agentive subjects account for fewer than 20% of both the lexical and pronominal subjects, despite the fact that the majority of both lexical and pronominal subjects have animate referents. For this reason, among others, Dubois's (1987) Given A constraint does not adequately capture the tendencies in our data. The Given A constraint states

that ergative-type arguments (i.e., subjects of transitive predications) tend to be topical, and thereby null instantiated or, as in English, pronominally expressed. However, as we have seen, pronominal expression is a feature not merely of *agentive* subject-referents but of subject referents in general. Therefore, although it would seem reasonable to assume that lexical subjects and lexical objects have common semantico-pragmatic properties, it appears that the object properties of undergoer status and focus status are neither necessary nor sufficient conditions on lexical subject-encoding. The former condition (undergoer status) is just as likely to result in pronominal as lexical encoding, and the latter condition (sentence focus) is virtually never in force. Like pronominal subjects, lexical subjects have animate, topical referents. There is, however, an important difference between lexical subjects and pronominal subjects: as we will show in the next subsection, the referents of lexical subjects strongly tend to be discourse-new.

3.2 Lexical subjects have denotata that are new to the discourse

As discussed earlier, we presume, along with a number of other theorists (Ladd 1996, Prince 1992, Lambrecht 1994, Lambrecht and Michaelis 1998), that a topical referent can in principle be a discourse-new referent. Because discourse-old referents will tend to be coded pronominally rather than lexically, one can reasonably predict that lexical subjects will in fact have discourse-new referents. In order to test this prediction, we used the scalar ANAPHORICITY measure employed by Gregory and Michaelis (2001) to code a stratified sample of approximately 280 of the lexical subjects in the corpus. This sample consisted of seven subsets of approximately forty tokens each (some tokens in each set had to be discarded for various reasons). The subsets were selected according to the morphosyntactic form of the subject NP, as follows: definitely determined NPs (e.g., *the teacher*), demonstratively determined NPs (e.g., *that place*), bare plural or mass nouns (e.g., *criminals*, *oil*), possessively determined nouns (e.g., *my uncle*), proper nouns (e.g., *Guns and Roses*), quantified nouns (e.g., *a lot of men*) and indefinitely determined NPs (e.g., *a turtle*). For each token in the sample, we examined the ten turns preceding the target utterance in order to determine whether there was (a) NO PRIOR MENTION of the lexical-denotatum, (b) PRIOR MENTION OF A SUPERSET containing the lexical-subject

denotatum, or (c) PRIOR MENTION of that denotatum. The passages in (12-13) exemplify each of the three respective points on the anaphoricity scale:

(12) No prior mention

B: Yeah, that's that's great. Well, when I was a kid my father had all different kinds of travel trailers.

A: Oh.

B: Whatever he liked he had a travel trailer, and he had a Winnebago. and he had a truck camper, and he was always buying new things like that and I hated it. I—I hated going camping and

A: Oh.

B: Wh[en]—when I got older I liked things like Caesar's Palace. You know, that's where I like to stay and and

B: Oh yeah and oh okay you can't you're gonna camp out at Caesar's Palace huh

B: Uh right exactly so this summer um **my boyfriend** lives in California

A: All right.

B: and he loves to go camping and he s[aid]: "Let's go camping", and I went.

(13) Superset mention

And it was uh it had a lot of **turtles** in it, and I got all ready to go and I was down in the water and and set to go and a and just as that boat took off **a turtle** bit me in the middle of the back.

(14) Prior mention

A: Yeah, I do too uh because doesn't make any difference uh uh whether it's a pistol or **a shotgun**, I suppose. They both do the same kind of job.

B: **Shotgun** hurts worse than a pistol does

The results of the coding for each group in the sample set are given in Table 3:⁴

⁴ Due to rounding, percentages in this and later tables add up to slightly more than 100 percent.

	No prior mention	Superset mention	Prior mention	Totals
Definite determiner	22	12	5	39
Demonstrative determiner	9	7	23	39
Bare plural/mass nouns	24	3	11	38
Possessive determiner	28	6	5	39
Proper nouns	22	1	16	39
Quantified nouns	31	4	2	37
Indefinite determiner	21	12	4	37
Totals	157 (60%)	45 (17%)	66 (25%)	268

Table 3. Anaphoricity counts for lexical subjects

These results suggest that the vast majority of the lexical subjects (77%) fall into the first two categories, and therefore have discourse-new referents. As indicated, however, we do find a relatively high percentage (25%) of lexical subjects which denote previously mentioned referents, as in (14). Why should a speaker resume lexical mention when pronominal reference could be used instead? One reason is AMBIGUITY AVOIDANCE, as in (14), where both shotguns and pistols are under discussion at the time of the relevant production, and lexical reference is thus necessary for the coherence of the comparison. Another source of high anaphoricity lexical mentions is a discourse strategy that Fox (1987) refers to as a RETURN POP: reinstatement of a reference following the intervention of some number of predications in which that referent does not play a role. The exchange in (15) gives an example of a return pop:

- (15) A: Do you have any children?
 B: Yeah, I have two, two boys, twelve and sixteen.
 A: Twelve and sixteen.

[Five turns intervene]

A: What, uh, do you feel like you have any time

B: [Laughter]
 A: to spend with them?

[Five turns intervene]

B: I have a kind of a fortunate situation, I think, right now with my sixteen year old. Um, I work two nights a week at a, I'm a librarian

A: Oh, uh huh.

B: and I work two nights a week at the Senior High School library. They keep it open, you know, for kids to do their work

A: Right.

B: and he, um, and then they, the school district hires an aide, you know, to work

A: Sure.

→ B: for me. Well, **my son** is my aide.

A: Oh, well, that's nice.

B: [Laughter] So, two nights a week for four hours we're stuck together.

In (15), speaker B introduces her two sons, begins to talk about her 16-year-old son, and then shifts to discussion of other referents (the library, the school administration, the school district) in order to describe her evening work situation at the school library. At the indicated turn, she re-establishes reference to her older son by using the full noun phrase *my son* in subject position in the predication *My son is my aide*.⁵ Such anaphoric lexical subjects are, however, relatively rare. As indicated in Table 3, the majority of the lexical subjects in the Switchboard corpus denote referents that are new to the discourse. In the next subsection, we will ask whether lexical-subject predications actually serve to establish their referents as topics.

3.3 The denotata of lexical subjects tend (weakly) to persist in the discourse

As established in the previous subsection, lexical subjects tend to denote discourse-new referents. In this respect, they are unlike prototypical topics, which denote evoked referents. The possibility remains, however, that lexical subjects are SWITCH TOPICS, and

that predications containing them have a topic-establishing function similar to that of English left dislocation, as described by Gregory and Michaelis (2001). An example of left dislocation discussed by Gregory and Michaelis is given in (16):

- (16) I like classical, but I can't deal with opera at all. And **heavy metal**,
uh, it's noisy.

Using hand-coded data from the Switchboard corpus, Gregory and Michaelis (2001) found that the preclausal NP in left dislocation (e.g., *heavy metal* in (16)) tends to denote a discourse-new and yet highly persistent referent, as compared to the preclausal NP of topicalization, a superficially similar fronting construction. They found that 65% of the preclausal-NP referents in left dislocation tokens persist in the following discourse, while the same was true of only 28% of the preclausal-NP denotata in topicalization tokens. They found a comparable, although inverted, asymmetry with regard to the anaphoricity scores of the preclausal NPs in the two constructions: only 38% of the denotata of the preclausal NPs in left dislocation were found to be anaphoric, while 75% of those in topicalization were found to be anaphoric. By combining these two tendencies, they showed that left dislocation, owing to its low anaphoricity and high topic-persistence indices, is appropriately viewed as a topic-establishing device.

If lexical-subject predications are, functionally speaking, 'compressed' instances of left dislocation, we would expect that their subject referents would have topic-persistence scores equal to those of the preclausal-NP denotata in left-dislocation tokens. In order to determine whether this was in fact the case, we used a measure of topic persistence to code a sample of lexical-subject predications. The composition of this sample was identical to that used for anaphoricity coding, as described in subsection 3.2 above. Topic persistence, according to Givón (1984:908), is "the number of times the referent persists as an argument in the subsequent ten clauses following the current clause". Because we wished to track only lexical-subject denotata and not argument

⁵ To facilitate exposition, we have chosen to ignore the pronominal mention in B's false start *and he, um, and then they, the school district hires an aide*.

denotata in general, the notion of topic persistence that we used was narrower than Givón's: we examined topic persistence as a property of only those referents that were initially introduced in canonical topic position. Further, because we were interested only in whether an entity appears again as a topic argument in subsequent utterances, and not in how long that entity persists as a topic, we used a scale with only four possible values: NO PERSISTENCE, LEXICAL PERSISTENCE, SUBSET PERSISTENCE and PRONOMINAL PERSISTENCE. This scale has one more category than that used for anaphoricity coding, as shown in Table 3: it contains an additional type of coreference link, lexical persistence. We added this category in order to determine whether lexical-subject predications might be defective topic proffers, requiring secondary introduction of the subject-denotatum in a subsequent turn. The passages in (17-20) give examples of each of the respective values on the topic-persistence scale.

(17) No persistence

- A: And then the power went off again. **Lightning** or something knocked it out, and we hooked it back up and now it's running ten or fifteen minutes fast again but on the east side of the house it runs fine
- B: Um. Oh boy, you do have a strange wiring problem there.

(18) Lexical persistence

W[ell], I had a friend who sat in on a or who was on a jury recently for a murder, but **the man** was not being tried for capital murder and so that was not even an option uh. The death penalty was not an option so and in in this case everyone on the jury felt that it should have been and they were very convinced **the man** had no redeeming uh qualities and couldn't be rehabilitated.

(19) Subset persistence

- B: Yeah I—I just have a scroll saw and a jig saw and I am really anxious to get a band saw and a router
- A: Yeah. **The band saw's** really nice I have **an Inca**. I don't know if you've ever heard of those.

(20) Pronominal Persistence

My dog was from the shelter and the idea was to get a Shepherd-Lab mix and **he** turned out to be about half that um **h[e]**—I just couldn't resist **him** at the uh **he** was well **he** was at least eighteen when **he** died because **he** was full grown when I got **him**.

The results of the topic-persistence coding for the sample are summarized in Table 4. As shown, a lexical-subject denotatum is slightly more likely to persist than not to persist: 51% of the lexical-subject denotata in the sample persisted in some fashion.

	No persistence	Lexical persistence	Subset persistence	Pronominal persistence	Totals
Definite determiner	22	4	3	10	39
Demonstrative determiner	22	2	3	12	39
Bare plural/mass nouns	19	4	1	14	38
Possessive determiner	18	4	0	17	39
Proper nouns	19	4	2	13	38
Quantified nouns	15	2	6	14	37
Indefinite determiner	15	5	2	15	37
Totals	130 (49%)	25 (9%)	17 (6%)	95 (36%)	267

Table 4. Topic-persistence counts for lexical subjects

This is not a strong tendency, and it is clearly weaker than the comparable tendency among left-dislocation tokens described by Gregory and Michaelis 2001. However, it is important to ask whether lexical denotata in general, be they subjects or preclausal NPs, persist to the same extent that pronominal denotata do. In order to answer this question, we coded a stratified, random sample of 30 nondeictic pronominal subjects for topic persistence. This sample was composed of five categories, each of which contained six tokens: the masculine singular pronoun *he*, the feminine singular pronoun *she*, the inanimate singular pronoun *it*, the plural pronoun *they*, the distal demonstrative singular pronoun *that* and the distal demonstrative plural pronoun *those*. (The proximal pronouns

this and *these* were not used because of their strong association with cataphoric constructions, e.g., *This is what I think.*) In order to control for the position of the pronominal subject in an (actual or potential) anaphoric chain, we coded only those pronominal subjects that immediately followed lexical mentions, that is, with no prior pronominal mentions intervening between the target pronoun and the most recent lexical mention. In addition, if the lexical reference before the target pronominal subject reference was a lexical subject, the target pronominal subject would be discarded in favor of another randomly selected token. In this way we ensured that we were not comparing topic-persistence ratings from overlapping sets of data (lexical subjects and pronominal subjects).

For the pronominal subject topic-persistence measure, we used a scale with five possible values: NO PERSISTENCE, LEXICAL PERSISTENCE, SUBSET PERSISTENCE, SUPERSET PERSISTENCE and PRONOMINAL PERSISTENCE. This scale contains one more value than that used for lexical-subject topic persistence in Table 4, that of superset persistence. We assigned this value to those discourse sequences in which the mention following the target pronominal subject denotes a set that includes the denotatum of the target pronominal mention, as in (24) below. In this example, Speaker A uses the pronoun *they* to refer to her own goldfish, while speaker B uses that same pronoun to refer to goldfish as a species. Examples of discourse sequences exemplifying each of the five respective categories are given in (21-25) below. In these examples, the lexical mention from which the target pronominal subject is traced is shown in bold, as are the target and subsequent mentions:

(21) No Persistence: *she* → no subsequent mention

A: But, um, my, again, **my mother back in Indiana**, we had a quilting frame in our basement, and **she** would—she would quilt whole quilts by herself.

B: Oh my god.

A: Um, yes, which I was always amazed. I have one of them, um, but, uh, you know to have the quilting frame and then to actually do that—it's a tremendous amount of work.

(22) Lexical Persistence: *he* → *this guy*

A, uh, fellow when he was much younger, uh, was tried and convicted and sentenced to death. Fortunately, in his case, the death penalty was revoked and, uh, so he served out his, his sentence until it was discovered by a fellow who was making a documentary called *The Thin Blue Line* that **this guy** had basically gotten railroaded through the judicial system.

(23) Subset Persistence: *they* → *the guy*

A: Because I know, I know **a couple of people here that work for, uh, the Army.**

B: Yeah. They civilians or, uh, military?

A: I think, I think **they**'re military. But I—

B: Uh huh

A: —mean, **the guy** is still in.

(24) Superset Persistence: *they* → *they*

A: We still have lots of little **goldfish.** **They** propagate pretty well, those little fish.

B: Yeah, **they** really do.

(25) Pronominal Persistence: *he* → *he*

A: Several of the things you mentioned were the things that, uh, **our son** has talked a lot about Texas A&M. He, but **he** thinks he wants to be a writer and I don't think that—

B: Well, then **he** should come to UT.

	No persistence	Lexical persistence	Subset persistence	Superset persistence	Pronominal persistence	Totals
<i>He</i>					5	5
<i>She</i>	1				4	5
<i>It</i>			1		4	5
<i>They</i>			1	1	3	5
<i>That</i>	1		1		3	5

<i>Those</i>	1	1	1		2	5
Totals	3 (10%)	1 (6%)	4 (13%)	1 (3%)	21 (70%)	30

Table 5. Topic-persistence counts for pronominal subjects

As shown in Table 5, the denotata of pronominal subjects persist in 90% of the discourse segments coded for the sample. Thus, it appears that pronominal subjects are far more likely to head anaphoric chains than are lexical subjects or left-dislocated lexical NPs. This fact, however, need not be taken as evidence against the view that lexical subject-encoding and left dislocation are topic-establishing strategies. As we have seen, both lexical subjects and left-dislocated NPs have discourse-new referents. And, after all, a topic proffer is still a topic proffer even when it is unsuccessful. Further, it simply stands to reason that a large percentage of topic proffers would in fact be unsuccessful: each potential topic referent might have many competing referents in a given conversation. By contrast, pronominal subjects represent already established referents: these referents are more likely to play a role in subsequent predications because established referents, like plants that have taken root, are more likely to persevere than those which have not. Thus, despite relatively weak persistence scores, lexical subjects are reasonably regarded as new topics. In the next section we will ask why new topics are not typically introduced by means of lexical-subject predications.

4 Constraints on subject position

4.1 The Principle of Separation of Reference and Role

Many researchers have observed that subject position is pragmatically constrained. For example, Prince (1992), using a small written corpus, finds that subjects tend to represent discourse-old information, whereas objects do not. Our results are consistent with this finding. However, our focus is upon the constraint that underlies this tendency, and upon the morphosyntactic form of productions that represent violations of this constraint. In

particular, we ask: what does this marked linguistic choice have to do with other kinds of marked linguistic behaviors as described by Grice (1975) and Horn (1984)?

The literature offers several candidate constraints. Chafe (1987) proposes that intonation units are aligned with information units in a one-to-one fashion. A corollary of this constraint is described by Chafe as the “light starting-point” principle: subject NPs do not constitute either intonation units or information units. Dubois (1987) proposes both the Given A constraint referenced in Section 3.1 and an information-load constraint similar to Chafe’s, whereby each predication can include at most one new argument. Lambrecht (1994) proposes the Principle of Separation of Reference and Role (PSRR), which he states in the form of a maxim: “Do not introduce a referent and talk about it in the same clause” (p. 185). For the model of markedness to be developed here, we will adopt Lambrecht’s PSRR, because, unlike competing constraints, it relates subject properties to a theory of topic, provides a model of what constitutes cooperative referring behavior and predicts (correctly) that referent introduction is typically accomplished by two-clause sequences like that exemplified in (26):

- (26) The, the procedure is utterly humiliating. You go in there with **the doctor**, he makes you take off all your clothes.

In (26) a referent, *the doctor*, is introduced in an oblique position, after which it is resumed pronominally as a subject. By means of this pattern, the two tasks, introducing the referent and predicating a property of it, are performed by two different clauses. Hearer burden is kept relatively low, as the hearer can open a file on a new referent prior to entering a property there.

The PSRR accounts for the existence of Prince’s 1996 “conspiracy of syntactic constructions” which insure pronominal expression of the subject role. These constructions include inversion constructions in Italian and Spanish (Ocampo 1993), the French presentational *y a* cleft (Lambrecht 1987b), subject-agreement morphology in Chichewa (Bresnan and Mchombo 1987), and pleonastic detachment constructions like English left dislocation (Prince 1981, Birner and Ward 1998, Ziv 1994, Gregory and Michaelis 2001). All such constructions provide for the introduction of a referent outside

of the clause in which it serves as distinguished argument. These constructions are thus the grammaticized analogs of the biclausal sequences for referent introduction exemplified in (26). The foregoing observations suggest that the PSRR has both synchronic and diachronic validity: it not only motivates aspects of conversational behavior but also explains the etiology of specific grammatical constructions in terms of recipient-design features. However, the fact that 9% of the subjects in declarative sentences are lexical NPs, and that most of the denotata of these NPs are both topical and discourse-new, indicates that the PSRR is a violable constraint.

4.2 Lexical subjects as PSRR violations

In light of the PSRR and similar constraints, passages like (27) pose a problem:

- (27) **Context:** A conversation about daycare.
 She sent him to kindergarten. As soon as he went there, **the teacher** took one look at him and he threw up again.

When discourse-new entities, e.g., *the teacher*, are used as clause topics, as in (27), we presume, by the logic of the PSRR, that the hearer burden is increased. As in cases of pragmatic accommodation described by Stalnaker (1974), the hearer must make inferences about the speaker's intentions in order to preserve the assumption that the speaker's referring behavior is cooperative. Scrutiny of examples like (27) leads us to ask two questions. First, what would drive a speaker to override the PSRR? Second, are violations of the PSRR constrained? We suggest that the lexical subjects in our data reflect the speaker's attempt to mediate a conflict between the two halves of the Gricean quantity maxim. The use of a lexical subject, like deletion up to recoverability, as described by Horn (1984), reduces speaker burden without compromising comprehension. The mediation involves the interplay between two halves of the Gricean quantity maxim as described by Horn:

- (28) Quantity 1. Hearer-based lower bound on information
 Say as much as you can.

- (29) Quantity 2. Speaker-based upper bound on information
Say no more than you must.

The second half of the quantity maxim, which sets the upper bound on information content, privileges the conflation strategy: it is more economical to convey in one clause what might otherwise be conveyed in two. The first half of the quantity maxim, which sets the lower bound on information content, prevents the speaker's economy from operating unchecked. This lower-bound constraint is similar to Clark and Haviland's (1977:4) GIVEN-NEW CONTRACT, in which "the speaker tries, to the best of his ability, to make the structure of his utterances congruent with his knowledge of the listener's mental world". We propose that the introduction of discourse-new referents as topics in subject position is motivated by the speaker's economy, Q2, and constrained by the speaker's adherence to the hearer's economy, Q1. Example (30) is indicative of the constrained nature of PSRR violations in the corpus:

- (30) I have an opportunity to go to uh Paris, France uh with my friend
in April. She is–**her family**, you know, lives there.

The speaker's economy motivates the use of a single clause to both introduce a new referent, *her family*, and predicate a property of that referent. However, the hearer-based lower bound prevents the speaker from introducing an unrecoverable referent in subject position. Recoverability is facilitated in (30) by use of the pronominal possessive determiner *her*, which links *her family* to an evoked discourse entity, *my friend*. The morphosyntactic coding patterns in our data indicate that speakers who choose to override the PSRR produce referents that are accessible and anchored. In the next section we will look at definite determination, possessive determination, and pronominal-subject relatives as measures of accessibility and anchoring.

5 Morphosyntactic coding of lexical subjects

Tendencies in the morphosyntactic coding of the lexical NPs in subject as against object position suggests that speakers who violate the PSRR choose referring expressions which denote referents that are either ACCESSIBLE via the speech context or ANCHORED to

referents that have already been evoked in the discourse. Table 6 provides an overview of the morphosyntactic properties of lexical subjects in the corpus data. It shows a comparison of determination patterns in subject NPs versus object NPs.

	Indefinite article	Definite article	Possessive	Undetermined	Demonstrative article
Subjects	65 (2%)	1,070 (37%)	724 (25%)	837 (30%)	171 (6%)
Objects	1,419 (29%)	784 (16%)	346 (7%)	1547 (31%)	825 (17%)

Table 6. Distribution of determiners for lexical subjects vs. lexical objects

In Table 6, the Undetermined category includes proper nouns, quantified plural NPs, and bare plural and mass NPs. While proper nouns can be considered semantically definite, in the sense that they pick out mutually identifiable referents, proper nouns lack morphosyntactic determination, and were therefore placed in the Undetermined group. Proper names account for about 8% of the subject tokens and about 2% of the object tokens; this asymmetry is probably a consequence of the markedness of animate object denotata (Aissen 2003). After quantified nominals, the largest percentage of tokens in the Undetermined group are bare nominals denoting types and unbounded sets, as in (31-32), respectively:

- (31) So **glass** has been recycled for a long time.
- (32) Hopefully, the next generation it won't even be an issue. I mean **people** will just look at you as a person and not as a man or a woman.

With Fillmore et al. (forthcoming: Ch. 2), we assume that bare nominal arguments are implicitly quantified, whether generically, as in (31), or partitively, as in (32). Thus, bare nominals have referential properties similar to those of weak and strong quantified nominals, e.g., *a lot of glass*, *most people*.

The Demonstrative article class contains both the (plural and singular) distal and proximal demonstrative articles. The majority of proximal demonstratives in both subject

and object position appear to be cataphoric, or referent-introducing, indefinite *this*, as described by Gernsbacher and Shroyer (1989).⁶ This use is exemplified for subjects and objects by (33-34), respectively:

- (33) I was getting my oil changed and I was sitting in the little lobby it's one of these you know five-minute change places and **this guy** comes storming into the lobby there and he says um
- (34) I see **this guy** in my building with some I think it's like a Golden Retriever or something like that.

Indefinite *this* is far more likely to occur in object position, as in (33), than in subject position, as in (34). This asymmetry is in turn responsible for the much higher incidence of demonstrative determiners among objects than among subjects.

For our purposes, the critical asymmetries are those involve which involve the incidence of possessive and definite articles. These two groups of articles form a natural class, in that both entail uniqueness and mutual identifiability of the NP's denotatum. As shown in Table 5, 62% of lexical subject NPs contain either a possessive or definite article, while only 2% contain an indefinite article. This asymmetry is reversed in the case of lexical objects: only 23% of lexical object NPs contain a possessive or definite article, while a full 30% contain an indefinite article. Thus, morphosyntactic definiteness is not a property of lexical NPs in general; it is specifically a property of lexical subjects. (The same observation holds, *ceteris paribus*, for morphosyntactic indefiniteness and lexical objects.) In the next two subsections, we will draw out the semantico-pragmatic

⁶ Anaphoric uses of the proximal demonstrative determiner are also found in subject position, as in the following passage:

- (a) When I first started it was like I saw **someone** at a Halloween party. **This lady** was from Turkey and she'd been belly dancing since she was four years old, you know.

In this passage, the demonstrative NP *this lady* refers to the individual introduced in the prior predication, *I saw someone at a Halloween party*.

implications of these morphosyntactic facts. In Section 5.1 we will discuss grammatical definiteness as an indicator of discourse accessibility. In Section 5.2 we will discuss the anchoring function of pronominal possessives and object-trace relative clauses.

5.1 Accessibility

There are a number of measures of the activation statuses of referents, including scales based upon FAMILIARITY (Prince 1981), IDENTIFIABILITY (Lambrecht 1994) and GIVENNESS (Gundel et al. 1993). We will focus here on the Givenness Hierarchy because it closely relates forms to cognitive states. The size of our data set makes such a system highly appealing, since it allows us to make inferences about the status of a referent from automatically searchable properties. The Givenness Hierarchy (Gundel et al. 1993) is based on the assumption that nominal encoding properties reflect the speaker's assumptions about the salience of the denoted referent in the mind of the addressee. The cognitive-status descriptors which are arrayed along the scale are conceived as necessary and sufficient conditions upon the use of each corresponding form of nominal reference. The Givenness Hierarchy can be exemplified as follows. The origin of the scale is equated with the lowest degree of cognitive salience that a referent may be presumed to have, TYPE IDENTIFIABLE status. This status licenses the use of indefinite referring expressions, which may denote either a specific referent or a nonspecific referent, as in (35a). The highest degree of cognitive salience, IN FOCUS status, licenses the use of an unstressed pronominal expression, as in (35b). UNIQUELY IDENTIFIABLE status falls between these two extremes; it licenses the use of definite determination, as in (35c). The definite determiner is used when the hearer can identify the referent on the basis of the NP alone.

- (35) a. You know, a, a, **a sixty-two year old guy** is less likely to be put on death row from what I've seen.
 b. **He, he** repairs it, gives it back to you, and takes your hundred dollars.
 c. **The, uh, Governor**, you know, has been trying to decide whether he's going to commute it or not.

Although the categories of the Givenness Hierarchy map to discrete attentional/memorial states, the hierarchy is implicational, in the sense that the conditions which license use of a particular referring form also license the use of any lower ranked form: “each status entails (and is therefore included by) all lower statuses, but not vice versa” (Gundel et al. 1993:276). An appropriate analogy comes from the set or cardinal numbers, an ascending halfline from zero to infinity: although the numbers mark discrete points on the scale, an assertion like *I have three dollars* entails that the speaker possesses all smaller amounts of currency (e.g., it entails the assertion *I have two dollars*); it is also upward compatible: it could be truthfully uttered if in fact I had five dollars, although, via Quantity 2, such a statement is likely to mislead. By the same token, the indefinite NP *a lady* can be used to implicate (via Quantity 2) a greater degree of cognitive salience of the denoted referent than that required to license use of indefinite NP. For example, the NP *a lady* could denote a referent that is currently in focus, as when a woman says of a suitor *He knows how to treat a lady*. In this instance, application of Quantity 1 would lead to the inference of disjunct reference (the NP *a lady* does not refer to the speaker). As in other instances of Gricean inference, only context will determine which of the two quantity-based inferences is applicable.

By allowing for inferentially based enrichments, the Givenness Hierarchy abandons strict reliance on morphosyntactic form, and comes to resemble the Familiarity Scale, as proposed by Prince 1981. Prince’s Familiarity Scale assigns labels to referential expressions based upon the relationship of the denoted entity to the discourse rather than upon the form of the referring expression. Category assignments are intended to capture the source of referent activation, and Prince’s scale thereby provides information that is not captured by the Givenness Hierarchy. For example, the Familiarity Scale captures two different means by which an NP-denotatum can achieve the discourse status that licenses the use of the definite article: the BRIDGING INFERENCE, in the sense of Clark and Haviland 1977, and NP-internal information (Prince 1981:237). Following Prince, we will refer to NPs which trigger bridging inferences as FRAME INFERRABLES. Frame inferrables include those NPs whose referents are identifiable by virtue of belonging to a semantic frame that is currently active. The passage in (36) provides an example of this class:

- (36) Uh, actually I lived over in Europe for a couple of years, I lived in Germany and in Germany they don't have the jury system. What they do is they have, uh, three judges, basically. And you get up there and **the prosecuting attorney** presents his evidence...

In (31), the NP *the prosecuting attorney* denotes an entity which, although new to the discourse, is highly recoverable by virtue of its relationship to the previously evoked frame, the litigation frame. Prince refers to those NPs which contain contextual links as CONTAINING INFERRABLES. Containing inferrables include both explicit and implicit partitive NPs. An example of a lexical subject that is an implicit partitive is given in (37):

- (37) You know, **the good ones** do cost maybe sixty dollars a week.

The subject NP in (37) contains implicit reference to a previously evoked set—that of daycare centers. In the present study, we made use of an expanded version of the containing inferrable category which included demonstratively determined nominal expressions that denote members of previously evoked sets, as in (38):

- (38) **This school** does, so uh it's been interesting.

The predication in (38) follows a discussion of colleges that offer tuition breaks to children of alumni; the subject denotatum is a member of this set. Another type of containing inferrable not explicitly recognized by Prince is that in which the bridge to uniquely identifiable status is a deictic or anaphoric referent contained in the NP. We will refer to containing inferrables of this type as ANCHORED INFERRABLES. Examples are given in (39-40):

- (39) But **our neighbor** decided he didn't like cats and shot one of them.

- (40) And at the same time, **the budget he sent to Congress** has tax and fee increases.

In (39), as in the majority of the possessive determiner subject NPs, the referent is linked to the discourse context through the use of a first person possessive determiner. In (40), the referent is linked to the discourse context through the use of an object relative clause, the subject of which is an active discourse entity.

Using the three classes of inferrables described above, as well as additional Familiarity-based categories, we hand-coded our stratified sample of lexical subjects in order to achieve a more detailed picture of activation status than that provided by form-based givenness coding alone. We will first discuss the application of the Givenness Hierarchy to the total set of lexical subjects, in which categorization was based solely on morphosyntactic form. We will then discuss the application of Familiarity-based coding to the stratified sample.

The distribution of morphosyntactic forms in subject and object position suggests that lexical subjects denote entities that are more accessible than those denoted by lexical objects. Table 6 above summarizes the asymmetric distribution of morphological forms for lexical subjects and objects. In total, 62% of lexical subjects are at least uniquely identifiable, compared to only 23% of the lexical objects. The contrast between subjects and objects with regard to type-identifiable status is also striking: only 2% of the subjects are potentially upper-bounded at type-identifiable status, as compared to 29% of the objects. Based on the correlations between morphological form and givenness status described above, we conclude that the referents of lexical subjects are typically at least uniquely identifiable.

Table 7⁷ summarizes the distribution of lexical NPs in the stratified sample according to a modified version of the Familiarity scale. This scale includes the three

⁷ Abbreviations used in Table 7 are as follows. In the left column, **Def**=definite NPs, **Dem**=demonstratively determined NPs, **Undet**=undetermined NPs (plural or mass), **Poss**=possessively determined NPs, **Proper**=proper nouns, **Quant**=quantified nouns, **Indef**=NPs containing the indefinite article, and **Tot**=totals. In the top row, **Anc**=anchored inferrable NPs, **Coi**=containing inferrable NPs, **Ev**=evoked NPs, **Fra**=frame inferrables, **Gen**=generic NPs, **Pred**=predicative NPs, **UnU**=unused NPs, **BN**=brand new NPs, **SE**=situationally evoked NPs, **BA**=brand new anchored NPs, and **Tot**=totals.

classes of inferrables discussed above, as well as Prince's Familiarity classes EVOKED, SITUATIONALLY EVOKED, UNUSED, BRAND NEW, and BRAND NEW ANCHORED. Examples of each of these statuses are given in (41-45), respectively:

- (41) My uh wife's grandmother had Alzheimer's and they were going to put her into a a nursing home, and **the nursing home** made them come and take her back because she was being a, a, you know, a, a nuisance.
- (42) **This country** seems a little behind on that.
- (43) They, they, they decided, you know, **George Bush**, who, who's the main owner of the Rangers, decided that, uh, they'd stay in Arlington.
- (44) And **this lady** was going to work she came home and everything in her house was total gone light fixtures everything and
- (45) **A friend of mine** gave me a clock kit.

We further expanded the Familiarity scale by including two categories of referring expressions whose denotational properties come not from discourse context but rather from the constructions in which each appears: PREDICATIVE NPS (46a-b) and GENERIC NPS (47a-b):

- (46) Predicative NPs
 - a. And **all we do** is metric stuff.
 - b. Sometimes, um, usually **the reason I will turn it on** is to hear the news.
- (47) Generic NPs
 - a. Well, they are evidently. But um I mean ev[en]—evidently **a normal cow** produces that much too.
 - b. **People** should be required to give a couple of years for the good of the country.

Predicative NPs are those NPs that are subjects of equative copular predications, and are also possible postverbal NPs (as in, e.g., *Metric stuff is all we do*). Generic NPs are those NPs that appear as arguments of predications which, typically by virtue of verbal aspect, are construed as describing stable properties of the world rather than episodes (Langacker 1996). Generic NPs have contingent or, equivalently, attributive reference; that is, they allow conditional paraphrases of the following form: ‘Should you find an instance of this type (e.g., normal cow), it will have the following property (e.g., producing a certain amount of waste)’.

	Anc	Coi	Ev	Fra	Gen	Pred	UnU	BN	SE	BA	Tot
Def	2	6	6	15	2	6	2	0	0	0	39
Dem	0	3	24	5	1	0	0	2	4	0	39
Undet	0	0	10	11	16	1	0	0	0	0	38
Poss	24	0	5	9	0	0	0	1	0	0	39
Proper	0	0	15	15	1	0	3	0	3	0	37
Quant	1	4	2	13	9	8	0	0	0	0	37
Indef	5	4	1	5	13	5	0	3	0	1	37
Tot	32 12%	17 6%	63 24%	73 27%	42 16%	20 8%	5 2%	6 2%	7 3%	1 0%	266

Table 7. Familiarity status of lexical subjects

The results of the familiarity coding are compatible with those reported for determiner type in Table 6. In that table, we saw that 62% of the lexical subjects contained definite determiners, corresponding to uniquely identifiable status. In Table 7, we see that 72% of the lexical subjects in the sample have either inferrable or evoked referents. Two of the trends reported in Table 7 are ostensibly puzzling. The first such trend is the relatively low percentage of anchored inferrables (12%). Given the fairly high percentage of lexical subjects whose heads denote kinship or social relationships (e.g., *my sister*, *my supervisor*), one would expect a higher percentage of anchored inferrables in the sample. However, the relatively low percentage of anchored inferrables is most likely an artifact of

the sampling method: because the vast majority of anchored inferrable NPs contain possessive determiners, and because only 14% of the lexical subjects in the sample contain possessive determiners, anchored inferrables will necessarily top out at the 14% level in the sample. The second such trend is the relatively high percentage of evoked referents. In light of the fact that, as reported in Table 4, 60% of the lexical subjects have discourse-new referents, the fact that 24% of the lexical subjects in the sample have evoked referents may appear surprising. However, this percentage is approximately the same as that reported for discourse-old referents in Table 4. As shown in that table, 25% of the lexical subjects denote previously mentioned referents.

The high percentage of lexical subjects that have inferrable denotata makes sense in light of the lower and upper bounds that constrain the use of lexical-subject denotata. As Prince (1992:305) observes, inferrable NPs, like *a friend of mine* in (45) represent HEARER-NEW referents, and therefore DISCOURSE-NEW referents. At the same time, as Prince (*ibid*) points out, inferrable referents share properties with HEARER-OLD and DISCOURSE-OLD referents, in that they are linked to the linguistic or extralinguistic context. Along these same lines, Givón (1983a:10) suggests that the denotata of some nominal expressions, such as kin terms, “are in the file permanently, and are thus always accessible to speakers/hearers as part of their *generic* firmament” (emphasis in original). Lambrecht (1994:114) similarly views inferrable status as the product of pragmatic accommodation. He argues that the speaker exploits the potential for easy activation of kin-term referent and “conveys a request to the hearer to act as if the referent of the NP were already pragmatically available”. Birner and Ward (1998) take a stronger position concerning the commonalities between hearer-old and discourse-old statuses. In their analysis of word-order inversion, they claim that “inferrable elements and explicitly evoked elements behave as a single class of discourse-old information for the purpose of word order inversion” (1998:178). In other words, inferrable NPs are ideal lexical subjects: they have the accessible denotata that are prototypical of lexical subjects and the discourse-new denotata that are prototypical of lexical NPs. In the next section, we will focus on a specific class of inferrable NPs—anchored inferrables—and, in particular, asymmetries in the use of anchoring devices in subject versus object position.

5.2 Anchoring

In this section, we examine referents that are rendered recoverable by virtue of a link to a discourse-active entity, in particular the speaker. According to the definition provided by Prince (1981:236) says, “A discourse entity is anchored if the NP representing it is linked by means of another NP or anchor properly contained in it to some other discourse entity.” We discuss two anchors here, possessive determiners and relative clauses. As seen in Table 5 above, pronominal determiners like *my* and *her* are more frequently associated with lexical subjects than with lexical objects. While 25% of lexical subjects contain possessive determiners, only 7% of lexical objects do. Example (47) illustrates a typical use of an anchored inferrable containing a deictic possessor:

- (47) A: I’m a single mother. I have three children.
 B: Oh, I see, uh huh.
 A: So, uh, right now, we’re on, we get, you know, aid from the state at this point because there’s no other way to do it. And **my ex-husband** just sort of took off and doesn’t pay child support.

In (47), the discourse-new denotatum of *my ex-husband* is anchored to the speaker through her use of the possessive pronoun *my*. The frame is deictically established in this case. We postulate that the high percentage of pronominally possessed subjects reflects the speaker’s drive to ease referent recoverability. A similar pattern can be discerned in Table 8, which shows the distribution of object-trace and subject-trace relative clauses by grammatical function.

	Subject relativization	Object relativization
Lexical Subject	102 (29%)	244 (71%)
Lexical Object	249 (60%)	164 (40%)

Table 8. Distribution of relative-clause types for lexical subjects and objects

Object relativization occurs in 71% of the lexical subjects that are postmodified with a relative clause. Why should this strong bias exist? As Fox and Thompson (1990) suggest, this type of relative clause anchors the discourse-new referent to a discourse-active frame, as in (48):

- (48) Our friend, **the President**, right now, says no new taxes. We should and especially, if anything, be cutting taxes now because of the recession and at the same time, **the budget he sent to Congress** has tax and fee increases, so uh I know the politicians uh aren't straightforward.

In (48), the pronominal reference to *the President* in the relative clause guides the hearer to relate the discourse-new referent of *the budget* to an entity in the context. By contrast, as shown in Table 8, the majority of the lexical objects that are postmodified by relative clauses have subject-trace relative clauses, as in (49):

- (49) We do oil well services. So, a lot of our clients are oil companies, big oil companies, and they go out to, we have **engineers who, uh, go out to the oil well, to the client's oil well, and work with a lot of heavy equipment and put tools down the oil well and stuff.**

In (49), the object nominal *engineers* is the subject of the relative clause. Its referent is also discourse-new, and introduced as the direct object of *have*. Why shouldn't the relative clause perform the same anchoring function that it does in (48)? The answer is that there is no need to anchor the denotatum of *engineers* in (49): while (48) is a violation of the PSRR, (49) is not. Thus, the use of an object relative is strongly indicated in the case of a marked association between the topic role and discourse-new status. As we have seen, patterns in the use of possessive determiners and relative-clause modifiers indicate that the denotata of lexical subjects are bound by recoverability constraints that do not constrain the denotata of lexical objects.

6 Conclusion

The data presented in this study demonstrate that (a) mappings to subject position in English conversation are constrained by the PSRR and (b) this constraint can be violated on the basis of Q2, the speaker's economy. However, as a function of Q1, the hearer's economy, violations of the PSRR are relatively constrained: speakers who use the conventionalized abbreviations that usage affords nonetheless work to ease the processing burden on hearers. Speakers accomplish this through morphosyntactic choice. Specifically, when violating the PSRR by the use of a lexical NP in subject position, speakers tend to minimize the potential impact on processing by choosing lexical NPs that are accessible via definite marking or anchored to the previous discourse by possessive personal pronouns or by object-trace relative clauses with pronominal subjects. Thus, we conclude that lexical subjects in conversational English are hybrids. As lexical NPs, they denote NEW referents. As subjects, they denote RECOVERABLE referents. This hybrid character reflects what Horn (1984) has shown to be the most fundamental dialectic underlying generalizations both about inference and about linguistic choice.

References

- Aissen, Judith. 2003. Differential Object Marking: Iconicity vs. Economy. *Natural Language and Linguistic Theory* 21.435-483.
- Basilico, David. 1998. Object Position and Predication Forms. *Natural Language and Linguistic Theory* 16.491-539.
- Battistella, Edwin L. 1990. *Markedness: The Evaluative Superstructure of Language*. Albany: State University of New York Press.
- Birner, Betty J., and Gregory Ward. 1998. *Information Status and Noncanonical Word Order in English*. Amsterdam: John Benjamins.
- Boersma, Paul. 1998. *Functional Phonology*. The Hague: Holland Academic Graphics.
- Bresnan, Joan and Sam Mchombo. 1987. Topic, Pronoun and Agreement in Chichewa. *Language* 63.741-782.

- Chafe, Wallace. 1987. Cognitive Constraints on Information Flow. Coherence and Grounding in Discourse, ed. by R. Tomlin, 21-51. Amsterdam: John Benjamins.
- Clark, Herbert and S. E. Haviland. 1977. Comprehension and the Given-New Contract. Discourse Production and Comprehension, ed. by R.O. Freedle, 1-40. Hillsdale, NJ: Lawrence Erlbaum.
- Du Bois, John. 1987. The Discourse Basis of Ergativity. *Language* 63.805-855.
- Fillmore, Charles J., Paul Kay, Laura A. Michaelis and Ivan A. Sag. forthcoming. Construction Grammar. Stanford: CSLI Publications.
- Fox, Barbara A. 1987. Discourse Structure and Anaphora: Written and Conversational English. New York: Cambridge University Press.
- Fox, Barbara and Sandra A. Thompson. 1990. A Discourse Explanation of the Grammar of Relative Clauses in English Conversation. *Language* 66.297-316.
- Foley, William A. and Robert D. Van Valin. Functional Syntax and Universal Grammar. New York: Cambridge University Press.
- Francis, Hartwell S., Michelle L. Gregory and Laura A. Michaelis. 1999. Are Lexical Subjects Deviant? Proceedings of the Thirty-fifth Regional Meeting of the Chicago Linguistics Society, Volume 1, 85-97. Chicago: Chicago Linguistics Society.
- Gernsbacher, Morton Ann and Suzanne Shroyer. 1989. The Cataphoric Use of the Indefinite *this* in Spoken Narratives. *Memory and Cognition* 17.536-540.
- Givón, Talmy. 1984. Syntax: A Functional Typological Introduction. Amsterdam: John Benjamins.
- Givón, Talmy. 1983a. Topic Continuity in Discourse: An Introduction. Topic Continuity in Discourse, ed. by T. Givón, 4-41. Amsterdam: John Benjamins.
- Givón, Talmy. 1983b. Topic Continuity in Spoken English. Topic Continuity in Discourse, ed. by T. Givón, 343-363. Amsterdam: John Benjamins.
- Givón, Talmy. 1984. Syntax: A Functional Typological Introduction. Amsterdam: John Benjamins.
- Godfrey J., E. Holliman and J. McDaniel. 1992. SWITCHBOARD: Telephone Speech Corpus for Research and Development. Proceedings of ICASSP-92, San Francisco. 517-520.

- Gregory, Michelle L. and Laura A. Michaelis. 2001. Topicalization and Left-Dislocation: A Functional Opposition Revisited. *Journal of Pragmatics* 33.1665-1706.
- Grice, H. Paul. 1975. Logic in Conversation. *Syntax and Semantics, Volume 3: Speech Acts*, ed. by P. Cole and J. L. Morgan, 41-58. New York: Academic Press.
- Gundel, Jeanette K. 1988a. Universals of Topic-Comment Structure. *Studies in Syntactic Typology*, ed. by M. Hammond, E. Moravcsik and J. Wirth, 209-239. Amsterdam: John Benjamins.
- Gundel, Jeanette K. 1988b. *The Role of Topic and Comment in Linguistic Theory*. New York: Garland Publishing, Inc.
- Gundel, Jeanette, Kathleen Houlihan and Gerald A. Sanders. 1983. Markedness and Distribution in Morphology and Syntax. *Markedness*, ed. by F. Eckman, E. Moravcsik and J. Wirth, 107-138. New York: Plenum Press.
- Gundel, Jeanette K., Nancy Hedberg, and Ron Zacharski. 1993. Referring Expressions in Discourse. *Language* 69.274-307.
- Halliday, M.A.K. 1970. Language Structure and Language Function. *New Horizons in Linguistics*, ed. by J. Lyons, 140-165. Baltimore: Penguin Books, Ltd.
- Horn, Laurence R. 1984. Toward a New Taxonomy for Pragmatic Inference: Q-Based and R-Based Implicature. *Meaning, Form and Use in Context: Linguistic Applications*, ed. by D. Schiffrin, 11-42. Washington, DC: Georgetown University Press.
- Ioup, Georgette. 1975. Some Universals for Quantifier Scope. *Syntax and Semantics 4*, ed. by J. Kimball, 37-58. New York: Academic Press.
- Jackendoff, Ray. 1972. *Semantic Interpretation in Generative Grammar*. Cambridge, MA: MIT Press.
- Keenan, Edward and Bernard Comrie. 1977. Noun Phrase Accessibility and Universal Grammar. *Linguistic Inquiry* 8: 63-99.
- Kuno, Susumo. 1972. Functional-Sentence Perspective: A Case Study from Japanese and English. *Linguistic Inquiry* 3. 269-320.
- Kuno, Susumo. 1991. Remarks on Quantifier Scope. *Current English Linguistics in Japan*, ed. by H. Nakajima, 261-287. Berlin: Mouton de Gruyter.

- Kuroda, S.-Y. 1972. The Categorical and the Thetic Judgment: Evidence from Japanese Syntax. *Foundations of Language* 9.153-185.
- Ladd, Robert. 1996. *Intonational Phonology*. Cambridge: Cambridge University Press.
- Lambrecht, Knud. 1987a. On the Status of SVO Sentences in French Discourse. *Coherence and Grounding in Discourse*, ed. by R. Tomlin, 217-261. Amsterdam: John Benjamins.
- Lambrecht, Knud. 1987b. Presentational Cleft Constructions in Spoken French. *Clause Combining in Grammar and Discourse*, ed. by J. Haiman and S.A. Thompson, 135-179. Amsterdam: John Benjamins.
- Lambrecht, Knud. 1988. There was a Farmer had a Dog: Syntactic Amalgams Revisited. *The Proceedings of the Fourteenth Annual Meeting of the Berkeley Linguistics Society*, ed. by S. Axmaker, A. Jaisser and H. Singmaster, 319-339. Berkeley: BLS, Inc.
- Lambrecht, Knud. 1994. *Information Structure and Sentence Form: Topic, Focus, and the Mental Representations of Discourse Referents*. Cambridge: Cambridge University Press.
- Lambrecht, Knud and Laura A. Michaelis. 1998. Sentence Accent in Information Questions: Default and Projection. *Linguistics and Philosophy* 21.477-544.
- Langacker, Ronald W. 1996. A Constraint on Progressive Generics. *Conceptual Structure, Discourse and Language*, ed. by A. Goldberg, 289-302. Stanford: CSLI Publications.
- Marcus, Mitchell, Beatrice Santorini and May Ann Marcinkiewicz. 1993. Building A Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics* 19.313-330.
- Mithun, Marianne. 1991. The Role of Motivation in the Emergence of Grammatical Categories: The Grammaticization of Subjects. *Approaches to Grammaticalization, Volume 2*, ed. by E. C. Traugott and B. Heine, 159-184. Amsterdam: John Benjamins.

- Ocampo, Francisco. 1993. The Introduction of New Referents in French and Spanish Discourse: One Constraint, Two Strategies. *Linguistic Perspectives on the Romance Languages*, ed. by W. J. Ashby, M. Mithun, G. Perissinotto, and E. Raposo, 351-362. Amsterdam: John Benjamins.
- Prince, Ellen. 1981. Toward a New Taxonomy for Given-New Information. *Radical Pragmatics*, ed. by P. Cole, 223-255. New York: Academic Press.
- Prince, Ellen. 1992. The ZPG Letter: Subjects, Definiteness, and Information Status. *Discourse Description: Diverse Linguistic Analyses of a Fund-Raising Text*, ed. by W. C. Mann and S. A. Thompson, 295-325. Philadelphia: John Benjamins.
- Raymond, William D. and Kristin Homer. 1996. The Interaction of Participant Role and Pragmatic Function in the Selection of Question Form. *The Proceedings of the Twenty-Second Annual Meeting of the Berkeley Linguistics Society*, ed. by J. Johnson, M. Juge and J. Moxley, 316-327. Berkeley: BLS, Inc.
- Roland, Douglas and Daniel Jurafsky. 2002. Verb Sense and Verb Subcategorization Probabilities. *The Lexical Basis of Sentence Processing: Formal, Computational and Experimental Issues*, ed. by S. Stevenson and P. Merlo, 325-346. Amsterdam: John Benjamins.
- Sapir, Edward. 1921. *Language: An Introduction to the Study of Speech*. New York: Harcourt, Brace and Company.
- Stalnaker, Robert. 1974. Pragmatic Presuppositions. *Semantics and Philosophy*, ed. by M.K. Munitz and P. Unger, 197-213. New York: New York University Press.
- Thompson, Sandra A. and Paul J. Hopper. 2001. Transitivity, Clause Structure, and Argument Structure: Evidence from Conversation. *Frequency and the Emergence of Linguistic Structure*, ed. by J. Bybee, 28-60. Amsterdam: Benjamins.
- Vallduví, Enric. 1991. The Role of Plasticity in the Association of Focus and Prominence. *Proceedings of the Eastern States Conference on Linguistics 7*: 295-306.
- Walker, Marilyn A. and Ellen F. Prince. 1996. A Bilateral Approach to Givenness: a Hearer-Status Algorithm and a Centering Algorithm. *Reference and Referent Accessibility*, ed. by T. Fretheim and J. K. Gundel, 291-306. Philadelphia: John Benjamins.

Ziv, Yael. 1994. Left and Right Dislocations: Discourse Functions and Anaphora. *Journal of Pragmatics* 22: 629-645.