# Why subject relatives prevail: Constraints versus constructional licensing

CECILY JILL DUFFIELD AND LAURA A. MICHAELIS*

*University of Colorado at Boulder*

*Abstract*

*Relative clauses containing subject relative-pronouns (e.g.* that go to Utah all the time*) are the prevalent type both across languages (Keenan and Comrie 1977) and in conversation, accounting for 65% of relative clauses in the American National Corpus (Reali and Christiansen 2007) and 67% of relative clauses in the corpus examined for this study, the Switchboard corpus. This fact appears attributable to parsing preferences, as per Hawkins (1999, 2004), Gibson (1998) and Gibson et al. (2005): subject extractions are the most local filler-gap dependency and therefore impose the lowest burden on short-term memory. This explanation, however, not only lacks strong psycholinguistic support but also fails to explain a major pattern in Switchboard: subject relatives are not preferred across the board but only as modifiers of postverbal (object and oblique) nominals. We propose that the preference for subject relatives is an effect not of general-purpose interpretive or encoding constraints but rather of constructional licensing: the subject relative belongs to an entrenched syntactic routine, the Presentational Relative construction, e.g.* I have friends that clip articles *(McCawley 1981; Lambrecht 1987, 1988, 2002). We investigate this hypothesis by examining the formal, semantic and pragmatic*

* Correspondence address: Laura A. Michaelis, Department of Linguistics, University of Colorado, Boulder, CO 80309, USA. Email: laura.michaelis@colorado.edu. This paper grew out of talks presented to the 2009 Annual Meeting of the Linguistics Society of America in San Francisco, CA in January 2009, the University of Colorado's Institute of Cognitive Science in March 2010 and several informal gatherings of faculty and students in CU's Department of Linguistics. We are grateful to audience members for comments and criticisms that have immensely improved this paper, in particular Ivan Sag, Al Kim, Bhuvana Narasimhan, Barbara Fox, Lise Menn, Jena Hwang, Les Sikos, Susan Brown, Michael Thomas, Steve Duman, Nick Williams, Archna Bhatia and David Harper. In addition, we are enduringly thankful to Gary McClelland for his generous help with statistical models and to Knud Lambrecht for early discussion and guiding insights. Finally, we thank an anonymous reviewer for insightful and constructive comments on an earlier draft. All remaining errors are, of course, our own.

*Language and Cognition* 3–2 (2011), 171–208
DOI 10.1515/LANGCOG.2011.007

1866–9808/11/0003–0171

*properties of relative-clause modifiers of postverbal nominals in the Switchboard corpus.*

*Keywords*
*syntax, Construction Grammar, relative clauses, presentational relative clauses, sentence processing, corpus linguistics, discourse pragmatics, locality*

## 1. Introduction

Relative clauses have played an important role in the development of syntactic theory, whether they are seen as the products of extraction rules of a universal nature (see Goodluck and Rochemont 1992, and Alexiadou et al. 2000 for an overview) or as patterns with specific communicative functions (Fox and Thompson 1990; Goldberg 2006: 146). A parameter that has proven significant in the analysis of relative-clause syntax is the grammatical function of the relative pronoun: it may be a subject, as in (1), an object as in (2) or an oblique, as in (3):

(1) I like cars that are designed with human beings in mind.
(2) I always like the letters that they read.
(3) They have choir that they go to.

The pattern illustrated in (1), which we will refer to as the *subject relative*, has a special status both across languages and in English usage patterns. Subject relatives are the dominant type in the accessibility hierarchy, a widely discussed implicational universal (Keenan and Comrie 1977). The accessibility hierarchy provides a set of contingent predictions concerning the grammatical function that relative pronouns can occupy in a given language. According to this hierarchy, if a language has object relative-pronouns, it will also have subject relative-pronouns, and if a language has oblique relative-pronouns, it will also have subject and object relative-pronouns. As a corollary, subject relatives are the least typologically marked relative-clause type. Subject relatives also appear to be the most frequently attested type in English-language corpora. In a study of relative clauses in the American National Corpus, a 22-million-word corpus containing a wide variety of spoken and written genres, Reali and Christiansen (2007) found that subject relatives account for 65% of relative clause types.

What accounts for the prevalence of subject relatives both across languages and in English corpora? In this paper, we will weigh two competing modes of explanation for this structural preference, each of which encapsulates a distinct view of the nature of syntactic generalizations. According to the first mode of explanation, which we will refer to as the *constraint-based view*, syntactic gen-

https://doi.org/10.1515/LANGCOG.2011.007 Published online by Cambridge University Press

eralizations are very general.[1] This view is closely connected to Chomsky's 'rule-free' conception of grammar (Chomsky 1989, 1995). Under the rule-free conception, traditionally recognized grammatical constructions (e.g. the imperative construction), and even the category-specific phrase-structure rules of early Transformational Grammar, are "taxonomic epiphenomena" (Chomsky 1989: 43)—predictable products of the interaction between universal constraints on phrase composition (most saliently the X′ template) and constraints on sentence processing (e.g. the Domain Minimization Constraint, as applied by Lohse et al. 2004). Proponents of the constraint-based view attribute the prevalence of subject relatives to constraints governing the processing of filler-gap dependencies, which effectively suppress other relative-clause configurations. Such accounts focus on the structural relationship between the relative pronoun and its coindexed trace, to the exclusion of other properties, including the lexical identity of the matrix verb and the discourse status of the NP referents in the relative-clause configuration (e.g. the head noun).

According to the second mode of explanation, which we will refer to as the *construction-based view*, syntactic generalizations range from the general to the very specific. This view is informed by Construction Grammar (Fillmore et al. 1988; Goldberg 1995, 2006; Michaelis and Lambrecht 1996; Kay 2002; Sag 2010a, *inter alia*), which in turn is closely connected to the exemplar-based model of grammar proposed by Bybee (2007) and others. According to Construction Grammar, the grammar is a set of constructions: phrasal patterns that people use to do things (e.g. issue an order, ask a question). Constructions mean what they mean in the same way that words do: via convention. Constructions combine syntactic, semantic, pragmatic and even phonological information, and may call for specific words or word classes. Under the construction-based view, the prevalence of subject relatives stems not from the action of general-purpose constraints that create a dispreference for competing relative-clause configurations, but rather from the communicative function served by subject relatives. It is the construction-based view that we will uphold in this paper. Based on a statistical study of the function of relative clauses in a corpus of conversational English, we will propose that the prevalence of the subject-

1. The term *constraint* is used here in a very general sense, to include both the 'soft' rules of Optimality Theory and the inviolable principles and parameters of Chomskyan theory. Whether based on 'soft' or 'hard' rules, constraint-based theories explain non-occurrences and structural preferences based on general configurational properties (e.g. the number of phrasal nodes or words that separate two terminal nodes). A well formed (or preferred) structure is simply one that has not incurred a constraint violation (or has incurred fewer violations than an alternative structure). In construction-based explanation, by contrast, an expression is well formed only if it represents a pairing of phonological form and meaning that is licensed by some construction of the grammar, however specific. See Zwicky (1994) and Malouf (2003) for a discussion.

https://doi.org/10.1515/LANGCOG.2011.007 Published online by Cambridge University Press

relative configuration is a fact not about its internal structure but about its usefulness—in particular its role in a highly entrenched grammatical construction, the Presentational Relative construction, e.g. *I have friends that go to Utah all the time*. As described by Lambrecht (1987, 1988, 2000, 2002) for both English and French, the Presentational Relative is a discourse-pragmatic 'short cut' that allows speakers to conflate the distinct functions of referent introduction and property predication into a single clause. We will support this constructionist explanation by demonstrating that the majority of subject-relative tokens in a conversational corpus bear formal, semantic and discourse-pragmatic hallmarks of presentational function.

The remainder of this paper will be structured as follows. Section 2 will provide a critical assessment of attempts to attribute the prevalence of subject relatives to general-purpose constraints on interpretation and production, while outlining the alternative, construction-based view of the structural preference at issue. Section 3 will describe the discourse-pragmatic motivations for the Presentational Relative construction, contrast this construction with other relative-clause patterns and explain the role of the subject-relative pattern within the Presentational Relative construction. Section 4 will concern methodology: it will outline the procedure by which we extracted, annotated and analyzed the relative-clause tokens used in this study. Section 5 will report the results of the study. Section 6 will offer concluding remarks and discuss further implications of this study for syntactic theory, in particular for the doctrine of syntactic locality.

## 2. Accounting for the distribution of subject relatives

Models based on the processing of filler-gap dependencies, and in particular the structural or linear distance between filler and gap, have provided powerful tools for explaining structural preferences within the domain of relative clauses. Prominent exemplars of this approach include Hawkins (1999, 2004), Gibson (1998) and Gibson et al. (2005). As Hawkins observes, fewer nodes intervene between the subject relative-pronoun and its trace than, say, between an object relative-pronoun and its trace, thus narrowing the structural domain within which the interpreter must find the antecedent of the trace. According to this view, subject relatives prevail because they place a lower burden on working memory than do other types of relative clauses. This hypothesis appears to be supported by psycholinguistic studies. For example, in a series of eye-tracking experiments, Traxler et al. (2002) found that sentences containing object-relative-clause modifiers of subjects, as in (4), were more difficult to process than sentences containing subject-relative clauses, as in (5), during the relative clause and the matrix verb:

https://doi.org/10.1515/LANGCOG.2011.007 Published online by Cambridge University Press

(4) The senator who the President insulted later apologized.
(5) The senator who insulted the President later apologized.

Wells et al. (2009), however, counter the argument that difficulty in processing object relative clauses is due to demands placed on working memory. They present an experience-based account, arguing that this effect may be due to the relative frequency and regularity of the structures encountered. Following MacDonald and Christiansen (2002), they describe object relatives as having a unique word order: they are less 'regular' than subject relative clauses, which display standard English SVO word order. This factor, in conjunction with the lower frequency of object relatives versus subject relatives, accounts for processing effects seen in object relative-clause comprehension. Moreover, Wells et al. demonstrate that difficulty in processing object relative clauses may be mitigated by experience. In a large-scale study lasting several weeks, adult participants presented with greater experience with relative clauses in reading tasks showed an increase in reading speeds for object relatives over subject relatives as compared to a control group presented with reading tasks containing no relative clauses. Thus, the ease (or difficulty) with which speakers process relative clauses can be attributed to experience, statistical learning and regularity.

The importance of contextual factors in relative-clause processing is also shown by Mak et al. (2008). They suggest that the processing advantage for subject relatives found in many studies in fact reflects a discourse-pragmatic disadvantage for object relatives—a disadvantage that is readily overcome. Based on the presumption that the most 'topicworthy' argument in a clause will tend to be expressed as the subject, they predict, in accordance with Traxler et al. (2002), that there will be a processing advantage for subject relative-clauses like (5). In (5), the senator is the topic of both the main clause and the relative clause, as indicated by the fact that this participant is realized as the subject of both clauses. By contrast, in (4), the subject of the relative clause does not refer to the main-clause topic (*the senator*), but rather to a different participant (*the President*), presumably leading to an increased processing burden. But Mak et al. also predict that the preference for subject relatives can be overridden in context: if another entity has greater topicality in context than the one referred to by the relative pronoun, the preference for a subject relative will be eliminated. For example, if (4) is uttered in a context in which the President has greater discourse salience than the senator, (4) will be easier to process than (5), especially if the NP *the President* in (4) is replaced by the pronoun *he* (Warren and Gibson 2002 make a similar prediction). Based on a reading-time study of Dutch relative clauses, Mak et al. confirm this prediction. They conclude that "[r]eaders choose the entity that is most topicworthy as the subject of the relative clause" (Mak et al. 2008: 181). These results suggest that

https://doi.org/10.1515/LANGCOG.2011.007 Published online by Cambridge University Press

there is in fact no general processing advantage for subject relatives, and they therefore undercut a processing explanation for the prevalence of subject relatives. Roland et al. (2007) provide a strong independent confirmation of the discourse effect observed by Mak et al. Using a participant-paced reading-time experiment, Roland et al. looked at the contribution of anaphoric linkage to the processing of object-relative-clause modifiers of subjects. They contrast neutral contexts like (6) to "discourse-appropriate" contexts like (7), in which a lexical subject within the relative clause (in this case, *the banker*) is anaphorically linked to prior context:

(6) There was a dinner party Saturday night. The lady that **the banker** visited enjoyed the meal.
(7) **The banker** was friendly. The lady that **the banker** visited enjoyed the meal.

They find that "[p]roviding an appropriate discourse context can eliminate the processing difficulty found in the relative clause region of sentences containing object relative clauses" (ibid.).

Further, if the prevalence of subject relatives were in fact due to the structural distance between filler and gap, structurally isomorphic filler-gap dependencies, in particular subject *wh*-questions (e.g. *Who won the game?*), would also be prevalent. In fact, however, subject *wh*-questions are strongly dispreferred, as indicated both by statistical usage trends and grammatical markedness phenomena. For example, in a study of questions in two English conversational corpora, Homer (2000) found that only 5% of tokens were subject questions of any kind (yes/no, indirect or *wh*-questions). Similarly, a wide variety of studies attest to the morphosyntactic markedness of subject extractions cross-linguistically. Ouhalla (1993), for example, demonstrates that subject extractions in null-subject languages trigger an 'Anti-Agreement Effect', in which the verb does not agree with the extracted subject but instead has invariant third-person singular form. Moreover, Aissen (1999) shows that focal agentive arguments require additional verb morphology in the Mayan languages Tzotzil and Tz'utujil. According to Van Valin and LaPolla (1997: 211), marked patterns of this nature result from violations of a "restriction on focal elements appearing preverbally"—violations that may be prevented by the use of 'work around' constructions for agent-focus interrogatives. Among these constructions are passive in the Sotho language Sesotho (Demuth 1989) and clefts in spoken French (Lambrecht 1994: Ch. 5). The spoken French examples in (8–9) illustrate, respectively, the preferred cleft strategy for an agent-focus interrogative and the less favored strategy, in which the focal word is the subject:

(8) C'est qui qui a téléphoné? (lit. 'It is who who has telephoned?')
(9) #Qui a téléphoné? ('Who has telephoned?')

https://doi.org/10.1515/LANGCOG.2011.007 Published online by Cambridge University Press

Such findings suggest that the statistical and typological prevalence of the subject-relative construction is not the product of a transconstructional preference for local filler-gap dependencies.

Moreover, facts of conversational English suggest that there is actually no across-the-board preference for subject relatives. For example, in a study of the Switchboard corpus of conversational English, Michaelis and Francis (2007) found that *object* relatives are the preferred modifiers of nominals in subject position: object relatives like that in (10) account for 71% of the relative clauses that modify nominals in subject position:

(10) Our friend the President right now says no new taxes [but] 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. (Michaelis and Francis 2007: example (48))

The preference for object relatives as modifiers of subject nominals has a straightforward discourse-pragmatic motivation—one that is tied to the dispreference for discourse-new subjects in conversation (Prince 1992; Lambrecht 1994: Ch. 4; Van Valin and LaPolla 1997: Ch. 5). As Michaelis and Francis (2007) observe, fewer than 10% of all subjects of declarative clauses in the Switchboard corpus are lexically headed expressions. This trend makes sense if we presume, following Lambrecht (1994), Mithun (1991) and others, that subjects are grammaticized clause topics. Object relative-clauses can be said to enhance the intelligibility of new mentions in the grammatical role conventionally reserved for topical (*qua* recoverable) referents; such relative clauses contain anaphorically linked arguments that serve to 'anchor' the discourse-new referent being modified to a previously mentioned one (Fox and Thompson 1990; Michaelis and Francis 2007). For example, in (10) above, the anaphoric subject *he* of the relative clause *he sent to Congress* links the discourse-new subject referent *the budget* to a previously introduced referent (*the President*).

In what grammatical contexts are subject relatives preferred? Subject relatives appear to be the most common relative-clause modifier of object and oblique nominals. According to Michaelis and Francis (2007), subject relatives account for 65% of the relative clauses that modify object or oblique nominals in the parsed portion of the Switchboard corpus. The current study, which also makes use of the parsed portion of the Switchboard corpus, confirms the Michaelis and Francis findings.[2] We assume here that the affinity between

2. The percentages reported in this study are slightly different from those reported by Michaelis and Francis (2007) in the relative-clause portion of their study, despite the fact that both studies are based on the parsed portion of the Switchboard corpus. This discrepancy may be due to the

https://doi.org/10.1515/LANGCOG.2011.007 Published online by Cambridge University Press

nonsubject nominals and subject relatives is what causes subject relatives to prevail overall. Our reasoning goes as follows. Object and oblique arguments are much more likely than subject arguments to be lexical NPs (e.g. *my neighbor*) than pronouns (Prince 1992; Michaelis and Francis 2007), and lexical expression is generally a prerequisite for modification, whether the modifier is a prenominal adjective or a postnominal relative clause. Thus, it stands to reason that the relative-clause type preferred by nonsubject nominals, the subject relative, will account for the majority of relative-clause tokens overall.

## 3. The presentational relative construction

We have seen that, counter to the predictions of a processing model, the prevalence of subject relatives has a local rather than global source: the structural configuration containing a nonsubject nominal head. But we must now address the question of why nonsubject nominals should prefer subject relatives in the first place. This preference lacks a discourse-pragmatic motivation akin to the 'anchoring function' adduced above for object-relative modifiers of subjects. We will argue here that what accounts for the prevalence of subject relatives, in those grammatical contexts in which they do prevail, is not a general-purpose discourse-pragmatic or processing constraint, but rather the role of the subject relative in an entrenched conversational routine: the presentational relative construction (PRC), referred to by McCawley (1981, 1988: 449–451) as the *pseudo-relative* construction. Examples of the PRC, taken from the conversational data collected for this study, are given in (11–13):

(11) But I have all these friends that, wherever you go, they, they sit down and the next thing you know they pull out of their uh bags some their most recent uh needle craft.
(12) They have a fish that's called an oyster cracker.
(13) There was a story of a woman last year who, who actually did slip on the ice and, and like sprained her ankle.

It is important to recognize that the PRC is distinct, both syntactically and semantically, from the restrictive-modification scheme that is typically used to exemplify relative clauses. The latter pattern is illustrated by the boldfaced portion of (14):

(14) And, you know, I want **a car that I can work on**, because I think it just costs too much even to get the oil changed anymore.

---

fact that the current study used a more restrictive set of relative-clause tokens. The relevant restrictions are discussed in Section 5.

https://doi.org/10.1515/LANGCOG.2011.007 Published online by Cambridge University Press

Following Lambrecht (1994: 51–56), we assume that a restrictive relative clause like that in (14) presupposes an open proposition. In the case of (14), this open proposition is presumably 'I can work on x'. That is, the speaker of (14) assumes that the hearer believes him capable of repairing *some* cars (in a previous turn he has discussed doing engine work on his 1970 Chevy). It is only by virtue of being mutually known that the proposition 'I can work on x' becomes a property useful for distinguishing the cars that he desires from those he does not. Crucially, the type of car being described in (14) belongs to a discourse-active set: the set of all cars. In this sense, the head nominal of a restrictive relative clause resembles the detached NP of a topicalization predication like (15):

(15) **Most** RAP, I don't LIKE.

In (15), the denotatum of the boldfaced preclausal NP (*most rap*) contrasts with other members of its set—in this case, the set of all rap-music genres (Prince 1984; Ward and Prince 1991). What makes a restrictive relative clause restrictive is that it limits the set of entities at issue to a subset of those under discussion.

We find, however, that the restrictive relative-clause model is of limited use in describing relative-clause productions in conversation. Most of these do not express mutually known properties, but instead assert properties. As described by McCawley (1981), the assertoric relative clause, illustrated by the Google examples in (16–17), is an essential part of the PRC:

(16) You will have friends, of course, who are in the same boat as you.
(17) I have my mother who is an Irish-Italian, and my father who is African, so I have the taste buds of an Italian and the spice of an African.

As McCawley points out (1981: 115), relative clauses like those in (16–17) convey assertions otherwise conveyed by matrix clauses (e.g. *Your friends will be in the same boat as you*, *My mother is Irish-Italian*). For its part, the matrix clause conveys the restriction on the range of the existential quantifier rather than an assertion. In (16), for example, the matrix clause conveys the restriction 'x is a friend'. Consequently, a pseudo-relative, unlike a restrictive relative, is a required rather than optional part of the clause in which it appears. This is shown by the fact that (17) cannot reasonably be construed as asserting 'I have my mother'. Why put the assertion in the position ordinarily reserved for a modifier? Speaking in evolutionary terms, one might say that the PRC is an exaptation from the existing schema for (restrictive) clausal modification. The PRC strategy enables the speaker to strike a balance between two countervailing discourse-pragmatic pressures—speaker-based effort conservation and hearer-based explicitness. By using a relative clause to express an assertion, the speaker can make a single sentence perform tasks that otherwise require two sentences: introducing a referent and predicating a property of that referent

https://doi.org/10.1515/LANGCOG.2011.007 Published online by Cambridge University Press

(Lambrecht 1987, 1988; Michaelis and Francis 2007). At the same time, the PRC strategy enables the speaker to avoid violating a hearer-based information-packaging constraint that Lambrecht (1994: 184–191) refers to as the Principle of Separation of Reference and Role (PSRR): "the grammatical principle whereby the lexical representation of a topic referent takes place separately from the designation of the referent's role as an argument in a proposition" (ibid.: 184). He describes the PSRR in the form of a maxim: "Do not introduce a referent and talk about it in the same clause" (ibid.). The PRC strategy enables the speaker to distribute the referent-introduction and predication functions over two clauses—the main and subordinate clauses, respectively. Thus, while (18) is a *prima facie* violation of the PSRR, the PRC version, in (17) above, is not:

(18) My mother is Irish-Italian.

It is here that we can see why the subject-relative pattern, but not the nonsubject-relative pattern, is a critical part of the PRC: relative clauses like *who is Irish-Italian* are in essence covert main-clause predicates, and, as (18) demonstrates, can be converted to such by the removal of the relative pronoun. The same cannot be said, for example, of the nonsubject relative clause in (2), repeated as (19):

(19) I always like the letters that they read.

If we were to convert the relative clause in (19) to a main clause, the result would be something like (20):

(20) They read the letters.

Clearly, (20) is an inadequate paraphrase of (19) because it does not mention either the speaker or the speaker's attitude toward the letters. Even if (20) were an adequate paraphrase of (19), it would not be a PSRR violation like the paraphrase in (18). To understand the PRC is to understand that it is an avoidance strategy: it is used when the less prolix, monoclausal encoding option is foreclosed for pragmatic reasons. Developmental data substantiate the primacy of the PRC as a conversational strategy: Diessel and Tomasello (2000) show that the PRC (which they define as a relative-clause token containing a copular matrix verb and intransitive relative clause) accounts for a large portion of children's early relative-clause productions. The conversational findings that we will describe suggest that adults are equally reliant on the PRC.

Semantico-pragmatic and syntactic diagnostics support the claim that the relative clauses of PRC tokens, unlike restrictive relative clauses, convey assertions. The relevant diagnostics involve negation, question-answer pairs, and insertion of parenthetical material. Let us discuss each of these tests in turn. Example (21) contains a restrictive relative clause. Because this relative clause is not part of the assertion, it is not within the scope of negation:

https://doi.org/10.1515/LANGCOG.2011.007 Published online by Cambridge University Press

(21) And certainly we don't have **the eye-stinging variety** that you get in the big cities.
⇒ You get the eye stinging variety in the big cities.

While the speaker of (21) denies having acrid smog in her area, this negation leaves intact the claim that those in the big cities get such smog. By contrast, when the main clause of a PRC predication is negated, as in (22), that negation has scope as well over the relative clause:

(22) If you don't have **one [viz. an aerobics instructor]** that's fun, [ . . . ] not acting like she's enjoying what she's doing, the class is not going to get out uh what they should get out of the class.
⇒ If *one's not fun*, not enjoying it . . . the class is not going to get out what they should get out of it.

Question-answer pairs can also be used to demonstrate that the relative clauses of PRC tokens express assertions. When the material inside the relative clause is asserted, as in PRC tokens, the entire matrix-clause utterance forms an appropriate answer to a question about information inside the relative clause. Consider the question-answer pair in (23), in which a PRC token forms an appropriate response:[3]

(23) Q: Where do your engineers go?
A: We have **engineers** who uh go out to the client's oil well.

By contrast, (24) shows that matrix-clause utterances containing restrictive relative clauses are not acceptable in such question-answer pairs:

(24) Q: What are people involved in?
A: #Uh you know, I feel bad for **the people** that are involved in that uh GM deal there in in Arlington.

Finally, as McCawley (1981: 106) proposes, the head nominal and relative clause of the PRC, unlike those of the restrictive-relative scheme, do not form a constituent, thus allowing them to be separated by adverbials like *of course*, as in (16) above, or parenthetical clauses, as in (25):

(25) There's a restaurant in, um, right outside of Reading, Pennsylvania, it's called Alfredo's, that does not look like a restaurant that you would really want to recommend to a lot of people.

By contrast, such 'intrusions' appear ungrammatical in restrictive relatives, as shown by (26–27):

3. In (23–24), the question is made up, while the 'answer' is a token attested in the Switchboard data.

https://doi.org/10.1515/LANGCOG.2011.007 Published online by Cambridge University Press

(26) ??I spoke to friends, of course, who liked the speech.
(27) ??Our hotel is across the street from a restaurant, it's called Alfredo's, that looks fairly decent.

Similarly, Lambrecht (1987, 1988, 2002) argues that PRC sentences like (11–13) and (16–17) contain ternary-branching VPs whose daughters are, respectively, a stative or depictive verb, a postverbal NP denoting a discourse-new referent and a subject-relative clause that predicates a property of this referent. We presume that the incidence of the PRC pattern in conversation is high, or at least high enough to account for the apparently strong affinity between object/oblique nominal heads and subject relatives, and thence the overall prevalence of subject relatives.

In order to substantiate the claim that the PRC is responsible for a preponderance of the subject relatives in conversation, we will show that a significant portion of the subject-relative tokens in a conversational corpus have formal, pragmatic and semantic properties that are symptomatic of presentational function. Like Michaelis and Francis (2007), we will base our analysis on the Switchboard Treebank corpus (Godfrey et al. 1992; Marcus et al. 1993), a syntactically parsed version of the Switchboard corpus of American English telephone conversations. Because of the overall rarity of lexical (and thus modifiable) subjects, and the rarity of subject relatives as modifiers of subjects, we have chosen to focus on those relative clauses that modify object and oblique nominal arguments of verbs.[4] Within this narrowed data set, we will compare object/oblique relatives (e.g. *who I met*) to subject relatives (e.g. *who knows you*) using three properties shown to be closely associated with the PRC in the literature. The first such property is definiteness. Since the function of the PRC is to introduce entities, and since discourse-new entities tend to be expressed

4. The criteria that we used to select object and oblique nominal arguments verbs were fairly inclusive. These criteria admit nominal heads that are direct objects of transitive verbs (e.g. *I caught my first bass that was actually big enough to keep*), second objects of ditransitive verbs (e.g. *She wasn't going to buy me something I was going to grow out of next week*), nominal heads that are complements of copular verbs and predicators like *be*, *be the same as* and *be like* (e.g. *It was a moving man pulled right up to her house*, *You're like the scout that goes ahead of the team*) and nominal heads that are governed by prepositions licensed by transitive and intransitive verbs (e.g. *I sure did have a mind lock about the movies I've seen*, *It depends on the crime that's been committed*). Also included were objects of adjective-licensed prepositions (e.g. *Murder is* ***hard*** *on the people that were related to the victim*, *I might have been more* ***sympathetic*** *with the person who got caught*). Such tokens were considered to have the matrix verb *be*. Finally, we included nominal complements of noun-licensed prepositions, when the licensing noun is itself the complement of a verb. An example of this type of token is given in (13): *There was a story of a woman last year who who actually did slip on the ice and, and like sprained her ankle*. Here, the nominal head of the relative clause is the daughter of a PP complement of a noun (*story*) that is the argument of a verb, *be*.

https://doi.org/10.1515/LANGCOG.2011.007 Published online by Cambridge University Press

by indefinite NPs, verbs in PRC predications will tend to have indefinite NPs as their direct or oblique second arguments. If the PRC is in fact responsible for the majority of subject-relative modifiers of objects, then objects modified by subject relatives will tend more strongly to be indefinite than objects modified by object and oblique relatives. The second property that we will view as symptomatic of PRC function is the lexical identity of the matrix verb. If the majority of subject-relative modifiers of objects are PRC tokens, then presentational matrix verbs (e.g. *have*, *be*) will be more likely to take an object or oblique argument that is modified by a subject relative than one that is modified by an object relative. The third and final property is discourse-pragmatic rather than lexical or morphological: the potential for monoclausal paraphrase. If in fact that majority of subject-relative modifiers of objects are PRC tokens, then these tokens will be more likely to admit of a monopropositional paraphrase than those tokens containing an object modified by an object or oblique relative. The logic here is that the biclausal PRC pattern is a 'stand in' for a monoclausal predication that, while semantically well-formed, violates the PSRR. An example of a successful monopropositional paraphrase is given in (18) above, while an example of an unsuccessful monopropositional paraphrase is given in (20) above.

In the following sections, we will describe the methods used to collect and code our data, and the results of our study.

## 4. Data and methods

As mentioned in Section 2, the relative-clause tokens analyzed in this study were extracted from the Switchboard Treebank corpus, a syntactically parsed portion of the Switchboard corpus of American English telephone conversations. The Switchboard corpus is composed of approximately 2,400 telephone conversations between previously unacquainted adults. Each conversation lasts about five minutes and concerns a pre-selected topic (e.g. pets, hobbies, cars). The participants in these conversations vary in age and represent a wide variety of American dialects (e.g. Western, New England, South Midland). The Switchboard Treebank corpus consists of 400 of these conversations, which were hand-parsed according to Treebank annotation conventions (Marcus et al. 1993). A user can retrieve all instances of a given syntactic pattern in this corpus by using tgrep search strings, in which regular expressions containing syntactic tags represent tree structures and their 'leaf' nodes.

Using a series of tgrep strings, we searched for all instances in the corpus in which a nominal is modified by a clause containing a trace, including multiple such instances within a given sentence or conversational turn. This search excluded tokens containing infinitival relative clauses (e.g. *very few people* ***to go with me***) but included all finite relative clauses (e.g. *anybody who bills you*

https://doi.org/10.1515/LANGCOG.2011.007 Published online by Cambridge University Press

*from Atlanta*), regardless of the grammatical function of the relative pronoun. The search strings used to retrieve object and oblique relative clauses looked for any trace tag (indicated as *T*) occurring postverbally within the relative clause; it therefore retrieved both tokens with argument-position gaps and tokens with adjunct-position gaps, and included tokens without a relative pronoun (e.g. *everything* ***you could imagine****, a job* ***I had to get a blood test for***). The search strings used did not distinguish between relative clauses in declarative sentences and those occurring in other sentence types (e.g. *Do you have any hobbies* ***that you like to do****?*) and included nonrestrictive relative clauses (e.g. *My dad lives in the state capital,* ***which is Pierre***). The distribution of finite relative-clause types retrieved by this search is summarized in Table 1, following conventions used by Geisler (1998):

Table 1. *The distribution of finite relative-clause types according to grammatical functions (matrix function by row and relative-pronoun function by column)*

| RP GF/Head GF | Object/oblique RP | Subject RP | Total |
|---|---|---|---|
| Subject head | 334 (60.7%) (SO: 12.6% of 2640) | 216 (39.3%) (SS: 8.2% of 2640) | 550 (20.8%) |
| Object/oblique head | 538 (25.7%) (OO: 20.4% of 2640) | 1552 (74.3%) (OS: 44.7% of 2143) | 2090 (76.4%) |
| Total | 872 (33%) | 1768 (67%) | 2640 |

In Table 1, the rows show the distribution of relative clause types by the grammatical function (GF) of the head nominal in the matrix clause—that is, according to whether the relative clause modifies a matrix subject or a matrix nonsubject (i.e. an object or oblique argument). The columns show the distribution of relative-clause types by the grammatical function of the relative pronoun (RP), i.e. according to whether the relative clause contains a subject relative-pronoun or a nonsubject relative pronoun.[5] The abbreviations in the cells label relative-clause type according to these grammatical functions—the first letter (S or O) refers to the grammatical function of the head nominal in the matrix clause while the second letter refers to the grammatical function of the relative pronoun.

The results reported in Table 1 might be taken to suggest that speakers prefer nonparallel structures to parallel structures—that is, that they prefer SO structures to SS structures and OS to OO structures, respectively. We do not,

5. We regard both *that* complementizers (e.g. *I found a book that I like*) and null relative pronouns (e.g. *I found a book I like*) as relative pronouns to which grammatical functions can be assigned.

https://doi.org/10.1515/LANGCOG.2011.007 Published online by Cambridge University Press

however, assume a preference for nonparallel structures, because we regard the similarity between the two low-frequency parallel structures SS and OO as coincidental: the rarity of the OO configuration (relative to OS) is not predicted by general linguistic principles, while the rarity of the SS configuration (relative to SO) can be attributed to the discourse-pragmatic factors discussed in Section 3. Further, a parallelism account might lead us to assume that the SO configuration is frequent, when in fact it accounts for only 12.6% of the relative-clause tokens in our sample. We simply do not find many relative clauses modifying subjects, and this fact too can be explained on discourse-pragmatic grounds. In English, only lexical nouns generally allow adjectival and relative-clause modifiers, and subjects containing lexical head nouns are rare in conversational English, a genre in which speakers tend to choose discourse-old (and thereby pronominally expressed) referents as subjects. One can therefore predict that more relative clauses have object or oblique head nominals than subject head nominals. This prediction is confirmed by the results reported in Table 1: only 20.8% (N = 550) of the relative-clause modifiers in our sample are modifiers of subjects. What is notable is the asymmetry within this set of 550 relative clauses: object and oblique relatives, as in (28), account for a full 60.7% of tokens, while subject relatives, as in (29), account for only 39.3% of tokens:

(28) The budget he sent to congress has taxes and fees. (SO)
(29) The people that run the system have given up on it. (SS)

This trend is statistically significant: a subject nominal is 0.22 (*95% CI: 0.18, 0.27*) times more likely to be modified by a object or oblique relative than by a subject relative (*Wald* $\chi^2$ *220.79,* $p < 0.0001$). It is explicable according to the anchoring function served by relative-clause modifiers of subject nominals: as discussed in Section 3, these relative clauses typically contain anaphoric subjects that link the discourse-new referent being modified to a previously mentioned one. Anchoring is critical for discourse-new referents in subject position because, as discussed above, speakers prefer subject arguments to be topical (and thus discourse-old) arguments. The SS configuration is rarer than the SO configuration because subject relatives do not perform the anchoring function that object relatives do. Indeed, a subject-relative modifier only serves to make an already heavy subject NP even heavier[6], without the compensation of enhanced referent-recoverability.

As Table 1 shows, once we move outside the argument position canonically reserved for topics (subject position), object relatives are no longer preferred. Object and oblique nominals instead prefer subject-relative modifiers, which

6. Heaviness is here to be understood in terms of word count.

https://doi.org/10.1515/LANGCOG.2011.007 Published online by Cambridge University Press

account for 74.3% of relative-clause modifiers of nonsubject nominals. In fact, the OS configuration, as in (1), repeated here as (30), accounts for the plurality of all relative clauses in the data set (47% of the total), while the OO configuration, as in (31), accounts for only about 20% of that total):

(30) I like cars that are designed with human beings in mind. (OS)
(31) I like those movies that you watch time and time again. (OO)

In statistical terms, an object or oblique nominal is 4.46 (*95% CI: 3.66, 5.43*) times more likely to be modified by a subject relative than by an object relative (*Wald* $\chi^2$ *220.79,* $p < 0.0001$.) Why should this be? We hypothesize that the OO configuration is simply less useful: when speakers modify an object or oblique argument, they are doing so in order to predicate an action or property of the referent—a referent that would be encoded as the subject of a matrix clause were it sufficiently familiar. In other words, while the OS configuration facilitates referent introduction, the OO configuration does not.

As Table 1 indicates, relative-clause modifiers of object and oblique nominals account for the lion's share of relative clauses overall: 77.5% of the relative clauses in the corpus modify an object or oblique nominal. We thus undertook to examine only those relative-clause modifiers of object or oblique nominal heads, on the assumption that the construction that accounts for the prevalence of subject relatives within this group would account for the prevalence of subject relatives across all grammatical contexts in spoken English. Our broader question is as follows: could speakers' reliance on the PRC strategy explain their preference for subject relatives over nonsubject relatives in the modification of nonsubject nominals? Answering this question in the affirmative will require us to show that the OS tokens in the corpus more consistently bear hallmarks of presentational function than do OO tokens. Accordingly, we will examine both OO and OS tokens for hallmarks of PRC function: the indefiniteness of the head nominal being modified, a semantically empty matrix verb and an assertoric relative clause. The annotation scheme used to identify each of these hallmarks will be described in the following subsections.

In order to create the subcorpus of OO and OS tokens for annotation, we hand-sorted the tokens in the set of 2090 object-headed, finite relative clauses retrieved from our initial search (the results of which are summarized in Table 1). Our goal was to ensure conformity with a basic type. This basic type was defined as containing (a) a finite relative clause, (b) an object or oblique head nominal (i.e. a lexical noun or pronominal expressions like *anything*, *something*, etc.) and (c) a relative pronoun from the set *who, whose, which, that* and *ø*. We included both matrix-clause tokens of this type, as in (30–31), and subordinate-clause tokens, as in (32), in which an OS token appears in a subordinate clause introduced by *because*:

https://doi.org/10.1515/LANGCOG.2011.007 Published online by Cambridge University Press

(32) And uh I'm holding out for *City Slickers* for the two of us because uh we had friends that went to see that.

Our hand-sorting procedure was designed to eliminate disfluent relative-clause tokens like (33), in which a word-substitution or other speech error affects semantic coherence, and tokens of the relative-clause types exemplified in (34–38):

(33) **Disfluent relative**: AIDS research is **something that that I think th-whether our country is putting enough money into it.**
(34) **Participial relative**: They're not quite the same as **kids going to the inner city schools**.
(35) **Headless relative**: But I could not foresee them severing **what they have with the US**.
(36) **Adverbial relative**: And I didn't want to run **an institution where that was the case.** That was **the reason why he cut all the players**. It's coming around at **the time when we're losing the most most of the forests**.
(37) **Appositive relative**: It was Buddy Ryan, **the one that I can't uh uh stand too much**.
(38) **Gapless relative**: In fact, we are in **a position that most of our friends why—wonder why we just don't go to a new car**.

The foregoing types were excluded because they did not allow for reliable determination of the head grammatical function, the relative-pronoun grammatical function or the matrix verb. For example, participial relatives, as in (34), typically occur in adverbial rather than verb-licensed positions; headless relatives, as in (35), conflate the nominal head and relative-pronoun, and appositive relatives, as in (37), have heads that are modifiers of arguments rather than arguments. We further excluded relative clauses modifying nominals not governed by a verb or verb-licensed preposition. Examples of these excluded types are given in (39–40):

(39) And they just cut them all up except for **one they kept for emergencies**.
(40) I can associate with some of the people in that movie because of **the young students I see over at the medical school**.

In (39), it is the adverb *except*, rather than a verb, that governs the PP dominating the nominal expression *one they kept for emergencies*. In (40), it is the adverb *because*, rather than a verb, that governs the PP dominating the NP *the young students I see over at the medical school*. Another class of tokens excluded were those that share the licensing predicator with a prior token, as in (41):

(41) Well, they're they're going to be cutting back so much on just, you know, **the number of troops we've got in Europe** and **the number of troops we have here**.

https://doi.org/10.1515/LANGCOG.2011.007 Published online by Cambridge University Press

In sentence (41), the two relevant NP tokens (indicated in boldface) are conjoined, and thus share a licensing predicator, *cut back on*. Because we must assume a unique matrix verb for the purpose of our matrix-verb labeling task, we admitted only the first relative-clause token in cases like (41).

Finally, we developed a labeling procedure for certain special-case structures: 'stacked' relative clauses, resumptive pronouns in relative clauses and relative clauses containing embedded gaps. Stacked relative clauses are tokens in which a relative clause modifies a nominal head that is already modified by a relative clause. One such example appears in the text below as (48): *They're talking about new rockets that they're designing now that, you know, are just like science fiction*. We labeled such tokens based on the grammatical function of the relative pronoun in the least embedded, outer relative clause. In the case of (48), this is the relative clause *that, you know, are just like science fiction*, a subject relative. Thus (48) was counted as an OS token. An example of a token containing a resumptive pronoun is (11) above: *But I have all these friends that, wherever you go,* ***they they*** *sit down [ . . . ]*. Here, a resumptive pronoun (*they*) replaces what would otherwise be a subject-position gap.[7] Such examples were labeled according to the grammatical function of the resumptive pronoun; for example, (11) was labeled an OS token. Example (42) illustrates a relative clause containing an embedded gap:

(42) We—we get the *Mercury* which I generally think is actually a pretty good newspaper.

While the matrix verb of the relative clause in (42) is *think*, the verb containing the gap is *is*. Since that gap is in subject position, (42) was labeled as an OS token.

From the approximately 1700 OO and OS tokens that remained after the hand-sorting procedure described above, we pseudo-randomly extracted a data set containing 500 OS tokens and 500 OO tokens. This number represents about 60% of the total sample. The annotation scheme used to label this set of 1000 tokens is described in the following three subsections.

### 4.1. *Hallmark 1: Indefiniteness of the head*

As discussed in the previous section, the proposed function of the PRC is to introduce a new entity, expressed by the head nominal, and then to predicate upon it. Accordingly, we presume that most object or oblique nominals in PRC

7. Resumptive pronouns were not limited to OS tokens; they occurred as well in OO tokens, e.g. *I talked to this person who I gathered from speaking to* ***her*** *that—that she and her family just didn't have much*. Here a resumptive-pronoun gap *her* occurs as the object of the preposition *to*, which in turn is licensed by the verb *speaking*.

https://doi.org/10.1515/LANGCOG.2011.007 Published online by Cambridge University Press

predications will be indefinite. The distinction between the 'given' and 'new' discourse statuses is not the same thing as definite versus indefinite form: as Gundel et al. (1993) point out, an indefinite NP can be used to implicate a higher discourse status, as in the constructed example (43):

(43) I am grateful for the kindness they showed to a young girl in need.

Sentence (43) could be used in a context in which the NP *a young girl in need* refers to the speaker, in which case it would contextually implicate a discourse status otherwise expressed by a pronoun, i.e. *me*. Leaving aside such rhetorical effects, however, discourse-active entities tend to be formally marked as definite, while discourse-new entities tend to be marked as indefinite (Prince 1992). If, as proposed here, the PRC accounts for the prevalence of subject relatives in conversation, then OS tokens will tend more strongly than OO tokens to have indefinite nominal heads. If instead the nominal heads of OO tokens are as likely as those of OS tokens to be indefinite, this is presumably an effect of object status in general, rather than an effect of the PRC.

All object and oblique nominal heads were annotated as indefinite or definite based on morphosyntactic form rather than perceived discourse status. Head nominals marked as indefinite included bare plural nouns (e.g. *engineers*), determinerless nominals modified by adjectives or cardinal numbers (e.g. *about forty kindergarteners*), bare mass nouns (e.g. *material*), nominals with weak quantifiers like *some* (e.g. *some companies*), indefinite pronouns (e.g. *somebody, anybody*) and nominals containing the indefinite article (e.g. *a fish*). Head nominals marked as definite were those that contained the definite article *the* (e.g. *the thing, the resources*), those that contained demonstrative determiners (e.g. *this tape recording*, *that grudge attitude*), those that contained possessive determiners (e.g. *my first bass*), those that contained strong quantifiers (e.g. *every story, each story, all these people*), demonstrative pronouns (e.g. *that, these*, *those*) and proper nouns (e.g. *Rockport, Albany*). Partitive nominal expressions with indefinite heads (e.g. *one of those things*, *a lot of people*, *a couple of things*, *some of my friends*) were labeled as indefinite irrespective of the definiteness of their complements. Because definiteness labeling was based on nominal morphosyntax rather than contextual factors, many of the nominal heads labeled as definite appeared in predications that intuitively qualify as PRC tokens. Examples (41–42) illustrate this point:

(44) And so, you know, he has **these stacks of Sunday newspapers that go unread**.

(45) This this guy had, both for our school district meeting and our town meeting, had **this proposal which, unfortunately, violates New Hampshire constitution**.

https://doi.org/10.1515/LANGCOG.2011.007 Published online by Cambridge University Press

The boldfaced NPs in (44–45) exemplify indefinite *this*, a cataphoric use of the proximate demonstrative determiner in which it acts like an indefinite article (Gernsbacher and Shroyer 1989). While a contextually based labeling procedure might have counted the head nominals in (44–45) as indefinite, we count each as definite solely due to the presence of a demonstrative determiner. This conservative labeling strategy ensures that our results are biased, if at all, toward the null hypothesis, according to which the PRC does not account for the majority of OS tokens.

### 4.2. *Hallmark 2: A semantically light matrix verb*

Following McCawley (1981), we assume that the matrix verb in the PRC does not convey an assertion, and that its purpose instead is to ensure that a discourse-new referent appears in postverbal position. Accordingly, the matrix verbs of PRC tokens tend to have low semantic weight, as in (46–48):

(46) Because you **see** the reliability and the types of problems they have.
(47) You **get** a guy down the street who comes up, uh, carrying a knife.
(48) And they **had** some guy that was uh defending himself.

The boldfaced matrix verbs in (46–48), which otherwise denote relations of perception, obtaining, and possession, respectively, here appear simply to 'set the stage' for their object referents. In other words, (46) does not assert that the addressee sees something, (47) does not assert that the addressee obtains something and (48) does not assert that some people possessed someone. Rather than predicating a property or action of the matrix subject, the matrix predications in (46–48) provide an explicit or inferred center of perspective from which to view the entity denoted by the head nominal (Koenig and Lambrecht 1999). Thus, our second hallmark of PRC function is the presence of a semantically empty matrix verb. The verbs considered to be semantically empty for the purposes of this annotation task include verbs of existence, perception and discovery. Some of these verbs (e.g. *find* and *see*) take NP second arguments, while others take PP second arguments headed by *of* or *about*, as in (49):

(49) They're **talking about** new rockets that they're designing now that, you know, are just like science fiction.

Table 2 lists all of the lexemes that were regarded as semantically empty for the purposes of this labeling task.

Note that we based the judgment of semantic emptiness solely on the lexical identity of the matrix verb: if a given token contained one of the matrix verb lexemes listed in Table 2, it was labeled as having a semantically empty matrix verb, irrespective of context. Because the verbs in Table 2 are frequent and highly polysemous, we expect that they will be used as matrix verbs not only

https://doi.org/10.1515/LANGCOG.2011.007 Published online by Cambridge University Press

Table 2. *Light matrix verbs*

| Verbs taking NP complements | Verbs taking PP complements |
|---|---|
| be | hear of |
| have | hear about |
| get | know about |
| find | know of |
| see | look at |
| know | talk about |
| hear | wonder about |

in presentational tokens like (43–45), but also in other types of tokens, such as (50–51):

(50) Seems like everyone that lives around us, ends up, you know, **hearing** every conversation that goes on outside with everyone.
(51) The driver **had** something on his belt that he scanned across the little bar code on our bin as soon as he took the stuff.

The boldfaced verbs in (50–51) have their basic meanings rather than 'bleached' meanings: in (50), *hear* means 'detect by ear' and in (51) *have* means 'possess'. The use of *have* in (51) contrasts with the bleached use in (48) above, while the use of *hear* in (50) contrasts with the bleached use in (52) below:

(52) I've even uh **heard** some people that have applied for credit cards with much less, uh rates and have paid off their, you know higher interest rate uh cards and just sent them back.

Tokens like (50–51) were labeled as containing semantically empty matrix verbs based on the lexical-identity criterion. Because the verbs in question are in fact semantically 'rich' in these contexts, it might appear that lexical identity is too inclusive a criterion for semantic emptiness. Use of the lexical-identity criterion does, however, eliminate a potential source of bias: it ensures that we do not make the semantic-weight judgment based on the perception that a given token is or is not an instance of the PRC. If in fact the majority of OS tokens in our data set represent PRC tokens, OS tokens will be more likely to contain a matrix verb from the list in Table 2 than OO tokens.

### 4.3. *Hallmark 3: Paraphraseability*

As discussed in Section 3, the PRC can be viewed as an avoidance strategy, in which speakers use a biclausal construction to replace a monoclausal one that would require a pragmatically suboptimal heavy subject. On this understanding, PRC productions are 'covert' monoclausal productions. Accordingly, our final labeling task involved a paraphrase test that we took to reveal whether or

https://doi.org/10.1515/LANGCOG.2011.007 Published online by Cambridge University Press

not the token in question contained an assertoric relative clause. Recall from Sections 3 and 4.2 that the PRC is an idiomatic construction in which the matrix predicator does not convey an assertion. Instead, the relative clause conveys the assertion (McCawley 1981; Goldberg 2006: 146; Lambrecht 2002: 172). Thus, we defined tokens as containing an assertoric relative clause just in case the propositional content of the utterance could be captured by a monopropositional paraphrase created by converting the main verb of the relative clause in the original production to the matrix verb of the sentence. For example, (39) above (*They had some guy that was defending himself*) can be paraphrased simply as 'Some guy was defending himself'.

All 1000 OO and OS tokens in our data set were subjected to the paraphrase test. Example (25), repeated as (53) below, illustrates the paraphrase procedure:

(53) There's a restaurant in, um, right outside of Reading, Pennsylvania, it's called Alfredo's, that does not look like a restaurant that you would really want to recommend to a lot of people.

As an instance of the existential *there*-construction, (53) is a prototypical PRC token, and it exhibits both hallmarks of the PRC described above: an indefinite nominal head and a light matrix verb (*be*). Following Goldberg (2006: Ch. 8), we can say that the relative clause in (53) conveys foregrounded rather than backgrounded information. As shown by (54), one can use a main clause to express the foregrounded content:

(54) A restaurant in Reading, Pennsylvania does not look like a restaurant that you would want to recommend to people.

While (54) sounds unnatural due to its indefinite subject NP, it captures the content of (53), and thus (53) passes the monopropositional-paraphrase test. Note that (54) is considered a monopropositional paraphrase of (53) despite the fact that both contain a subordinate clause, the relative clause *that you want to recommend to people*. By the same token, (11), repeated here as (55), was assessed as having a monopropositional paraphrase, (56), despite the fact that (55) contains conjoined clauses:

(55) But I have all these friends that, wherever you go, they, they sit down and the next thing you know they pull out of their uh bags some their most recent uh needle craft.

(56) All these friends of mine sit down and the next thing you know pull out of their bags some of their most recent needlecraft.

It may appear inappropriate to refer to paraphrases like (54) and (56) as *monopropositional* paraphrases, because each contains more than one clause. What is critical for our purposes, however, is that such paraphrases were produced by the same procedure used to create the other paraphrases: (a) eliminating the

https://doi.org/10.1515/LANGCOG.2011.007 Published online by Cambridge University Press

matrix verb and (b) changing the relative-clause predication to the matrix predication.

The validity of a paraphrase may depend on context, and therefore we used the 10 lines of context immediately preceding the target turn, and any within-turn context immediately preceding or following the target token, to confirm whether a monopropositional paraphrase was appropriate.[8] Example (57) will be used to illustrate this procedure:

(57) Usually, you know, you'll find a, a woman that's keeping like six children [ . . . ] in the home.

Example (57) was produced during a conversation about day care, in which the speaker is describing the typical home day-care operation. If the speaker were explaining, for example, how one locates good day-care facilities, the relative clause might be interpreted as restrictive: 'If you want good day care, you should go find a woman—not just any woman, but a woman who is keeping six children in the home'. However, the context surrounding this production, shown in (58), suggests that this restrictive interpretation is not the appropriate one. It reveals that the speaker (Speaker B) is describing the day-care services used by friends, and that the target production (shown in boldface) is a generalization about day-care operations:

(58) Speaker A: So, I, I hope it helps. It seems to help the new mothers not have to come back full-time.
Speaker B: Oh that's good.
Speaker A: Because that's hard.
Speaker B: That's good to know. I have a couple of friends that have, have found the, uh, you know, a pri-, a private home to take their children to when they're young until they hit the preschool age and they
Speaker A: Uh huh.
Speaker A: Uh huh.
Speaker B: **Usually, you know, you'll find a, a woman that's keeping like six children or four to six children in the home**, and my future, future sister-in-law's mother does that too full-time.

As shown by the context in (58), Speaker B is not predicting that one will encounter a woman in a particular situation, but rather describing the typical day care operation as one in which a woman keeps six children in her home. Accordingly, we concluded that (58) has a monopropositional paraphrase, given in (59):

(59) Usually a woman's keeping six children in the home.

8. Paraphraseability judgments were made by the first author, while the second author provided adjudication of ambiguous cases.

https://doi.org/10.1515/LANGCOG.2011.007 Published online by Cambridge University Press

Our contextual checking procedure ensured that in judging the validity of a monopropositional paraphrase we were not unduly influenced by the presence of one or both of the formal hallmarks described above, e.g. the presence of a matrix-verb lexeme from the list in Table 2. For example, (60) and (61) each contain an OO token (shown in boldface) whose matrix verb is *get*—a lexeme that appears on the list of semantically empty matrix verbs in Table 2. However, while the OO token in (60) has a valid monopropositional paraphrase, the OO token in (61) does not:

(60) I, I think it should be i- it should go to the to the heart of the matter though and say okay guy y- y- everybody gets you know **everybody gets five pounds of garbage that they can throw away** you know uh but more than that every week uh you've got to pay by the pound.
⇒ Everybody can throw away five pounds of garbage.
⇏ Everybody receives five pounds of garbage that they can throw away.

(61) Knew what you were getting when you voted uh yeah yeah they say **you get the government you deserve**.
⇏ You deserve the government.
⇒ You receive that government that you deserve.

Put differently, the matrix verb *get* is dispensable in the OO token in (60) but not in the OO token in (61). The reason is that in (60), as against (61), the verb *get* is contentful, and thus, as indicated, it is synonymous with the verb *receive*.

It is important to note at this juncture that paraphraseability and conformity to the OS pattern are two different properties. If they were not, the paraphrase task would simply 'redescribe' the feature that we are trying to predict: the presence of a subject relative. In fact, however, an OO token can pass the paraphrase test, as shown by (60) above, while an OS token can fail it as shown by (62–63) below:

(62) I like cars that are designed with human beings in mind. (OS) (= (1), (26))

(63) *Cars are designed with human beings in mind.

Sentence (62) fails as a paraphrase of the OS token in (63) because it does not capture the speaker's stance toward the particular class of cars (expressed by the matrix verb *like*), or the implicit contrast between these cars and others. Thus, while we expect many OO tokens to fail the paraphrase task, and many OS tokens to pass it, paraphraseability is not the same thing as having the OS configuration. If we are correct that the PRC licenses the majority of OS tokens in our sample, a significantly higher number of OS tokens will yield valid monopropositional paraphrases than will OO tokens.

The PRC is a linguistic gestalt; however, the annotation scheme that we have described in this section focuses on the distinct properties that comprise

https://doi.org/10.1515/LANGCOG.2011.007 Published online by Cambridge University Press

it. By separating those properties into three distinct annotation tasks, we have not only operationalized the impressionistic functional and formal characterizations of the PRC given by McCawley, Lambrecht and others but also embraced a prototype-based model of the PRC. According to this model, the PRC is not a single construction but a family of related constructions, of the type described in Lakoff's (1987) study of English *there*-constructions, Goldberg's (1995) study of the English ditransitive construction, Kay's (2002) study of English tagged sentences and Sag's (2010a) study of English filler-gap constructions. The best example of the PRC is that which satisfies all of the three criteria described above: it has a light matrix verb whose direct object is an indefinite nominal and its meaning can be captured by monoclausal paraphrase. At the same time, we recognize that some tokens, while intuitively qualifying as PRC instances, fail to satisfy one or more of these criteria. Sentence (64) is a case in point:

(64) I have a friend who was telling me about her brother who gets high all the time.

Example (64) intuitively satisfies the monoclausal-paraphrase criterion: its meaning seems to be captured by the sentence *My friend's brother gets high all the time*. At the same time, (64) lacks a light matrix verb (*tell about*) and its direct object contains a definite rather than indefinite nominal head (*her brother*). We assume therefore that (64) is a PRC token, but that the PRC construction that licenses it is less constrained than those that license more prototypical instances. Conversely, there are tokens in our data set that do not satisfy the monoclausal-paraphrase criterion (the third feature) but display the other two hallmarks of PRC function. An example is given in (65):

(65) And we have a paper recycling program that is, uh, company wide.

Example (65) contains a light matrix verb (*have*) and an indefinite nominal head (*a paper recycling program*). However, when examined in context, (65) does not appear to satisfy the monoclausal paraphrase criterion. That is, its meaning is not captured by the paraphrase *Our paper recycling program is company wide*, since in the discourse context, the matrix clause is not dispensable: it asserts possession, or more specifically that the speaker's company, unlike other organizations, does, in fact, have a recycling program. The relevant context is shown in (66):

(66) Speaker A: Texas is not one of [the places that recycles glass] see so I have to throw [bottles] away cause there is no place to take glass.
Speaker B: Yeah.
Speaker A: So.
Speaker B: Yeah. No, I don't think that there is enough being done. Now, I work at JC Penny (sic) at their corporate headquarters.

https://doi.org/10.1515/LANGCOG.2011.007 Published online by Cambridge University Press

The context shown in (66) makes clear that Speaker B, the producer of (65), intends to contrast Texas, with its lack of a recycling program, to the organization with which she is affiliated, JC Penney. Thus, the matrix clause in (65) is informative, insofar as it asserts possession and is contrastive. Such examples again demonstrate the relevance of context to the monopropositional paraphrase test. Insofar as (65) bears two hallmarks of PRC function, it can be regarded as illustrating a subtype of the PRC construction.

If some PRC constructions lack some of the three properties, then each of the three properties is a potentially significant predictor of relative-clause type. Thus, the statistical analyses reported in Section 5 will examine the three properties both separately and in conjunction. The results of these analyses will determine whether we can uphold a central claim of this paper—that the prevalence of the subject-relative pattern is due to the role that it plays in an entrenched conversational routine, the PRC. If we are correct, each of the three properties will be significantly more likely to occur in OS tokens than in OO tokens.

## 5. Results

We used a logistic regression analysis to determine whether or not the three hallmarks of the PRC described in Section 4 (i.e. an indefinite head, a semantically empty matrix verb, and a valid monopropositional paraphrase) were significant predictors of relative-clause type. By using a logistic regression analysis we were able to test the significance of each feature as a predictor of relative-clause type within the context of the other two (for further discussion of logistic regression, see Diessel (2008: 478–479) and references therein). In this way, we could determine whether there were confounds within our feature set. If, for example, the predictive effect of paraphaseability is instead attributable to the presence of a light matrix verb, the logistic regression analysis will show this. Table 3 presents a summary of our results.

When examined in conjunction, the three features significantly predicted the presence of a subject relative clause (*Wald* $\chi^2$ *69.30,* $p < 0.0001$). This result is

Table 3. *Results of logistic regression analysis*

| Factor | Regression coefficient | Wald $\chi^2$ | df | p | Odds ratio | 95% Wald Confidence Limits | |
|---|---|---|---|---|---|---|---|
| PRC features model | | 69.30 | 3 | <0.0001 | | | |
| Indefinite head | 0.60 | 17.09 | 1 | <0.0001 | 1.81 | 1.37 | 2.41 |
| Matrix verb | −0.22 | 2.18 | 1 | 0.1399 | 0.80 | 0.60 | 1.07 |
| Paraphraseability | 1.49 | 41.04 | 1 | <0.0001 | 4.42 | 2.80 | 6.96 |

https://doi.org/10.1515/LANGCOG.2011.007 Published online by Cambridge University Press

shown in the first row of Table 3, labeled *PRC features model*. The PRC features model represents the cluster of properties that collectively define the PRC. The result confirms that OS tokens are significantly more likely to display this cluster of properties than are OO tokens. However, our prototype-based model of the PRC also allows us to recognize as PRC instances tokens that lack one or more of the features that define the central case. Accordingly, we also used the regression model to examine whether each feature individually was a significant predictor of relative-clause type. When matrix verb and paraphraseability are controlled for, the presence of an indefinite head significantly increased the odds of a subject relative: indefinite heads were 1.81 (*95% CI: 1.37, 2.41*) times more likely than definite heads to be modified by subject relative clauses (*Wald* $\chi^2 = 17.09, p < .0001$). Paraphraseability was also a significant predictor of subject relative clauses, over and above both definiteness and matrix-verb identity: paraphraseable tokens were 4.42 (*95% CI: 2.80, 6.96*) times more likely than non-paraphraseable ones to contain a subject relative (*Wald* $\chi^2 = 41.04, p < .0001$).

By contrast, the presence of a semantically light matrix verb was found not to be a significant predictor of relative-clause type (*Wald* $\chi^2 = 2.18, p = 0.1399$). This outcome might be the result of our very inclusive criterion for matrix-verb lightness. Recall from Section 4.2 that we used lexical identity, rather than sense in context, to label each matrix verb as heavy or light. That is, a matrix verb was counted as a light verb if it appeared on the list in Table 2, regardless of its meaning in the sentence at hand. While all of the verbs on this list were assessed as having a reasonable incidence of presentational use, many, as already noted in Section 4.2, also have nonbleached literal uses. For example, while verb *find* in (67) has a presentational function in which it could reasonably be viewed as semantically light, in (68) it carries its literal meaning, i.e. 'discover':

(67) And, when you **find** someone like that, that you know is guilty, he confessed already to killing eleven, I, I'd, you know, I guess I have a hard time feeling merciful toward him.

(68) And if he gets in there and starts rooting around and **finds** something in there that's really tremendously wrong with it, then he eats it [sc. the repair cost].

The matrix-verb labeling procedure that we used erased the meaning distinction between these two tokens: both (64) and (65) were labeled as containing a semantically light matrix verb by virtue of containing the matrix-verb lexeme *find*. In fact, as it turned out, the set of OO tokens and the set of OS tokens were about equally likely to contain a verb lemma from Table 2: 42.4% of OO tokens (212/500) contained a light matrix verb by our definition, as against 51.8% of OS tokens (259/500). However, the true incidence of light matrix verbs among these 500 OS tokens might have been far higher than among the comparable

https://doi.org/10.1515/LANGCOG.2011.007 Published online by Cambridge University Press

set of OO tokens. It is thus reasonable to ask whether matrix-verb type would prove to be a significant predictor of relative-clause type if we had replaced the lexical-identity diagnostic with one based on sense in context. It also may be the case that while matrix verb type and paraphrasability are separate features (as demonstrated by (61–62)), the likelihood of the two features co-occurring may have been such that when paraphrasability is controlled for, matrix verb type simply is not a significant predictor of subject relatives. To further investigate the significance of semantically light verbs as predictors of subject relatives, it would likely be necessary to annotate each token for the sense of the matrix verb, in addition to lexical identity.

As discussed above, we assume a prototype-based model of the PRC, in which a token need not bear all three hallmark features in order to be regarded as a member of the PRC class. Our results show that of the 500 subject relative-clause (OS) tokens, 112 (22%) display all three hallmark features, while 226 (45%) display at least two of the three hallmark features. By contrast, of 500 object relative-clause (OO) tokens, only 27 (5%) display all three features, and 161 (32%) display at least two. Do these findings explain why an object or oblique nominal is more likely to be modified by a subject relative than an object relative? In other words, is the PRC sufficiently frequent to explain why we find so many more OS tokens than OO tokens in our larger data set? In this connection, recall from Table 1 that of 2090 object- and oblique-headed relative clauses, 1552 (74%) contain subject relative-clauses while only 538 (26%) contain object/oblique relative-clauses. This asymmetry is represented in Table 4.

Figure 1 shows that the probability of an object-headed relative-clause token's being an OS token is 48% greater than the probability of its being an OO token. In order to determine whether the PRC (or rather, its rate of use) accounts for the 48% edge enjoyed by subject relatives, we must first estimate, based on our 1000-token sample, how many OS tokens and how many OO tokens qualify as

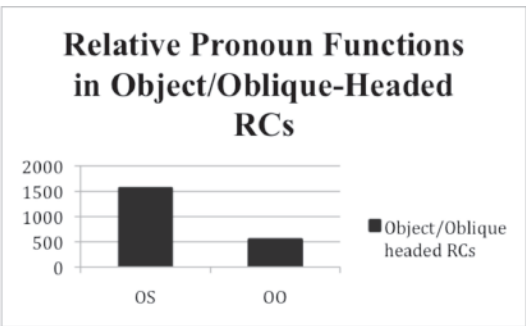


Figure 1. *Subject-relative vs. object-relative tokens among all object- and oblique-headed relative-clause tokens (n = 2090)*

https://doi.org/10.1515/LANGCOG.2011.007 Published online by Cambridge University Press

PRC tokens, among the 2090 object-headed relative-clause tokens in our data set. If a large portion of the OS tokens can be viewed as PRC exemplars, and a much smaller proportion of the OO tokens can be so viewed, we can reasonably 'blame' the PRC for the preponderance of subject relatives in the larger data set. But while we can easily make the required estimate based on the distribution of PRC features in our sample, we must first determine how many PRC features, and potentially which PRC features, we will require a given OO or OS token before we deem it a PRC token. We could, for example, stipulate that any token which has *any two* of the PRC features (an indefinite head, a light matrix verb, a monopropositional paraphrase) is a instance of the PRC. This criterion is not feasible, however, for reasons related to the matrix-verb criterion. While the presence of a semantically light matrix verb is an *intuitively* important property of the PRC, recall that our statistical analysis showed verb type to be an insignificant predictor of relative-pronoun grammatical function. We will therefore disregard matrix-verb identity when determining which relative-clause modifiers of postverbal nominals are exemplars of the PRC. For our purposes here, only those tokens that both contain an indefinite head nominal and yield monopropositional paraphrases are members of the PRC class.

While far fewer tokens are estimated to qualify as PRC exemplars under this more restrictive criterion for PRC membership, we still find an asymmetry in the expected direction: of the 500 subject relative-clause (OS) tokens in our sample, 112 (22.4%) display the two hallmark features indefiniteness and paraphraseability, while of the 500 object relative-clause (OO) tokens, only 28 (5.6%) display these two features. If we then project these percentages onto the larger data set, 347 of 1552 OS tokens will be PRC exemplars, and 30 of the 538 OO tokens will be PRC exemplars. This projection is represented in Figure 2:

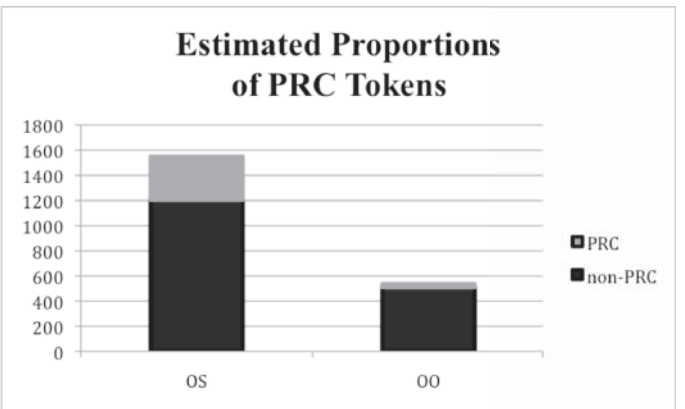


Figure 2. *Proportions of OS versus OO tokens estimated to be PRC tokens in the larger sample*

https://doi.org/10.1515/LANGCOG.2011.007 Published online by Cambridge University Press

Thus, the probability of an OS token being an exemplar of the PRC class is about 17% higher than the probability of an OO token being a member of the PRC class. Admittedly, the 'PRC advantage' enjoyed by the OS tokens accounts for less than half of the numerical advantage seen in Figure 1. We believe, however, that the proportion of PRC tokens within the OS set would rise significantly were we to replace the lexical-identity criterion with a stricter method of determining whether the matrix verb of an object-headed relative clause is semantically empty. While the lexical-identity criterion enabled us to avoid excessively subjective labeling, it is clearly an overly inclusive criterion: as revealed by linguistic analysis in Section 4.2, not all tokens labeled as containing a semantically empty matrix verb actually contained one (see, e.g. the examples in (64–65) above involving matrix verb *find*). This finding in turn suggests that we need a reliable way to annotate matrix verbs according to sense rather than lexical identity. Such a method would offer enhanced discrimination, and perhaps increase the distance between the OS tokens that qualify as PRC exemplars and the OO tokens that do. If so, we could potentially account for a larger portion of the OS 'edge' seen in Figure 1.

Accordingly, we sought to test the feasibility of sense-based matrix-verb labeling, using verb-sense annotation criteria developed for computational applications. Specifically, we applied OntoNotes verb senses (Hovy et al. 2006; Duffield et al. 2007) to a subset of our annotated data. This subset comprised all OO and OS tokens containing the matrix verb *have* (approximately 20% of the annotated data set, or 211 tokens). OntoNotes sense groupings are based on WordNet (Fellbaum 1998). Human annotation efforts using the OntoNotes sense inventories have achieved high inter-coder agreement rates, yielding effective training data for automatic word-sense disambiguation in natural-language processing applications (Chen et al. 2007; Dligach and Palmer 2008).

Recall that all PRC tokens containing *have* qualified as semantically empty according to the lexical-identity criterion used in the main study. It was our belief that a more sensitive labeling scheme, based on the OntoNotes *have* sense groupings, would reveal a significant correlation between the presence of a semantically light matrix verb and the presence of a subject- (as against object-) relative clause, thus supporting the constructional account. We tested this prediction in the following way. After tagging each of the 211 tokens containing matrix *have* with an OntoNotes sense, we determined which of the OntoNotes *have* sense groupings could be considered 'semantically empty', in failing to impute literal possession to the subject referent.[9] Two of the

9. The complete OntoNotes *have* sense groupings can be found at http://verbs.colorado.edu/html_groupings/have-v.html.

https://doi.org/10.1515/LANGCOG.2011.007 Published online by Cambridge University Press

Table 4. *Results of logistic regression analysis for tokens containing the matrix verb have with more restrictive coding of the matrix verb.*

| Factor | Regression coefficient | Wald $\chi^2$ | df | p | Odds ratio | 95% Wald Confidence Limits | |
|---|---|---|---|---|---|---|---|
| PRC features model for *have* tokens | | 26.91 | 3 | <0.0001 | | | |
| Indefinite head | 1.02 | 6.99 | 1 | 0.0082 | 2.76 | 1.30 | 5.87 |
| Paraphraseability | 0.75 | 4.75 | 1 | 0.0293 | 2.12 | 1.08 | 4.17 |
| Matrix verb | 0.98 | 10.02 | 1 | 0.0015 | 2.65 | 1.45 | 4.85 |

OntoNotes sense groupings met this criterion: Sense 2: "Have something inalienable" (e.g. *She has a beautiful voice, We had a great idea*) and Sense 5: "Hold in a certain relationship" (e.g. *She has two sisters, The company has three employees*). The Sense 2 grouping includes all abstract uses of *have* in which the object argument denotes a property or feature rather than something that can be transferred, while the Sense 5 grouping comprises abstract senses of *have* in which the object argument denotes an animate entity that bears a social, kinship or institutional relationship to the subject argument. Accordingly, all PRC tokens whose matrix verbs were found to belong to either Sense 2 or Sense 5 were tagged as presentational (i.e. 'empty'), while all those whose matrix verbs that were found not to belong to either of these two sense groupings were labeled as non-presentational. We performed a logistic regression analysis of the 211 relabeled tokens, using the same features as in the previous analysis (indefinite head, paraphraseability, and (presentational) matrix verb), with the final feature redefined as described. Results are shown in Table 4.

The analysis summarized in Table 4 demonstrates that when a more restrictive definition of semantic emptiness is used to label matrix verbs, the presence of a subject relative-clause is significantly predicted by all three hallmark features of the PRC (*Wald* $\chi^2$ *26.91,* $p < 0.0001$). Let us examine each feature in turn. The presence of an indefinite head is a significant predictor of a subject relative, over and above both matrix-verb type and paraphraseability: indefinite heads are 2.76 (*95% CI: 1.30, 5.87*) times more likely than definite heads to be modified by subject relative clauses (*Wald* $\chi^2 = 6.99$, $p = .0082$). Paraphraseability is also a significant predictor of a subject relative, over and above both definiteness and matrix-verb type: paraphraseable tokens are 2.12 (*95% CI: 1.08, 4.17*) times more likely than non-paraphraseable ones to contain a subject relative (*Wald* $\chi^2 = 4.75$, $p = .0293$). Finally, in a departure from our original results, the presence of a semantically light matrix verb was *also* found to be a significant predictor of a subject relative-clause, over and above the other two features: tokens containing a semantically light sense of the verb

https://doi.org/10.1515/LANGCOG.2011.007 Published online by Cambridge University Press

*have* are 2.65 (*95% CI: 1.45, 4.85*) times more likely to contain a subject relative than those that do not (*Wald* $\chi^2 = 4.75$, $p = .0293$).

We leave to future research a full reexamination of the matrix verbs of all relative-clause tokens based on semantic criteria rather than lexical identity. However, the reanalysis reported here does suggest that all three hallmark features of the PRC are significant predictors of the presence of a subject relative-clause, thus supporting the claim that the PRC is responsible for the prevalence of subject relatives in English conversational data.

## 6. Conclusion

Why do subject relatives prevail where they do? As discussed, processing-based accounts have looked for explanation within the relative clause. They can thus be considered *localist* explanations, in that they focus on the relationship between the relative-pronoun and its clausal sister, the latter of which can be characterized as containing a particular kind of gap (e.g. an 'object gap' in the case of an object relative clause).[10] According to the doctrine of syntactic locality, the only dependency relationship that a syntactic rule can describe is that which holds between a mother and its daughter node(s) or between two syntactic sisters, e.g., a lexical head and its complement(s) or a head and its specifier (Sag 2010b). Locality is a foundational assumption of context-free phrase-structure grammar: phrase-structure rules (e.g. VP → V PP) describe only the immediate daughters of a phrasal node; a phrase-structure rule that, for example, expands a VP to a V followed by a PP that itself contains a PP daughter is not permitted. That is, no phrase-structure grammar would allow a rewrite rule of the form *VP → V (PP → P PP).

Early conceptions of Construction Grammar (see, e.g. Fillmore et al. 1988; Zwicky 1994), abandoned localist assumptions in favor of a more flexible conception of phrase-structure rules. According to this conception, phrase-structure rules may represent niece- and granddaughter-dependencies (e.g. the requirement that the PP complement of the verb *hope* have the preposition *for* as its head), in addition to the strictly local trees of classic phrase-structure grammar. Rules of the latter sort include that which defines the relative-clause

10. Admittedly, the filler-gap dependency is traditionally characterized as a *non-local* dependency, meaning that there is a potentially unlimited set of branching nodes intervening between the left-isolated phrase and the 'trace' in the following clause. However, in declarative models like Sag's Sign-Based Construction Grammar (Sag 2010b), 'extraction' phenomena that count as long-distance dependencies in transformational approaches are instead represented as cascades of local dependencies via the upward percolation of a *gaps* feature, whose value is a phrasal sign co-indexed with the filler daughter. It is the latter 'feature-passing' approach that we assume here.

https://doi.org/10.1515/LANGCOG.2011.007 Published online by Cambridge University Press

pattern, a subtype of the more general filler-gap schema (Sag 2010a). The relative-clause rule is local in that it defines a dependency relationship between syntactic sisters: it consists of a 'filler' daughter (the relative pronoun or complementizer) and the clausal head-daughter containing a coindexed gap (Sag 2010a). Our contention is, however, that the prevalence of subject relative clauses cannot be attributed to properties of this local tree, in particular, the structural distance between the filler element and the gap in the following clause.

If structural distance were dispositive, we would expect the psycholinguistic evidence to support a processing advantage for subject relatives. In fact, as we have argued, the evidence is equivocal, and at least two studies (Roland et al. 2007 and Mak et al. 2008) suggest that *object* relatives actually hold the processing advantage under certain (realistic) discourse circumstances. Further, the structural-distance account fails to recognize that the phenomenon for which explanation is sought is in fact a contextually restricted one. Subject relatives do not prevail overall, but only as modifiers of *postverbal* nominals. Why would this be? Our answer is that the prevalence of the subject-relative pattern is a reflex of its role in a bigger pattern: the PRC. Crucially, this construction defines a non-local dependency. The immediate daughters of the phrase type licensed by the PRC are, respectively, a verb and a PP or NP complement; if the complement is a PP, as in the case of presentational matrix verbs like *talk about* and *look at*, then this PP must be defined as having a NP daughter (itself an instance of niece-licensing). Whether or not the postverbal NP is the daughter of a PP, however, this NP must contain a relative-clause daughter as sister to the head noun. This configuration is exemplified, using a basic version of X′ syntax, in Figure 3, where the relative-clause daughter is represented as an S′.

As Figure 3 shows, the S′ (relative clause) is not a sister to the matrix verb, as would be required by the locality doctrine; this S′ is instead sister to an N′ ($N'_2$) that is the head daughter of an NP sister to the matrix verb ($NP_2$). Thus, the PRC represents an instance of niece-licensing, in which a verb requires its NP sister to contain a particular complement daughter: a relative clause containing a subject-position gap. The explanatory gains made here by assuming this configuration appear to validate the constructionist conception of grammar as an inventory of local and nonlocal trees (rather than an inventory of phrase-structure rules), while offering a challenge to the doctrine of strict locality.[11]

11. If should be noted that if we were to adopt the ternary-branching model of the PRC proposed by Lambrecht (1987, 1988, 2002) and discussed in Section 1 above, the PRC would not in fact challenge a strict view of locality. However, we choose to remain agnostic about the best constituent-structure analysis for the PRC, as the analysis illustrated in Figure 1 has support as well. Notice, for example, that coordinate-NP examples like *I have friends that clip*

https://doi.org/10.1515/LANGCOG.2011.007 Published online by Cambridge University Press

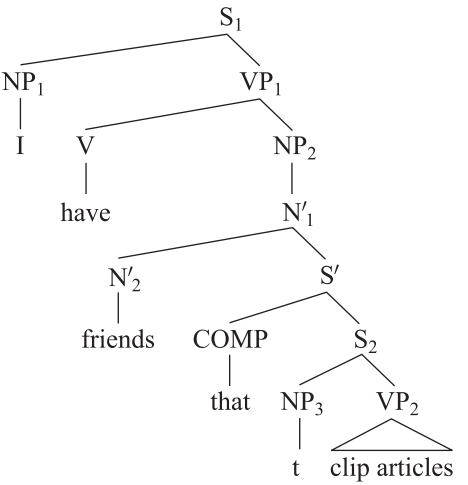


Figure 3. *The PRC as a non-local dependency*

A closing disclaimer is in order here: while our findings suggest that certain large-scale linguistic patterns are the effects of constructional licensing, there are a number of reasons why construction-based explanation cannot simply replace constraint-based—and, in particular, processing-based—explanation. Below we will discuss three of these reasons.

First, the scope of explanation offered in this paper is relatively limited, in that it relies on a particular construction in a particular language. If the PRC is to explain the prevalence of subject relatives across languages, it must be present in a wide array of languages. Lambrecht (2000) argues that it is; he points to biclausal presentational structures in a diverse set of languages including French, Romansch, Welsh, Mandarin, Egyptian Arabic and the Cushitic language Boni (see also Sasse 1987, who refers to such structures as *split structures*). However, many of these constructions (including the serial-verb constructions described by Lambrecht) do not contain relative clauses. Because of the relative paucity of typological studies within Construction Grammar, the processing explanation currently has stronger claims to cross-linguistic validity. As Lin and Bever (2006) point out, the parsing preference for subject relatives has been demonstrated in a diverse array of languages. At the same time, they observe, studies of languages with head-final relative clauses, including Mandarin, Japanese and Korean, reveal more varied parsing preferences than the

*articles and friends that clip coupons* provide support for the binary-branching analysis, in which nominal head and relative clause jointly form a NP constituent, as in the traditional X′ analysis of relative-clause modification.

https://doi.org/10.1515/LANGCOG.2011.007 Published online by Cambridge University Press

processing model would predict. For example, subject-relatives appear to be preferred in Japanese and Korean despite the fact that a subject gap is more distant from the final head noun than is an object gap. Thus, neither constructionist nor processing accounts yet provide a complete story of subject-relative prevalence across languages.

Second, it may be that the constructionist and processing models described here target different aspects of the language faculty. The processing account focuses on a comprehension bias (i.e. a parsing preference), while the constructionist account focuses on a production bias—the tendency of speakers of conversational English to use the OS structure more frequently than the OO structure. Language producers do not necessarily anticipate the needs of recipients during utterance planning (Ferreira and Dell 2000; Wardlow Lane and Ferreira 2008), and thus we cannot assume that speakers choose subject relatives in order to ease the hearer's processing burden. However, there is reason to regard production and comprehension biases as two sides of the same coin. According to Gennari and MacDonald's (2009) *Production-Distribution-Comprehension* framework, the distributional patterns of certain structures in a language—themselves the result of production choices—become probabilistic constraints guiding the comprehension system of listeners exposed to such input (ibid.: 2). Similarly, Pickering and Garrod (2007) argue that highly predictable linguistic elements are more easily processed by listeners, and suggest that the processing of structural elements is facilitated by predictability to an even greater degree than is the processing of lexical items (ibid.: 106, Box 3). If this general approach is correct, the processing advantage enjoyed by subject relative clauses may be attributable to their comparative frequency/predictability rather than to their comparative structural simplicity. This is not to say, however, that relative structural simplicity could not influence speakers' production choices. Language production requires computation and storage resources, perhaps even more so than comprehension (Hartsuiker and Barkhuysen 2006: 183). Thus, we might postulate that the relative ease of processing of subject relatives encourages speakers to use the PRC. In any case, production and comprehension biases seem inextricably interlinked in the domain of relative clauses.

Third, explaining usage patterns within a corpus may require *both* processing-based and construction-based explanation. The combined approach is exemplified by Francis and Michaelis (2010); its authors examined the conditions governing the use of the discontinuous dependency known as relative-clause extraposition (RCE), e.g. *[New sets] soon appeared [that were able to receive all the channels]*. They found that RCE use conditions include both the subject-predicate weight ratio and discourse factors. Comparison of RCE and non-RCE sentences from the ICE-GB corpus revealed a strong preference for RCE when the weight ratio was less than 0.2 (i.e. when the RC was more than five

https://doi.org/10.1515/LANGCOG.2011.007 Published online by Cambridge University Press

times longer than the VP), and a strong dispreference for RCE when the weight ratio was greater than 0.8. However, they also found that when weight ratios were between 0.2 and 0.8, discourse factors became operative: the choice of structure was determined primarily by definiteness and predicate type. Specifically, RCE was preferred for tokens with presentational characteristics: indefinite subjects and unaccusative or passive verbs. What this suggests is that some RCE tokens, like some subject relatives, are the reflexes of a presentational construction, while others are the by-products of processing constraints.

The moral of the story is that while constructional accounts may not supplant processing-based accounts, constructions can no longer be dismissed as "taxonomic epiphenomena", as in Chomsky's memorable formulation; they are instead fundamental tools for linguistic explanation.

## References

Aissen, J. 1999. Agent focus and inverse in Tzotzil. *Language* 75(3). 451–485.

Alexiadou, A., P. Law, A. Meinunger & C. Wilder. 2000. *The syntax of relative clauses*. Amsterdam: John Benjamins.

Bybee, J. 2007. *Frequency of use and the organization of language*. Oxford: Oxford University Press.

Chen, J., D. Dligach & M. Palmer. 2007. Towards large-scale high performance English verb sense disambiguation by using linguistically motivated features. *First IEEE International Conference on Semantic Computing Proceedings*, 378–388.

Chomsky, N. 1989. Some notes on economy of derivation and representation. *MIT Working Papers in Linguistics* 10. 43–74.

Chomsky, N. 1995. *The minimalist program*. Cambridge, MA: MIT Press.

Demuth, C. 1989. Maturation and the acquisition of the Sesotho passive. *Language* 65(1). 56–80.

Diessel, H. 2008. Iconicity of sequence: A corpus-based analysis of the positioning of temporal adverbial clauses in English. *Cognitive Linguistics* 19(3). 465–490.

Diessel, H. & M. Tomasello. 2000. The development of relative clauses in spontaneous child speech. *Cognitive Linguistics* 11(1/2). 131–151.

Dligach, D. & M. Palmer. 2008. Improving verb sense disambiguation with automatically retrieved semantic knowledge. *Proceedings of the Second IEEE International Conference on Semantic Computing (ICSC)*, 182–189.

Duffield, C. J., J. D. Hwang, S. W. Brown, D. Dligach, S. E. Vieweg, J. Davis & M. Palmer. 2007. Criteria for the manual grouping of verb senses. *Proceedings of the Linguistic Annotation Workshop held in conjunction with ACL-2007*, 49–52.

Fellbaum, C. (ed). 1998. *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Ferreira, V. S. & G. S. Dell. 2000. Effect of lexical ambiguity on syntactic and lexical production. *Cognitive Psychology* 40(4). 296–340.

Fillmore, C., P. Kay & M. C. O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: The case of *let alone*. *Language* 64(3). 501–538.

Fox, B. & S. Thompson. 1990. A discourse explanation for the grammar of relative clauses in English conversation. *Language* 66(2). 297–316.

Francis, E. J. & L. A. Michaelis. 2010. Combining weight and discourse factors to predict relative clause extraposition in English. Poster presented at the Annual Meeting of the Linguistics Society of America, Baltimore, MD, January 2010.

Geisler, C. 1998. Infinitival relative clauses in spoken discourse. *Language Variation and Change* 10(1). 23–41.

Gennari, S. P. & M. C. MacDonald. 2009. Linking production and comprehension processes: The case of relative clauses. *Cognition* 111(1). 1–23.
Gernsbacher, M. A. & S. Shroyer. 1989. The cataphoric use of the indefinite *this* in spoken narratives. *Journal of Memory and Cognition* 17(5). 536–540.
Gibson, E. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition* 68(1). 1–76.
Gibson, E., T. Desmet, D. Grodner, D. Watson & K. Ko. 2005. Reading relative clauses in English. *Cognitive Linguistics* 16(2). 313–353.
Godfrey, J., E. Holliman & J. McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. *Proceedings of ICASSP-92*, 517–520.
Goldberg, A. 1995. *Constructions: A construction grammar approach to argument structure*. Chicago: Chicago University Press.
Goldberg, A. 2006. *Constructions at work*. Oxford: Oxford University Press.
Goodluck, H. & M. Rochemont (eds.). 1992. *Island constraints: Theory, acquisition and processing*. Dordrecht: Kluwer Academic.
Gundel, J., N. Hedberg & R. Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language* 69(2). 274–307.
Hartsuiker, R. J. & P. N. Barkhuysen. 2006. Language production and working memory: The case of subject-verb agreement. *Language and Cognitive Processes* 21(1/2/3). 181–204.
Hawkins, J. 1999. Processing complexity and filler-gap dependencies across grammars. *Language* 75(2). 244–285.
Hawkins, J. 2004. *Efficiency and complexity in grammars*. Oxford: Oxford University Press.
Homer, K. 2000. *A discourse constraint on subject information questions*. Boulder, CO: University of Colorado dissertation.
Hovy, E., M. Marcus, M. Palmer, L. Ramshaw & R. Weischedel. 2006. OntoNotes: The 90% solution. *Proceedings of HLT-NAACL 2006*, 57–60.
Kay, P. 2002. English subjectless tag sentences. *Language* 78(3). 453–481.
Keenan, E. & B. Comrie 1977. Noun phrase accessibility and universal grammar. *Linguistic Inquiry* 8(1). 63–69.
Koenig, J. P. & K. Lambrecht. 1999. French relative clauses as secondary predicates. In F. Corbin, C. Dobrovie-Sorin & J.-M. Marandin (eds.), *Empirical issues in formal syntax and semantics* 2, 191–214. The Hague: Thesus.
Lakoff, G. 1987. *Women, fire and dangerous things: What categories reveal about the mind*. Chicago: University of Chicago Press.
Lambrecht, K. 1987. Presentational cleft constructions in spoken French. In J. Haiman & S. A. Thompson (eds.), *Clause combining in grammar and discourse,* 135–179. Amsterdam: John Benjamins.
Lambrecht, K. 1988. There was a farmer had a dog: Syntactic amalgams revisited. *Proceedings of the Fourteenth Annual Meeting of the Berkeley Linguistics Society,* 319–339.
Lambrecht, K. 1994. *Information structure and sentence form*. Cambridge: Cambridge University Press.
Lambrecht, K. 2000. When subjects behave like objects: An analysis of the merging of S and O in sentence-focus constructions across languages. *Studies in Language* 24(3). 611–682.
Lambrecht, K. 2002. Topic, focus, and secondary predication: The French presentational relative construction. In C. Beyssade, R. Bok-Bennema, F. Drijkoningen & P. Monachesi (eds.), *Romance languages and linguistic theory 2000,* 171–212. Amsterdam: John Benjamins.
Lin, C. J. C. & T. G. Bever. 2006. Subject preference in the processing of relative clauses in Chinese. *Proceedings of the Twenty-Fifth West Coast Conference on Formal Linguistics,* 254–260.
Lohse, B., J. Hawkins & T. Wasow 2004. Domain minimization in English verb-particle constructions. *Language* 80(2). 238–261.
MacDonald, M. C. & M. H. Christiansen. 2002. Reassessing working memory: Comment on Just & Carpenter (1992) and Waters & Caplan (1996). *Psychological Review* 109(1). 35–54.

Mak, W., W. Vonk & H. Schrieffers. 2008. Discourse structure and relative clause processing. *Memory and Cognition* 36(1). 170–181.
Malouf, R. 2003. Cooperating constructions. In E. J. Francis & L. A. Michaelis (eds.), *Mismatch: Form-function incongruity and the architecture of grammar,* 403–424. Stanford: CSLI Publications.
Marcus, M., B. Santorini & M. A. Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn treebank. *Computational Linguistics* 19(2). 313–330.
McCawley, J. 1981. The syntax and semantics of English relative clauses. *Lingua* 53(2/3). 99–149.
McCawley, J. 1988. *The syntactic phenomena of English*, Volume 1. Chicago: University of Chicago Press.
Michaelis, L. A. & H. S. Francis. 2007. Lexical subjects and the conflation strategy. In N. Hedberg & R. Zacharski (eds.), *Topics in the grammar-pragmatics interface: Papers in honor of Jeanette K. Gundel*, 19–48. Amsterdam: John Benjamins.
Michaelis, L. A. & K. Lambrecht. 1996. Toward a construction-based model of language function: The case of nominal extraposition. *Language* 72(2). 215–247.
Mithun, M. 1991. The Role of motivation in the emergence of grammatical categories: The grammaticization of subjects. In E. C. Traugott & B. Heine (eds.), *Approaches to grammaticalization*, Volume 2, 159–184. Amsterdam: John Benjamins.
Ouhalla, J. 1993. Subject extraction, negation and the anti-agreement effect. *Natural Language and Linguistic Theory* 11(3). 477–518.
Pickering, M. J. & S. Garrod. 2007. Do people use language production to make predictions during comprehension? *Trends in Cognitive Science* 11(3). 105–110.
Prince, E. 1984. Topicalization and left-dislocation: A functional analysis. *Annals of the New York Academy of Sciences* 433. 213–225.
Prince, E. 1992. The ZPG letter: Subjects, definiteness, and information status. In W. C. Mann & S. A. Thompson (eds.), *Discourse description: Diverse linguistic analyses of a fund-raising text,* 295–325. Amsterdam: John Benjamins.
Reali, F. & M. Christiansen. 2007. Processing of relative clauses is made easier by frequency of occurrence. *Journal of Memory and Language* 53. 1–23.
Roland, D., C. O'Meara, H. Yun & G. Mauner. 2007. Processing object relative clauses: Discourse or frequency? Poster presented at the CUNY sentence processing conference, San Diego.
Sag, I. 2010a. English filler-gap constructions. *Language* 86(3). 486–545.
Sag, I. 2010b. Feature geometry and predictions of locality. In G. Corbett & A. Kibort (eds.), *Features: Perspectives on a key notion in linguistics*, 236–271. Oxford: Clarendon Press.
Sasse, H. J. 1987. The thetic/categorical distinction revisited. *Linguistics* 25(3). 511–580.
Traxler, M., R. Morris & R. Seely. 2002. Processing subject and object relative clauses: Evidence from eye movements. *Journal of Memory and Language* 47(1). 69–90.
Van Valin, R. D. & R. J. LaPolla. 1997. *Syntax*. Cambridge: Cambridge University Press.
Ward, G. & E. Prince. 1991. On the topicalization of indefinite NPs. *Journal of Pragmatics* 16(2). 167–177.
Wardlow Lane, L. & V. S. Ferreira. 2008. Speaker-external versus speaker-internal forces on utterance form. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 34(6). 1466–1481.
Warren, T. & E. Gibson. 2002. The influence of referential processing on sentence complexity. *Cognition* 85(1). 79–112.
Wells, J., M. Christiansen, D. Race, D. Acheson & M. C. MacDonald. 2009. Experience and sentence comprehension: Statistical learning and relative clause comprehension. *Cognitive Psychology* 58(2). 250–271.
Zwicky, A. 1994. Dealing out meaning: Fundamentals of grammatical constructions. *Proceedings of the Twentieth Annual Meeting of the Berkeley Linguistics Society,* 611–625.