Lexical Variation in Relativizer Frequency

Thomas Wasow, T. Florian Jaeger, and David M. Orr*
Stanford University

0. Introduction

The notion of exception presupposes that of rule; as Webster puts it, an exception is "a case to which a rule does not apply". Linguistic rules (and, more recently, constraints, principles, parameters, etc.) are usually taken to be categorical, at least in the generative tradition. Quantitative data like frequency of usage are widely considered irrelevant to grammar, and gradient theoretical notions like degrees of exceptionality have remained outside of the theoretical mainstream.

This antipathy towards things quantitative probably has its origins in Chomsky's early writings, which dismissed the significance of frequency data and statistical models (see, e.g., Chomsky 1955/75, pp. 145-146; 1957, pp. 16-17; 1962, p. 128; 1966, p. 35-36). But recently, the availability of large on-line corpora and computational tools for working with them has led some linguists to question the exclusion of frequency data and non-categorical formal mechanisms from theoretical discussions (for example, Wasow, 2002, and Bresnan, et al, 2005). Moreover, corpus work has revealed that natural-sounding counterexamples to many purportedly categorical generalizations can be found in usage data (Bresnan and Nikitina, 2003).

If categorical rules are replaced by gradient models, what becomes of the notion of exceptionality? The paradigmatic instance of an exception is a lexical item that satisfies the applicability conditions of a (categorical) rule, but cannot undergo it. (When rules are categorical, so are exceptions). The obvious analogue for a non-categorical generalization would be a lexical item whose frequency of occurrence in a given

---

* This paper is dedicated to Professor Günter Rohdenburg of Paderborn University, whose sixty-fifth birthday coincides with the completion of the paper. Professor Rohdenburg's seminal studies on English usage and structure have been an inspiration to many data-oriented students of language, ourselves included.

We received help and advice on this work from many people. Paul Fontes did essential work on the maximum entropy predictability model described at the end of section 3. Sandy Thompson was generous in sharing an early version of Fox and Thompson (in press) with us, and in giving us very useful feedback on earlier versions of this work. Additional help and advice was provided by at least the following people: David Beaver, Joan Bresnan, Brady Clark, Liz Coppock, Vic Ferreira, Edward Flemming, Ted Gibson, Jack Hawkins, Irene Heim, Dan Jurafsky, Rafe Kinsey, Roger Levy, Chris Manning, Tanya Nikitina, Doug Rohde, Doug Roland, Neal Snider, Laura Staum, Michael Wagner, and Annie Zaenen. Thanks also to Heike Wiese and Horst Simon, for organizing the workshop at which this material was originally presented, and to the audience at that workshop for their excellent comments.

environment is dramatically different from that of other lexical items that are similar in relevant respects.

For example, whereas about 8% (11,405/146,531) of the occurrences of transitive verbs in the Penn Treebank III corpora (Marcus et al., 1999) are in the passive voice, certain verbs occur in the passive far more frequently, and others far less frequently. Among the former is *convict*, which occurs in the passive in 33% (25/76) of its occurrences as a verb; the latter is represented by *read*, fewer than 1% (6/788) of whose occurrences as a transitive verb are passive.[1]

Such skewed distributions, which we will call "soft exceptions", are by no means uncommon. For grammarians who make use of non-categorical data and mechanisms, soft exceptions constitute a challenge. Simply recording statistical biases in individual lexical entries may be feasible and useful in applications to language technologies. But it is theoretically unsatisfying: we would like to explain why words show radically different proclivities towards particular constructions.

The remainder of this paper examines one set of soft constraints and offers an explanation for them in terms of a combination of semantic/pragmatic and psycholinguistic considerations.

1. Background

The particular phenomenon we examine is the optionality of relativizers (*that* or *wh*-words) in the initial position of certain relative clauses (RCs). This is illustrated in the following examples:

(1) a. That is certainly one reason (why/that) crime has increased.
b. I think that the last movie (which/that) I saw was *Misery*.
c. They have all the water (that) they want.

We have been exploring what factors correlate with relativizer occurrence in RCs, using syntactically annotated corpora from the Penn Treebank III. The results presented below have been carried out using the Switchboard corpus, which consists of 650 transcribed telephone conversations between pairs of strangers (on a list of selected topics), totalling approximately 800,000 words.

Certain factors make relativizers obligatory, or so strongly preferred as to mask the effects of other factors. As is well-known (see Huddleston and Pullum, 2002; 1055), if the RC's gap is the subject of the RC, then the relativizer cannot be omitted:[2]

---

[1] These numbers are based on searches of the parsed portions of the *Wall Street Journal*, Brown, and Switchboard corpora, looking at the ratio of passive verb phrases to the total number of VPs directly dominating the verb in question and an NP (possibly a trace).

[2] There are dialects that permit relativizer omission in some RCs with subject gaps, as in the children's song, *There was a farmer had a dog….*

(2) I saw a movie *(that) offended me.[3]

We have excluded these from our investigations, concentrating instead on what we will call non-subject extracted relative clauses, or NSRCs. We have also excluded examples involving what Ross (1967) dubbed "pied piping", as in (3):

(3) a. a movie to *(which) we went
b. a movie *(whose) title I forget

Non-restrictive relative clauses are conventionally claimed (Huddleston and Pullum, 2002; 1056) to require a *wh*-relativizer, and this seems to be correct in clear cases:

(4) a. *Goodbye Lenin*, which I enjoyed, is set in Berlin
b. **Goodbye Lenin*, (that) I enjoyed, is set in Berlin

The converse – that *wh*-relativizers may not appear in restrictive RCs – is a well-known prescription (e.g., Fowler 1944; 635), though it does not appear to be descriptively accurate. Evaluating these claims is complicated by the fact that the boundary between restrictive and non-restrictive modifiers seems to be quite fuzzy. Instead of trying to identify all and only non-restrictive RCs, we excluded all examples with *wh*-relativizers. This decision was also motivated in part by our observation that disproportionately many of the examples with *wh*-relativizers were questionable for other reasons (e.g. some embedded questions were misanalyzed as RCs). Thus, our results are based on the comparison between NSRCs with *that* relativizers and those with no overt relativizer.[4]

In addition, we excluded reduced subject-extracted and infinitival RCs, since they never allow relativizers (except for infinitival RCs with pied-piping – where the relativizer is obligatory):

(5) a. a movie (*that) seen by millions
b. a movie (*that) to see
c. a movie in *(which) to fall asleep

After these exclusions, our corpus contained 3,701 NSRCs, of which 1,601 (43%) begin with *that* and the remaining 2,100 (57%) have no relativizer. A variety of factors seem to influence the choice between *that* and no relativizer in these cases. These include the length of the NSRC, properties of the NSRC subject (such as pronominality, person, and number), and the presence of disfluencies nearby. We discuss these elsewhere (Jaeger & Wasow, in press; Jaeger, Orr, & Wasow, 2005; Jaeger, 2005), exploring interactions

[3] An asterisk outside parentheses is used to indicate that the material inside the parentheses is obligatory.
[4] The studies were replicated including the NSRCs with *wh*-relativizers. The results are qualitatively the same, though the numbers are of course different.

among the factors and seeking to explain the patterns on the basis of processing considerations.

The focus of the present paper is on how lexical choices in an NP containing an NSRC can influence whether a relativizer is used. We show that particular choices of determiner, noun, or prenominal adjective may correlate with exceptionally high or exceptionally low rates of relativizers. We then propose that this correlation can be explained in terms of the predictability of the NSRC, which in turn has a semantic/pragmatic explanation.

2. Lexical Choices and Relativizer Frequency

Early in our investigations of relativizer distribution in NSRCs we noticed that relativizers are far more frequent in NPs introduced by *a* or *an* than in those introduced by *the*. Specifically, *that* occurs in 74.8% (226/302) of the NSRCs in *a(n)*-initial NPs and in only 34.2% (620/1813) of those in *the*-initial NPs. Puzzled, we checked the relativizer frequency for NSRCs in NPs introduced by other determiners. The results are summarized in Figure 1, where the numbers in parentheses indicate the total number of examples.

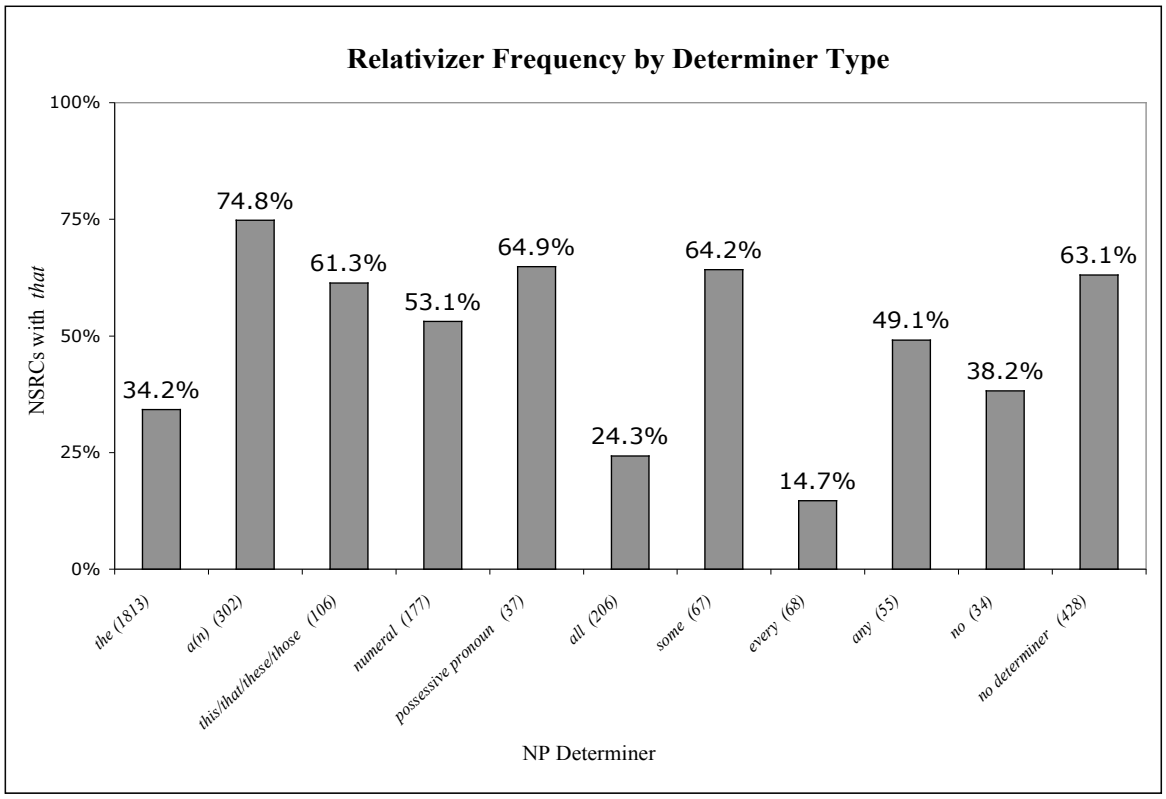

Figure 1

The variation in these numbers is striking, but it is by no means obvious why they are distributed as they are. Curious whether other lexical choices within NPs containing

NSRCs might be correlated with relativizer frequency, we compared rates of relativizer occurrence for the nouns most commonly modified by NSRCs. Again, we found a great deal of variation, with no obvious pattern.

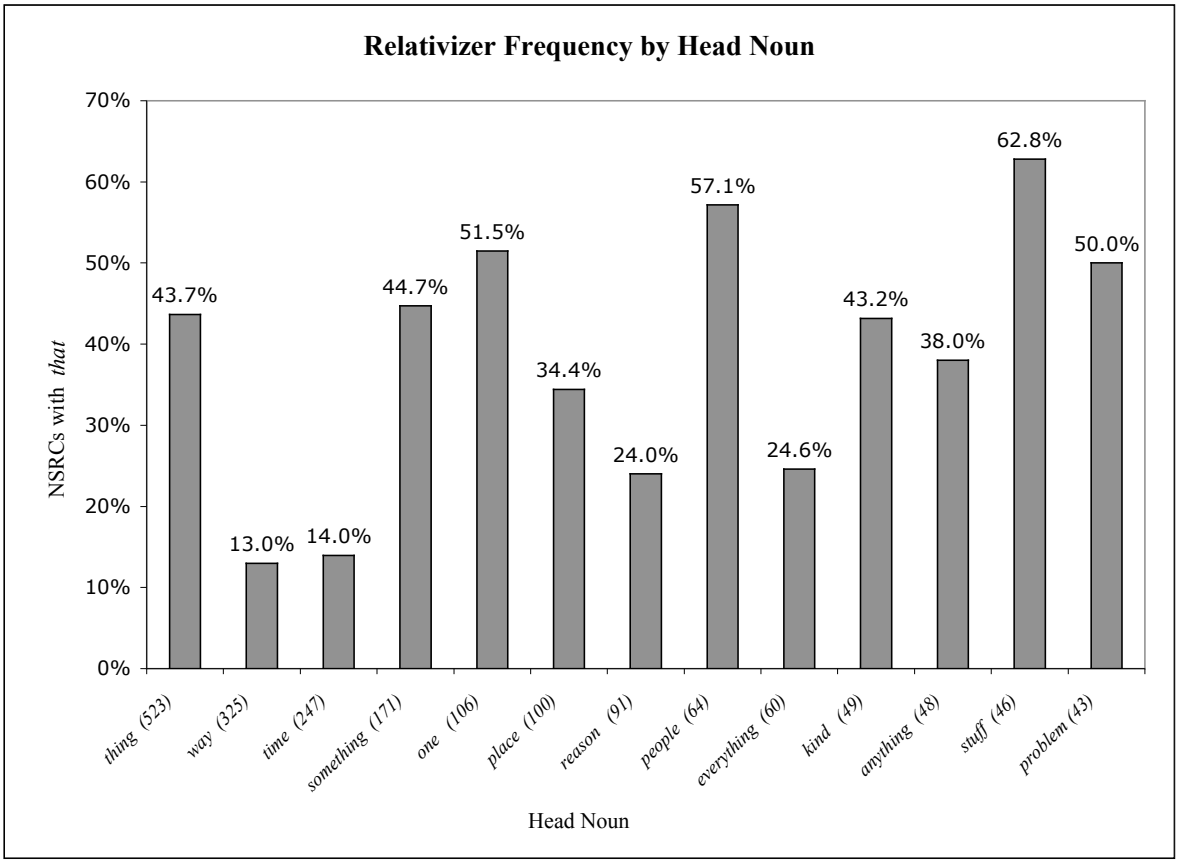

Figure 2

If individual determiners and head nouns are correlated with such highly variable rates of relativizer presence, we reasoned that the words that come between determiners and head nouns – namely, prenominal adjectives – might show similar variation. And indeed they do: Figure 3 shows the relativizer frequencies for the prenominal adjectives that occur most frequently in NPs with NSRCs.

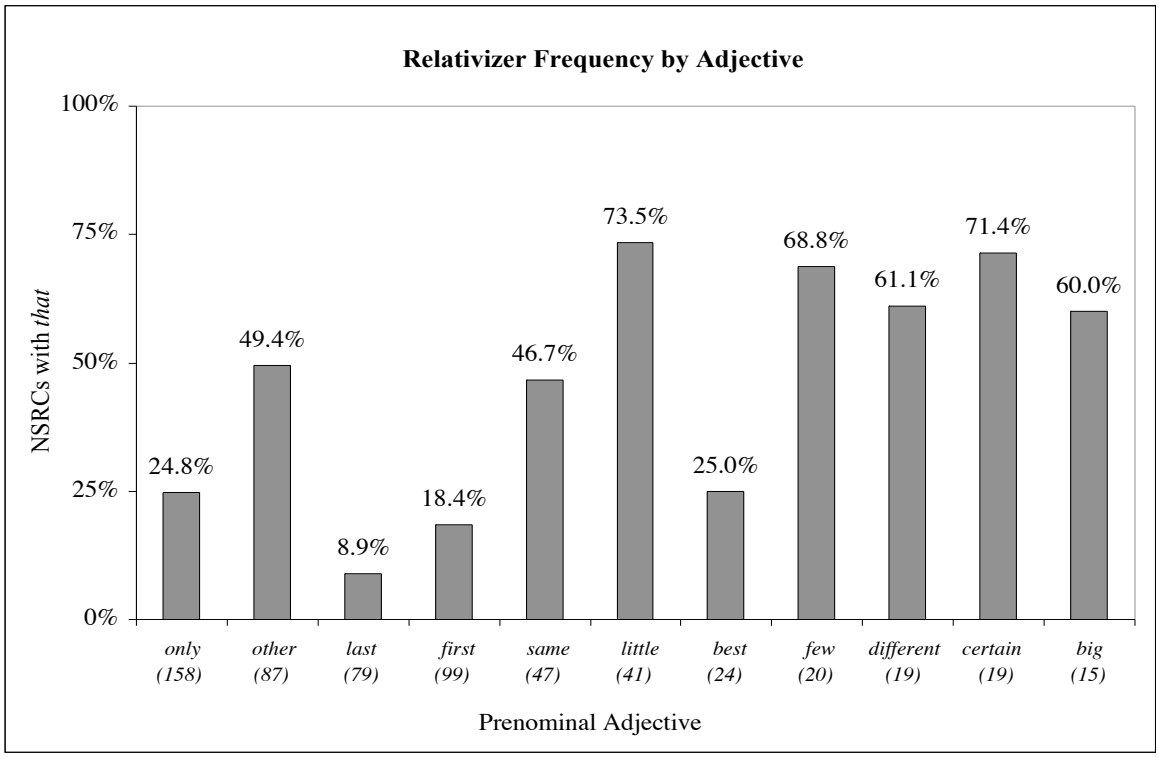

Figure 3

The differences in relativizer frequency based on properties of the modified NP are immense. For example, NSRCs modifying NPs with the adjective *little* are on average over eight times more likely to have a relativizer than NSRCs modifying NPs with the adjective *last*. These differences are not due to chance (the sample size is large enough, see Figures 1-3).

Why should lexical choices in the portion of an NP preceding an NSRC make such a dramatic difference in whether the NSRC begins with *that* or has no relativizer? How can we explain such soft exceptions to the optionality of *that* in NSRCs as *a(n), every, stuff, way, little,* and *last*?

3. Predictability

An example from Fox and Thompson (in press) provided a crucial clue. They observed that the following sentence sounds quite awkward with a relativizer. [5]

(6) That was the ugliest set of shoes (that) I ever saw in my life.

Moreover, the sentence seems incomplete without the relative clause:

[5] Fox and Thompson's account of the preference for no relativizer in (6) is based on the claim that (6) falls at the monoclausal end of a "continuum of monoclausality to bi-clausality". We discuss this idea in section 5 below.

(7) That was the ugliest set of shoes.

(7) would be appropriate only in a context in which some comparison collection of sets of shoes is clear to the addressee.

These observations led us to conjecture that the strong preferences in (6) for a relative clause in the NP and for no relativizer in the relative clause might be connected. Looking at *the* vs. *a(n)* in our corpus (the contrast that first got us started on this line of inquiry), we found that, of the 30,587 NPs beginning with *the*, 1813 (5.93%) contain NSRCs, whereas only 302 (1.18%) of the 45,698 NPs beginning with *a(n)* contain NSRCs. This difference ($\chi^2 = 812$, p=0) lent plausibility to our conjecture.

Hence, we propose the following hypothesis:

(8)) **The Predictability Hypothesis:** In environments where an NSRC is more predictable, relativizers are less frequent.

This formulation is somewhat vague, since neither the notion of 'environment' nor of 'predictability' is made precise. Our initial tests of the hypothesis use simple operationalizations of these notions: the environments are the NPs containing the determiners, nouns, and adjectives described in the previous section, and an NSRC's predictability in the environment of one of these words is measured by the percentage of the NPs containing that word that also are modified by an NSRC.

Figures 4-6 plot cooccurrence with NSRCs against frequency of relativizer absence in NSRCs. The points in Figure 4 represent the eleven determiner types given in Figure 1; the points in Figure 5 represent the thirteen head nouns given in Figure 2; and the points in Figure 6 represent the eleven adjectives given in Figure 3.[6] We regressed the mean NSRC predictability given a specific determiner, adjective, or head noun against mean frequency of relativizer absence (other tests showed that the trend is indeed linear and not of a higher order). The correlation between NSRC cooccurrence and relativizer absence is significant for all three categories. Correlating the predictability of NSRCs for all 35 words (the determiners, adjectives, and head nouns in our sample) against frequency of relativizer absence was also significant (adjusted r^2=.36, F(1,33)=19.9, p<.001).[7]

[6] The mean plots in the three figures represent rather different sample sizes. Determiners are a closed class, so Figure 4 includes almost all NSRCs, whereas Figures 5 and 6 are based on just the head nouns and adjectives that cooccur most frequently with NSRCs. And since almost all NPs include a head noun but most do not have prenominal adjectives, the sample size in Figure 6 is far lower than in Figure 5

[7] After removing two extreme outliers, the adjusted r^2=.56, F(1,31)=36.1, p<0.001.

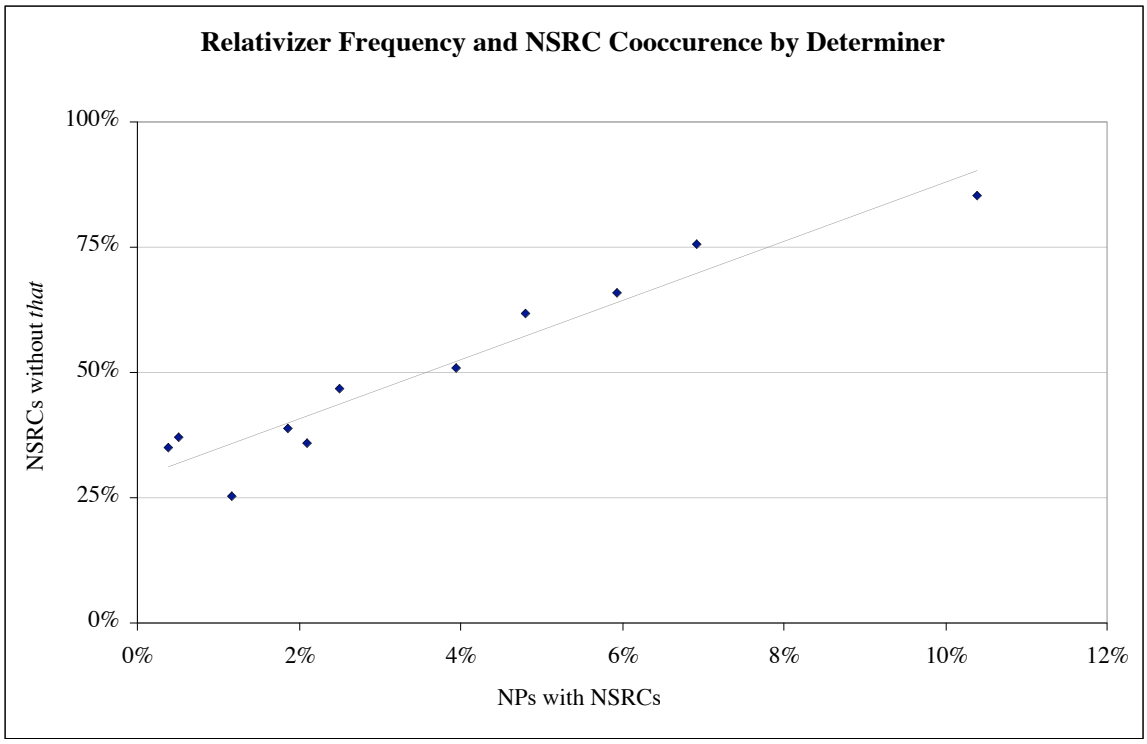

Figure 4
adjusted r^2= .91[8]
F(1,9)=105.1, p<.001

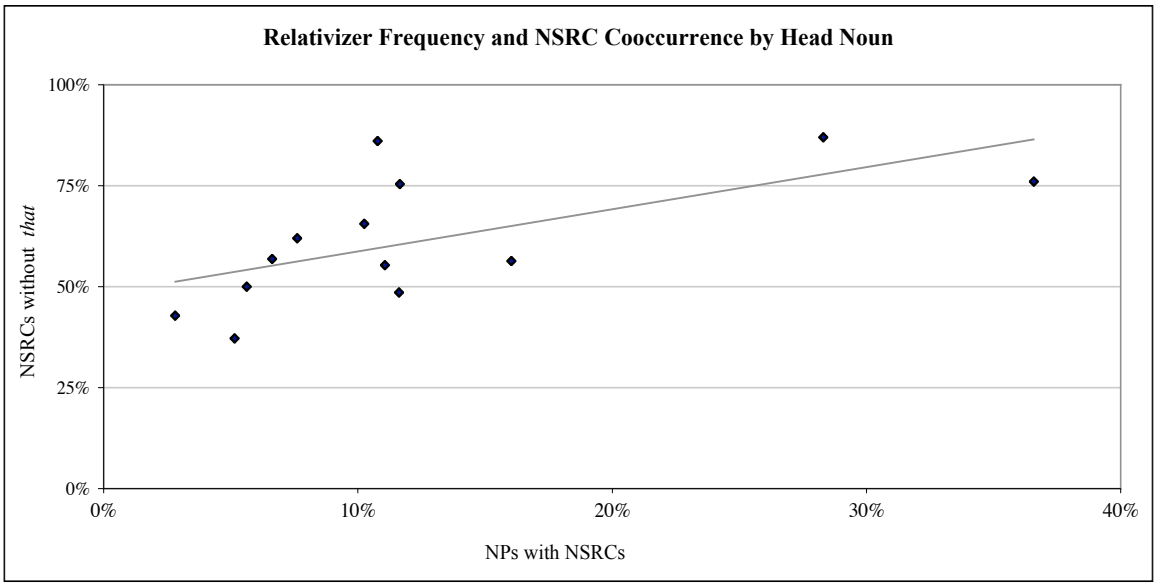

Figure 5

[8] Adjusted r^2s provide a more reliable measure of the goodness of fit of the model compared to normal, unadjusted r^2s, which usually are too optimistic. Generally, r^2s estimates the amount of variation in the data accounted for by the model, e.g. an r^2 of .92 means that the model accounts for 92% of the variation.

adjusted r^2= .35
F(1,11)=7.4, p=.02

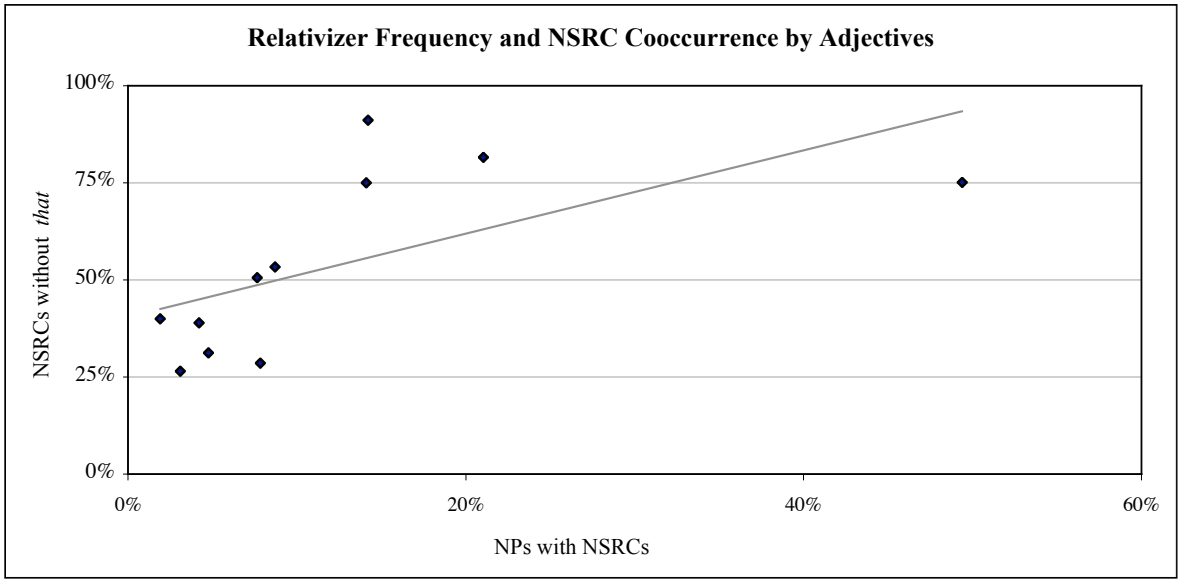

Figure 6
adjusted r^2= .32
F(1,9)=5.8, p=.04

These results support the Predictability Hypothesis: on average, if a determiner, prenominal adjective, or head noun within an NP increases the likelihood that the NP will contain an NSRC, then it also increases the likelihood that an NSRC in the NP will lack a relativizer.

The evidence presented above supports the Predictability Hypothesis, but the predictability measures employed are rather simple. We used one word at a time in the modified NP to estimate the predictability of an NSRC, and, we only used the most frequent types of determiners, adjectives, and head nouns.[9] There are several ways to develop more sophisticated models of an NSRC's predictability that (i) take into account more than one word in the NP at a time, and (ii) are not limited to the most frequent types. We present one such approach, using a machine learning technique. This approach would also enable us to include information relevant to NSRC predictability that is not

[9] Furthermore, we used means to predict means (i.e. we used the mean predictablity of an NSRC given a certain word in the modified NP and correlated that against the mean relativizer likelihood for NSRCs modifying those NPs). This method arguably inflates our r^2s (i.e. the measure of how much of the variation in relativizer omission is captured by predictability). Elsewhere (Jaeger, Levy, Wasow, & Orr, 2005), we address this issue by using binary logistic regressions that predict the presence of a relativizer based on the predictability of the NSRC on a case-by-case basis.

due to lexical properties of NPs (such as their grammatical function), but the study we report on here is limited to lexical factors.[10]

We created a maximum entropy classifier (see Ratnaparkhi, 1997), which used features of an NP to predict how likely a relative clause[11] was in that NP. Features included the type of head noun, any prenominal adjectives, and the determiner, as well as some additional properties, such as whether the head noun was a proper name, and whether the modified NP contained a possessive pronoun. Based on these features, the classifier assigned to each NP in the corpus a probability of having an RC, which we will refer to as its "predictability index". We then grouped NPs according to these predictability indices, and examined how the relativizer likelihood in an NSRC varied across the groups.[12]

Before checking on relativizer presence, however, we needed to test the accuracy of the predictability indices our classifier assigned. We did this by comparing the predictability index range of each of the groups with the actual rates of RCs in the NPs in the groups. As can be seen in Figure 7, the occurrences of RCs in the NPs in each group were consistently within or close to the range assigned by the classifier. This indicates that the classifier was producing reasonable values for the predictability index of the NPs.

---

[10] Studies involving non-lexical factors are in progress.

[11] This study differs from the earlier ones in that considered the predictability of any relative clause, not just of non-subject relative clauses. This broader criterion provided the classifier with more data on which to base its classifications; the narrower criterion would have required a larger corpus in order to get reliable classifications. So this study is testing for a slightly different correlation than the one stated in the Predictability Hypothesis. However, since the probability that an NP will contain an NSRC and the probability that an NP will contain an RC are highly correlated, a correlation between RC predictability and relativizer absence still supports our claims (cf. also footnote 14). Future research may determine which of the two measures is the better predictor of relativizer frequency.

[12] Here we present the result of a classifier trained on the Switchboard corpus, similar results were found for the parsed Wall Street Journal (Penn Treebank III release).

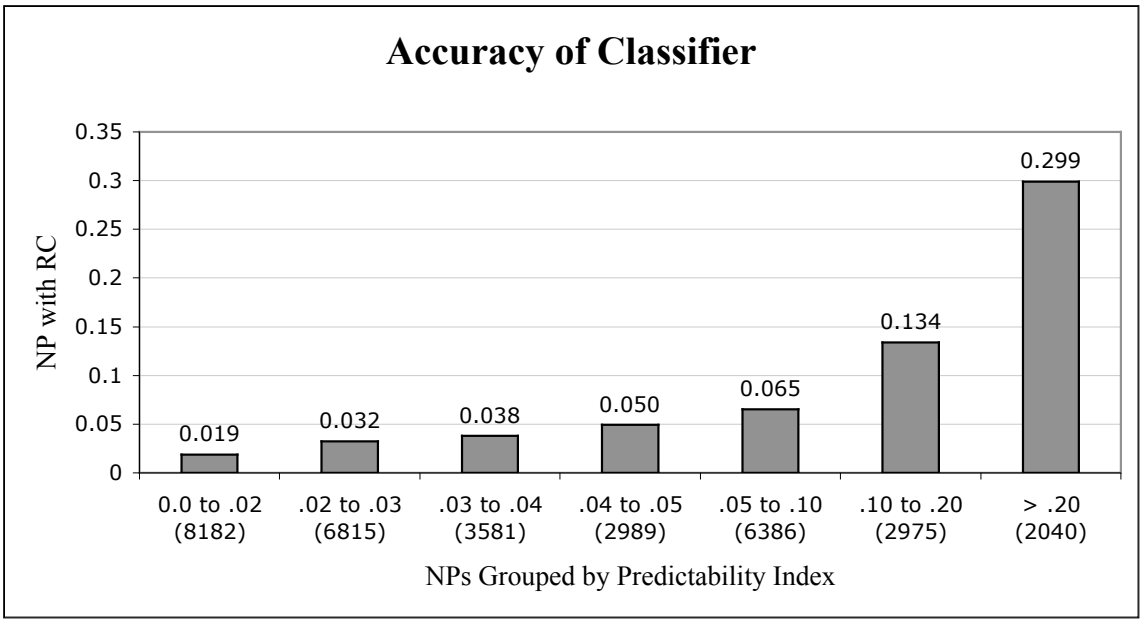

Figure 7

For the NPs containing NSRCs, we then used the classifier's predictability indices to test whether relativizers are less frequent where RCs are more predictable. We did this by examing the rates of relativizer absence for each of our groupings of NPs. As Figure 8 shows, the results are similar to what we found looking at the most frequent determiners, adjectives, and nouns separately: NSRCs in NPs whose features make them more likely to contain RCs are less likely to have relativizers.

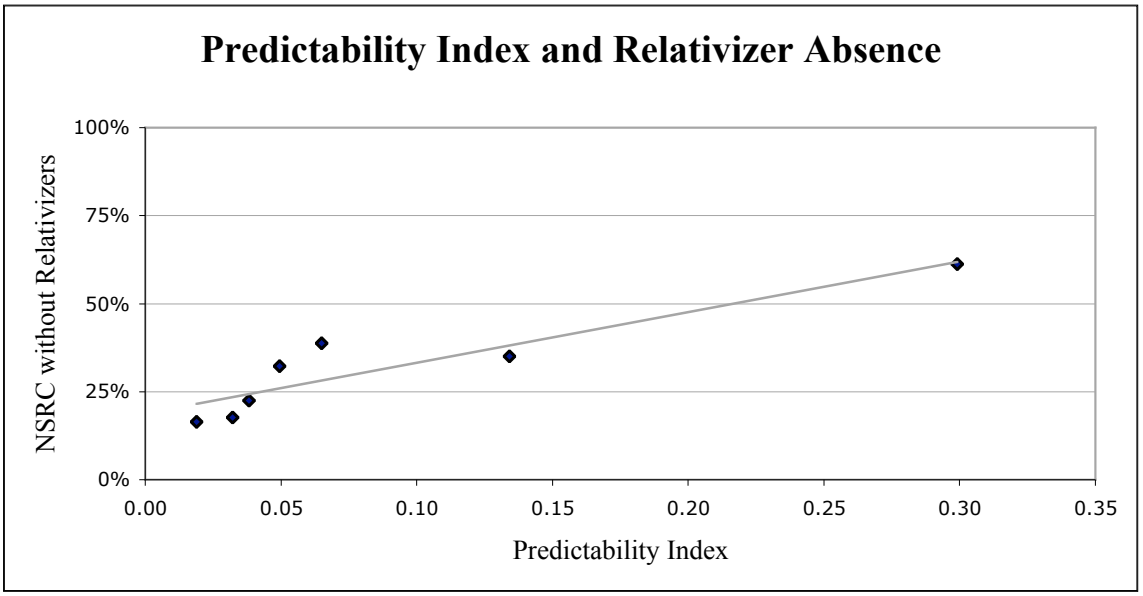

Figure 8
adjusted r^2= .86

$F(1,5)=36.9$, $p=.002$

This result provides more support for the Predictability Hypothesis. Furthermore, the fact that a simple maximum entropy classifier provides reasonable measurements of the predictability of relative clauses suggests that predictability in this sense is learnable. In other words, it is reasonable to assume that speakers have access to estimates of how likely an RC is in a given context.

4. Explaining the Correlation

The Predictability Hypothesis seems to be correct: NSRCs evidently begin with *that* less frequently in environments where an NSRC (or any RC) is more likely to occur. But we have still not answered our original question: Why do different lexical choices correlate with such large differences in relativizer rates? Our answer involves two steps. First, we suggest a processing explanation for the correlation between NSRC predictability and relativizer absence. Second, we argue that there are semantic/pragmatic reasons why certain determiners, head nouns, and adjectives tend to cooccur with NSRCs relatively frequently. Put together, these will constitute an account of why those lexical choices lead to low relativizer rates.

Explaining the presence vs. absence of relativizers in NSRCs in terms of processing can involve considerations of comprehension, production, or a combination of the two. Relativizers could facilitate comprehension by marking the beginning of a relative clause and thereby helping the parser recognize dependencies between the head NP and elements in the NSRC (see Hawkins, 2004, for an account along these lines). Relativizers could facilitate production, e.g. by providing the speaker with extra time to plan the upcoming NSRC (see Race and MacDonald, 2003, for an account along these lines). Both types of explanation predict that relativizers should occur more frequently in more complex NSRCs (though the factors contributing to comprehension complexity and production complexity might not be identical). Teasing apart the predictions of these different kinds of processing explanations is by no means straightforward (see Jaeger, 2005, for much more detailed discussion of this issue).

Whatever kind of processing explanation one adopts, it can be employed to explain why predictability of the NSRC influences relativizer frequency. In a context in which an NSRC has a relatively high probability, the listener gets less useful information from having the beginning of the NSRC explicitly marked. Hence, relativizers do less to facilitate comprehension where NSRCs are predictable. And in environments where NSRCs are likely, speakers would begin planning the NSRC earlier (on average) than in environments where they are less likely. Consequently, they would be less likely to need to buy time by producing a relativizer at the beginning of the NSRC. In short, the correlation between predictability and relativizer absence follows from the hypothesis that relativizers aid processing.

But why do certain lexical choices early in an NP have such a strong effect on the likelihood of there being an NSRC later in the NP? To answer this, it is useful to

consider the semantic function of restrictive relative clauses. As the term "restrictive" implies, such clauses characteristically serve to limit the possible referents of the NPs in which they occur. For example, in (8), the NSRC *that I listen to* restricts the denotation of the NP to a proper subset of music, namely, the music the speaker listens to; without the NSRC, the NP could refer to any or all music.

(8) music that I listen to.

Certain determiners, nouns, and adjectives have semantic properties that make this sort of further restriction very natural or even preferred.

Consider, for example, the determiners *all* and *every*, which express universal quantification. Universal assertions are generally true of only restricted sets[13]. Thus, (9a) is true for many more VPs than (9b).

(9) a. Every linguist we know VP
b. Every linguist VP

More generally, universal assertions are more likely to be true if the quantification is restricted, and NSRCs are one natural way to impose a restriction.[14] Hence, in order to avoid making excessively general claims, people frequently use NSRCs with universal quantifiers.

Notice that the opposite is true for existentials: (10a) is true for many more VPs than (10b), since (10a) is true if VP holds of any linguist, whereas (10b) is true only if it holds of a linguist we know.

(10) a. A linguist VP
b. A linguist we know VP

So while restricting a universally quantified assertion increases its chances of being true, restricting an existentially quantified assertion reduces its chances of being true. Correspondingly, *every* and *all* cooccur with NSRCs relatively frequently (10.40% and 6.92%, respectively), whereas *a(n)* and *some* rarely cooccur with NSRCs (1.18% and 2.10%, respectively).

---

[13] Students in elementary logic classes are taught that sentences beginning with a universal quantifier almost always have a conditional as their main connective. The antecedent of this conditional is needed to restrict the set of entities of which the consequent is claimed to hold. That is, for a sentence of the form $\forall xP(x)$ to be true, P should include some contingencies. In natural language, NSRCs are one way of expressing such contingencies.

[14] Other kind of restrictive modifiers such as subject-extracted relative clauses, prenominal restrictive adjectives, and postnominal PPs are also options. Whenever there is a need to restrict the reference of an NP, each of these options becomes more likely. For the current purpose, it only matters that NSRCs constitute one of these options.

The definite determiner generally signals that the referent of the NP it is introducing is contextually unique – that is, the listener has sufficient information from the linguistic and non-linguistic context to pick out the intended referent uniquely. But picking out a unique referent often requires specifying more information about it than is expressed by a common noun. NSRCs can remedy this: for example, there are many situations in which (11a) but not (11b) can be used to successfully refer to a particular individual.

(11) a. the linguist I told you about
b. the linguist

Even when *the* is used with plural nouns (e.g. *the linguists*) a contextually unique set of individuals is the intended referent. Hence the denotation of the head noun often needs to be restricted, and NSRCs are consequently relatively common.

The pragmatic uniqueness associated with the definite article is very often a result of the fact that the referent of the NP introduced by *the* has recently been mentioned or is otherwise contextually very salient. In these cases, no restriction of the noun phrase is needed, so NSRCs would not be expected. And while *the* cooccurs with NSRCs at about three times the baseline rate for all (nonpronominal) NPs, the vast majority – about 94% – of NPs beginning with *the* have no NSRC.

Certain adjectives, however, involve a uniqueness claim for the referent of NPs in which they appear, and these cooccur with NSRCs at far higher rates[15]. The most frequent of these is *only*; others are superlatives like *first*, *last*, and *ugliest*. Our arguments for the relatively high rate of cooccurrence of *the* with NSRCs applies equally to these adjectives. And since superlatives make sense only with respect to some scale of comparison, the reference set that the scale orders often needs to be explicitly mentioned. Consequently, it is not surprising that these words cooccur with NSRCs at a very high rate. Indeed, we noted in connection with example (6) (following Fox and Thompson, in press) that NPs containing these adjectives sometimes sound incomplete without a modifying relative clause.

The dark bars in Figure 9 show that NPs with the "uniqueness adjectives" *only* and superlatives have far higher rates of cooccurrence with NSRCs than NPs with other adjectives. And, as the Predictability Hypothesis leads us to expect, the same applies to relativizer absence in those NSRCs (see the lighter bars in Figure 9).

[15] This was pointed out by Fox and Thompson (in press). As noted above, it was their discussion of this observation that led us to the Predictability Hypothesis.

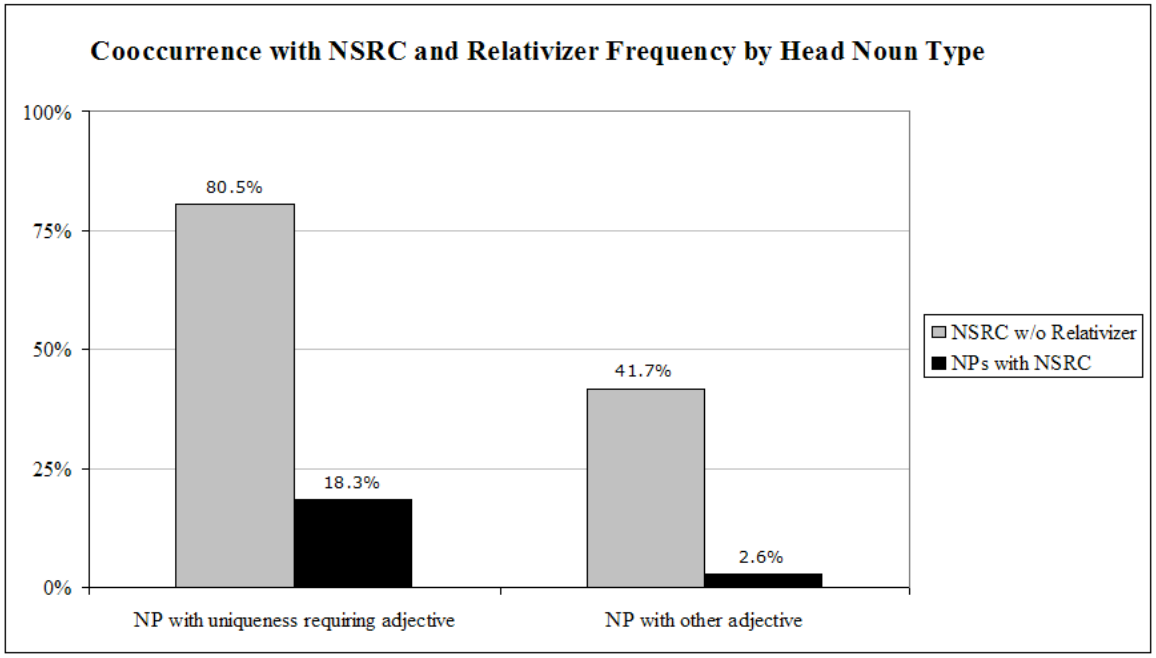

Figure 9

Turning now to the head nouns, one striking fact about the ones that cooccur with NSRCs most frequently is their semantic lightness – that is, nouns like *thing, way, time,* etc. intuitively seem exceptionally non-specific in their reference[16]. Again, there is a semantic/pragmatic explanation for why semantically light nouns would cooccur with NSRCs more than nouns with more specific reference. In order to use these nouns successfully to refer to particular entities, some additional semantic content often needs to be added, and an NSRC is one way of doing this. For example, saying (12a) is less likely to result in successful communication than saying (12b):

(12) a. The thing is broken.
b. The thing you hung by the door is broken.

Testing this intuition requires some basis for designating a noun as semantically light. As a rough first stab, we singled out the non-*wh* counterparts of the question words, *who, what, where, when, how,* and *why*. That is, we looked at how often NSRCs occur in NPs headed by *person/people, thing, place, time, way,* and *reason*, and compared the results to the occurrence of NSRCs in NPs headed by anything else. And, of course, we also compared the frequency of relativizerlessness in those NSRCs. The results, shown in Figure 10, are as we expected, with a far higher percentage of NSRCs in the NPs headed by the light nouns and a far lower percentage of NSRCs introduced by *that*.

[16] This was noticed independently (and first) by Fox and Thompson (in press).

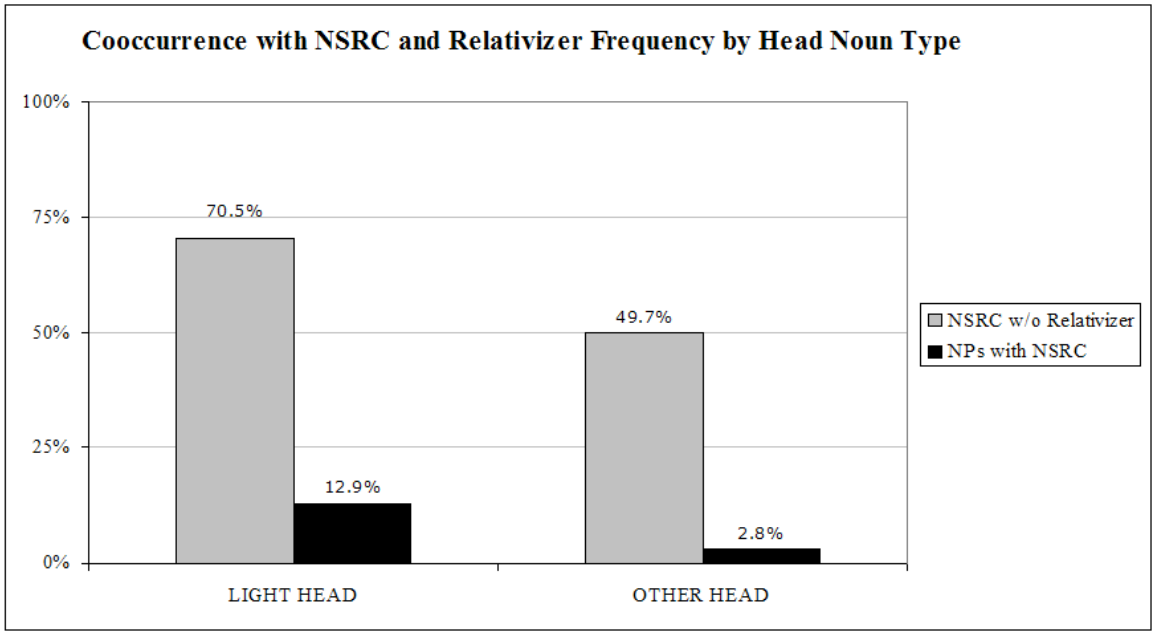

Figure 10

5. Concluding Remarks

Summing up, the variation in relativizer frequency associated with particular lexical choices of determiners, prenominal adjectives, and head nouns in NPs with NSRCs can be explained in terms of two observations. First, whether a word is likely to cooccur with an NSRC depends in part on the semantics of the word and on what people tend to need to refer to. Second, the more predictable an NSRC is, the less useful a relativizer is in utterance processing. Thus, determiners, adjectives, and nouns that increase the likelihood of a following NSRC decrease the likelihood that the NSRCs following them will begin with relativizers.

Our focus has been on how lexical choices influence relativizer frequency. But many non-lexical factors are also known to be relevant. Ideally, a theory of this phenomenon would bring all of these together and explain variation in relativizer use in terms of a single generalization.

One attempt at a unified account of several diverse factors influencing relativizer frequency is Fox and Thompson (in press). They conducted a detailed analysis of a corpus of 195 NSRCs from informal speech, identifying a variety of factors that correlate with relativizer presence or absence. Adapting a suggestion from Jespersen (1933), they argue that their examples fall at different points along "a continuum of monoclausality", with more monoclausal utterances being less likely to have relativizers. Among the factors contributing to monoclausality, in their sense, are semantic emptiness of the clause containing the NP that the NSRC modifies (which subsumes semantic lightness of the head noun), simplicity of the head NP, and shortness of the NSRC.

The idea of a one-dimensional scale combining various factors relevant to relativizer omission has obvious appeal, particularly if it can be characterized precisely. However, we have two reservations about Fox and Thompson's notion of "monoclausality". First, their characterization is rather vague, and they give no independent way of assessing degree of "monoclausality". Second, the terminology is confusing, since even the most "monoclausal" of their examples contain (at least) two clauses, in the sense that they have two verbs and two subjects. Nevertheless, we share the intuition that the contents of the two clauses in the more "monoclausal" examples are more closely connected.

We believe that the notion of predictability might provide a precisely definable scale that can do the work of Fox and Thompson's "monoclausality". Predictability has the further advantages that its influence on relativizer absence can be explained in processing terms and that it is often possible to explain why some NSRCs are more predictable than others, as we did above.

Some of the utterances Fox and Thompson consider the most monoclausal are stock phrases or frequently used patterns (e.g. *the way it is*), which they suggest may be stored as units. Stock phrases are by definition highly predictable, so they fit well with our account. Some higher-level grammatical patterns[17] might not be covered by a simple, lexically-based characterization of predictability like the ones we employed. If so, it would suggest that more sophisticated metrics of predictability should be explored. In short, the Predictability Hypothesis of relativizer variation provides testable questions for future research. Next we briefly mention some of them.

First, we believe it is important to investigate what information speakers use to determine the predictability of an NSRC. For examples, does the grammatical function of the modified NP matter? Or do speakers only use 'local' information to predict NSRCs (i.e. lexical properties of the NP).[18] More specifically it will be relevant for our understanding of predictability to see whether the factors investigated in this paper interact. In other words, do speakers use simple heuristic like the association of a particular lexical item with the likelihood of an NSRC, or do speakers compute the overall predictability of an NSRC given the combination of lexical items in the modified NP? A further question that deserves attention is whether speakers use some sources of information more than others to compute the predictability of a construction (here: NSRCs). As we have seen in Section 3 predictability information related to determiners seems to correlate much more strongly with the relativizer absence than information related to adjectives and the head noun of the modified NP. This may simply be due to the larger sample size available for the estimation of the mean for each of the words. But it is also possible that probability

[17] We know of no clear cases of such patterns that don't have any identifying lexical items associated with them. One possible one is *the* X-*er* S_1, *the* Y-*er* S_2, as in *The bigger they are, the harder they fall*. But it is not clear that the two Ss (*they are* and *they fall*) should be analyzed as relative clauses here.

[18] In this context, it is interesting that research on the effect of predictability on phonetic reduction (e.g., Bell, et al., 2003) finds that the best measures of predictability are also the most local (i.e. bigrams).

distributions for closed class items (like determiners) are easier to acquire or are more efficient to use, since there are fewer items in those classes. We hope future research will discover generalizations that go beyond the particular phenomenon discussed here. Ongoing research that addresses some of the above issues and investigates a related phenomenon, complementizer omission, is presented in Jaeger, Levy, Wasow, & Orr (2005).

Finally, let us return to the theme of this volume: exceptions. We have shown that the notion of exception can be generalized from hard (categorical) to soft (probabilistic) rules. We explored some soft exceptions to the optionality of relativizers in NSRC, ultimately concluding that they could be explained in terms of the interaction of the semantics of the "exceptional" words, the pragmatics of referring, and processing considerations.

Those who question the use of gradient models in syntax might suggest that this illustrates an important difference between hard and soft generalizations, namely, that the latter reflect facts about linguistic performance, not competence, and will hence always be explainable in terms of extra-grammatical factors, like efficiency of communication. In contrast, they might argue, many categorical generalizations are reflections of linguistic competence, and hard exceptions to them may be as well.

We would respond that it is always preferable to find external explanations that tie properties of language structure to the functions of language and to characteristics of language users. Such explanations should be sought for both hard and soft exceptions. We know of no reason to believe that they will always be possible for the soft cases, but not the hard cases.

References

Alan Bell, Daniel Jurafsky, Eric Fosler-Lussier, Cynthia Girand, Michelle Gregory, and Daniel Gildea. 2003. Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *Journal of the Acoustical Society of America* **113** (2), 1001-1024.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina, and Harald Baayen (2005) "." Royal Netherlands Academy of Science Workshop on Foundations of Interpretation.

Bresnan, Joan and Tatiana Nikitina (2003) "On the Gradience of the Dative Alternation". Available at http://www-lfg.stanford.edu/bresnan/download.html.

Chomsky, Noam (1955/75) *The Logical Structure of Linguistic Theory*. Chicago: University of Chicago Press.

Chomsky, Noam (1957) *Syntactic Structures*. The Hague: Mouton.

Chomsky, Noam (1962) "A Transformational Approach to Syntax". *Third Texas Conference on Problems of Linguistic Analysis in English*, 124-169. Austin: The University of Texas.

Chomsky, Noam (1966) *Topics in the Theory of Generative Grammar*. The Hague: Mouton.

Fowler, H. W. (1944) *A Dictionary of Modern English Usage*. Oxford: Oxford University Press.

Fox, Barbara A., and Sandra A. Thompson.(in press) "Relative Clauses in English conversation: Relativizers, Frequency and the notion of Construction". To appear in *Studies in Language*.

Hawkins, John A. (2004) *Efficiency and Complexity in Grammars*. Oxford: Oxford University Press.

Huddleston, Rodney and Geoffrey K. Pullum (2002) *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Jaeger, T. Florian (2005) "Optional *that* indicates production difficulty: Evidence from disfluencies". Workshop on Disfluencies in Spontaneous Speech. Aix-en-Provence.

Jaeger, T. Florian, Roger Levy, Thomas Wasow, and David Orr (2005) "The Absence of 'that' is Predictable if a Relative Clause is Predictable". Architectures and Mechanisms of Language Processing conference. Ghent, Belgium.

Jaeger, T. Florian, David Orr, and Thomas Wasow (2005) "Comparing and combining frequency-based and locality-based accounts of complexity". Poster presented at the 18th CUNY Sentence Processing Conference. Tucson, Arizona.

Jaeger, T. Florian and Thomas Wasow (in press) "Processing as a Source of Accessibility Effects on Variation". *Proceedings of the 31st meeting of the Berkeley Linguistics Society*.

Jespersen, Otto (1933) *Essentials of English Grammar*. London: Allen & Unwin.

Marcus, Mitchell P., Beatrice Santorini, Mary Ann Marcinkiewicz, and Ann Taylor (1999) *Treebank III*. Linguistic Data Consortium, University of Pennsylvania.

Race, David and Maryellen MacDonald (2003) "The use of 'that' in the production and comprehension of object relative clauses." 26th Annual Meeting of the Cognitive Science Society.

Ratnaparkhi, Adwait (1997) "A Simple Introduction to Maximum Entropy Models for Natural Language Processing". Technical Report 97-08, Institute for Research in Cognitive Science, University of Pennsylvania.

Ross, John R. (1967) *Constraints on Variables in Syntax*. MIT Dissertation.

Wasow, Thomas (2002) *Postverbal Behavior*. Stanford: CSLI Publications.