

Lexical Variation in Relativizer Frequency

*Thomas Wasow, T. Florian Jaeger, David M. Orr**

Abstract

An exception to a non-categorical generalization consists of a lexical item that exhibits the general pattern at a rate radically different – either far higher or far lower – from the norm. Lexical differences in noun phrases containing non-subject relative clauses (NSRCs) correlate with large differences in the likelihood that the NSRC will begin with *that*. In particular, the choices of determiner, head noun, and prenominal adjective in an NP containing an NSRC may dramatically raise or lower rates of *that* in the NSRC. These lexical variations can be partially explained in terms of predictability: more predictable NSRCs are less likely to begin with *that*. This generalization can be plausibly explained in terms of processing, assuming *that* facilitates processing and/or signals difficulty. The correlations between lexical choices in the NP and the predictability of an NSRC can, in turn, be explained in terms of the semantics of the lexical items and the pragmatics of reference.

0. Introduction

The notion of exception presupposes that of rule; as Webster (<http://www.m-w.com/dictionary>) 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: 145-146, 1957: 16-17, 1962: 128, 1966, 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 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.ⁱ

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 exceptions 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 about 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:ⁱⁱ

- (2) I saw a movie *(that) offended me.ⁱⁱⁱ

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.^{iv}

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, and Wasow 2005, Jaeger 2005), exploring interactions 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 pronominal 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 Table 1, where the numbers in parentheses indicate the total number of examples.

Table 1: NSRC *that* Rate by NP Determiner

| DETERMINER (FREQUENCY) | NSRC WITH <i>THAT</i> |
|---|-----------------------|
| <i>a</i> or <i>an</i> (302) | 74.8% |
| Possessive pronoun (37) | 64.9% |
| <i>some</i> (67) | 64.2% |
| No determiner (428) | 63.1% |
| <i>this</i> , <i>that</i> , <i>these</i> , <i>those</i> (106) | 61.3% |
| Numeral (177) | 53.1% |
| <i>any</i> (55) | 49.1% |
| <i>no</i> (34) | 38.2% |
| <i>the</i> (1813) | 34.2% |
| <i>all</i> (206) | 24.3% |
| <i>every</i> (68) | 14.7% |

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.

Table 2: NSRC *that* Rate by NP Head Noun

| HEAD NOUN (FREQUENCY) | NSRC WITH <i>THAT</i> |
|------------------------|-----------------------|
| <i>stuff</i> (46) | 62.8% |
| <i>people</i> (64) | 57.1% |
| <i>one</i> (106) | 51.5% |
| <i>problem</i> (44) | 50.0% |
| <i>something</i> (171) | 44.7% |
| <i>thing</i> (523) | 43.7% |
| <i>kind</i> (49) | 43.2% |
| <i>anything</i> (48) | 38.0% |
| <i>place</i> (99) | 34.4% |
| <i>everything</i> (60) | 24.6% |
| <i>reason</i> (91) | 24.0% |
| <i>time</i> (247) | 14.0% |
| <i>way</i> (325) | 13.0% |

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.

Table 3: NSRC *that* Rate by Prenominal Adjective

| ADJECTIVE (FREQUENCY) | NSRC WITH <i>THAT</i> |
|-----------------------|-----------------------|
| <i>little</i> (41) | 73.2% |
| <i>certain</i> (19) | 68.4% |
| <i>few</i> (20) | 65.0% |
| <i>different</i> (19) | 63.2% |
| <i>big</i> (15) | 60.0% |
| <i>other</i> (87) | 49.4% |
| <i>same</i> (47) | 46.8% |
| <i>best</i> (24) | 25.0% |
| <i>only</i> (158) | 24.7% |
| <i>first</i> (99) | 18.2% |
| <i>last</i> (79) | 8.9% |

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; chi-square tests on all three of these distributions are highly significant.

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 soft exceptions to the optionality of *that* in NSRCs. That is, why do the presence of words like *a(n)*, *every*, *stuff*, *way*, *little*, and *last* correlate with exceptionally high or low rates of *that* in NSRCs that follow them within an NP?

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.^v

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

Moreover, the sentence seems incomplete without the relative clause:

(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 1-3 plot cooccurrence with NSRCs against frequency of relativizer absence in NSRCs. The points in Figure 1 represent the eleven determiner types given in Table 1; the points in Figure 2 represent the thirteen head nouns given in Table 2; and the points in Figure 3 represent the eleven adjectives given in Table 3.^{vi} The lines represent linear regressions – that is, the lines represent the best (linear) generalization over the data points in that the total squared distance between the points and the lines is minimized (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 is also significant (adjusted $r^2=.36$, $F(1,33)=19.9$, $p<.001$).^{vii}

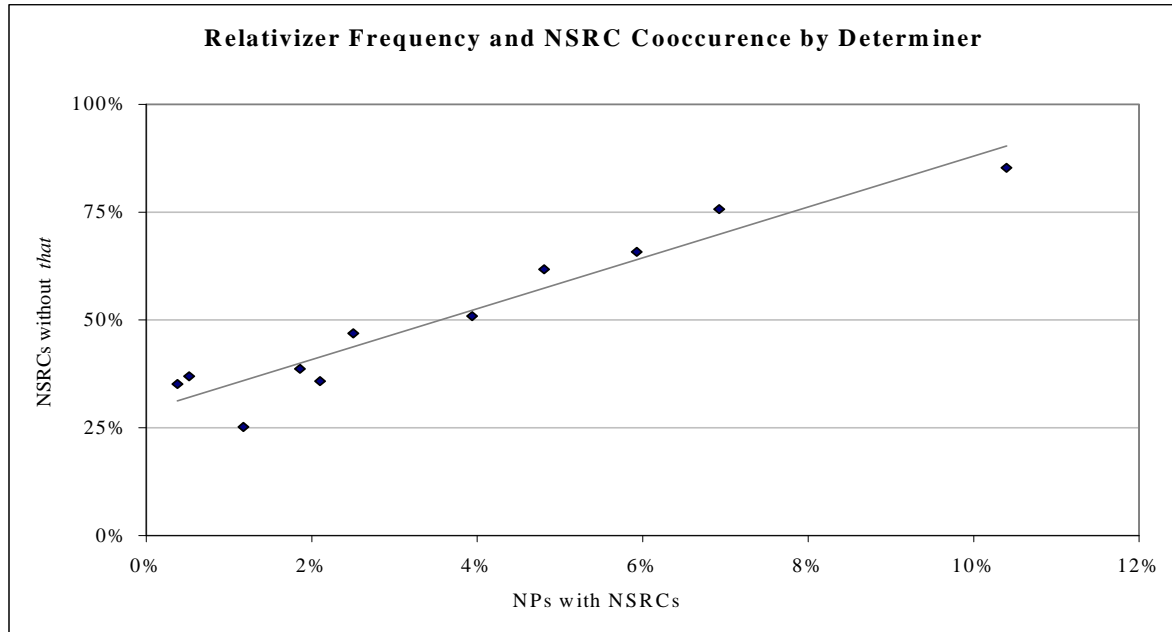


Figure 1: Relativizer Frequency and NSRC Cooccurrence by Determiner
 adjusted $r^2 = .91^{\text{viii}}$
 $F(1,9) = 105.1, p < .001$

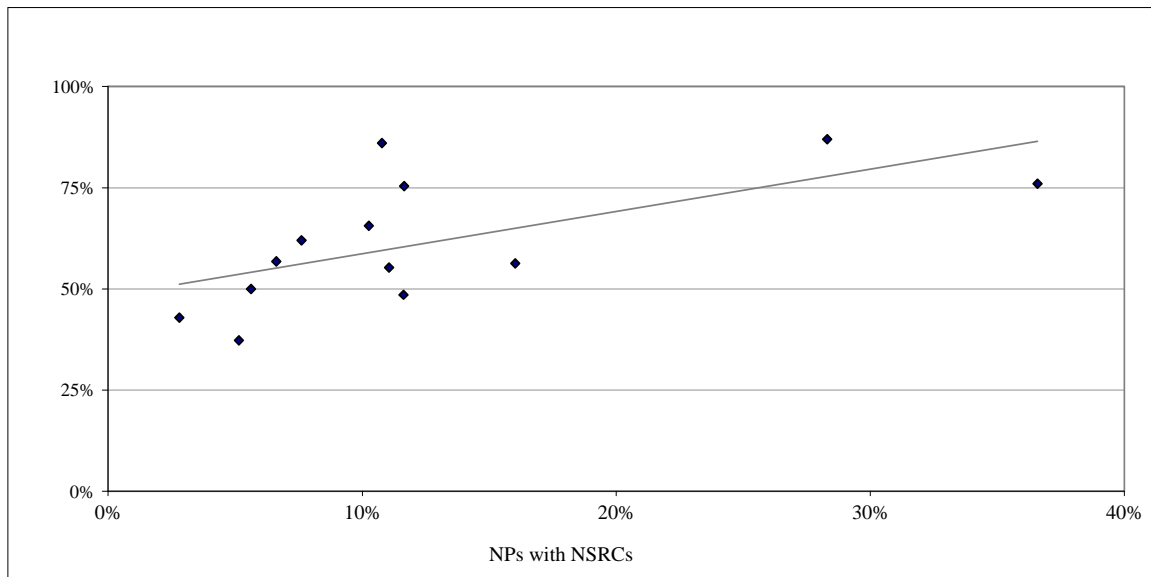


Figure 2: Relativizer Frequency and NSRC Cooccurrence by Head Noun
 adjusted $r^2 = .35$
 $F(1,11) = 7.4, p = .02$

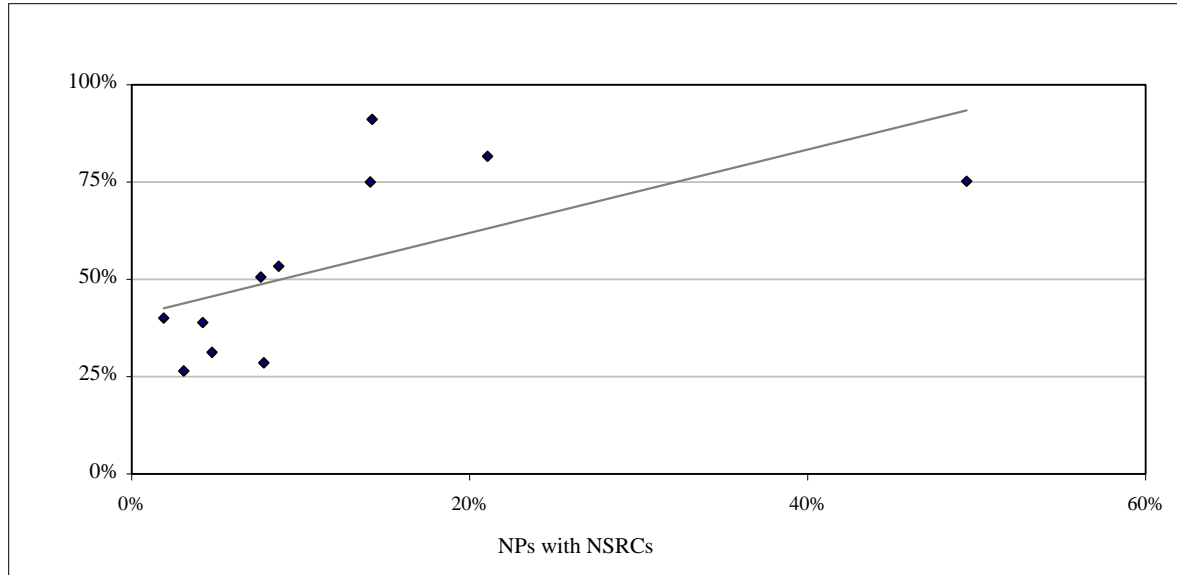


Figure 3: Relativizer Frequency and NSRC Cooccurrence by Adjectives

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.^{ix} 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 due to lexical properties of NPs (such as their grammatical function), but the study we report on here is limited to lexical factors.^x

We created a maximum entropy classifier (see Ratnaparkhi, 1997), which used features of an NP to predict how likely a relative clause^{xi} 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.^{xii}

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. That is, we compared the fraction of the NPs in each group that contained an RC with the range of predictability indices the group represented. As can be seen in Figure 4, 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 predictability indices that the classifier was assigning to the NPs were generally reasonable estimates.

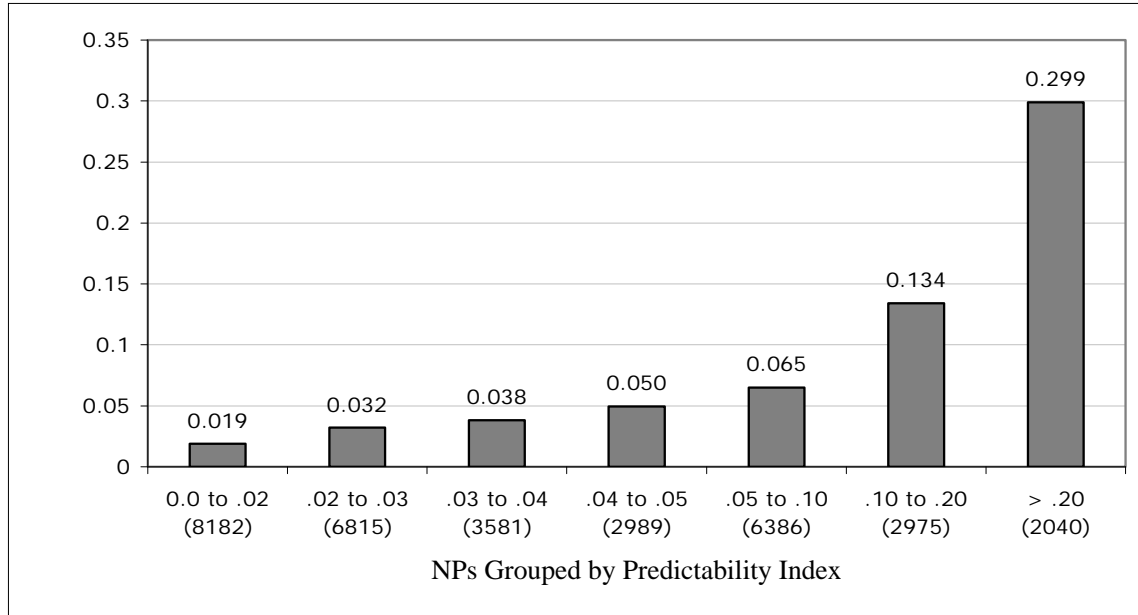


Figure 4: Accuracy of Classifier

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 examining the rates of relativizer absence for each of our groupings of NPs. As Figure 5 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.

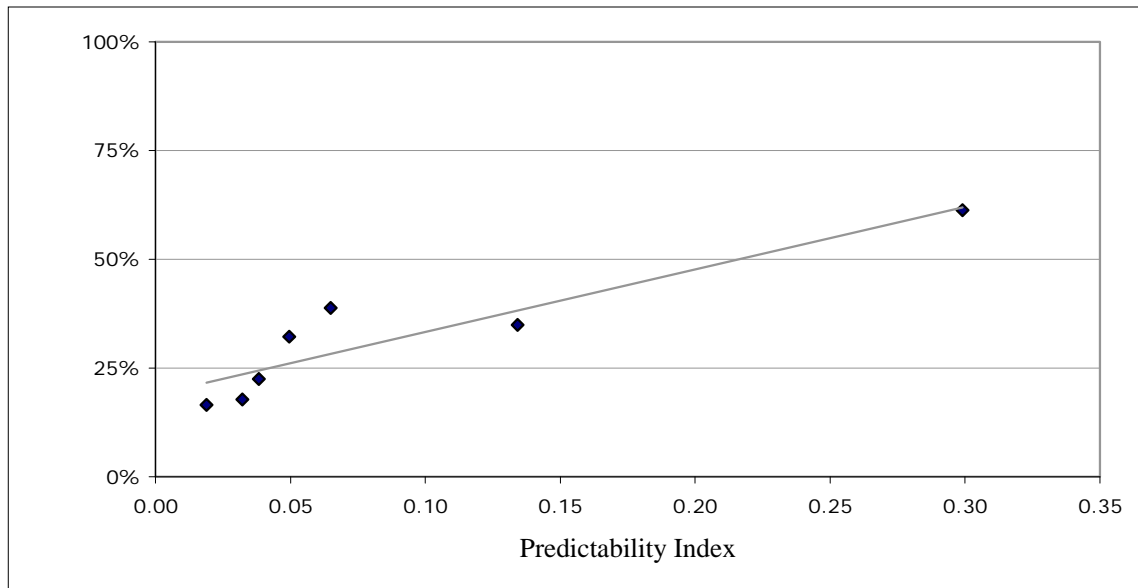


Figure 5: Predictability Index and Relativizer Absence
 adjusted $r^2 = .86$
 $F(1,5) = 36.9$, $p = .002$

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 can be computed by means of a standard machine-learning method. Hence, 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^{xiii}. 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.^{xiv} 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).

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^{xv}. 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 6 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).

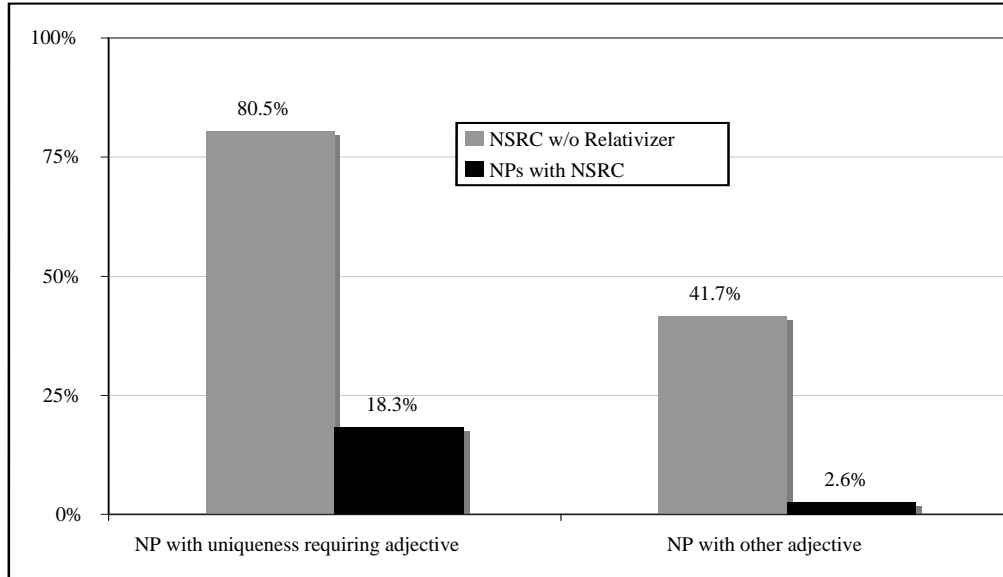


Figure 6: Cooccurrence with NSRC and Relativizer Frequency by Adjective Type

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^{xvi}. 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 relativizers in those NSRCs. The results, shown in Figure 7, 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*.

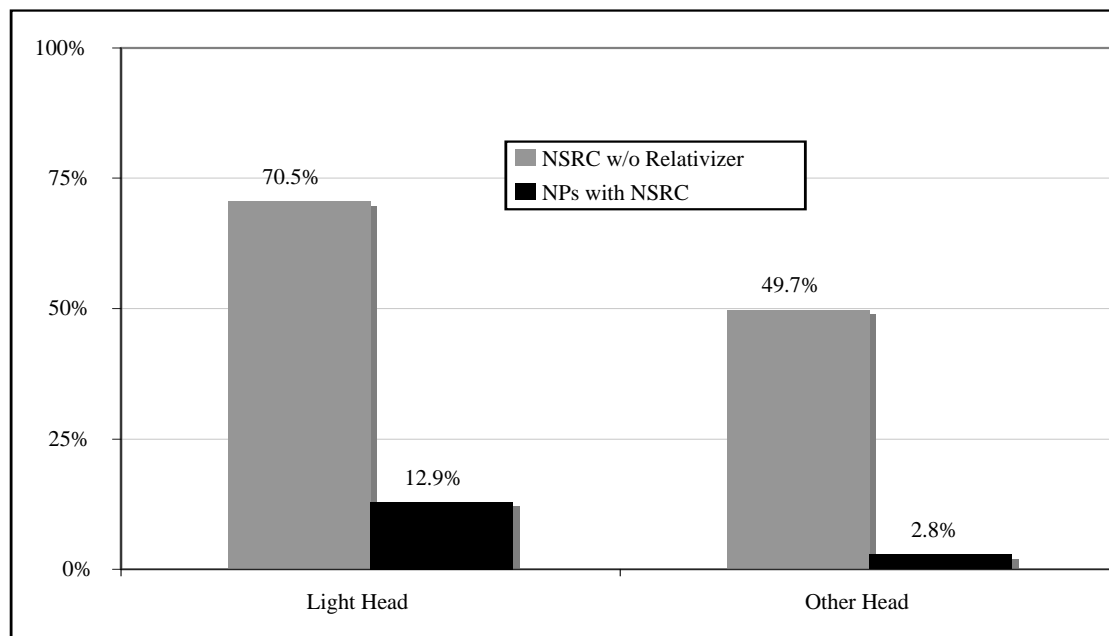


Figure 7: Cooccurrence with NSRC and Relativizer Frequency by Head Noun Type

5. Concluding Remarks

Summing up, the variation in relativizer frequency associated with particular lexical choices of determiners, pronominal 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^{xvii} 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).^{xviii} 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 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, et al (2005) and Jaeger (2006).

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. There is no basis for bifurcating linguistic phenomena a priori into those that are and those that are not amenable to external explanation. In particular, 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.

Notes

* This paper is dedicated to Professor Günter Rohdenburg of Paderborn University, whose sixty-fifth birthday coincided with the completion of the first draft 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. Special thanks also to Heike Wiese and Horst Simon, first for organizing the workshop at which this material was originally presented, and for comments on the written version.

ⁱ 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).

ⁱⁱ 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...*

ⁱⁱⁱ An asterisk outside parentheses is used to indicate that the material inside the parentheses is obligatory.

^{iv} The studies were replicated including the NSRCs with *wh*-relativizers. The results are qualitatively the same, though the numbers are of course different.

^v 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.

^{vi} The mean plots in the three figures represent rather different sample sizes. Determiners are a closed class, so Figure 1 includes almost all NSRCs, whereas Figures 2 and 3 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 3 is far lower than in Figure 2

^{vii} After removing two extreme outliers, the adjusted $r^2 = .56$, $F(1,31) = 36.1$, $p < 0.001$.

^{viii} Adjusted r^2 s provide a more reliable measure of the goodness of fit of the model compared to normal, unadjusted r^2 s, which usually are too optimistic. Generally, r^2 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.

^{ix} Furthermore, we used means to predict means (i.e. we used the mean predictability 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^2 s (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.

^x Studies involving non-lexical factors are in progress.

^{xi} This study differs from the earlier ones in that it 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.

^{xii} 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).

^{xiii} 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.

^{xiv} Other kinds 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.

^{xv} 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.

^{xvi} This was noticed independently (and first) by Fox and Thompson (in press).

^{xvii} 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 S₁, the Y-er S₂*, 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.

^{xviii} 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).

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 "<http://www-lfg.stanford.edu/bresnan/qs-submit.pdf>" 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.
1957 *Syntactic Structures*. The Hague: Mouton.
1962 "A Transformational Approach to Syntax". *Third Texas Conference on Problems of Linguistic Analysis in English*, 124-169. Austin: The University of Texas.
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.
2006 *Probabilistic Syntactic Production: Expectedness and Syntactic Reduction in Spontaneous Speech*. Stanford University dissertation.
- 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.

Corpus evidence and the role of probability estimates in processing decisions

Ruth Kempson
King's College London

January 25, 2007

Wasow, Jaeger and Orr (WJO) address the phenomenon of exceptions from a background of increasing interest in models of language where generalizations about natural languages are made on the basis of probabilistic generalizations, rather than on categorical distinctions.¹ What they provide is a case for a concept of gradient exceptionality, expressed in terms of what is unlikely to occur - the other side of the coin from what does occur with high predictability. The example is the correlation between the predictability of a given determiner, or adjective, or noun occurring with a relative clause and the likelihood of that relative occurring without a relativizer: an expression which is likely to occur with a relative is unlikely to occur with a relativizer. In this demonstration and the consequences they draw from it, the window of focus is deliberately narrow, with subject relatives, relatives with a *wh* relativizer, non-finite relatives, and pied-piping constructions all left on one side from the corpus cull they make, as displaying different idiosyncracies which detract from the primary issue of what makes the relativizer preferred or dispreferred when it is essentially optional.

It is a little disappointing that the variety of relative-clause types considered is so narrow, since the distinction between restrictive and nonrestrictive relatives, one of the primary features supposedly distinguishing *that*- and *wh*- marked relatives in English is, as they note, not clearcut. Relatives with *that* exceptionally allow nonrestrictive construals, particularly if they occur second in a sequence of relatives :

- (1) There was that man at the party that you had introduced me to, that annoyed me enormously by his pompous posturing.
- (2) Last week I bought this game pie for the party, that went bad on me before the end of the week.
- (3) I am thinking of buying a piece of land, that I hope you like.

So, of the finite relative clauses, it is only relativizer-less relatives which REQUIRE restrictive construal.

Despite the restrictions on their corpus cull, WJO provide what are nonetheless fascinating tables displaying how individual determiners vary in their likelihood of co-occurrence with a *that* relativizer, and adjectives, and also nouns, with in the determiner class the indefinite *a* being the determiner to occur most frequently with an accompanying *that* relativizer, in the noun class it is the indefinite *stuff* which is the most common, *way* that is least common;

¹I am grateful to Jieun Kiaer, Nancy Kula and Lutz Marten for comments on this note, and the issues which the WJO paper raise.

and in the adjective class it is the adjective *little* that strikingly comes highest in the list, being an average of over eight times more likely to have a relativiser than nonsubject relatives modifying NPs with the adjective *last*. Some of these distributions seem more puzzling than others; however in all cases, as WJO demonstrate, there is a regular correlation between predictability in the corpus of the particular word being associated with a relative, overall phrasal predictability of relative clause modification, and predictability of the relativiser. On these gradience lists, the quantifiers present perhaps the least obvious distribution, with *a* at the top of the list with the highest proportion of relative clauses with *that*, with *some* coming lower in the list, but nevertheless twenty-five percent more than the numerals to occur with the relativiser. On the other hand, *every* and *all* come out bottom, with *any* displaying three times the proportion of *that*-specific relative clauses than *every*, and double that of *all*. This makes any simple-minded account of quantification based exclusively on quantificational properties seem unlikely. Based on these differential probabilities of occurrence with relative clause, W et al provide a measure of the cumulative predictability of relative clause construal of the determiner-adjective-noun sequences so collected; and from this basis, they pose the claim central to their paper, the so-called Predictability Hypothesis: the more predictable a non-subject relative, the less frequent is its co-occurrence with the relativiser *that*. Whatever the surprises there may be in the probability estimates associated with individual words, this is an intuitive result; and it is extremely good to see this properly quantitatively confirmed, buttressing what is otherwise no more than an intuition.

The question, then, is why there should be such a strong correlation between predictability and lack of relativiser? And this is where the interest of probability-based results arises: should such correlations be explicable solely in terms of the interaction of other pragmatic, semantic, or processing-oriented considerations – with probability assessments themselves playing no part in the explanation; or does the predictability itself have a role to play? The starting point for the analysis which W et al provide is their conjecture that because the “maximal entropy classification” which they used to provide the accumulated measurements of the predictability of relative clauses can be computed by standard machine-learning methods, it is plausible to assume that “speakers have access to estimates of how likely a relative clause is in a given context.” They go on from there to explore the composite effect of parsing and or production considerations in conjunction with the intrinsic content of the various determiners/adjectives/nouns, and from there consider how these might in part explain the probability distributions. One such factor is that both determiners and nouns which allow anaphoric, context-dependent interpretations will not need a relative clause modifier whenever they can be so identified. The other is that those nouns which are semantically “light” but do not allow anaphoric forms of construal almost must occur with a relative clause modifier. The processing explanation they offer is that in such cases, the presence of the relativizer has relatively low functional load. In this connection, one factor which might influence processing considerations over and above occasion-specific functional-load considerations is the effect of routinization. That is, where there is common co-occurrence of determiner/adjunctive/noun and the presence or absence of the relativizer, eg in the predictability of *way* and lack of relativizer and predictability of the indefinite article *a* and presence of the relativizer, such co-occurrence might become stored as a routinized strategy associated with that particular item, thereby accentuating the frequency distribution results for cases at either end of the continuum (see Kempson and Cann 2006 for arguments that routinization is a force in syntactic change).

WJO note with approval the Fox and Thompson observation of ‘monoclausality’ of relativizer-less relatives, but without exploring any semantic analogue to this, they suggest

that their account in terms of predictability might take the place of this “rather vague ” mono-clausal notion; and they proceed to set out explanations that might confirm such a stance. However, this move is too swift. Rather than simply seeking to replace this observation altogether, the authors might have considered the semantic analogue to the Fox and Thompson observation. This is that relatives can be used either to build up a complex restrictor for a quantifying expression within a single clause, i.e. a restrictive relative clause construal, or, conversely, they can be used to provide an adjunct, independent structure, a nonrestrictive relative clause construal. Indeed, whatever the difficulties of formally characterising nonrestrictive relative clause construal (see Potts 2004 for a re-analysis in terms of ‘supplements’ for which he gives a conventional-implicature analysis), it is not in question that, unlike in restrictive relative clause construals, the two clauses give rise to two independent propositions, and in some analyses, the distinctiveness of the two is made explicit (Potts 2004, Kempson et al 2001, Cann et al 2005).² This distinction is often reported to be only disambiguated by intonation. In writing, where relativizers may play the role of defining a clausal edge but cannot disambiguate between restrictive and nonrestrictive construals, it is only the lack of any such indicator that can unambiguously indicate a restrictive construal. The “monoclausal” observation of Fox and Thompson thus has a natural counterpart in a semantic characterisation of relative clause construal: relativizer-less relatives uniquely identify a single overall assertion, a distinctive attribute of restrictive relative-clause construal which has independent syntactic and semantic motivation. If, then, a speaker is planning a relative clause sequence indicating a restrictive construal, they may not even consider the possibility of using a form which would allow the alternative form of construal: certainly the most secure way of ensuring the appropriate construal is to select a form which precludes it, and of the finite forms, only the relativizer-less form definitively does so. Hence the more likely the form is to be associated with a restrictive form of construal, the less likely it is to be introduced with a form which allows for any other form of construal. By comparison, the move from the demonstration of the statistical correlation between probability of a relative-clause and inverse probability of the relativizer, to the assumption that calculations of probability might drive production decisions, is a leap which needs substantial independent argumentation. At the very least, there is a well-motivated alternative to be aired.

There is in any case linguistic evidence from other languages which tends to favour the explanation of relativizer-less relatives in terms of definitively indicating the singleton status of the over-all propositional structure. In some languages boundary marking of structure can be made by tone. One such is Bemba. Bemba is a tone language which marks relative clauses in one of two ways, by tone or by pronominal marking. Relative clauses marked by tone alone are exclusively associated with restrictive construal and have to coincide with what is called the conjoint form of the verb, the low tone of the conjoint verb-form determining that the noun head and verb initiating the relative clause will be processed as a single prosodic unit. In consequence the construal of the relative as an integral part of the containing structure is unambiguously indicated. Relative clauses involving pronominal marking, being morphologically marked, can be construed restrictively or nonrestrictively, these construals being distinguished by use of the conjoint form (low tone) and the disjoint verb (high tone).³ The

²Some authors have argued that nonrestrictive relatives are presuppositional, but there are many examples to the contrary:

(i) John ignored Mary, who burst into tears.

³There are differences between object and subject marking, with morphological marking of object relatives taking several forms but with restrictions on the availability of the tonal strategy. However, all that is relevant here is that the low-tone strategy, which is the conjoint form of the verb, is invariably associated with a

striking aspect of the two strategies, tonal vs pronominal, is that they do not distribute in a complementary fashion. Rather, just one of those strategies provides unambiguous indication that the producer is continuing immediately with construction of a complex restrictor, i.e. a restrictive relative clause. Thus it is the use of the conjoint verb form with its low tone that forces restrictive construal in Bemba, analogous to the relativizer-less relatives of English. Such parallels from analysis of one language to another have to be treated with some caution, of course. In principle, the correlation between Bemba tone and morphologically explicit relative-clause marking might well be characterisable in terms of probability of co-occurrence. Nevertheless, the explanation of such conjoint low tone in terms of phonological indication of the mode of compositionality seems much more consistent with orthodox assumptions about how to explain encoded properties of natural language (see Cheng and Khula 2006 for independent arguments of the feeding relation between phonological marking and Bemba relative-clause structure). And this, by analogy, favours the explanation of the distribution of relativizerless relatives in English in terms of their unambiguous correspondence with restrictive relative clause construal.

WJO are careful to keep the Predictability Hypothesis as a claim restricted only to relative clauses of a particular type, and applied only to English. However, they end by asking questions of a much more general nature that presume the relevance of predictability weightings in the making of speakers' decisions. They ask how do English speakers determine the predictability of a non-subject relative clause; and do the speakers compute over-all predictabilities or do they rather manipulate locally available heuristics of particular items? Further questions might be whether there are speed-up or conversely lengthening phenomena associated with presence or absence of complementizer choice. Are there also any correlations between how many average words follow after each determiner and whether this affects the occurrence of the relativizer? There are also more general questions. Predictability correlations are string-based observations and not category-specific, and one might expect that if they can be manipulated constructively by speakers, they should provide a basis for explaining distributions in other cases where two options are apparently equally available. This raises fascinating new research questions. Is it the case, by analogy with these cases, that in cases where two alternative forms are possible, but one much more probable than the other, that the morphologically more marked form is less likely to be chosen, being unnecessary? One such case is structural vs prosodic indication of question-hood. Questions, incidentally like nonrestrictive relatives, are invariably marked by intonation, a para-linguistic marking characteristically recognisable early on in a parse sequence. By analogy, if such prosodic form is so reliably associated with question construal, one might expect that speakers of a language might deem it inessential to provide morphologically explicit forms of interrogative; and indeed in many languages they commonly do indeed use declarative rather than interrogative forms, relying solely on the prosody. However, as the authors are well aware, the relevance of probability results has to be treated with caution: probability of occurrence cannot in general be a guide as to whether or not a simpler form will be used. Take the case of approaching an information desk in an airport. The speaker has two ways of asking a yes-no question, either the declarative form (without auxiliary) or an inverted form with an auxiliary. Does the very fact that you are highly likely to be construed by the person at the desk as asking a question influence your decision to present it in one form rather than another, with a tendency to choose the simpler declarative form? One might seek empirical

restrictive construal (see Cheng and Kula 2006 for details). These observations are due to Nancy Kula; and I am grateful to both her and Lutz Marten for discussing these data with me and reminding me of their relevance to this issue.

test of this prediction, but intuition would surely suggest the answer is “No”.

The moral to be drawn from this fascinating setting out of data and probability assignments thus seems to be two-fold. It is clear on the one hand that probability distributions over corpus evidence, if reliably replicable, provide fascinating new data which anyone facing up to the challenge of articulating grammar interfaces will be interested in mulling over. On the other hand, the conclusion that speakers manipulate probability estimates as input to the decisions as to how to say what they do would seem to be as yet premature. While there are clear probabilistic distributions to be culled from language data to great effect, providing new impetus for theoretical explanations of a subtlety most frameworks do not make provision for, it remains far from obvious that probabilistic distributions constitute part of the explanation. The test of such putative explanations will be their generalisability to explain optional distributions on a broad cross-linguistic basis.

References:

- Cann, R. Kempson, R. and Marten, L. 2005. *The Dynamics of Language*. Oxford: Elsevier.
- Cheng, L. and Kula, N. 2006. Syntactic and phonological phrasing in Bemba relatives. ZASP Papers in Bantu Linguistics.
- Fox, B. and Thompson, S. A. forthcoming. Relative clauses in English conversation: relativizers, frequency and the notion of construction. *Studies in Language*.
- Potts, C. 2002 *The Logic of Conventional Implicatures*. Oxford: Oxford University Press.

Response to Kempson's Comments*

Thomas Wasow, Stanford University
T. Florian Jaeger, University of Rochester
David Orr

Kempson's interesting commentary raises two important points. First, while extolling the value of probabilistic corpus data, she is not ready to accept "that speakers manipulate probability estimates as input to the decisions as to how to say what they do". Second, she suggests an alternative to our attempt to explain the correlation between predictability of non-subject relative clauses and the absence of *that* in such clauses. We discuss these points in reverse order and raise some additional questions for future research.

Our proposed explanation of the correlation, which is admittedly somewhat programmatic, is that more predictable NSRCs are easier to produce and/or comprehend than less predictable ones, and hence do not need the extra function word. Kempson's alternative explanation is based on the fact that relative clauses without a relativizer must be interpreted as restrictive, whereas non-restrictive construals are often possible when a relativizer is present. She suggests that relativizer omission is used as a way of disambiguating the intended construal of the relative clause.

She points out that another method of disambiguation can be intonation. Since intonation is not marked in writing, her reasoning predicts that relativizer omission should be more common in writing than in speech. As we first noted in Jaeger & Wasow (2005), this does indeed seem to be the case. NSRCs in the parsed portions of the Wall Street Journal (WSJ) and the Brown corpus (BC) are significantly less likely to have a *that* relativizer (24% and 11%, respectively) than NSRCs in the parsed Switchboard corpus (SWBD: 43%; $\chi^2=453.0$, $p < 0.0001$). This difference decreases but prevails even when all relativizer types are counted (WSJ: 47%, BC: 36%, SWBD: 52%; $\chi^2=79.8$, $p < 0.0001$) and after other factors influencing *that* are controlled for (see also Fox & Thompson, 2006, who report 60% relativizer rate for NSRCs in informal conversations). Note in particular that NSRCs in the two written corpora are on average 21-36% longer than NSRCs in the Switchboard. A priori, this would suggest the opposite of the observed pattern since longer NSRCs are more likely to contain a relativizer (Race & MacDonald, 2003; Jaeger, 2006). The observed distributional differences between speech and written texts hence are in line with Kempson's hypothesis (see Jaeger & Wasow, 2005, for an alternative explanation based on the hypothesis that relativizer mentioning is driven by production pressures).

As intriguing as it is, Kempson's ambiguity avoidance hypothesis leads to a prediction that is inconsistent with the data discussed in our paper. The problem with the ambiguity account is related to the link between predictability and restrictiveness. Kempson does not discuss this link. The discussion in section 4 of our paper, on the other hand, provides a natural link between restrictiveness and predictability: when the content of an NP minus its relative clause is insufficient to pick out the intended referent, some kind of additional modifier is likely to be included; an NSRC is one of the options, so the probability of an

NSRC is relatively high. To be more precise, it is the probability of a *restrictive* NSRC that is relatively high in such contexts. After all it is restrictive NSRCs rather than non-restrictive NSRCs that serve to provide additional information necessary to identify a referent. In other words, the need for sufficient identifiability influences the distribution of restrictive NSRCs and hence is a *cause* for increased predictability of restrictive NSRCs in such contexts. Note that there may be other reasons why RCs are more predictable in some context than in others. Here and in our paper, we focus on increases in NSRC predictability due to the pragmatically motivated need for certain referents to be identifiable. Crucially, it is not restrictiveness that causes greater NSRC predictability.

If the need for identifiability is one of the major factors determining NSRC predictability, this means that more predictable NSRCs are likely to be restrictive. The predictable NSRCs discussed here occur in contexts where they will naturally be interpreted as restrictive, irrespective of whether a relativizer is present. If disambiguation between restrictive and non-restrictive construals is one of the functions of relativizer omission, then we should expect omission to occur most when the possibility of a non-restrictive interpretation is greatest. By much the same reasoning that led to the prediction of more relativizer omission in writing than in speech, Kempson's disambiguation account would predict that *less* predictable restrictive NSRCs would have higher rates of relativizer omission. Since restrictive NSRCs in contexts that don't require further identifying information are more likely to be misconstrued as non-restrictive, speakers should be more likely to omit the relativizer to guarantee the intended (restrictive) reading. And this is of course the exact opposite of our central empirical finding.

The point here is that the correlation between predictability of an NSRC and absence of a relativizer seems natural from a processing perspective, but not if relativizer omission is thought of as a disambiguation strategy along the lines Kempson suggests. There is at least preliminary evidence that relativizers facilitate processing. There is some debate as to whether relativizers help production or comprehension (or both). On the one hand, relativizer presence has been shown to facilitate comprehension (e.g. Race & MacDonald, 2003). On the other hand, there is evidence that relativizer omission is correlated with production complexity (Jaeger & Wasow, 2005; see also Ferreira & Dell, 2000 for complementizers), but also that relativizers do not seem to *alleviate* production difficulty (Jaeger, 2005). While future studies are necessary to test whether speakers insert relativizer to facilitate production or comprehension, there is an established link between relativizer presence and processing (for further discussion, see Jaeger, 2006; Levy & Jaeger, 2006). Similarly, high-predictability of a parse can alleviate or avoid comprehension difficulties (see Jurafsky, 2003 for references). Thus providing relativizers for less predictable NSRCs seems like a reasonable hypothesis, although, admittedly, future work is necessary to test it.

The discussion of Kempson's proposal brings up another interesting point. Our work so far does *not* show that there is a *direct* causal link between NSRC predictability and relativizer omission. Could it be that it is the need for identifiability that directly causes relativizer omission? While we are not aware of any theory that would predict that this, it is a testable question that should be addressed in future research. As Harry Tily also

points out to us, it would be worth investigating to what extent variance in the predictability of NSRCs is explained by the need for identifiability, and to what extent other factors determine NSRC predictability. If other factors influence NSRC predictability and if increases in NSRC predictability due to these other factors correlate with relativizer omission, this would provide strong evidence for a *direct* causal link between NSRC predictability and relativizer omission.

Turning to the question of whether “speakers manipulate probability estimates”, we are puzzled why Kempson seems so reluctant to think that they do. In many other areas of cognitive science, including motor control (Trommerhäuser et al. 2005), visual inference (Kersten, 1999), concept learning (Tenenbaum, 1999), and reasoning (Anderson, 1990), there is little controversy over the fact that human information processing involves access to probabilistic distributions. Why should language be so different? Indeed, research over the past few years has revealed many cases of probabilistically-conditioned language production. For example, predictable syllables (Aylett & Turk, 2004) and more predictable words (Bell et al, 2003) are pronounced shorter and with less articulatory detail. Similarly, vowels that are more predictable given the preceding segments in a word are produced shorter and with less distinct formants (van Son & Pols, 2003). And cases of probabilistically-condition reduction are not limited to the phonetic level. Jaeger (2006) provides evidence that complementizer omission is correlated with predictability of a complement clause. Even phrasal omission has been linked to probabilistic distributions (see Resnik, 1996, on the distribution of implicit objects as in “John ate (dinner) before Mary arrived”). For a more detailed discussion of these phenomena as well as an information-theoretic account that links probabilistically-conditioned reduction to efficiency and successful information transfer, see Jaeger (2006: Chapter 6).

As far as we can tell, the widespread assumption that knowledge of language must consist of categorical mechanisms is a legacy of half a century dominated by grammatical theories built with tools borrowed from logic. That assumption was generally accepted for many years in part because the computations needed to develop serious quantitative models of language were infeasible with the technologies of the time. Over the past twenty years or so that has changed, and there is now a wealth of interesting results on language built with the tools of statistics and probability.

This has led to a vigorous debate within linguistics over the role of probabilistic findings in the theory of language; see, for example, Newmeyer (2003, 2006), Gahl and Garnsey (2004, 2006), and Jaeger (2007), among others. Kempson does not actually commit herself to one side or the other in this debate, but she makes it clear which side she thinks bears the burden of proof.

There are, however, other passages in her comments in which her prose suggests just the opposite. One example is her suggestion that a correlation between a lexical item and either presence or absence of relativizers “might become stored as a routinized strategy associated with that particular item”. The examples she gives (*way* is associated with relativizer absence and *a* with relativizer presence) are not categorical constraints, as our corpus studies demonstrate; so the stored strategies she posits would have to be

probabilistic.ⁱ Similarly, in arguing for the role of restrictiveness in the correlation between predictability and relativizer omission, she writes the following:

“... the more likely the form is to be associated with a restrictive form of construal, the less likely it is to be introduced with a form which allows for any other form of construal.”

This is a manifestly probabilistic claim.ⁱⁱ Since relativizer absence categorically entails restrictiveness, she needs a probabilistic formulation in order to avoid the false prediction that all restrictive NSRCs lack relativizers. But if speakers do not “manipulate probability estimates”, Kempson needs to explain where the non-categorical nature of this correlation comes from.

We hasten to add that we are in broad agreement with much of what Kempson says. She is quite right that the focus of our paper is narrow, and that much is to be learned from broader investigations looking at the effects of predictability in a wider range of constructions and languages (for examples of such work, see Jaeger, 2006 on English complementizer omission; Jaeger, 2007 on reduced subject relatives – so-called *whiz*-deletion; and ongoing work on relativizer omission in Danish). We also agree that restrictiveness may be an important factor in relative clause structure. In both connections, her discussion of Bemba is fascinating and illuminating.

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.
- Anderson, J. R.
1990 *The adaptive character of thought*. Hillsdale, NJ: Lawrence Erlbaum.
- Aylett, Matthew and Alice Turk
2004 “The Smooth Signal Redundancy Hypothesis: A Functional Explanation for Relationships between Redundancy, Prosodic Prominence, and Duration in Spontaneous Speech”. *Language and Speech* *47*(1), 31-56.
- Ferreira, Victor S. and Gary S. Dell
2000 “The effect of ambiguity and lexical availability on syntactic and lexical production”. *Cognitive Psychology* *40*, 296–340.
- Fox, Barbara A., and Sandra A. Thompson
2006 “Relative Clauses in English conversation: Relativizers, Frequency and the notion of Construction”. **To appear in *Studies in Language***.
- Gahl, Susanne and Susan Garnsey
2004 “Knowledge of grammar, knowledge of usage: Syntactic probabilities affect pronunciation variation”. *Language* *80*(4), 748–775.
2006 “Knowledge of grammar includes knowledge of syntactic probabilities”, *Language* *82*(2), 405-410.
- Jaeger, T. Florian
2006 *Redundancy and Syntactic Reduction in Spontaneous Speech*. Stanford University dissertation.
2007 “Usage or Grammar? Comprehension and production share access to same probabilities”. *Paper presented at the 81st Annual Meeting of Linguistic Society of America (LSA), Anaheim*.
- 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 and Thomas Wasow
in press “Processing as a Source of Accessibility Effects on Variation”. *Proceedings of the 31st meeting of the Berkeley Linguistics Society*.
- 2005 “Production-complexity Driven Variation: Relativizer Omission in Non-Subject-extracted Relative Clauses”. The 18th CUNY Sentence Processing Conference, Tucson, AZ.
- Kersten, Daniel
1999 “High-level vision as statistical inference”. In: M. S. Gazzaniga (ed.): *The New Cognitive Neurosciences*. Cambridge, MA: MIT Press. 2nd edition. 353–364.
- Levy, Roger and T. Florian Jaeger
2006 “Speakers optimize information density through syntactic reduction”. *Proceedings of Neural Information Processing Systems (NIPS) 2006, Vancouver, B.C.*
- Newmeyer, Frederick J.
2003 “Grammar is Grammar and Usage is Usage”. *Language* 79(4), 682-707.
2006 “On Gahl and Garnsey on usage and grammar”. *Language* 82(2), 399-404.
- 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.
- Resnik, Philip
1996 “Selectional constraints: An information-theoretic model and its computational realization”. *Cognition* 61, 127-159.
- Tenenbaum, Joshua B.
1999 “Bayesian modeling of human concept learning”. In: M. S. Kearns, S. A. Solla, and D. A. Cohn (eds.): *Advances in Neural Information Processing Systems 11*. Cambridge, MA: MIT Press.
- Trommershäuser, Julia, Sergei Gepshtein, Laurence T. Maloney, Michael S. Landy, and Martin S. Banks
2005 “Optimal Compensation for Changes in Task-Relevant Movement Variability”. *The Journal of Neuroscience* 25(31), 7169–7178.
- van Son, Rob J. J. H. and Louis C. W. Pols
2003 “How efficient is speech?”. *Proceedings of the Institute of Phonetic Sciences* 25, 171–184.

* This reply benefited immensely from the feedback by Harry Tily, whose challenging comments led us to entertain additional alternatives to our hypothesis that NSRC predictability drives *that*-omission.

ⁱ Incidentally, the discussion in Jaeger (2006: Chapter 6.2.3) contains control studies suggesting that the effect of predictability on relativizer omission holds beyond a few conventionalized tokens.

ⁱⁱ A clarification may be in order. We use the term probabilistic to refer to events that are conditioned by a probability of another event. This use ‘probabilistic’ is different from its use in, for example, the Stochastic OT literature, where an event is called probabilistic when it occurs with a certain probability. In the latter sense, the claim that relativizer omission is probabilistic is almost trivially true. Even if all variation in relativizer omission were determined by absolutely categorical contrasts – which is extremely unlikely given that the *same* speaker will sometimes say the *same* sentence with and sometimes without a relativizer – the resulting distribution would still be binomial (with *p* being either 0 or 1, depending on the context). Our question here is different. Our specific hypothesis is that it is the probability of an RC that influences relativizer omission. But we can ask more generally whether probabilities are part of the predictors of relativizer omission (cf. Jaeger, 2006, 2007).