When Stylistic and Social Effects Fail to Converge: a variation study of complementizer choice

1. The Question

Labov observes in *The Social Stratification of English in New York City* that "in general, a variant that is used by most New Yorkers in formal styles is also the variant that is used most often in all styles by speakers who are ranked higher on an objective socio-economic scale" (1982, 279). This observation, that stylistic effects tend to mirror social effects, has been considered a guiding principle of variationist sociolinguistics for long enough that when a variable shows evidence of stylistic conditioning, we may expect or even assume that social conditioning is also present. Indeed, there are good theoretical reasons to believe that this relationship is anything but a coincidence; the exploration of this and related ideas has led to a much deeper understanding of the nature of social meaning, among other things. However, this relationship between stylistic and social conditioning is not a logical necessity, but depends crucially on the specific social meanings associated with the variables in question, and if the mirroring relationship does not always hold, then we need to account somehow for the existence of a set of variables that have so far been largely ignored in the literature.

Because the positive relationship between social and stylistic stratification has already been demonstrated for several different types of variables, and we have no reason to believe that it does not hold in these cases, the most straightforward way to further investigate this relationship is to look for evidence of new, un- or under-studied variables for which it does not hold. Identifying a variable that shows either social or stylistic stratification, but not both, would challenge the traditional view that this relationship is somehow inherent to variation and also provide some insight into why it exists in the first place. The most obvious place to begin such an endeavor is with a variable that is widely thought to vary on one of these axes and not necessarily the other; in this study, I will investigate complementizer choice as an example of such a variable.

2. The Variable

The complementizer *that* introduces sentential verbal complements; however, many of these complements may also have a null complementizer, and in these cases the null complementizer may be said to be in free variation with *that*:

> *We understand a lot of hard work goes into making music. We also understand that listening is a subjective thing, but we can't put everything we get online.*
> http://www.npr.org/programs/asc/help/index.html - top

As the above example shows, the same writer/speaker using the same subject and verb can use the null complementizer in one sentence and the overt complementizer (*that*) in the next. The obvious question raised by a situation of "free variation" is that of whether the variation is truly free, or there are non-categorical factors contributing to the presence or absence of the overt complementizer. Fortunately, there is a fairly large body of work oriented toward answering this question for this phenomenon; most recently (and perhaps most comprehensively) Roland, Elman, and Ferreira (to appear) have addressed this question, using a very large automatically parsed corpus of data from the British National Corpus to establish a set of factors that contribute to the presence or absence of *that* to introduce sentential complements.

In addition to the internal factors that have been established to contribute to complementizer realization, several descriptions of this usage suggest that factors relating to style contribute to the choice between *that* and zero, as well. Few if any of the existing studies make any mention of the possible effect of social factors on the realization of the complementizer, making it a good choice for our investigation into the relationship between stylistic and social conditioning. As mentioned above, it has been well established by a large body of sociolinguistic literature that variables that are stylistically conditioned are often socially conditioned as well; one common explanation for the existence of stylistic variation is that the variable in question has some degree of social meaning for speakers, and this social meaning can be used by speakers to index aspects of their identities as well as aspects of their communicative situations. Thus, if complementizer realization really is stylistically conditioned, it would be reasonable to expect that we might also find some of the traditionally investigated global social factors conditioning its use. If these variables do not emerge as conditioning factors, then it would be very helpful to us as variationist sociolinguists to develop an understanding of social and stylistic conditioning of variables that can account for how and when this relationship breaks down.

The object of the current study is to determine whether social factors such as gender, age, geographical origin, and level of education play a part in the conditioning of realization of the complementizer. Using data from the Switchboard corpus, which contains the above social information about the speakers involved, I built a logistic regression model of the factors contributing to the presence/absence of the overt complementizer *that* to determine which internal and external factors are the most significant in conditioning complementizer realization.

In addition to the work specifically on complementizer realization, there is a fairly large body of work addressing similar questions regarding relativizer realization. Much of this literature draws similar conclusions to those drawn by researchers working on complementizer variation, but there are some studies on relativizer variation that go beyond what has been done for complementizers. Though there are more than two possible items that may appear in the relativizer position, this area of variation is intuitively similar to variation in complementizer realization, and some of this work (especially Sigley 1997 and 1998 and Jaeger and Wasow 2005) is quite relevant to the interpretation of the results of the current study. For this reason, I will occasionally refer to this literature in addition to the studies that directly address the question of factors conditioning the *that/zero* variable.

3. The Factors

In the literature that touches on the phenomenon of *that* presence/absence, there is a fairly high degree of agreement on what factors may encourage or discourage the presence of the complementizer; what is perhaps most surprising about this agreement is that some of it seems to be based on very little evidence, if any. In addition to recent publications on the phenomenon described, there are also descriptions of and comments on this phenomenon to be found in grammars of the English language, as well as in manuals on grammar, style, and usage in English. These sources provide many of the same insights into the factors determining complementizer presence/absence as the recent articles on the subject, despite the fact that only a few of them involve quantitative data.

Among the factors generally agreed to discourage the presence of the overt complementizer are the following: a) high frequency matrix verb (especially of cognition or communication), b) co-referential subjects between matrix clause and complement clause, and c) a personal pronoun as the subject of the complement clause (Biber et al. 1999, 680; Huddleston and Pullum 2002, 953). An "extraposed subject in a matrix clause containing *be* and a short predicative complement" (e.g. *It's a good thing (that) we showed up*) is also cited as a factor disfavoring the overt complementizer (Huddleston and Pullum, 952).

Factors generally agreed to encourage the presence of the overt complementizer include a) the presence of material that intervenes between the verb of the matrix clause and the subject of the complement clause, b) passive voice in the main clause or an extraposed subject of a transitive main clause, and c) coordination of complement clauses. It is also generally accepted that complements of nouns almost always require the overt complementizer (Biber et al., 680; Huddleston and Pullum, 953).

Another view that is supported by a few studies (Biber et al., Roland et al.) is that those factors that are known to encourage the presence of *that* are most relevant in contexts where the "baseline" level of that production is quite low (such as with verbs like *think*, according to Roland et al.), while those factors that are known to encourage the absence of that are most relevant in contexts where the "baseline" level of *that* production is rather high (such as in news registers, according to Biber et al., 680-683).

There are also some factors that are more controversial, including non-linguistic factors, such as style; some other factors relating to syntactic position; specific effects of the verb lemma, in addition to frequency; and semantic factors. In the sections below, I will attempt to summarize the arguments and comments from the literature relating to each type of factor, after which I will present my own results as they relate to the established and postulated factors affecting speakers' choice between the overt and zero complementizers.

A. Syntactic Position

Among the things that are fully agreed upon regarding complementizer realization is the restriction that "when the that-clause is object or complement (or delayed subject), the conjunction *that* is frequently omitted in informal use, leaving a 'zero that clause'. When the clause is subject, *that*… cannot be omitted,"(Quirk 734). This restriction is stated in similar terms by Fowler (632), Kruisinga and Erades (114), Bolinger (11), and Huddleston and Pullum (953), among others. Kruisinga and Erades also explain that "clauses dependent on nouns…are always introduced" (110). Huddleston and Pullum also describe some other conditions under which the overt complementizer is obligatorily present:

a) when the content clause is subject or otherwise precedes the matrix predicator
b) when the content clause is adjunct (*He appealed to us to bring his case to the attention of the authorities that justice might be done*)
c) when the content clause is complement to comparative *than/as* (*I'd rather that he hired a taxi than that he drove my car*)
d) appositives (*This motion, that John be fired, was defeated*) (probably obligatory)

The overt complementizer is, on the other hand, obligatorily omitted when "the content clause is embedded within an unbounded dependency construction in such a way that its subject is realized by a gap (*She thinks that Max is the ringleader. Who does she think is the ringleader? Max is the one she thinks is the ringleader.*)" – this has been called the "*that*-trace effect" (Huddleston and Pullum, 953).

Interestingly, among the factors relating to syntactic position, none of them are associated with moderately high or low rates of that presence; that is, they are all associated with categorical or nearly categorical presence or absence of the overt complementizer.

Because these "factors" have a nearly categorical effect, they (and others like them) are not of particular interest to the current study. For the current purpose, we are more interested in those factors that have a non-zero, non-categorical effect on the realization of the complementizer. Regarding such non-categorical factors, Huddleston and Pullum say, "the relative likelihood of dropping the *that* depends largely on the structure of the matrix clause but also on that of the content clause itself" (953). Below we will discuss the effects of different aspects of the structures of both clauses.

B. Verb

Many different attributes of the matrix verb have been suggested as possible conditioning factors.

1) Verb Lemma

Roland et al. used the identity of the verb lemma as a factor, and showed that the specific verb was a useful predictor of *that*/zero realization. In this vein, the *that* literature is full of various sorts of claims regarding the preferences of individual verbs for taking complements marked with *that* or unmarked ones. For example, Fowler (632) claims that the overt complementizer is absolutely required in the following circumstances:

> assert that X
> I am abashed to see that X
> My view is that X
> point out that X
> State that X

Google, however, says otherwise:

> I **assert** you can't logically do that. www.macaddict.com/forums/topic/56824/2 (April 18, 2005)
>
> **My view is** you will do best with the things you love to do. www.businessownersideacafe.com/cyberschmooz/startupstew/6185.html (April 18, 2005)
>
> Wyneth was **abashed to see** the plastered figure wore a toga.
> www.fictionwise.com/ebooks/eBook25186.htm (April 18, 2005)
>
> Amir, may I **point out** you have fallen into the trap many muslims do.
> forum.onlineopinion.com.au/thread.asp?article=2811 (April 18, 2005)

> I would just like to **state** you do not have to believe the world is full of dualities: good and evil, right and wrong, black and white, salt and pepper. lenus.blog-city.com/read/15279.htm (April 18, 2005)

Fowler (632) also provides lists of specific verbs that exhibit non-categorical "preferences" for overt expression or omission. According to Fowler, the following verbs prefer *that*:

agree assert assume aver calculate
conceive hold learn maintain reckon
state suggest

The following are verbs that Fowler claims prefer omission:

believe presume suppose think

Fowler also claims that the following verbs vary according to "the tone of the context":

be told confess consider declare grant
hear know perceive propose say
see understand

It certainly seems to be the case that for each of these lists, it is possible for the verbs to take introduced or unintroduced complements:

> You **agree** you will not copy, distribute, publish, transmit, modify, display or create derivative works from or exploit the contents of this Site in any way. www.getcosi.com/terms_use.asp (April 18, 2005)

> You **agree** that if you post any information to this Site, you are acting as on your employer's behalf, publishing such information on your employer's behalf and you confirm that such information is not confidential to you or your employer. https://www.online.nokia.com/public/shared/acceptable_use_policy.htm (April 18, 2005)

> Fourth, if you **believe** people are basically good, you, of course, **believe** that you are good -- and therefore those who disagree with you must be bad. www.townhall.com/columnists/ dennisprager/dp20021231.shtml (April 18, 2005)

> I **am told** that you intended to sell half of the 240 undamaged tablets, at a small profit. www.courts.sa.gov.au/sent_remarks/ sr/0322_cetojevic_petar_kris.htm (April 18, 2005)
>
> I **am told** you have no memory of the events leading up to the accident and the accident itself. http://www.courts.sa.gov.au/sent_remarks/sr/0330_cullen_nicole_terri.htm (April 18, 2005)

While it is clear that Fowler's levels of distinction are not sufficient, and his comments leave much to be said with regard to those verbs he does not list, his taxonomy of verbs does raise an interesting question: could it be that "free variation" only exists in the context of certain verbs, or even that this variation only surfaces in certain stylistic contexts?

2) Factivity

Cofer (299-306) considers whether the factivity of the predicate of the matrix clause could be a factor[1]. Only two factive predicates occurred in his data five times or more, unfortunately, and after a quantitative analysis, Cofer determined that it was in fact frequency of the verb lemma that accounted for the differences that he was originally attributing to factivity; no researcher has found an effect based on factivity alone.

3) Epistemicity

A similar quality that has been attributed to verb lemmas is epistemicity, which Thompson and Mulac define as "degree of speaker commitment" (243). According to Roland et al., "more epistemic" verbs are those that create less of a distinction between main and subordinate clauses, making *that* less likely; this presumes that the overt complementizer somehow creates more distinction between the clauses than the null complementizer, which is certainly possible but not strictly obvious. Thompson and Mulac make some interesting but controversial claims regarding highly epistemic examples; in short, Thompson and Mulac's argument is that when there is no overt complementizer, "the main clause subject and verb function as an epistemic phrase, not

[1] According to Cofer, "factive predicates occur when the speaker pre-supposes that the sentential subject or object of the predicate is true; non-factive predicates generally occur when the speaker only asserts or believes that the embedded sentence is true" (299).

as a main clause introducing a complement" (241). Part of this argument involves an evaluation of which main verbs are "characteristically associated with epistemicity", which is far from clear-cut (242).

Thompson and Mulac use corpus data to support many of their claims, but they do not attempt to find any way of quantitatively operationalizing epistemicity. While their idea (which has been taken up by many subsequent researchers including Roland et al.) is an intriguing one, the lack of any sensible way to quantitatively determine whether the epistemicity of a matrix verb might be a predictor of *that* /zero realization is highly problematic; for example, as Roland et al. point out, the same verb lemma can appear in an epistemic usage and a non-epistemic one, as in the following examples:

> Epistemic: I believe it's going to rain.
>
> Non-epistemic: I believe the death penalty is wrong.

However, other researchers have also argued that the epistemicity of a construction is a determining factor in whether or not the overt complementizer appears (Yaguchi 2001, 1133). In addition, epistemicity has been cited as a possible factor in relativizer realization, as well; more epistemic relative clauses (ie *That's the way (?that) it works*) often do not take an overt relativizer. I will return to the question of how to include the notion of epistemicity in a quantitative model in the Methods section.

C. Attributes of Subject of Embedded Clause

There are many attributes of the subject of the subordinate clause that have been suggested as factors influencing the realization of the complementizer.

1) Semantic features:

Roland et al. cite animacy and abstractness of the embedded subject among the semantic factors that help to predict the presence/absence of the complementizer. In their study, these factors were second only to factors related to the verb lemma in their effectiveness at improving the model. Specifically, they found that animate and pronominal NPs (*I, he*) predicted the null complementizer, while inanimate, more abstract full NPs (*the problem,* *the two reasonable interpretations*) predicted the overt complementizer. As they point out, the characteristics associated with those embedded subjects that predicted the overt complementizer are also characteristics of common direct objects, suggesting that the overt complementizer might be acting to disambiguate between an embedded subject and a matrix object; however, they also note that there is other evidence (such as case marking on pronouns) that suggests that this is not the case. Thus Roland et al. suggest that these semantic factors, rather than pointing to ambiguity resolution as a motivation for including the overt complementizer, support Thompson and Mulac's theory of epistemicity as a major determining factor.

2) Person/number/pronominality:

In line with Roland et al.'s findings, McDavid 1964 observes that *that* disappears more often when the subject of a clause is a pronoun. However, some non-semantic explanations for this preference have been suggested; Bolinger posits that the overt complementizer is not necessary before pronominal subjects because subordinate clauses without full NP subjects do not create "distracting noun-noun combination(s)", which would appear in the case of a double object verb with an unmarked clausal complement that had an NP subject (e.g. *I told my <u>dad cheesecake</u> was going to be dessert*) (13). This is essentially a member of the class of ambiguity avoidance arguments, which Roland et al. have suggested are not empirically supported.

Thompson and Mulac claim that first and second person pronouns have a higher level of epistemicity than other pronouns, thereby explaining why the presence of these pronouns in the matrix clause discourages the presence of *that*. Presumably this could also have an effect in the subordinate clause, although Thompson and Mulac do not discuss this possibility.

3) Demonstrative *that*:

Demonstrative *that* as the embedded subject seems to decrease the rate of complementizer *that*. The most obvious explanation for this phenomenon is that it is essentially a kind of weak repetition avoidance or OCP effect: because the phonological forms of demonstrative *that* and complementizer *that* are the same, this combination (in which demonstrative *that* directly follows the complementizer in linear order) is disfavored (Walter and Jaeger to appear).

4) Coreference with the matrix subject:

Coreference between the subjects of the two clauses may decrease the distinction between the two clauses, following Thompson and Mulac, thereby increasing the epistemicity of the matrix clause and decreasing the likelihood of the overt complementizer. For example, saying "I think I'm going to be sick," is more highly epistemic than saying "I think it's going to rain tomorrow", because the speaker is likely more committed to a projection about his own behavior than that of the elements. Ferreira and Dell 2000 alternatively propose that the omission of *that* is licensed by an effort to allow the early mention of some previously mentioned material; in this case, if the pronouns are coreferential, then the second reference constitutes a repetition of previously mentioned material, and Ferreira and Dell predict that the complementizer might not appear to promote "early mention" of the repetition (321). Roland et al. explain that because the pronouns that are most often coreferential between the two clauses are first and second person singulars, it is "difficult to distinguish between epistemicity as a cause for the lack of complementizer, and Ferreira and Dell's early mention" (27). Thus, coreference between the matrix and embedded subjects could quite reasonably discourage the appearance of the overt complementizer, but this fact does not lend itself easily to interpretation.

D. Ambiguity Resolution

Though it is not structured as a measurable factor in the same way that the above factors are, several researchers have suggested that one of the motivating forces behind the presence of the overt complementizer is to resolve potential ambiguities. Bolinger explains, "If one function of *that* is to identify its clause as a constituent, anything that makes it difficult to identify constituents will make *that* more necessary" (33). In addition, with reference to a situation in which *that* is considered completely obligatory, he suggests that "it appears that the reason why *that* as a subject cannot normally be omitted is not because it is a subject but because without it [*that*] the constituents are too hard to identify" (1972).

Sigley similarly argues, for relative clauses, that "The fact that [the null relativizer] appears influenced by formality in writing just where there is no other immediate signal of the clause boundary suggests writers and editors are acting… to reduce ambiguity" (Sigley 1997, 28). This claim is similar in nature to the findings of Roland et al., who establish that the embedded subjects of complement clauses introduced by *that* are very similar to direct objects of the same verbs (the subjects of introduced clauses are object-like), whereas they do not share as many properties with the embedded subjects of unintroduced complement clauses (the subjects of unintroduced clauses are not object-like). That is to say, *that* seems to appear when there is a danger that the embedded subject would be locally ambiguous with a direct object (a garden path type of ambiguity).

E. Frequency

As mentioned briefly above, it is widely held that higher frequency verbs may be less likely to take the overt complementizer. In addition to this, Roland et al. have suggested that higher frequency subjects of the subordinate clause may have the same effect. One way to interpret this, if it is true, is by means of an ambiguity resolution explanation; that is, that the subcategorization frames of higher-frequency forms are more salient, so there is less difficulty in identifying the sentential complement as such when it appears, thereby somewhat obviating the need for the overt complementizer. However, Bolinger notes that "Though high frequency forms are not necessarily restricted to relaxed speech, it is a fact that informal expressions tend to be high in frequency" (22). This observation leads us to ask whether the "frequency" effect is really a hidden effect of formality. Conversely, could the frequency effect account, instead, for the purported effect of formality? We will return to this question in the Analysis section.

F. Phonological Factors

According to Cofer, phonological factors "apparently do not affect *that* deletion to any great extent." None of the other *that* literature makes any comment on the subject, but there is some indication from some of the literature on double- *is* constructions (Isis) that intonation in these examples is related to the presence or absence of the overt complementizer (c.f. Brenier and Michaelis 2002). While it is not clear that intonation is directly influencing this variable, it may at least be indicative of a semantic difference that would be relevant to the complementizer realization. There doesn't, however, seem to be any evidence regarding whether this influence would also pertain to other *that*-complementation constructions, including the construction that is represented by the data in the current study.

G. Intervening Material

Cofer reports of McDavid's study that "single object verbs were more favorable to deletion than double object verbs like *tell*". One analysis of this conclusion is that it is an effect of intervening material – double object verbs are extremely likely to have an indirect object intervening between the verb and the complement clause (which is generally the direct object), whereas single object verbs can only have adverbials intervening in this position. Because intervening material has been suggested to make "retention" (that is, the appearance of *that*) more likely, double object verbs would naturally be "less favorable to deletion" than single object verbs.

On the other hand, it is well documented, as discussed above, that the verb lemma contributes quite strongly to the likelihood of "deletion" or "retention", so this observation of McDavid's could be motivated by what is really a "verb" effect – that is, single-object verbs (such as *think, know*) could share some of the other characteristics (such as semantic characteristics, as suggested by Roland et al.) that are thought to discourage the appearance of the overt complementizer, whereas double-object verbs (such as *tell, bet*) could share some of the other characteristics that are thought to encourage its appearance.

This effect could of course also be based on the principle of ambiguity resolution; that is, since many instances of "intervening material" are objects of double object verbs (*I told my cousin Bertha was coming*), they may create Bolinger's "distracting noun-noun combinations" (in this example, *my cousin Bertha*) which could possibly be interpreted as compounds or restrictive appositives; in the case of intervening adverbial material (*I decided* ***immediately*** *that it would be over*), the overt complementizer could help the listener determine to which clause the adverbial belongs.

H. Processing Factors

Ferreira and Dell (2000) propose that the "principle of immediate mention" may contribute to whether or not the overt complementizer is selected:

> "The principle of immediate mention makes a straightforward prediction for sentence complement structures with optional complementizers, like *The coach knew (that) you missed practice*. Assume a speaker has already selected the lemmas for *coach* and *know*, so that the next word in the sentence will be the complementizer *that* or the embedded subject *you*. If the *you* lemma becomes available quickly, then according to the principle of immediate mention, a sentence complement structure without a *that* should be used, since only such a structure permits immediate mention of *you*. If the *you* lemma becomes available more slowly, then a sentence complement structure with a *that* can be used, perhaps to maintain the impression of fluency despite the relatively greater difficulty (i.e., the *that* operates as a grammatical "um"). More generally, if the embedded subject of a sentence complement is selected quickly, then a *that*-less sentence complement structure should be used to accommodate immediate mention of that quickly selected embedded subject. (299)

They conducted an experiment showing that speakers produce *that* less often when the embedded subject is a repetition of the matrix subject (p. 317-18), which they interpret to mean that the greater availability of these subjects made that less necessary; however, as noted above, this result could also be explained by the greater epistemicity of these examples (smaller distinction between the main and subordinate clauses). Their subsequent experiments control for this by using "recall cues" that did not involve coreference between the main and embedded subjects, and found that the more "available" still favored the null complementizer, supporting their "principle of immediate mention". Jaeger and Wasow (2005) have suggested that relativizer variation can also partly be explained by processing constraints, specifically accessibility. These processing factors seem to be at odds with ambiguity resolution, making complementizer/relativizer production fertile ground for the investigation of production vs. comprehension motivations for linguistic behavior (Jaeger and Wasow, 2005).

I. Style

Several studies, from some of the earliest to some of the most recent, have made reference to the style, genre, or register of the text(s) in question as a possible factor in *that*/zero realization. Unfortunately, these references have often not been specific about exactly which of these is in question; Fowler (1954) states, “the use or omission of the *that* of a substantival clause depends partly on whether the tone is elevated or colloquial”, leaving the reader with very little idea exactly what an “elevated tone” or a “colloquial tone” might refer to (632).

One of the largest quantitative investigations of this question was a corpus study done by Fries (1940) of different types of written materials. He distinguished between “Standard English Materials” and “Vulgar English Materials” for his corpus work. Fries identifies 414 of the total 994 conjunctions (complementizers) as *that* in SEM, and 185 of the total 959 conjunctions as such in the VEM (207-8). This might seem to indicate that in certain styles *that* -complementation itself is simply more common; however, he assesses that some of the “total conjunctions” he counts are instances of the null complementizer that could have contained the overt complementizer, and thus argues that it is not the rate of *that* -complementation but the rate of overt complementizer use that differs between the two sets of materials:

> “*That* appears to be used much more frequently in Standard English than in Vulgar English. In this connection one should point to the figures for those clauses in which no function word appears but in which a *that* might be used. There were 414 instances of *that* in the Standard English letters and but 185 instances of *that* in those of Vulgar English. On the other hand there were in the Standard English letters only seventy eight instances of clauses without a function word in which *that* might have been used, as against 206 such instances in the Vulgar English letters. If these figures are put together, one would have 492 for Standard English and 391 for Vulgar English – not a very significant difference.” (209-210)

While this conclusion does seem reasonably robust, on the basis of the numbers he provides, the materials themselves may cast the validity of the entire study into doubt. These materials are described as “letters written by US citizens to the government” and the groups are distinguished on the basis of which of them “conformed most closely to standard English", as opposed to being "vulgar English" (Cofer 1972). This implies that the two groups are actually distinguished from each other by some aspect(s) of their language use. Cofer postulates that these differences might lie in whether the writers had a "distinct written style" or they "used the same patterns in writing as in speech." This methodology seems fraught with problems, not the least of which is that if the groups are subjectively distinguished, it becomes nearly impossible for us to know whether a subjective impression of *that* use was one of the things that indicated this difference to Fries. Beyond this, as we will discuss in the analysis section, this sort of distinction lends itself much more easily to an interpretation based on register than on style. While the results of this early study are quite interesting and suggestive, the methodology seems to be fatally flawed. Despite this, Cofer strongly supports the idea that *that* is stylistically conditioned, citing results from Graf 1962, McDavid 1964 and Jespersen 1909 for further evidence of stylistic effects in writing.

Cofer also looks to fiction for support, and cites Storms in suggesting that *that* carries "stylistic and emotional overtones" and that "the presence or absence of *that* reflects the emotional tone of the situation being portrayed" in fiction, implying that the same may be true in speech.

More recently, both Huddleston and Pullum and Biber et al. support style as a conditioning factor for the *that* /zero variable. Huddleston and Pullum argue that "the default case is the one where *that* is present as a marker of the subordinate status of the clause. Departures from this default case, declaratives without *that*, are more likely in informal than in formal style" (953). As they do not give any explanation for their choice of the overt complementizer as the "default," we can only assume that in the domain in which they are working (standard written English), *that* appears more often than the null complementizer. Rather than claiming one default, Biber et al. provide the generalization that "in conversation, the omission of *that* is the norm, while the retention of *that* is exceptional. At the opposite extreme, retention of *that* is the norm in academic prose" (680). Their analysis of these facts (which were gleaned from a study of the Longman Spoken and Written English corpus) is that the differences are due to different "production circumstances" and "communicative purposes". This suggests that Biber et al. believe that the traditional stylistic interpretation of modality differences (speaking vs. writing) as reflecting formality differences may not be the whole story. Specifically, it opens the door to an interpretation involving processing effects on *that*/zero choice, in addition to allowing for other social/discourse motivations beyond the simple formal/casual distinction. However, Biber et al. do not make any attempt to specify what these effects and motivations might be, leaving this for further research.

J. Social Factors

In general, most studies on complementizer realization make no mention of social factors; however, there has been at least one quantitative study involving social factors, done by Cofer in 1972. In his study of a Philadelphia speech community, Cofer found "no clearcut correlation with differences in class or race". But with "less common predicates" (i.e. not *think* or *guess*), he finds "some tendency for working class informants and informants in speech groups I and II to delete more." Though he does not claim statistical significance, this finding is interesting in its exceptionality – it is suggestive of social conditioning of the *that*/zero variable, an idea that has largely been overlooked in the literature.

As interesting as it is, analyzing this suggestive finding presents a challenge, because the specific predicates involved may have a very strong effect on the results, among other reasons. Adamson (1992) presents results of a similar study of relativizer use, which also suggest some kind of social conditioning; specifically, he found that the upper-class speakers in his sample produced fewer zero relatives than the working-class speakers did (128-129). These studies were both based on data from traditional sociolinguistic interviews, which differs significantly from corpus data (the kind used in almost all the other studies cited) in a variety of ways, some of which may become relevant in our analysis. We will return to the question of whether style and/or social factors may affect complementizer presence/absence and what it might mean if they do in the Results section.

4. The Methods

The current study was based on corpus data from the parsed version of the Switchboard corpus, which contains some amount of social information about each speaker (specifically the speaker's self-reported birth year, gender, region of origin, and level of education). I chose to extract sentences with first person singular subjects of matrix verbs that were in the simple present or simple past tense (e.g. *I think/I thought*). This reduced the number of factors that could possibly affect the usage of the complementizer, which could have included the number and/or person of the matrix subject, and the mood and/or aspect of the matrix verb. While this information could potentially have been of interest in some way, it seemed unlikely that it would add to an understanding of how non-linguistic factors contribute to *that* /zero realization, and reducing the number of factors and factor groups in the statistical analysis (and the number of examples to be dealt with) seemed like a worthwhile consequence of this choice. To a large extent this probably turned out to be true; however, in the Results section, I will consider whether including a larger number of examples in general could have improved the analysis.

From each sentence that was selected, the following information was automatically extracted[2]: verb lemma of the main clause verb; subject of the embedded clause; number of intervening phrases between the verb of the main clause and the subject of the embedded clause; tense of the matrix verb; gender, age, region, and educational level of the speaker.

I also calculated a few measures of frequency of the verb lemmas; the first is based only on the frequency with which each verb lemma appears in my data. The second is based on how frequently each verb lemma appears in the Switchboard Corpus in toto. The third measure is based on the frequency with which each verb lemma appears in the British National Corpus (which is independent of the Switchboard Corpus, and much larger). The log transformations of all three measures of frequency were included as factors, in order to compress a very large range of frequencies into a more manageable set of values, in keeping with the intuitive judgment that exponential differences in frequency (1 vs. 10,

[2] The extraction was done via a combination of tgrep and Perl scripts.

10 vs. 100, etc) are approximately equivalent. In general, unless otherwise mentioned, when I refer to "frequency" as a factor, it refers to the first measure described, but it should be noted that all three measures behave very similarly in the models.

Based on these measures of frequency, I also calculated the conditional probability that any given instance of a verb is an instance of the type I collected. This is simply the first measure of frequency divided by the second. In the absence of any evidence that the specific set of sentences I selected should represent a different proportion of the instances of *that* -complementation for one verb than for any other, this measure should serve as a good proxy for the overall conditional probability that any given instance of a verb will have a *that* -complement[3].

I also attempted to operationalize the notion of epistemicity (as discussed in Thompson and Mulac and Roland et al.) to some degree of approximation by determining how often each relevant verb is used parenthetically (e.g. *He said, I think, that he would be late*.). Using an API script[4], I extracted from Google all examples of a particular parenthetical usage (*is, I* verb, *a*)[5] and then calculated a "parentheticality" measure by dividing the number of parenthetical examples of each verb by its frequency in my corpus (the first measure of frequency). This is not intended to measure epistemicity itself, but to measure the likelihood of a verb to appear in a certain type of epistemic usage. Below is a table listing all of the factors included in the final analysis and the possible values for each factor.

[3] While it is possible that, for example, certain verbs are more likely to appear with a first person singular subject than others, and therefore will achieve a higher conditional probability than other verbs, this does not constitute a problem because the purpose of the conditional probability measure is to determine how likely the utterance is to be an instance of *that* -complementation once the verb is known. Since the information about the subject is already available by the time the verb appears, a measure of probability that includes this information does not inaccurately model what a speaker and/or hearer knows when producing the *that* -complement.

[4] The script was written by Liz Coppock, 2005.

[5] Because Google does not recognize punctuation in its own search strings, searching for ",*I* verb[,.]", as one might want to, does not yield anywhere near the total number of parenthetical uses in the first 1000 results, leading to a plethora of complications that render the measure useless; to solve this problem, I restricted the measure to a consistent subset of the parenthetical uses, all of which appear in the first 1000 results, allowing for direct comparison among the different verbs.

Table 1. Factor Values

Factor	values
Verb lemma	agree, assume, believe, bet, decide, doubt, expect, feel, figure, find, guess, hear, hope, imagine, know, mean, note, notice, (would) rather, read, realize, recognize, remember, say, see, suppose, suspect, swear, take (it), tell, thank, think, understand, wish
Verb tense	past, present
Log of Verb Frequency	0-1
Conditional probability	0-1
Intervening material	0,1,2,3
Subject	I, you, that, other pronoun, full NP
Parentheticality	0-1
Gender	female, male
Birth year	up to 1940, 1941-1950, 1951-1960, 1961 on
Region	mixed, northeast, north midland, north, NYC, south midland, south, west
Education	0,1,2,3, did not respond

After collecting the data and coding for the relevant factors, I built a logistic regression model of the data using SPSS to determine which of the factors contributed significantly to explaining the variation. Because the number of examples for many of the verbs was quite low, following Roland et al., I built several models using different subsets of the verbs. The results of these models appear below.

5. The Results

The first regression model I built included all the examples I had collected, adding the variables stepwise. The p-values associated with each variable represent the significance of the change in -2 log likelihood of the model when the variable is removed from the model (i.e. how much worse the model gets without the information provided by the factor group). Below are the significant results from the all-inclusive model:

Table 2. Results of regression model including all examples

Variable		Model Log Likelihood	Change in -2 Log Likelihood	df	Sig. of the Change
Step 6	Intervening material	-1436.728	83.583	1	.000
	Subject of subordinate clause	-1409.103	28.334	4	.000
	Region (binary)	-1398.956	8.040	1	.005
	Conditional probability (binned)	-1411.491	33.110	1	.000
	Parentheticality	-1437.094	84.315	1	.000
	Log frequency (1)	-1402.097	14.322	1	.000

As the table shows, there were six significant factor groups for this set of examples: the conditional probability that the verb took a *that*-complement, the presence of intervening material between the verb and the subject of the complement, the type of subject the complement had, the log of the frequency of the verb within my data, the frequency of parenthetical usages of the verb, and a binary version of the region variable[6].

Though the "verb" factor did turn out to be highly significant, I have left it out of this analysis because the sample included many verbs with very few examples, so it obfuscates many other effects. However, it must be considered that because verb lemma is such a strong predictor of *that* vs. zero realization, other factors that are determined by the verb (such as the frequency, conditional probability, and epistemic factors) may act as stand-ins for the verb lemma and could show spurious effects. One way of controlling this type of error is to put these continuous variables into bins, so that instead of individual values, each verb gets a score of low, medium, or high (for example). Below are the results of a regression model using the binned versions of conditional probability, frequency, and epistemic frequency:

[6] Of the regions reported by speakers, only the Northeast showed any divergence from the mean, so the binary version of this variable distinguishes speakers from the Northeast from speakers from all other regions.

Table 3. Results of regression model including all examples (binned)

Variable		Model Log Likelihood	Change in -2 Log Likelihood	df	Sig. of the Change
Step 6	Intervening material	-1485.696	82.674	1	.000
	Subject of subordinate clause	-1467.524	46.329	4	.000
	Region (binary)	-1447.830	6.942	1	.008
	Conditional probability (binned)	-1458.661	28.603	1	.000
	Log frequency (1)	-1446.552	4.385	1	.036
	Parentheticality	-1475.506	62.293	1	.000

Looking at the significance of the change in -2 log likelihood when we remove the binned version of the log frequency measure, we can see that it has decreased, but is still below the 0.05 level. The other two binned factors are still highly significant.

While the results of the all-inclusive model are quite interesting, it is important to recognize that the different verbs vary greatly in frequency, and because the verbs have a strong effect on the complementizer realization, this bias toward some verbs might create the impression that certain factors are or are not significant in general when they are in fact only significant or not in the context of certain verbs. With this in mind, I created a model that included only verbs of "medium" frequency – that is, greater than 25 and less than 100 examples appearing in my data[7]. The table below shows the results of this regression:

Table 4. Results of regression model including medium frequency verbs

Variable		Model Log Likelihood	Change in -2 Log Likelihood	df	Sig. of the Change
Step 5	Intervening material	-211.411	13.221	1	.000
	Subject of subordinate clause	-211.391	13.181	4	.010
	Conditional probability (binned)	-243.508	77.416	1	.000
	Log frequency (1)	-231.477	53.354	1	.000
	Parentheticality	-221.082	32.563	1	.000

[7] See appendix for a table of verbs and their frequencies, conditional probabilities, and parentheticalities.

This table shows that for the medium frequency subset of verbs, the following factors significantly improve the model: intervening material, subject of the complement, conditional probability (the binned version), epistemic frequency, and log frequency[8]. Interestingly, one factor that was significant in the previous model (the binary region variable) does not appear to be significant in this model. This suggests that it may have been some of the more frequent verbs, which were not included in this model, that caused this factor to achieve significance in the first model. I created individual models for the three most frequent verbs (*think, guess*, and *know*) to see whether region appeared to provide useful information to any of these one-verb models. Because so few of the *guess* examples had a complementizer, the verb *guess* as a factor has a nearly categorical effect, thereby making it impossible for any other factors to have a significant effect. Possibly because there were fewer than 400 relevant examples, only intervening material reached significance in the model for *know* (p=.001). Because there were more than 2500 examples with the verb *think*, however, the model for this verb provided some insight into our previous results. The table below presents the results of the logistic regression model for all examples containing the matrix verb *think*:

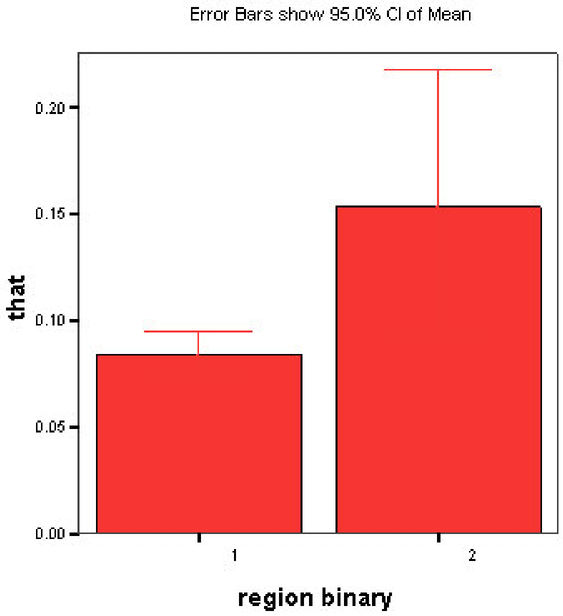
Table 5. Results of regression model including only the verb *think*

Variable		Model Log Likelihood	Change in -2 Log Likelihood	df	Sig. of the Change
Step 3	Intervening material	-801.355	52.533	1	.000
	Subject of subordinate clause	-781.623	13.068	4	.011
	Region (binary)	-777.961	5.746	1	.017

The fact that the binary region variable contributes significantly here (p=.017) but not in the medium frequency subset suggests that the effect we saw in the all-inclusive model may have been primarily due to the high frequency of the *think* examples. As the graph below shows, being from the Northeast apparently makes a person more likely to produce an overt complementizer after the verb *think*.

[8] The binned version of frequency was not included in this model because all the verbs in the model were in the same frequency bin.

Figure 1. Mean *that* by region (1= other, 2= Northeast)

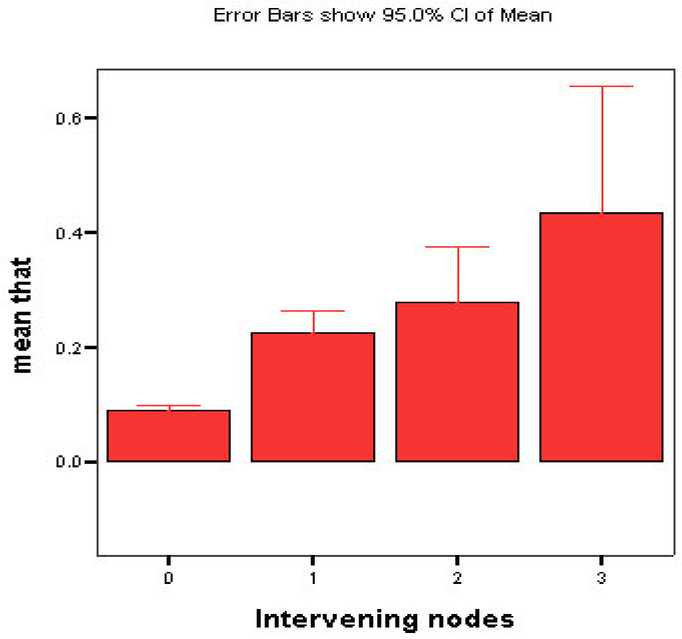


This result brings us back to Cofer's question about the relationship between individual verbs and social factors: is it possible that complementizer realization is socially conditioned at the level of the verb, not as a single variable? Another way to think of this is that it amounts to an interaction effect between a linguistic factor (verb lemma) and a non-linguistic factor (region); Sigley (2003) points out that these types of interaction effects have long been underestimated or ignored. In addition to the question of why this region effect does appear with the verb *think*, in the next section I will address the question of why other social effects do not appear.

It is worth noting that there were several factors that appeared to be significant as expected. The following linguistic factors contributed significantly to one or more of the models:

A. As expected, intervening material between the verb of the matrix clause and the subject of the complement clause increased the likelihood of an overt complementizer.

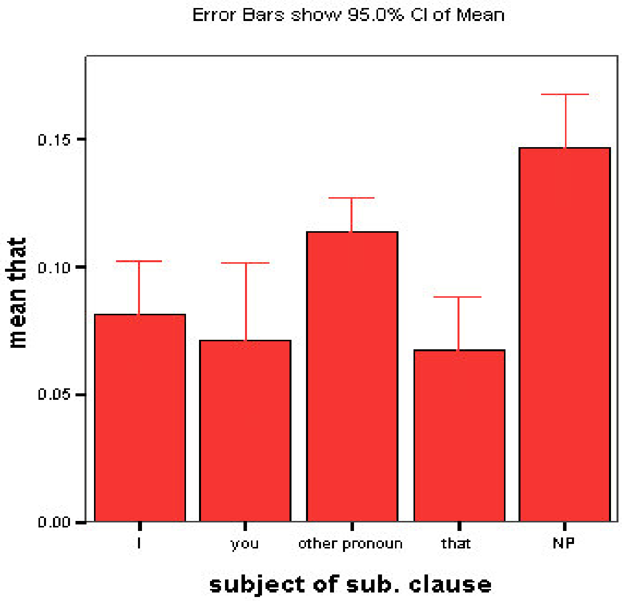
Figure 2. Mean *that* by number of intervening nodes.



It is interesting that this effect does not seem to increase linearly with the number of intervening words/phrases. There seem to be three distinct levels of intervening material: none, 1-2 nodes, and 3-4 nodes. The most significant difference, notably, is the difference between zero interveners and a non-zero number of intervening nodes.

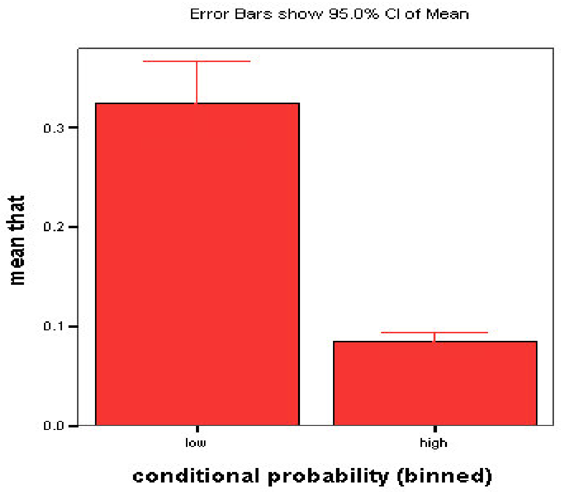
B. The subject of the complement clause also had the expected effect; as the graph below shows, the most significant difference appears between full NPs and pronominal subjects, with full NPs strongly encouraging the presence of the overt complementizer (cf Roland et al.), but there are also suggestive (but not significant) differences among the pronominal subjects, with first and second person singular pronouns and the demonstrative *that* all discouraging the presence of the overt complementizer.

Figure 3. Mean *that* by subject of subordinate clause.



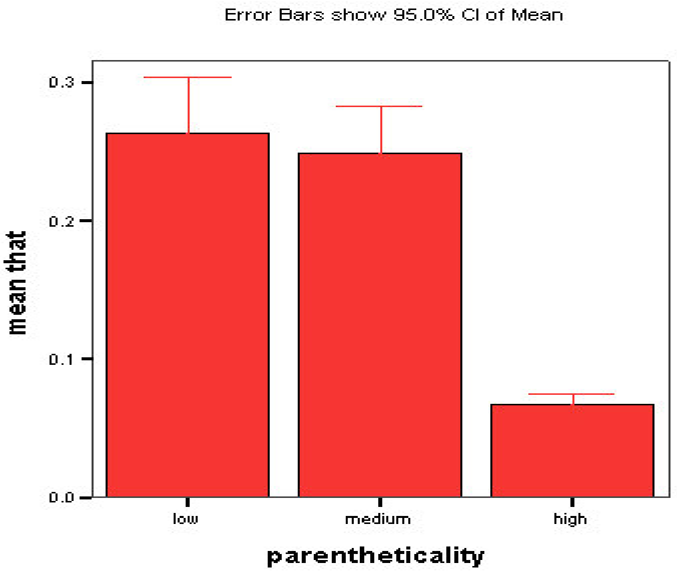
C. The results for conditional probability are also generally as expected, and may suggest some sort of ambiguity resolution. The likelihood of an overt complementizer decreased as the conditional probability increased.

Figure 4. Mean *that* by conditional probability of the clause given the verb (binned).



D. As predicted by Thompson and Mulac, as the epistemicity of the verb increases, the likelihood of the overt complementizer decreases. As the graph shows, however, it is only those examples with the highest epistemicity (those in the highest bin) that show a true decrease in overt complementizer presence.

Figure 5. Mean *that* by parentheticality of the verb (binned).



It is important to note that those examples with the highest parentheticality also have the highest frequency (these two measures are highly correlated):

Table 6. Correlation between Parentheticality and Frequency

		Parentheticality	Log of Frequency
Parentheticality	Pearson Correlation	1	.667(**)
	Sig. (2-tailed)		.000
	N	4770	4770
Log of Frequency	Pearson Correlation	.667(**)	1
	Sig. (2-tailed)	.000	
	N	4770	4770

** Correlation is significant at the 0.01 level (2-tailed).

Thus it is extremely difficult for the model to isolate the contributions of these two factors. This suggests that a measure of epistemicity that is tied to the individual verb may be flawed, and some measure of epistemicity for each individual example might help to untangle the effects of epistemicity and frequency. However, such a measure is difficult to envision for natural speech, and this question may best be investigated in a laboratory setting.

6. The Analysis

While the results for the linguistic factors did not provide many surprises, the results relating to non-linguistic factors (both what appeared to be significant and what didn't) require some explanation. There are two main questions raised by these results: first, what might explain the appearance of a regional effect within the context of the verb *think*, and second, why didn't any of the other non-linguistic variables included in the study produce an effect? Another way to phrase this second question is the following: how do we conceptualize a variable that seems to show stylistic effects without showing effects of any of the traditional global social variables?

A. The first question is raised by any positive result: why did this result obtain and what might it mean? First, for the sake of discussion, we will assume that the region effect is real and reproducible, and therefore needs some interpretation. There are a few possible interpretations of the region effect, all of which are unfortunately somewhat difficult to evaluate. One possibility is that region may interact with some other variable. For example, it's possible that New Englanders are using fewer epistemic uses of *think* than people from other regions, which would be an explanation of sorts for why they use *that* more often - if *think* has an inherently less epistemic meaning to people from the Northeast, then we would expect it to take an overt complementizer more often (according to Thompson and Mulac, creating a greater distinction between the two clauses). Unfortunately, not only is it extremely difficult to determine to what extent a given usage is epistemic, but if we choose this explanation, we are left with the question of why New Englanders would produce fewer epistemic uses, which is a similarly difficult question.

A second possibility is that the social meaning associated with the variable may differ regionally. This assumes, first, that the variation we see is motivated by social meaning, which is an idea that will be discussed at length below. Beyond that, however, it is not an overly plausible explanation for the current situation, because it would imply not only that this variable has social meaning for New Englanders (and possibly other people), but that this social meaning differs from the social meaning that people from other regions understand it to have only in the context of one verb. This is not an impossible state of affairs, but it is far-fetched enough to be an appealing possibility only with the addition of some evidence of the existence of these meanings.

It should also be mentioned here that as this effect of region was not predicted before the study was undertaken, it should have the general caveats of any post-hoc conclusion. While the effect achieves statistical significance at the traditional level of $p < 0.05$, there is always the potential for Type II (experimentwise) error, and this effect could be spurious. There is little precedent for the idea of regional or dialectal conditioning of this type of variable. Tagliamonte et al. 2005 discusses possible dialectal conditioning of relativizer variation, but provides robust evidence only in the case of WH-relativizers, not in the case of *that* vs. zero variation. Perhaps greater consideration of its possible interpretations should be given if a similar effect appears in an independent but similar set of data; this route of investigation will be discussed in the section on Further Work.

B. The second question is raised by the combination of the standard logic that relates stylistic stratification to social stratification and the state of affairs in the case of the complementizer variable. To restate slightly, the trouble we are faced with here is that stylistic stratification and social stratification are generally understood to be derived from social meaning in different ways, such that the attachment of a certain social meaning to a variable will have an effect on both what kinds of people use this variable and when/how these people choose to use it. While there have been examples in the literature of variables that show one kind of stratification and not the other, such as relativizer choice, which appears to show stylistic effects but not social ones (Sigley 1998), we lack a theory of what kinds of social meanings these variables may have such that they appear to be used differently in different styles but not by different sets of people or vice versa. Since we are faced with a variable here that appears to behave this way, we must find a way to account for this behavior. There are two possible answers to this question. First, the effect that people have attributed to style could be due to something else, or second, the style effect might not translate directly into a social effect.

The evidence on which previous researchers seem to have based their claims regarding style is also consistent with a few other stories. Specifically, most of these claims rest on a difference between speaking and writing or a difference between written text types such as academic texts vs. non-academic. Sigley 1998 points out, based on his work on relativizers, that "the simple equation of speech with informality and writing with formality yields contradictory results, demonstrating that 'channel' may act separately from 'formality'" (Sigley 1997, 30). Below we will consider how modality/channel, register, and genre may contribute to our understanding of the "stylistic" effects observed.

Register is generally understood to describe conventionalized ways of speaking/writing associated with a specific speech situation/text type. For example, there is a well-known "recipe" register, which includes features such as object-drop and a lack of articles ("Add eggs. Mix well."); these features are associated with recipes, but they are not associated with a particular level of formality. Understanding this distinction, we could ask ourselves why people might interpret a register difference as a stylistic one. One possible reason is that people may notice that in certain types of texts, e.g. academic prose, there is a higher level of *that* (this could even be based on the types of verbs used in these texts). It is then possible to make an analogy based on the fact that many aspects of academic prose are attributable to the writer's attempt to come across as educated, authoritative, etc., which leads to the interpretation of *that*-usage as one of these resources rather than an aspect of the academic register. Because this kind of ambiguity exists, it is not totally straightforward to determine which variables are stylistic resources and which are features of a register.

Nagamine 2002 claims to have found genre differences in *that* usage in relative clauses. Taking a syntactic environment where *that* /zero was the only choice, she found that zero was preferred in letters published in *The Humanist*, whereas *that* was preferred in scientific articles in *Scientific American*. These results, however, pose several problems: first, how did she determine that this was a genre effect rather than a register one? This is a thorny issue: the difference between letters and articles is a genre difference, but one of the differences between the genres could be that one has a specific register (in this case, probably the scientific articles) and the other doesn't. In addition, there could be a variety of differences between the genres involving people's rates of usage of individual sociolinguistic variants in them, and the difference in *that* /zero usage could be one of these (thereby reducing the genre difference to a stylistic difference).

After considering the possible contributions of modality, register, and genre, the only conclusion that it is possible to come to without further quantitative analysis is that these aspects of a text interact with style in complex ways, and it is quite difficult to isolate the contributions of each. That said, this suggests that one way of analyzing the behavior of a variable like complementizer realization is that there are in fact neither stylistic effects nor social effects, and what has been observed and attributed to style is rather due to one of the other text attributes discussed above. In this case, because an effect can be observed at the level of different types of "articles", for example, attributing the observed differences in complementizer usage to register seems to be the most compelling, but other variables that exhibit similar behavior may not follow this pattern and could show different behavior only for different modalities, for example.

There are a few other ways that previous researchers could have come to the conclusion that this variable was stylistically stratified without actual stylistic stratification. They could have concluded that complementizer realization is stylistically conditioned by a direct intuitive analogy – other usages that are more explicit and/or more verbose are interpreted as more formal (such as the use of “within” instead of “in”), so the knowledge that *that* is the more explicit, longer way of speaking (as opposed to the zero complementizer) could cause people to assume that it is more formal. We should also consider the possibility that this is a prescriptive issue, in that some people may have been taught to include *that* in their writing in school, presumably on the grounds of ambiguity avoidance or some similar principle. Both of these explanations, while reasonable in principle, call into question the actual observations of researchers, and since some works (for example Biber et al.) provide data to back up their claims, they are not by themselves enough to rid us of the original problem.

There is one final way in which some other effect could masquerade as a stylistic effect. Bolinger points out that “though high frequency forms are not necessarily restricted to relaxed speech, it is a fact that informal expressions tend to be high in frequency,” (Bolinger 22), which allows that the stylistic effects proposed by others may be a “reanalysis” based on frequency, with people analyzing the higher level of overt complementizer usage in formal settings as a function of the formality, rather than as a coincidental effect of the low-frequency forms. On the other hand, we should also consider that the frequency effect may really be an artifact of formality. This distinction could presumably be made reasonably well in a study that controlled both formality and frequency (though what constitutes the right kind of “formality” for this case is far from clear).

Another way that we might tease apart stylistic effects from global social effects is to problematize the relationship between formality and currency on the standard language market. We are accepting here the Labovian style continuum that defines style as level of formality, as this is most likely to be the notion of style referred to in other researchers' claims about style; however, it is not clear that this requires that all resources that are more "formal" are understood as more standard by speakers. That is to say, more formal does not necessarily mean more standard, and the mere fact that something is used in a more formal setting doesn't imply that it acquires the meaning of standardness. If this "standard" meaning does not attach to the more formal resource, then it does not necessarily acquire the bundle of meanings that tends to associate with more standard resources, such as "educated," "intelligent," "successful," etc., and it is this bundle of meanings (and people's desire to be associated with them) that is often credited with determining the social distribution of more formal variables.

One last thing that should be considered is the possibility that global social categories do indeed have an effect that simply did not show up in the current study. Adamson 1992 found that socioeconomic class correlated with zero relativizer usage, which is somewhat in conflict with the current study's assertion that education, which is often a reasonable substitute for SEC, does not have any effect whatsoever on complementizer realization. While it seems at first glance to be an indication that the requisite social meanings do exist for relativizer variation, we could analyze Adamson's finding as an effect of class-differential understandings of the interview genre. If "upper class" respondents had more experience with white collar job-type interviews, they might interpret the interview experience as a more formal event than the "working class" respondents, or they might have a more nuanced understanding of an "interview register" that includes more overt relativizers. Sigley 1998's corpus-based findings on stylistic and social effects on relative clauses support the current study in demonstrating that this type of variable does not generally show any effect of global social variables, and Adamson's findings have yet to be replicated.

7. Further Work

There are several interesting, related avenues of research that were not in the scope of this study, some of which derive directly from the conclusions of this paper. For example, on the basis of the somewhat unexpected nature of the conclusions regarding regional effects on complementizer realization, it would be interesting to use another, similar corpus to try to confirm (and possibly extend the range of) some of the current study's conclusions. Now that we know that it is possible that region might affect speakers' usage of *that* vs. zero complementizer, a study applying the same principles to another, preferably larger corpus would yield firmer conclusions. In addition, a corpus with more examples would allow us to investigate the individual verbs better – if we had a corpus the size of Roland et al.'s, each verb could have its own regression analysis as in their study. This would allow us to see if the region effect appears with any verbs other than *think*, for example. To this end, it might be possible to mine Switchboard for examples in environments ignored for this study (non-first person matrix subjects, other verb aspects/moods) to constitute a comparable corpus to the one used in this study. If this provides too few examples, the Fisher corpus could be automatically parsed (following Roland et al.) to provide additional examples. One additional benefit of including verbs in other aspects or moods would be that it appears that some verbs that were categorical predictors (e.g. *guess*) in the data collected here might exhibit more variation in a wider set of data:

> I guess (*that) he's not coming.
>
> I'm guessing (that) he's not coming.

These differences could help establish a more nuanced measure of epistemicity than the verb lemma-based one used in the current study.

Another reason that it might be useful to look at other types of *that* -complementation is that some of the stylistic/regional conditioning of these types could be different. For example, as discussed earlier, presentational type usages such as the T-i construction (e.g. *The problem is that* X) and Isis (or double *is*) examples (e.g. *The problem is is that* X) behave differently from one another regarding complementizer realization depending on their structure – the Isis examples appear to take complementizers less often than simplex examples (cf Brenier and Michaelis). Because some of these types (such as the Isis type) are not produced by all speakers, the differences produced by different choices of presentationals could cause social effects to appear for *that/*zero realization.

Also, if we analyze the regional differences that appear with the verb *think* as related to epistemicity, discourse factors, or any other factors that depend at least partially on the construction in which the complement appears, then finding (or not finding) the same differences in the context of other constructions involving that-complementation could provide some indication of which of these factors might be contributing to the effect.

Another possibly fruitful avenue for research is the analysis of examples of the following type: *I told him* ***that*** *for sure* ***that*** *Dan and I would be around.*

The existence of examples that contain more than one complementizer but only one complement raises a variety of questions, the most obvious of which is whether the conditioning of the presence/absence of additional *that*s is the same as that of the first *that*. Do the same internal and external factors identified above contribute to this phenomenon? Do the factors that encourage the presence of the first *that* also encourage the presence of additional *that*s? Also, this type of example tends to sound qualitatively different to native speakers from the way they hear examples including zero or one instance of the overt complementizer, and many speakers call them nonstandard – this suggests that even if the same internal factors encourage the presence of additional complementizers, the situation may be quite different with respect to external ones. Specifically, if it is indeed widely considered non-standard, we might expect it to be socially stratified on the basis of some stigma that is attached to it.

In addition to the interesting possibilities they present for traditional variationist study, these examples provide ample fodder for processing-based inquiry. The suggestion that the additional *that*s may appear to give the speaker additional processing time for the rest of the sentence is supported by the following example, in which one of the additional complementizers appears in a place where it cannot actually be fulfilling its presumed grammatical function of marking the following constituent as part of a verbal complement:

1. We can't say **that** for certain **that** if you don't have a complete DNA match **that** it came from a particular individual.

In this example, the first *that* is ambiguous between a demonstrative and a complementizer; without hearing the utterance spoken, it is impossible to determine which of these it is. The second *that*, however, cannot be a demonstrative, which leaves only the possibility that it is a complementizer. This is interesting because the constituent it appears to be marking, *if you don't have a complete DNA match*, does not seem to be part of the complement; that is, it seems to apply to the main clause of the sentence (*we can't say)* rather than to the subordinate clause (*it came from a particular individual*). An investigation of these multiple-that examples could provide some insight into both processing and social factors, and possibly the relationship among them.

Finally, given the knowledge that neither complementizer choice nor relativizer choice seem to be socially conditioned, it seems natural to want to look for other variables that are thought to be "stylistically" conditioned but have not been studied socially. If there is indeed a whole set of variables that exhibit stylistic differences in usage but are not conditioned by traditional social categories, it would greatly enhance our understanding of the relationship between variation and social meaning to increase our knowledge of what these variables are and how they work.

The way in which we study these variables, however, may need to be different from how traditional sociolinguistic variables are studied. That is, studies of such variables need to be contextualized better, both in the discourse sense and in the social environment. Following Campbell-Kibler's work on the role of the listener in variation, it may be useful to directly question the meaning of these variables through matched guise studies and other attitude-assessing techniques.

In addition to studying variables that, like complementizer and relativizer choice, seem to be stylistically but not socially conditioned, it would greatly enhance this research program to determine whether or not variables exist that are socially but not stylistically conditioned. If such variables do indeed exist, our understanding of social meaning needs to be drastically revised. If they do not, then some of the possible interpretations of the “style only” variables become more likely (those that provide an explanation for the “stylistic” effects that is independent of social meaning).

References

Adamson, H. D. 1992. Social and Processing Constraints on Relative Clauses. *American Speech*. 67 (2). Duke University Press.

Bakovic, Eric and Edward Keer. To appear. Optionality and Ineffability. *Optimality Theoretic Syntax*. Geraldine Legendre, Jane Grimshaw and Sten Vikener (eds.).

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. Pearson Education Limited.

Bolinger, Dwight. 1972. *That's that*. Mouton: The Hague.

Boskovic, Zeljko and Howard Lasnik. 2003. On the Distribution of Null Complementizers. *Linguistic Inquiry*: 34 (4), 527-546.

Brenier, Jason and Laura Michaelis. To appear. Optimization via Syntactic Amalgam: Syntax-Prosody Mismatch and Copula Doubling. *Corpus Linguistics and Linguistic Theory*, 1.

Cacoullos, Rena Torres and James A. Walker. 2003. *Taking a Complement… Variably*. Presented at NWAVE 32. Philadelphia, PA.

Campbell-Kibler, Kathryn. 2004. I know she can say her Gs: Conceptualizing variation in context. NWAV 33. Ann Arbor, Michigan.

Cheshire, Jenny, Paul Kerswill and Ann Williams. 2005. Manuscript. On the non-convergence of phonology, grammar, and discourse.

Cofer, Thomas Michael. 1972. *Linguistic Variability in a Philadelphia Speech Community*. University of Pennsylvania, dissertation.

Dor, Daniel. 1995. *Representations, Attitudes and Factivity Evaluations: an epistemically-based analysis of lexical selection*. PhD Thesis, Stanford University.

Elsness, J. 1984. That or Zero? A Look at the Choice of Object Clause Connective in a Corpus of American English. *English Studies* 65.

Fanego, Teresa. 1990. Finite Complement Clauses in Shakespeare's English. I. *Studia Neophilologica* 62: 3-21.

Ferreira, Victor S. and Gary S. Dell. 2000. Effect of Ambiguity and Lexical Availability on Syntactic and Lexical Production. *Cognitive Psychology* 40, 296-340.

Fowler, H. W. 1954 *A Dictionary of Modern English Usage*. Oxford P.

Francis, W. Nelson. 1958. *The Structure of American English*. The Ronald Press Company: New York.

Fries, Charles Carpenter. 1940. *American English Grammar*. D. Appleton-Century Company: New York.

Gibson, Edward. 1998. Linguistic complexity: locality of syntactic dependencies. *Cognition* 68 1-76.

Grodner, Daniel, Edward Gibson, and Susanne Tunstall. 2002. Syntactic Complexity in Ambiguity Resolution. *Journal of Memory and Language* 46, 267-295.

Hiroe, Akira. 1999. Mood and Complementizer Deletion. *English Linguistics* 16:1, 55-77.

Huddleston, Rodney and Geoffrey Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge UP.

Jaeger, T.F. and Tom Wasow. 2005. *The role of referential accessibility hierarchies in language production*. BLS 31.

Kaltenbock, Gunther. 2004. *That* or no *that*? – that is the question: on subordinator suppression in extraposed subject clauses. *Vienna Working Papers in Linguistics*. Vol. 13.1.

Kawagushi, Chihiro. 2001. The Change of Use of the *that* complementizer in *that*-Clauses during the Thirty Years between the Early 1960s and the Early 1990s – A Study Based on Brown, LOB, Frown and FLOB Corpora. *Studies in English Corpus Linguistics* 2. Edited by Toshio Saito and Yamazaki Shunji. Dept. of English, Daito Bunka University: Tokyo.

Kruisinga, E. and P.A. Erades. 1953. *An English Grammar. Volume I: Accidence and Syntax, First Part*. P. Noordhoff N.V.: Groningen.

Labov, William. 1982. *The Social Stratification of English in New York City*. Third printing. Center for Applied Linguistics: Washington, DC.

Lopez Couso, Maria Jose. 1996. A Look at *That/Zero* Variation in Restoration English. *English Historical Linguistics 1994: Papers from the 8th International Conference on English Historical Linguistics (8 ICEHL, Edinburgh, 19–23 September 1994)*. Edited by Derek Britton. John Benjamins.

McDavid, Virginia. 1964. The Alternation of "That" and Zero in Noun Clauses. *American Speech* 39 (2), 102-113.

Mizumura, Tomoko. 2001. Retention vs. Omission of the *That* Complementizer in British and American English – A Study Based on the Brown, LOB, Frown and FLOB Corpora. *Studies in English Corpus Linguistics* 2. Edited by Toshio Saito and Yamazaki Shunji. Dept. of English, Daito Bunka University: Tokyo.

Moralejo-Garate, Teresa. 2000. *That/Zero* Variation in an Early Modern English Corpus of Private and Non-Private Letters. *Southwest Journal of Linguistics*. Linguistic Association of the Southwest.

Nagamine, Toshinobu. 2002. A Preliminary Corpus-Based Study on Genre-Specific Features in Restrictive Relative Clauses. *Journal of Language and Linguistics*. Vol. 1, No. 3.

Noonan, Michael. 1985. Complementation. *Language typology and syntactic description: Complex constructions*. Edited by Timothy Shopen. Cambridge UP: London.

Olofsson, Arne. 1981. *Relative Junctions in Written American English*. Acta Universitatis Gothoburgensis: Goteborg, Sweden.

Otsu, Norihiko. 2002. On the Absence of the Conjunction *That* in Late Middle English. *Language and Computers*. 38 (1), 225-234.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1972. *A Grammar of Contemporary English*. Seminar Press: New York.

Ransom, Evelyn. 1986. *Complementation: Its Meanings and Forms*. John Benjamins: Amsterdam.

Rissanen, Matti. 1991. On the history of *that*/zero as object clause links in English. *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. Edited by Aijmer, K. and B. Altenberg. 272-289. London: Longman.

Roland, Douglas W., Jeffrey L. Elman, and Victor S. Ferreira. 2003. Why *that*? Predicting *that* presence and ambiguity resolution. Poster Presentation, Architectures and Mechanisms for Language Processing 2003, Glasgow, Scotland.

--. In press. Why is *that*? Structural prediction and ambiguity resolution in a very large corpus of English sentences. *Cognition*.

Sigley, Robert. 1997. The influence of formality and channel on relative pronoun choice in New Zealand English. *English Language and Linguistics*. Vol. 1, No. 2. 207-232.

--. 1998. Interpreting social variation using stylistic continua: the strange case of relativiser choice. *Wellington Working Papers in Linguistics*. Vol. 10. 64-103.

--. 2003. The importance of **interaction** effects. *Language Variation and Change*. Vol. 15, 227-253.

Tagliamonte, Sali, Helen Lawrence, and Jennifer Smith. 2003. *That* or no *that* in English dialect corpora: Grammaticalization, frequency and complexity in the emergence of grammar. Presented at NWAVE 32 in Philadelphia, PA.

Tagliamonte, Sali, Jennifer Smith and Helen Lawrence. 2005. No taming the vernacular! Insights from the relatives in northern Britain. *Language Variation and Change* 17 (1). 75- 112.

Thompson, Sandra A. and Anthony Mulac. 1991. The discourse conditions for the use of the complementizer *that* in conversational English. *Journal of Pragmatics* 15 237-251.

Walter, Mary Ann and T. Florian Jaeger. To appear. Constraints on complementizer/relativizer drop: A strong lexical OCP effect of *that*. *Proceedings of the 41st Annual Meeting of the Chicago Linguistics Society.*

Yaguchi, Michiko. 2001. The function of the non-deictic *that* in English. *Journal of Pragmatics* 33, 1125-1155.

Zwicky, Ann D. and Arnold M. Zwicky. 1986. The Thing Is, Some That's Aren't There at All. *American Speech* 61 (2) 182-183.

Appendix

code	Verb	Swbd	Swbd lemma	BNC	Parentheticals
A	agree	17	264	23497	0
B	assume	12	33	11044	0
C	believe	69	279	34603	3
D	bet	39	118	2292	0
E	decide	20	187	24380	0
F	doubt	3	14	2565	0
G	expect	2	94	27221	0
H	feel	49	692	62185	1
I	figure	28	139	3301	0
J	find	53	680	98899	0
K	guess	955	1561	3920	92
L	hear	30	587	36575	0
M	hope	56	122	21763	0
N	imagine	20	122	8300	1
O	know	386	495	185534	4
P	mean	8	2200	66556	1
Q	note	2	5	6405	0
R	notice	13	97	9663	0
S	would rather	2	149	42341 (adv)	0
T	read	5	561	28216	0
U	realize	15	118	5849	0
V	remember	34	317	26748	1
W	say	23	1977	333518	0
X	see	10	2029	191661	1
Y	suppose	36	258	14482	1
Z	suspect	13	24	3983	0
1	swear	4	11	2290	0
2	take (it)	6	1292	179220 (take)	0
3	tell	15	535	77245	0
4	thank	2	38	13531	0
5	think	2747	5355	153881	196
6	understand	13	236	24252	0
7	wish	81	116	16647	0
8	recognize	2	26	9316	0