On the persistence of grammar in discourse formulas: a variationist study of *that**

RENA TORRES CACOULLOS AND JAMES A. WALKER

Abstract

This article provides evidence that, just as lexical meaning is retained in grammaticization, grammatical conditioning persists in fixed discourse formulas. Despite their high frequency and formulaic status, such formulas are not completely autonomous from the productive constructions from which they emerge. This evidence comes from a variationist analysis of that *and zero complementizer in a corpus of spoken Canadian English. Testing syntactic, semantic, and discourse-pragmatic factors proposed to account for the variation, we focus on claims that frequent collocations have developed as discourse formulas. Multivariate analysis shows that, although the variation is largely lexically constrained,* that *serves to demarcate the boundaries of two clauses with lexical content, while zero tends to occur when the clauses function like a single unit. Moreover, the linguistic conditioning of* that *in frequent collocations that behave like discourse formulas parallels its conditioning in the general construction. These findings suggest that the principle of semantic retention or persistence should be extended to grammar.*

1. Introduction

Although grammar and the lexicon have been considered discrete modules, and language change viewed as abrupt, the principle of *persistence* (Hopper 1991) or *retention* (Bybee and Pagliuca 1987) predicts that grammatical forms will retain features or nuances of meaning from their lexical source. Studies have demonstrated that lexical and grammatical uses of the same form coexist synchronically and, in particular, that the lexical history of a form is manifested in its current patterns of distribution (see, e.g., Schwenter 1994; Poplack and Tagliamonte 1996, 1999; Torres Cacoullos 2001; cf. Hopper and Traugott 2003: 3). Beyond

Linguistics 47–1 (2009), 1–43
DOI 10.1515/LING.2009.001

0024–3949/09/0047–0001

grammaticization, recent literature has identified general frequency effects in language change (e.g., Bybee and Thompson 1997). *Autonomy* (Bybee 2003, 2006; cf. Hopper and Traugott 2003: 127) applies to collocations becoming units: as particular instances of constructions increase in frequency and develop into new linguistic resources, they lose internal structure and become disassociated from their cognates.

This paper extends the principle of persistence: not only are semantic nuances of the lexical source retained in grammatical morphemes (Bybee et al. 1994: 16), but discourse formulas also retain traces of their erstwhile grammatical conditioning. While the literature on grammaticization has focused on SEMANTIC retention, we propose GRAMMATICAL retention, in that vestiges of the linguistic conditioning of productive grammatical constructions persist even in fixed discourse formulas originating from those constructions. In other words, constraints on the source construction persist in the discourse formula.

The focus of this study is the variable use of the complementizer *that* to introduce a subordinate clause, illustrated in (1) and (2), which is a widely attested feature of all varieties of English, both synchronically and diachronically.

(1) a. And I let it slip *that* Darth Vader was Luke's father.
(071.468)[1]
b. I can't even believe *Ø* I just said that.
(059.1840)

(2) a. She said *that* her father was the rector of St. Michael's Church.
(003.162)
b. She said *Ø* she used to play in- in Sillery- in the Brewar swamp.
(003.164)

This variability has received attention in studies within different frameworks and from different approaches (e.g., Bolinger 1972; Finegan and Biber 1995, 2001; Doherty 2000; Ferreira and Dell 2000), which have uncovered a number of factors conditioning the variation. The most intriguing proposal, in our view, is that certain frequent collocations of main-clause subjects and verbs, such as *I think* and *I guess*, have become conventionalized (or "grammaticized") as discourse formulas that function more as epistemic or evidential adverbial phrases than as main-clause propositions (Thompson and Mulac 1991a, 1991b; Traugott 1995; Aijmer 1997; Diessel and Tomasello 2001; Thompson 2002). However, we still lack empirical tests: that is, we do not know how the patterning of fixed (discourse) formulas differs quantitatively from that of productive complement-taking predicates.

In any account of complementation with or without *that*, variation remains an unresolved issue, as the examples in (1) and (2) demonstrate. In light of competing accounts, the question we ask in this study is: how much of *that*-variation is lexically constrained, and how much of it is grammatically productive? If it is productive, how do we disentangle the effects of syntactic structure, semantics and discourse or pragmatics? Answering these questions requires not only observing rates of *that* (as in other corpus-based studies), but also, crucially, using multivariate analysis to extricate the linguistic factors conditioning the variation and gauge their relative magnitude of effect. In this paper, we perform such an analysis on a corpus of spoken Canadian English.

We begin by reviewing the diachronic trajectory of the complementizer *that* and its treatment in the prescriptive grammatical tradition and contemporary linguistic approaches in Section 2. In Section 3, we discuss the data and explain the variationist method, detailing our classification and coding of the data to test the competing claims of different accounts via multivariate analysis. Section 4 presents the general results: while *that*-variation is lexically constrained, nevertheless *that* serves to demarcate two clauses which each have lexical content. Section 5 focuses on frequent collocations of first-person subjects and verbs (such as *I think* and *I guess*), the prime candidates for discourse formulas. Multivariate analysis of these collocations shows that despite slight rates of *that*, the linguistic conditioning parallels the general construction.

Thus, frequent collocations retain traces of the conditioning of their source construction. Although discourse formulas develop their own discourse-pragmatic characteristics, their autonomy is incomplete. Our findings support the view of a dynamic relationship between fixed formulas on the one hand and productive constructions on the other, or a gradual rather than discrete relationship among discourse, the lexicon and grammar (cf. Hopper 1987, 1998; Bybee 1988).

2. What's *that*?

2.1. *The downs and ups of* that

Historical linguists generally agree that the complementizer *that* originated in Old English (OE) from a neuter demonstrative pronoun (e.g., Gorrell 1895; Ellinger 1933; Mitchell 1985; van Gelderen 2004), apparently by the reanalysis of juxtaposed clauses ("parataxis") as subordinate ("hypotaxis"). For example, in a construction such as (3), the demonstrative object of the first clause, which refers to the second clause, would have been reinterpreted as introducing a subordinate proposition (Hopper and Traugott 2003: 190–194; Harris and Campbell 1995: 287–289; Heine and Kuteva 2002). In OE, *þæt* was one of a number of complementizers, often occurring alongside the subordinating particle *þe* (Mitchell 1985: 13; Moulton 2002; van Gelderen 2004).

(3) *He geseah þæt. Hit wæs god.*
'He saw that: It was good.'
(Mitchell 1985: 13)

The origin of the zero complementizer is more controversial. Jespersen (1967: 76–77) attributes it to Scandinavian influence, but Kirch (1959) provides evidence of its existence in early OE. As Rissanen (1991) notes, the existence of both *that* and zero in the earliest written texts makes it inaccurate to speak of "deletion". However, the use of the zero complementizer has not remained constant across time. Studies of OE and early ME report very few occurrences (Ogura 1979; Rissanen 1991), and only in late ME did it begin to occur with any frequency, though even here its rate of occurrence is very low. For example, Warner's (1982) analysis of the Wyclifite sermons (written ca. 1390) reports a rate of only 3.6% (9.3% if ambiguous cases are included). Not until late ME and Early Modern English (EModE) did the zero complementizer become more frequent. Fanego (1990: 5) reports a rate as high as 78% for three of Shakespeare's (1564–1616) plays, and López Couso (1996: 272) reports 63% for Dryden's (1631–1700) prose. Figure 1 summarizes the results of a number of corpus-based studies of late ME and EmodE. As the gray line (which indicates the average within each time period) shows, zero complementizer increased steadily between the 14th and 17th centuries, then reversed course in the 18th century (Rissanen 1991; Palander-Collin 1999; Finegan and Biber 1985; Suárez Gómez 2000).

Interestingly, this reversal coincides with the rise of prescriptivism in the English-speaking world (Rissanen 1991; Haugland 1995; Bex and Watts 1999; Milroy and Milroy 1999; Wright 2000), suggesting that the strong prescriptive pressures that began to apply in written English in this period may have been reflected in a decrease of zero complementizer in writing.[2]

Any generalization based on the results shown in Figure 1, however, should be taken with caution. As the graph shows, rates of zero vary widely, not only from time period to time period, but also within the same time period according to corpus and genre or register. In Rissanen's (1991) finer breakdown of the frequencies (not shown in Figure 1), zero occurs more frequently in speech-like genres (e.g., trial records, sermons, fiction, comedies) than in institutional genres (e.g., legal and educational

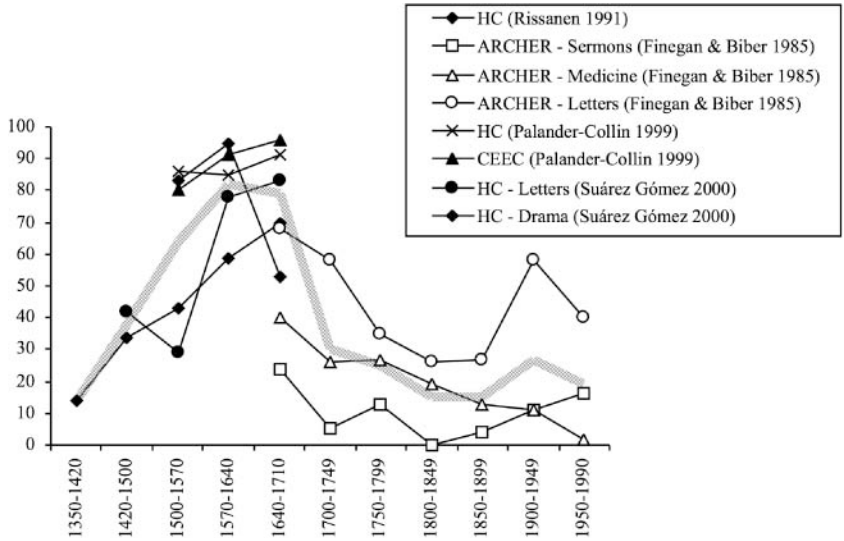

Figure 1. *Rate of zero complementizer across selected historical corpora, 1350–1990. (Corpora include the Helsinki Corpus [HC], the Corpus of Early English Correspondence [CEEC] and A Representative Corpus of Historical English Registers [ARCHER].)*† *The gray shaded line indicates the average for each time-period*

†Finegan and Biber's (1985: 282) 1650–1699 period has been subsumed under the Helsinki Corpus EmodE2 (1640–1710) period. Palander-Collin's (1999: 207) percentages have been recalculated from her figures for zero vs. *that*.

writing). Similarly, in Finegan and Biber's (1985) analysis of the ARCHER corpus (shown on the right side of Figure 1), formal registers (such as sermons and writings on medicine) have strongly preferred *that* since 1650, while the more informal register of personal letters prefer zero until 1750, after which they prefer *that*. Thus, it appears that there has always been a great deal of variability in the use of zero, with register or genre a primary consideration.[3]

Another consideration which vitiates any generalization based on Figure 1 is the different sampling criteria applied in each of the studies.[4] For example, Rissanen (1991) considers only a small number of main-clause verbs which admit variation (*know*, *think*, *say*, *tell*), whereas other authors consider a wider number of main-clause verbs and adjectives. Furthermore, Finegan and Biber (1985: 252) note that the individual styles of particular authors can affect the overall frequency of zero. Such differences may be responsible for the widely different rates of zero within each time period, and even within the same corpus.

2.2. *The prescriptivist grammatical tradition and* that

The complementizer *that* has attracted little censure from traditional grammarians, as indicated by Kirkby's (1746: 126) remark that "The Adverb *that* is often understood; as *I beg* (that) *you wou'd tell him*."[5] As we saw above, a principal consideration is the difference between written and oral language, or genre and register. For example, Bayly (1772: 89–90) accepts that "*That* may be omitted in the subjunctive mood on some occasions, as 'I desire, he may, you, he would come; see, thou tell no man'" (where presumably the subordinate verbal morphology signals the relationship between the two clauses). Nevertheless, he finds it "improper to omit it in writing before the indicative or infinitive, as it often is in discourse, because it makes the parts of speech equivocal; as, I am glad you are come".

The grammarians' main argument for retaining *that* is to ensure clarity. Lowth (1762: 147) and Murray (1795: 133) both disapprove of Bacon's sentence "it is reason the memory of their virtues remain to posterity", and Ash (1763: 128) and Brittain (1788: 83) object to similarly complex sentences where the relationship between clauses is not overtly signaled. In contrast, the sentences where grammarians allow omission of *that* are relatively simple syntactically, as in Murray's (1795: 133) example *See thou do it*. This concern for clarity foreshadows recent work on ambiguity avoidance (Temperley 2003). Similarly, the grammarians presage modern analyses of complement-taking predicates as discourse-pragmatic markers: their example sentences without *that* usually involve performatives (4a) or first-person subjects with epistemic verbs (4b)–(4c).

(4) a. *I beg ye would come*
(Murray 1795: 133)
b. *I know it was*
(Elphinston 1765: II.27)
c. *I think you are mistaken*
(Hodgson 1763: 157)

2.3. *Contemporary linguistic analysis*

In contemporary linguistics, complementation constitutes a prolific area of research in both formal and functionalist approaches, the subject of numerous articles, monographs, and edited volumes (e.g., Rosenbaum 1967; Bresnan 1972; Noonan 1985; Horie 2000). There is fairly uniform agreement that complement clauses are clausal arguments of predicates

(Noonan 1985: 42). Complementation in English occurs in a range of syntactic forms: predicate adjuncts (5a), *-ing* forms or bare infinitives (5b)–(5c), *to*-infinitives (5d), and *that* clauses (5e) (Verspoor 2000: 206).[6]

(5) a. I find [the room cold].
b. I saw [her going].
c. I saw [her go].
d. I want [to go].
e. I knew [that she was going].

Complementation with *that* (5e) has been studied as a formal syntactic phenomenon since the early years of generative grammar (e.g., Rosenbaum 1967; Bresnan 1972; Stowell 1981; Kayne 1981), though the alternation between *that* and zero has received little attention until relatively recently. In some accounts, *that* indicates the presence of a Complementizer Phrase (CP) and its absence indicates an Inflectional Phrase (IP) (Doherty 2000; see also Hudson 1995), while in other accounts the complementizer is always present but may not be pronounced (Pesetsky 1998). In any case, the thorny issue of "optionality", or variability, remains unresolved (see also Grimshaw 1997).

Quantitative studies of spoken and written data have converged on a number of factors influencing the variation, though each study tends to privilege syntactic, semantic, or discourse-pragmatic explanations.

The use of *that* has been correlated with syntactic complexity, either in terms of the size or "weight" of constituents (e.g., Elsness 1984; Rohdenburg 1996: 163–164) or the presence of material intervening between the two clauses (e.g., Warner 1982; Rissanen 1991: 279; Finegan and Biber 1985: 253–254; Rohdenburg 1996: 160–161; Suárez Gómez 2000: 194). In fact, virtually every study identifies the latter factor, such that intervening material favors *that* in order to "preserve the identity of the clause" (Bolinger 1972: 38; cf., e.g., Warner 1982; Elsness 1984: 524; Thompson and Mulac 1991b: 247; Yaguchi 2001: 1147). Complexity effects have been attributed to ease of parsing and/or the avoidance of ambiguity (e.g., Rissanen 1991: 286–287; Rohdenburg 1996: 160; Temperley 2003), though they may also support a discourse-pragmatic explanation, since syntactic complexity detracts from formulaic use.

Different complement structures have also been argued to correspond to different semantic nuances (e.g., Wierzbicka 1988). In *that*-variation, the main semantic consideration is semantic proximity or congruence. For example, Bolinger (1972: 38) suggests that *that* is more likely when the polarity (negative or affirmative) of the two clauses "clashes" than when they agree in polarity.[7] Some have attributed to *that* a conceptual distance-marking meaning, reflecting either the strength of the main verb on a hierarchy of clause-binding (Givón 1980) or the iconic separation of the two clauses (Langacker 1991; Givón 1995). In Givón's (1980: 358, 368–370) binding hierarchy, which includes overlapping scales of success vs. nonsuccess (implicativity), strong vs. weak emotion, and epistemic attitude, the complementizer encodes lower rungs of the hierarchy. In other words, *that* encodes the semantic independence of the complement clause (Givón 1980: 371). For example, in (6a), Event 1 is cotemporal with and implicates Event 2, while in (6b) Event 1 does not implicate Event 2 (Givón 1993: 2, 27; cf. Verspoor 2000: 202–203).

(6) a. [She made]$_{\text{Event 1}}$ [him leave]$_{\text{Event 2}}$.
b. [She told him]$_{\text{Event 1}}$ [*that* he should leave]$_{\text{Event 2}}$.

Langacker (1991: 447), in turn, proposes that complementizer *that* "serves to objectify" (in terms of Cognitive Grammar) the conception of the proposition by structurally distancing the clauses. For example, he argues (Langacker 1991: 450) that (7a) is more likely if Susan learned the results of consumer reaction tests, whereas (7b) implies that she tried the bed herself.

(7) a. Susan found [*that* the bed was uncomfortable].
b. Susan found [the bed uncomfortable].

Langacker's proposal resembles Givón's argument for greater semantic distance in *that* complements, as well as more general semantic analyses such as Dor (2005: 375–377) (see also Verspoor 2000).

Discourse explanations center on two (not unrelated) lines of argumentation: considerations of information flow and the development of frequent collocations as discourse-pragmatic formulas. Explanations based on information flow, which has to do with the status of discourse entities as "*given* or *new*, *thematic* or *topical*, *foregrounded* or *backgrounded*, and the like" (Chafe 1992: 215), begin with work by Bolinger (1972),[8] who proposes that *that* has an anaphoric-deictic meaning, stemming from its origins as a demonstrative pronoun (cf. Yaguchi 2001). Under this view, *that* indicates that the content of the complement clause is known or given. Along the same lines, Underhill's (1988, cited in Thompson and Mulac 1991a, 1991b) study of newspaper articles reports that *that* is absent when the writer asserts the complement clause and is present when the writer doubts, denies, or is noncommittal about it. Underhill also more explicitly relates *that* to the relationship between clauses, suggesting that *that* is less likely to occur when the subject of the complement clause is the topic of the utterance. This observation is relevant to the hypothesis of discourse formulas, the focus of this study.

It is well known that particular subject-verb combinations, such as *you know* and *I mean*, function in conversation more as pragmatic or discourse markers than as main clauses (cf. Schiffrin 1987). Such constructions may occur in positions other than before, or even in the vicinity of, a likely complement clause, usually constitute a separate phonological or intonational phrase, and need not be interpreted literally (Quirk et al. 1985: 1112; Thompson 2002: 143–145). For example, the expression *I don't know* has a range of pragmatic functions beyond literally not knowing (Scheibman 2000). Thompson and Mulac (1991a, 1991b) propose that certain main-clause subject-verb collocations, in particular *I think* and *I guess*, have been reanalyzed or conventionalized as epistemic parentheticals (see also Traugott 1995: 38–39; Aijmer 1997; Thompson and Hopper 2001: 38–39). Diessel and Tomasello (2001) call them performative and formulaic uses. Thompson (2002) argues further that such structures involve not two clauses but rather combinations of fragments which serve as frames of speaker stance with finite indicative clauses: that is, certain complement-taking predicates are really modals and epistemic/evidential adverbs, while their putative subordinate clauses carry out "the work that the utterance is doing" (Thompson 2002: 155). In her view, such "complement constructions" are in fact single clauses that include a speaker-stance frame (Thompson 2002: 142).

The hypothesis of discourse formulas predicts that frequent collocations of a main-clause first- (or second-) person subject and an epistemic verb should show far lower rates of *that* than less frequent (but more grammatically productive) complement-taking predicates. Furthermore, since such frequent collocations have been conventionalized as formulas, not only should *that*-variation be much more limited (if it exists at all), but the linguistic factors conditioning the variation should also differ from that of complement-taking predicates in general. In other words, we expect the general patterns affecting *that*-complement structures not to apply to fixed discourse formulas.

Thus, we set ourselves two tasks: first, to determine which of the proposed constraints affect the use of *that*; and, second, to determine whether these constraints operate the same way on frequent subject-verb collocations.

3. *That*: a variationist approach

The variationist method reflects a "scientific interest in accounting for grammatical structure IN DISCOURSE" and a "preoccupation with the polyvalence and apparent instability IN DISCOURSE of linguistic form-function relationships" (Sankoff 1988a: 141, emphasis in original). Only by examining corpora of spoken discourse can we observe patterns of behavior, reflected in FREQUENCIES of (co)occurrence rather than simple presence or absence (Sankoff 1988a: 141). Our interest lies not in "grammaticality" or "possibility" but in likelihood of occurrence in actual language use. In this view, grammatical structure is manifested in recurrent patterns, or "a series of parallel occurrences (established according to structural and/or functional criteria) occurring at a non-negligible rate in a corpus of language use" (Poplack and Meechan 1998: 129; see also Poplack and Tagliamonte 2001: 89).

For variable features, the patterns of speaker choices in discourse are largely inaccessible to introspection or psycholinguistic testing (Sankoff 1988a). Since the attribution of intention to the speaker by the analyst (or even by the speaker) may be an *a posteriori* artifact of theoretical bias or normative pressures, the only access we have to the speaker's intention in the choices of different forms is through their speech (Sankoff 1988a: 154; Poplack and Tagliamonte 1999: 321–322; cf. DuBois 1987: 811–812). As Labov (1972: 199) notes, "intuition is less regular and more difficult to interpret" than the data of "unreflecting" speech production. Thus, the variationist method requires large samples of unreflecting speech: that is, recordings of lengthy conversations, preferably between speakers of the same vernacular, or recordings of natural interactions, rather than formal interviews (see, e.g., Labov 1984).

3.1. *Data*

The data for the present study were taken from recordings of naturalistic conversations conducted with native speakers of English in Quebec City in 2002, part of a larger project investigating grammatical variation and change in the English spoken in the Canadian province of Quebec (Poplack et al. 2006). These recordings were transcribed in several passes to faithfully render the informants' speech (cf. Poplack 1989; Poplack et al. 2006). Selecting a representative sample of 34 informants, we worked with over 70 hours of recorded speech, corresponding to 870,484 words of running transcribed text.

3.2. *Variable context*

The first step in the analysis involves circumscribing the "variable context": that is, the place in discourse where the speaker has a choice between variant forms (Labov 1972; Guy 1993; Poplack and Tagliamonte 2001: 90), in this case, between *that* and zero. For the purposes of this study, we define the variable context as finite declarative complements in object position, including not only complements of main-clause verbs, as in (1) and (2), but also objects of predicate adjectives (8a), "extraposed subject" clauses (or "predicate nominals") (8b) and "extraposed" clauses with *it* (8c).

(8) a. We *were delighted Ø* we didn't have to go to school anymore.
(006.123)
b. It*'s sad* that the New York skyline is destroyed.
(044.1449)
c. *It seems* to me *Ø* it was every night.
(003.1045)

We note, however, that these last three types do not comprise a substantial portion of the data. Predicate adjectives account for 90 tokens (overwhelmingly (*I'm*) *sure* (N = 54), with the remainder distributed among 14 other lexical types). There are 27 predicate nominal tokens, spread over 16 different lexical types. Extraposed clauses add up to 30 occurrences, mostly *it seems* and *it turns out*.[9]

Since we have limited the study to complements in object position, we exclude appositives (e.g., *the fact that*...), relatives (e.g., *I have a cat that* ...), and adverbial conjunctions (e.g., *so that*...). We also exclude a number of contexts that do not admit variation, including medial or final parentheticals (9a), quotative verbs introducing direct speech (9b), and false starts or hesitations (9c). For established discourse markers *you know* and *I mean*, for which (near-) categorical rates of zero have been reported (99% [535/541] and 100% [428/428], respectively [Tagliamonte and Smith 2005: 299]), we exclude all medial (9d) and virtually all sentence-initial occurrences (discourse-marker status was flagged in the transcription, based on prosodic cues), though we retain a handful of cases (9e). These protocols yielded a total dataset of 2,820 tokens.

(9) a. Oh, reading, *I think*, is very important.
(007.1200)
b. ... stuff I can look back on and *say*, "Hey, that was like that."
(004.534)
c. But uh- the- it was really nice *that* — it's nice to see when they're you know, they're all on your back.
(007.410)
d. That started everything, like, *you know*, the boys went overseas.
(002.110)

e. But you know *that* they're not from here.
(007.1589)

3.3. *Operationalizing hypotheses: coding of tokens*

For each token, we noted whether *that* was present or absent (the dependent variable) and coded for a number of "factor groups" (independent variables), each of which tests a hypothesis or finding reported in the literature on *that*-variation, as discussed above. In this study, we examine the role of the following considerations:

- frequency and semantic class;
- semantic proximity (operationalized as subject coreferentiality, event cotemporality, and harmony in polarity);
- subjects;
- intervening material; and
- other measures of complexity (operationalized as adverbial modification, verbal morphology, and transitivity).

3.3.1. *Frequency and semantic class.* The role of frequency in linguistic variation and change has received increasing attention of late (e.g., Haiman 1994; Bybee and Thompson 1997; Bybee 2001, 2003; papers in Bybee and Hopper 2001). In particular, frequency has been argued to propel the reduction to discourse formulas of collocations such as *I think* and *I guess* (Thompson 2002: 140; cf. Bolinger 1972: 22). Our initial measurement of frequency is the frequency of the lexical type in the data, for example, the token count of all forms of *think*. Lexical types were coded as high, medium, or low frequency.[10]

The classification for semantic class follows that of Noonan (1985: 110–133) and Quirk et al. (1985: 1180–1183). "Propositional attitude" predicates include *assume*, *be certain*, *believe*, *doubt*, *suppose*, and *think*. Verbs such as *ask*, *promise*, *report*, *say*, and *tell* were coded as "utterance" verbs. "Knowledge" predicates, such as *discover*, *dream*, *find out*, *know*, and *realize* (as well as *see* and *hear* when not used perceptually) correspond to Terrell and Hooper's (1974) semifactives. "Commentatives" (or factives), such as *be ironic*, *be rare*, *be sad*, and *be sorry*, express emotional reaction or evaluation. Although English makes little or no use of the indicative/subjunctive distinction, in languages in which this distinction reflects degree of assertion, the complements of commentatives appear in the subjunctive. "Suasive" predicates such as *make sure*, *decide*, and *suggest* are said to be followed by a mandative subjunctive complement (10a) or to commonly occur with an NP + infinitive complement

(10b). Finally, we combined extraposed-subject constructions involving verbs such as *appear*, *happen*, *seem* and *turn out*.

(10) a. They suggested [*that* he buy a farm].
(002.724)
b. They intended [the news to be suppressed].
(Quirk et al. 1985: 1182–1183)

The hypothesis tested by the semantic class coding is that attitude predicates are most likely to be used parenthetically to express degrees of commitment to the proposition in the complement clause and to develop as stance markers and hence should disfavor *that* (cf. Noonan 1985: 114). More generally, the semantic class factor group tests semantic accounts of *that*-variation (e.g., Dor 2005).

3.3.2. *Semantic proximity.* Three factor groups attempt to test the notion of semantic "proximity" between the main and subordinate clauses.

First, we examine the coreferentiality of main-clause and subordinate-clause subjects, as in (11), to investigate findings that *that* is disfavored if the two subjects refer to the same entity (Elsness 1984: 526; Finegan and Biber 2001: 262).

(11) a. My dad$_i$ said Ø he$_i$ kept certain animals there.
(061.279)
b. We$_i$ thought *that* maybe they$_j$ found him dead.
(025.1097)

A second measure of semantic distance is cotemporality, since we reason that events further removed in time (12a) are conceptually more distant than cotemporal events (12b).

(12) a. My little brother insists *that* he just never wanted to go back reffing.
(050.1131)
b. The doors were closed until Dad said Ø it was okay to come in.
(071.1179)

Finally, we coded for whether the clauses agreed (13a) or differed (13b) in polarity (cf. Bolinger 1972: 32).

(13) a. Everyone thinks Ø I'm from Montreal.
(067.1778)
b. Anybody that comes here knows *that* I don't speak it.
(057.1408)

If the semantic proximity hypothesis is correct, non-coreferential subjects, non-cotemporal events, and clauses which lack harmony in polarity should all favor *that*.

3.3.3. *Subjects.* We coded the subject of both the main and complement clauses. Thompson and Mulac (1991b: 242) find that *I* and *you* (14a) disfavor *that* more than other main-clause subjects (14b), which they explain by the higher frequency of *I* and *you* in discourse and their capacity to express epistemicity or subjectivity.

(14) a. I'm like, "Well, how do you know *Ø* I'm from Canada?"
(067.1706)
b. Less than a week, the teacher realized *that* I was reading English at the same level as the other kids.
(071.248)

In the complement clause, full NP subjects (15a) have been found to favor *that* more than pronominal subjects. Elsness (1984: 527) attributes this effect to "a desire [for] syntactic clarity", while Thompson and Mulac (1991b: 248; Thompson 2002: 155) attribute it to the high discourse topicality of pronouns: the use of the erstwhile main clause as a stance marker correlates with the topicality of the subject of the complement, which is the clause that carries the burden of the utterance. Among non-NP subjects, we distinguished *I* (15b), other pronouns (15c), and expletive *it* or *there* (15d).[11]

(15) a. Some people say *that* their French culture is endangered.
(071.1326)
b. And she- she thinks *Ø* I'm such a baby.
(066.792)
c. And she opens it up and she sees *Ø* it's a cell phone.
(058.980)
d. I would say *Ø* it was very French Canadian.
(004.715)

3.3.4. *Intervening material.* Two factor groups measure the effect of material intervening between the clauses.

First, intervening elements may consist of verbal arguments, such as indirect objects (whether pronominal [16a] or NP [16b]) and prepositional phrases (16c).

(16) a. I told him *Ø* if they're going to Saskatchewan, I will never visit them.
(059.1707)

b. When I went to tell Michael *Ø* I was getting married.
(006.414)
c. I've just announced to the entire world *that* if anybody wants to buy Dad's car for over five hundred dollars it'd be a good deal, you know.
(058.1607)

Other intervening elements include single-word or phrasal adverbials (17a)–(17b), clauses (17c), parentheticals, hesitations, and fillers (17d)–(17e), or combinations of the above (17f).

(17) a. I expected maybe *that* we would be talking about it.
(009.490)
b. I remember *Ø* on Saturday, my mother used to- I used to go with my mother to uh- Morrissette's- Morrissette's in Sillery.
(002.1297)
c. Uhm, I remember *Ø* when I was in secondary-two, I thought it was hell.
(009.298)
d. But I knew, *Ø* you know, there wasn't much going on with CTV.
(004.33)
e. Some people do not believe *that* uh- that we are more or less bilingual here.
(002.11)
f. We put a thermometer in h- in him so that we make sure *that*, you know, when he's older, he'll get used to having a thermometer.
(009.1006)

These factor groups test the hypothesis of adjacency, which predicts that *that* will be favored when material intervenes between the main and subordinate clauses.

3.3.5. *Other measures of syntactic complexity.* Besides the abovementioned factor groups of intervening material and subjects, we coded for three other considerations which might increase the syntactic complexity of the clauses.

First is the presence of adverbial modification in the main clause (cf. Thompson and Mulac 1991b: 247). We distinguished between phrasal adverbials (18a) and adverbials which preceded (18b) or followed (18c) the subject.

(18) a. At the beginning, we told the guy *that* we were gonna both-each have our own.
(061.1469)
b. Now I find Ø like, even adults use slang words.
(067.2007)
c. I totally thought Ø he was a big jerk.
(048.793)

Second, syntactic complexity may also be implicated in the morphosyntactic realization of the main-clause predicate. Periphrastic and nonfinite forms (19a)–(19b) should favor *that*, either because both are more likely to detract from the treatment of the subject-verb as a unitary formula than simple forms (19c)–(19d) (cf. Thompson and Mulac 1991b: 246), or because they involve more syntactic complexity.[12]

(19) a. You would tell Ø they were really darn happy to leave.
(023.1774)
b. Somebody's either gonna go there and tell them *that* the police are in- in the village.
(023.2188)
c. Yeah, I already told her Ø I was going.
(059.1261)
d. But they tell me *that* I speak differently.
(013.1106)

Similarly, periphrastic complement-clause verbs (20a) should favor *that* more than simple indicative affirmatives (20b). We single out modals, such as *can*, *should* and *must* (20c), because modals not only involve greater syntactic complexity, but may also signal greater semantic independence of the clauses (cf. Givón 1980), and hence should favor *that* over zero.[13]

(20) a. I don't think Ø I'm gonna like really go to school for it.
(076.267)
b. I figure Ø I went- I went enough when I was younger.
(067.2287)
c. And then I was like telling my mother *that* she should get that.
(050.2068)

A final measurement of syntactic complexity in the complement clause is whether the verb is intransitive (21a) or transitive (accompanied by verbal objects or complements (21b)–(21d)). We predict that the latter, which involves greater syntactic complexity, should favor *that*.

(21) a. I think *Ø* Brian went to McGill.
(027.478)
b. I decided *that* at eighty-four, you'd had it.
(027.35)
c. I think *Ø* I was counting securities funnily enough.
(027.160)
d. I gather *that* mo- most people didn't s- find that they uh- had a job that was worthwhile here.
(027.265)

3.4. *Multivariate analysis*

As we have seen, there is no dearth of explanations (syntactic, semantic, and discourse-pragmatic) for *that*-variation. Particularly intriguing is the proposal that absence of *that* is a measure of the development of discourse formulas. We must first address the methodological question of how to decide among these competing explanations.

Claims about change in *that*-complementation have largely relied on the overall number of occurrences of *that* or zero, or the frequency of *that* relative to zero. But overall rates are subject to fluctuation from a number of considerations, not all of which are relevant to the linguistic system (cf. Rickford and McNair-Knox 1994). This is acutely so in the present case: as we have seen, rates of *that* and zero vary widely not only from time period to time period but also by register or genre within the same time period.

In addition, all previous quantitative studies have examined factors one by one. But we know that factors can act together in various ways (antagonistically or synergistically), limiting the effectiveness of univariate analysis (e.g., Guy 1988; Sankoff 1988b). Several studies of *that*-variation lament this very problem (e.g., Elsness 1984: 31; Rydén 1979: 12; López Couso 1996: 274), most succinctly expressed by Thompson and Mulac (1991b):

> Readers might wonder whether each of the variables discussed [...] independently influences the occurrence of *that*. That is, in assessing the role of a given variable, should we not also control for the possible influence of another co-occurring variable? (1991b: 249–250)

As a solution to this problem, they consider "cleansing" the data so that utterances contain only one of the variables they considered, which of course greatly reduces their data (by 80%). They raise the methodological question:

> [...] do we learn more about the phenomenon we are investigating if we report findings from a greatly diminished data base in which variables are considered strictly independently, or do we learn more if we report findings from a full data base with variables confounded such that it is indeterminate which one is responsible for the occurrence of *that*? (Thompson and Mulac 1991b: 249–250)

Multivariate analysis offers a straightforward solution to this dilemma, since it assumes that each of the competing variants occurs at greater or lesser rates depending on contextual (linguistic and social) features. Each context is decomposed into a specific configuration of conditioning factors (Sankoff 1988b: 984; Poplack and Tagliamonte 2001: 91) and the contribution of these factors to speaker choice is modeled probabilistically by operationalizing competing hypotheses about constraints in the choice process as factor groups in multivariate analysis (Poplack and Tagliamonte 2001: 91).

The primary analytical tool is variable-rule analysis ("VARBRUL"), which uses a multiple-regression procedure to determine which environmental factors contribute statistically significant effects, and how strongly, to the choice of variant when all are considered simultaneously (Sankoff 1988b; Rand and Sankoff 1990; Paolillo 2002). Despite its name, variable-rule analysis does not (necessarily) involve "rules" and is independent of assumptions about underlying forms (Sankoff 1988b: 984; Sankoff and Rousseau 1989; Guy 1993: 237–248; Paolillo 2002: 191–196). We refer to the configuration of factors affecting the choice of variants as the conditioning of variability, which remains constant despite differences in the rates of occurrence of variant forms attributable to extralinguistic considerations such as register or genre (e.g., Rickford and McNair-Knox 1994; Poplack and Tagliamonte 2001: 92). All of the factors discussed in the preceding section were analyzed individually and together using GoldVarb 2.1 (Rand and Sankoff 1990), a variable-rule application for the Macintosh.

4. The grammar of *that* usage

4.1. *Lexical types and frequent collocations*

Corpus-based studies have revealed differences in the register and lexical (and grammatical) associations of complementation structures. For example, Biber (2000: 298–299) finds that, compared to *to*-clauses, *that*-clauses are confined to a relatively small number of verbs, three of which (*think*, *say*, *know*) occur more frequently in conversation than in writing. We consider the number of different complement-taking predicates (lexical

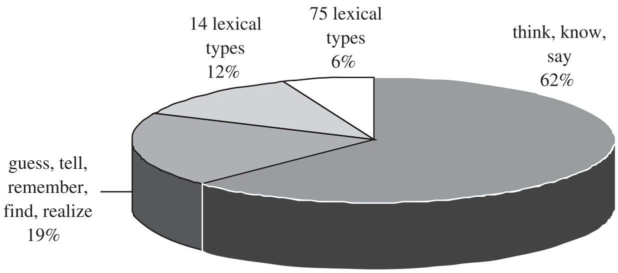

Figure 2. *Frequency of main-clause lexical types*

types) to which *that*-complementation applies, or the type frequency of the structure (cf. Bybee 2003: 604–605), in combination with the distribution of the lexical types, to constitute a measure of the limited productivity of the construction. Figure 2 shows the remarkable skewing of the distribution of main-clause lexical types in our data: *think*, *know*, and *say* account for 62% of all the data, while five additional verbs (*guess*, *tell*, *remember*, *find*, and *realize*) account for a further 19%: that is, just eight lexical types make up 81% of the data. Fourteen predicates make up another 12% and the remaining 6% is distributed over 75 lexical types.

Table 1 shows the distribution of complement-taking predicates by rate of *that* and proportion of all data. Predicates with nine or fewer occurrences, which together make up 6% of the data, show an overall rate of 60%, in line with their low frequency, while the single-most frequent verb, *think*, which alone makes up over 40% of the data, shows a low rate of 7%. However, there is no one-to-one correlation between frequency and rate of *that*. In predicates with 10–50 occurrences, *wish*, *hope*, and *suppose* have rates of 0%, 11%, and 0%, respectively, while *realize*, *seem*, and *understand* show 59%, 62%, and 75%, respectively. Among the mid-frequency predicates (50–200 occurrences), the highest frequency verb (*guess*) has a rate of 3%, but the second-highest (*tell*) has a higher than average 43%. Finally, among the high-frequency predicates, although *know* and *say* occupy roughly the same proportion of the data (9–10%), *know* shows 34% *that* while *say* has 27%, both at rates higher than several much lower frequency predicates. Evidently, then, it is not the frequency of the lexical type *per se* that correlates with zero.

However, if we modify our view of frequency to include collocations of main-clause subjects and verbs, we notice a substantial difference between *think*, *know*, and *say*, which present quite disparate rates of *that*. Unlike *know* and *say*, *think* is largely restricted to first-person singular and the present tense.[14] As Table 2 shows, 61% of all occurrences of *think* are *I think*. Six other subject-verb collocations also make up an average of

Table 1. *Distribution of main-clause lexical verbs/adjectives by rate of complementizer* that *and proportion of all data*

Lexical verb/adjective	N	% *that*	% of all data
Verbs or adjectives with more than 200 occurrences:			62 (N = 1761)
think	1212	7	43
know	287	34	10
say	262	27	9
Verbs or adjectives with 50–200 occurrences:			19 (N = 523)
guess	165	3	6
tell	120	43	4
remember	94	5	3
find	90	33	3
be + sure	54	15	2
Verbs or adjectives with 10–49 occurrences:			13 (N = 363)
realize	49	59	2
decide	36	53	1
believe	33	52	1
figure	32	25	1
make sure	25	40	1
notice	24	50	1
find out	23	17	1
feel	22	55	1
see	21	52	1
wish	20	0	1
hope	19	11	1
seem	13	62	0
mean	13	23	0
understand	12	75	0
suppose	11	0	0
be happy	10	60	0
Other verbs or adjectives			
	173	60	6
Total	2820	21	

80% of their respective lexical type: *I guess* (99%), *I remember* (96%), *I find* (66%), *I'm sure* (74%), *I wish* (85%) and *I hope* (79%). In contrast, all other subject-verb combinations that make up a substantial proportion of their respective verb types present an average of only 19%.

This sharp difference in relative frequency clearly correlates with rates of *that*: the average rate of *that* for the frequent collocations is a bare 8%, in contrast to 31% for the infrequent collocations. These frequent collocations are the prime candidates for conventionalization as discourse formulas: they are frequent, they appear with first-person subjects, they occur in adverbial positions within the complement clause (where we find other parentheticals), and they correlate with low rates of *that*.

Table 2. *Main-clause subject-verb collocations by proportion of lexical type and rate of complementizer* that

Frequent collocations	N	% Lexical type	% *that*
I think	734	61	5
I guess	163	99	3
I remember	90	96	4
I find	59	66	24
I'm sure	40	74	10
I wish	17	85	0
I hope	15	79	7
Total	1118	80	8
Other collocations	216	19	31

This raises an important question: if the highly frequent collocations are discourse formulas that express speaker stance, are they better analyzed as belonging to the lexicon as fixed or frozen (discourse-pragmatic) units or to a productive grammar as instantiations of a construction with open-class positions? To answer this question, we first examine the patterns of *that*-variation in the remaining 1,552 tokens of nonfrequent combinations of main-clause subjects and predicates.[15] We can then determine whether these patterns are the same for frequent subject-verb collocations.

4.2. *Extricating the effects of lexical type, frequency and semantic class*

As we have seen, use of *that* has been linked to the frequency and semantic class of the main-clause predicate. Table 3 compares the results for lexical frequency, semantic class, and lexical type in three independent variable-rule analyses, each of which also included the other factor groups considered (subjects, intervening material, other measures of syntactic complexity, etc). In each case, the factor group of interest was selected as statistically significant by the stepwise multiple regression procedure.[16] In the lexical frequency factor group (the third column), low- and medium-frequency main-clause predicates favor *that* (.64 and .65, respectively), while high-frequency lexical types disfavor (.39). Thus, in the aggregate, despite the lack of a one-to-one correlation between frequency and rate of *that* (Table 1), the expected frequency effect does appear to hold, such that the frequency of main-clause predicates in the *that*-complement construction correlates (inversely) with the rate of *that*, even when frequent collocations (*I think*, *I guess*, *I remember* etc.) are not considered.

Table 3. *Comparison of variable-rule analyses of factors contributing to the occurrence of complementizer* that *including semantic class, lexical type, and lexical frequency (other factor groups not shown remain constant across analyses)*

	Semantic class		Lexical type		Lexical frequency	
Input:	.323		.309		.320	
	Comment	.72	*realize*	.79	Medium	.65
	Extraposition	.71	*find*	.70	Low	.64
	Suasive	.64	Other	.65	High	.39
	Knowledge	.60	*tell*	.59		
	Utterance	.44	*know*	.57		
	Attitude	.41	*say*	.40		
			remember	.31		
			think	.23		
Log likelihood:	−870.994		−833.587		−861.503	
		$\chi^2 = 74.814$ df = 2 p < .001		$\chi^2 = 55.832$ df = 5 p < .001		

The results also confirm a correlation between the semantic class of the main-clause predicate and the occurrence of *that*, as shown in the first column of Table 3. As predicted, least favorable to *that* are attitude predicates (.41), which are the most likely to be "inverted" (Bolinger 1972: 23) or to be used parenthetically to express degrees of commitment to the proposition in the complement clause (Noonan 1985: 113–115) (recall that these data do not include the highly-frequent first-person collocations). The favoring effect of commentatives (.72) is expected, for a number of reasons. In an information-flow explanation, the content of the complement clause is presupposed with commentatives, hence they should favor *that* as marking old information (Bolinger 1972). There is another, syntactic consideration: since commentatives are largely adjectives, their relatively greater complexity (*be* + AdjP vs. V) might favor *that* (though we found no overall difference between verbs and adjectives in the rate of *that* [see Note 9]). More compelling, in our view, is the distribution across different complementation structures: comment predicates also occur with infinitival complements (*be important to*, *be sad to*), which may detract from the predictability of the complement structure. From a processing perspective, reduced (syntactic) predictability due to usage patterns should favor *that* as an explicit complement clause marker (cf. Hawkins 2002: 106).

At first glance, the results for semantic class appear to support the hypothesis of semantic differences between clauses with and without *that*, each of which is compatible with different main-clause predicates. In the grammaticality judgments reported by Dor (2005), "speech act" (e.g., *say*, *tell*) and "belief, knowledge, and conjecture" (e.g., *believe*, *remember*) predicates, both of which entail the notion of "truth claim", allow bare clauses, whereas "emotive" (e.g., *be amazed*, *be pleased*) predicates (largely our "Comment", in Table 3), which imply presupposition of the truth claim, prefer *that*.

However, as the middle column of Table 3 shows, the separation of semantic and frequency classes into individual lexical types is also selected as significant, with *realize* and *find* most favorable, *tell* and *know* somewhat less favorable, and *say*, *remember*, and *think* least favorable to *that*.[17] Note that the lexical effect obtains despite similar frequency counts (compare *find* (N = 90), which favors *that*, and *remember* (N = 94), which disfavors) and semantics (compare "epistemic" or "belief/knowledge/conjecture" *realize*, which highly favors *that*, and *think*, which highly disfavors). These results suggest that apparent effects of both frequency and semantic class may mask effects which are purely lexical (see also Poplack 1992: 255; Bybee and Thompson 1997: 384).

Since frequency, semantic class, and lexical type are interrelated, we use multivariate analysis to disentangle their effects to determine which best accounts for the observed variation. For each of the three variable-rule analyses in Table 3, we show the "log likelihood", a measurement of how well the configuration of factors fits the observed data, with figures closer to zero representing a better fit (see Sankoff 1988; Guy 1993; Paolillo 2002). As shown, the analysis with lexical type features the log likelihood closest to zero, indicating best fit. Moreover, a comparison of the log likelihoods of all three analyses shows that this difference is statistically significant.[18] Therefore, we conclude that frequency and semantic-class effects in this case reflect, at least in part, the specific effects of particular main-clause predicates, some of which are strongly associated with *that* and others with zero. This empirical result provides evidence that in addition to generalizations based on frequency and semantic class, structures may be lexically specific (cf. Bybee 2006).

4.3. *The confluence of syntactic and discourse effects*

Having isolated the effects of lexical type, we are now in a position to gauge the relative contribution of syntactic, semantic, and discourse-pragmatic factors. Table 4 shows the factors contributing to the occurrence of *that* in a multivariate analysis including all the factor groups we adapted from proposals or findings in earlier studies (Section 3.3).[19]

4.3.1. *Lack of evidence for semantic proximity.* Neither clausal cotemporality nor subject coreferentiality, our operationalization of semantic

Table 4. *Factors contributing to the occurrence of complementizer* that *in Quebec City English (excluding frequent main-clause subject-verb collocations)*

Total N:	1552		
Corrected mean:	.328		
		%	N
Complement-clause subject			
Noun phrase	.65	48	250
Other pronoun	.52	36	707
I	.42	28	390
it/there	.38	23	159
Range:	27		
Adjacency: intervening material			
Present	.72	57	273
Absent	.45	29	1279
Range:	27		
Main-clause subject			
Noun phrase	.68	52	326
Pronoun	.45	30	1226
Range:	23		
Main-clause adverbial			
Post-subject	.65	48	141
Phrasal	.59	42	165
None	.47	32	1191
Pre-subject	.45	28	54
Range:	20		
Adjacency: intervening verbal arguments			
Present	.60	46	138
Absent	.49	33	1414
Range:	11		
Main-clause verbal morphology			
Nonfinite	.58	41	388
Finite	.47	32	1164
Range:	11		
Complement-clause transitivity			
Transitive	.56	39	488
Intransitive	.47	32	1054
Range:	9		

Factors not selected as significant: Subject coreferentiality, harmony of polarity, complement-clause mood/morphology, co-temporality

Table 5. *Cross-tabulation of main-clause and complement-clause subjects*

	Main-clause subject					
	I		*You*		Other	
Complement subject:	% *that*	*N*	% *that*	N	% *that*	N
I	25	212	16	25	35	153
You	23	39	36	25	49	68
Other	30	430	41	76	40	524

proximity, turns out to be statistically significant. In light of the coreferentiality effects reported in other studies (e.g., Elsness 1984: 526; Finegan and Biber 2001: 262), the lack of significance for coreferentiality deserves discussion. Since we coded coreferentiality by semantic reference, we considered the possibility that the effect reported in other studies has more to do with formal isomorphy (*I* and *I*, *you* and *you*, etc.). If so, we would expect the rates of *that* in subject pairs to resemble each other, while simultaneously differing from their rates when subjects are different (that is, *I*+*I* should have a rate similar to that of *you*+*you*, *she*+*she*, etc., but different from that of *I*+*you*, *I*+*she*, etc.). However, when we examine the rates of *that* in such subject pairs, we find that they are quite different. As Table 5 shows, *I*+*I* features a rate of 25%, while *you*+*you* is 36%. At the same time, when *I* occurs with other subjects, the rate of *that* remains low, whether the *I* is the subject of the complement-taking predicate or of the complement clause. *I*-*you* shows 23% and *I*-Other 30%, while *you*-*I* shows 16% and Other-*I* 35%. We conclude that the reported effect of coreferentiality follows not from the co-occurrence of identical subject pairs, but from the individual effects of one pronoun, *I*.

The fact that polarity was also not selected as significant is surprising, since our preliminary findings for main-clause polarity showed that negation correlates highly with zero, contradicting the prediction that negation should favor *that*, since it often involves *do* support or adverbials such as *never* and may render a clause syntactically more complex or less formulaic.[20] Again, we considered the possibility of a lexical-type effect, but while main-clause negation is highly correlated with five verbs (*believe*, *find*, *mean*, *realize*, and *think*), these negated verbs do not disfavor *that* more than average. The exception is *I don't think*, which accounts for 54% (139/256) of the negated tokens and has a near-categorical absence of *that*, at 4%. This comes as no surprise, since this relatively frequent collocation is another potential discourse formula (also, negative-raising may be taken to indicate a unitary or monoclausal proposition; see Horn 2001: 308–330). Thus, as with subject coreferentiality and

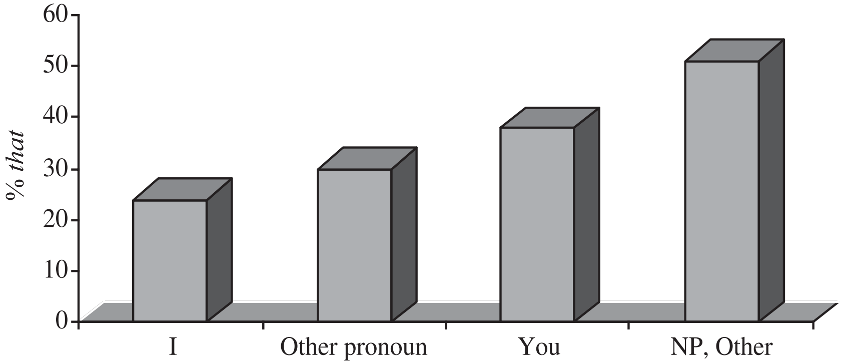

Figure 3. *Rate of* that *complementizer by main-clause subject*

clause cotemporality, we find no support for considerations of semantic proximity.

4.3.2. *Syntactic complexity.* Four of the five factor groups that measure syntactic complexity are significant: main-clause subject, main-clause adverbial, main-clause verbal morphology, and complement-clause transitivity. Meanwhile, the presence of periphrastic forms or modals in the complement clause is not significant.[21] Three of these four significant factor groups concern the main clause.

For the main-clause subject, as Figure 3 shows, while *I* does disfavor *that* the most, the range of effect is actually greater between pronouns and NPs than it is between *I* or *you* and other subject types. The effect reported in Thompson and Mulac (1991b: 242) for *I* can be attributed to their inclusion of formulaic *I think* and *I guess*.[22] When we set aside the highly frequent collocations, as in Table 4, we see a relatively strong effect, as indicated by the *range* (23) for this factor group, with full NPs favoring *that*, as opposed to pronominal subjects, which disfavor.

The presence of an adverbial in the main clause also favors *that*, but only if the adverbial occurs after the subject, as in (22a). Notably, adverbials occurring before the subject (22b) disfavor *that* as much as no adverbial. Thus, although syntactic complexity is relevant, as indicated by the favoring effect of phrasal adverbials, the difference between pre- and post-subject adverbials suggests that it is not merely a question of constituent size or weight. Rather, pre-subject adverbials are more likely to have scope over both the main and complement clause as a unitary proposition than adverbials appearing between the main clause subject and verb.

(22) a. They still think *that* it's gonna happen and like- it's gonna be perfect.
(067.1371)

b. So like I had to open the door, so already they think *Ø* this is bizarre as heck.
(050.2230)

For main-clause verbal morphology, nonfinite forms (19b) significantly favor *that* (cf. Thompson and Mulac 1991b: 246). In the complement clause, while the difference between simple and periphrastic forms (including modals) does not achieve significance, the presence of an object or complement, as opposed to an intransitive verb, favors *that*. These two syntactic effects are the weakest, as indicated by relatively low ranges (11 and 9, respectively).

4.3.3. *Intervening material.* The effect of intervening material, together with the complement-clause subject, contributes the greatest effect to *that*-variation, as indicated by the relatively large range (27). The presence of intervening constituents highly favors *that*. Figure 4 shows that the rate of *that* increases correspondingly as intervening constituents become more complex, lowest with one-word adverbs and highest when more than one category intervenes. On the other hand, if the intervening material may be viewed as an argument of the main clause verb (as in [16]), the effect is much smaller (with a range of 9).

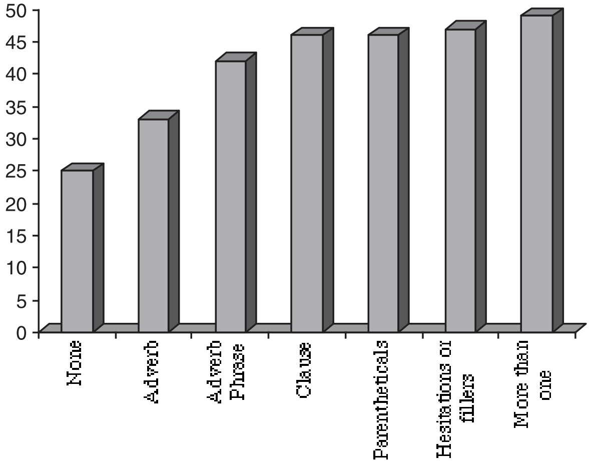

Figure 4. *Rate of complementizer* that *by material intervening between the main clause and complement clause*

4.3.4. *Complement-clause subject.* Of all the factor groups selected as significant, the complement-clause subject exerts the strongest effect, together with intervening constituents (both with a range of 27). However, contrary to what might be expected if *that* retained anaphoricity from its demonstrative origins (Bolinger 1972), the difference is not between indefinite NPs (which are more likely to refer to new information) and pronouns and definite NPs (which are more anaphoric).[23] In fact, here the difference between definite NPs (36%) and indefinites (43%) is not significant. Rather, most favorable to *that* is any kind of lexical subject.

The higher discourse topicality of pronouns is another information-flow explanation for the effect of the complement-clause subject. If the (erstwhile) main clause is reduced to a discourse formula, the discourse topic is really the subject of the complement clause (Thompson and Mulac 1991b: 248). The ordering of factors shows that pronominal subjects do favor *that* less than NP subjects. Since cross-tabulations of main clause verb and complement subject confirm that this effect is not related to frequent co-occurrence with particular complement-taking predicates, we conclude that the pronominal subject effect is not lexically specific.

Nevertheless, *that* is disfavored not by pronouns in general, but rather by one pronoun, *I*, which is the single most frequent complement-clause subject (390 tokens, or 25% of the data). Yet it is not evident that subject topicality explains the complement-clause subject effect, since this general constraint operates even when we exclude the frequent subject-verb collocations. Least favorable to *that* are expletive *it* and *there*, as in (23), which are neither anaphoric nor topical, and may not be referential at all.[24] Rather, the ordering of complement-clause subjects from least to most favorable to *that* (*it*/*there* < *I* < other pronoun < NP) appears to be in order of increasing referentiality.[25]

(23) a. Everybody knows Ø it's a joke.
(050.875)
b. They found out Ø it was cancer.
(025.1049)
c. And he said Ø it would be okay to supervise the activities.
(031.480)
d. And he thought Ø there was a skunk inside of it.
(058.1388)

4.3.5. *That's* that. We are now in a position to assess the relative contribution of syntactic, semantic, and discourse explanations to the occurrence of *that*. The lack of coreferentiality, cotemporality, and polarity harmony effects does not support a semantic distancing explanation, as proposed in iconicity accounts (e.g., Langacker 1991; Givón 1993). Nor have we found evidence for the claim of an anaphoric function for *that* (Bolinger 1972) in the complement-clause subject constraint. Rather, the constraint with the greatest magnitude of effect (intervening material) is congruent with the suggestion that *that* serves to mark the boundary between the two clauses.

However, straightforward syntactic complexity does not account for the effects. All of the significant factors can be related to complexity, but not so much in terms of the size or weight of constituents *per se*. For example, in the strong and clear main-clause adverbial constraint, it is not the mere presence of an adverbial that favors *that* but rather its post-subject position. Consider together the factors that favor *that*: first, increasing subject referentiality and the presence of objects or complements in the complement clause; and second, in addition to post-subject adverbials, lexical versus pronominal subjects, complex verbal morphology, and arguments in the main clause. This configuration of factors indicates that *that* is used not merely to demarcate two clauses, but to demarcate two clauses that both have (lexical) content. In contrast, zero occurs when there is less semantic or propositional content overall and the two clauses behave more like a single proposition (see Fox and Thompson 1990 for a similar argument for English relative clauses).[26] Note that, even when we set aside frequent subject-verb collocations, most factors concern the main clause: NP subjects, post-subject adverbials, complex verbal morphology, and verbal arguments. This confluence of factors suggests that the burden of proof of "clausehood" lies with the main clause, in line with the proposal that complement-taking predicates tend to reduce to fragments or formulas (Thompson 2002).

5. Persistence of grammar: *that* and zero in frequent collocations

We now return to an examination of the frequent subject-verb collocations (*I think*, *I guess*, *I remember*, *I find*, *I'm sure*, *I wish*, and *I hope*), which we set aside in the previous analysis.[27]

5.1. *Discourse formulas*

We note that there is some disagreement as to whether the conventionalization of *I think*, *I guess*, and other collocations as discourse formulas constitutes an example of "grammaticization" or is more aptly characterized as "pragmaticization" (or even "lexicalization"). Thompson and

Mulac (1991a: 324) first proposed that the high rates of zero complementizer with *I think* reflect grammaticization, in that "grammaticization always involves the reanalysis of one kind of pattern [...] as another kind of pattern"; in this case, "a frequent collocation of subject and complement-taking verb" as "an epistemic phrase with no syntactic complement-taking properties". Aijmer (1997: 3; cf. Erman and Kotsinas 1993) argues that this process is better characterized as pragmaticization, since pragmatic elements, unlike grammatical ones, cannot be analyzed in truth-conditional terms. Others view it as an instance of "subjectification" in grammaticization (Traugott 1989: 35, 1995: 31), which Traugott (1995: 32) defines as "the development of a grammatically identifiable expression of speaker belief or speaker attitude to what is said." This process occurs through a mechanism of "pragmatic inferencing", such that implicatures become conventionalized with repeated use (Traugott 2003: 635).

Exactly how subjectification relates to grammaticization is a matter of some debate, especially as it concerns discourse markers (cf. Palander-Collin 1999: 47–60). The development of discourse markers may follow the principle of *decategorialization* (Hopper 1991; Hopper and Traugott 2003: 104–115), which posits that forms shed the morphosyntactic trappings of their source constructions as their discourse role shifts: "forms undergoing grammaticization tend to lose or neutralize the morphological markers and syntactic privileges characteristic of the full categories Noun and Verb ..." (Hopper 1991: 22). Indeed, in a study of verb-based discourse markers in Spanish, Company (2006) proposes that the degree of subjectivity correlates with "syntactic cancellation", or the loss of syntactic relations: as discourse markers, erstwhile verbs do not take arguments, complementation, or any kind of modification. Modification is precisely what the frequent subject-verb collocations we have identified here appear to lose in their use as discourse formulas: specifically, their complement-taking properties (Thompson and Mulac 1991a: 324).

However, syntactic isolation does not characterize more canonical instances of grammaticization. Bybee (2006) argues that new constructions can arise without necessarily undergoing a process of grammaticization, which involves the creation of a new GRAMMATICAL morpheme (our emphasis) that can become obligatory. Nevertheless, Traugott (2003: 642–643; cf. Hopper and Traugott 2003: 207–209) argues that discourse markers or parentheticals, like adverbs, are part of the grammar, and "if we focus not on lexical item > grammatical item, but [...] on lexemes undergoing change in the context of constructions", the development of discourse markers manifests more similarities than differences with canonical grammaticization.

Evidently, the site of disagreement is the concept of grammaticization or, more precisely, the boundaries between discourse-pragmatics, grammar, and the lexicon. This issue of nomenclature does not concern us here, since we argue precisely for the absence of discrete boundaries and the gradualness of change in the evolution of linguistic resources, as properties, both semantic and syntactic, persist from the previous stage.

One of the proposed frequency effects on mental processing and storage is the developing autonomy of new units, whether these result from grammaticization or not (Bybee 2006). In defining grammaticization as "the process by which a frequently used sequence of words or morphemes becomes automated as a single processing unit," Bybee (2003: 603) draws attention to the conventionalization of usage patterns involving collocations. She proposes that a frequent collocation gains autonomy in two ways. First, its erstwhile individual constituents weaken their association with other instances of the same constituents. For example, as (*be*) *going to* reduces to *gonna*, *go* loses its association with the uses of *go* as a verb of motion (Bybee 2003: 618). Second, the collocation dissociates from other instances of the construction. For example, (*be*) *going to* loses its association with the more general motion-purposive schema, as in *I am going/traveling/riding to see the king* (Bybee 2003: 603). In this perspective, *I think* becomes autonomous from other forms of *think* as well as from the general *that*-complement construction. According to Bybee (2003: 618), increasing opacity of internal structure and autonomy enables single processing units to gain new discourse-pragmatic functions.

5.2. *Testing persistence and autonomy*

If the frequent collocations we identified in Table 2 are fixed units, autonomous from their composite verbs and other instances of the *that*-complement construction (cf. Bybee 2003, 2006), exactly how do they differ from other instances of the *that*-complement construction? If they are extrapropositional formulas that develop from erstwhile syntactically active elements (cf. Company 2006), are they better treated as belonging to the lexicon or to the grammar? Rather than merely relying on overall rates of *that* (cf. Hopper and Traugott 2003: 209), which are very low (8%; see Table 2), we impose the stricter criterion of parallel ranking of the linguistic constraints on *that*-variation.

We can use multivariate analysis to compare the linguistic conditioning of variability in frequent subject-verb collocations with complement-taking predicates composed of infrequent combinations of subjects and verbs (see Table 4). If the highly frequent subject-verb collocations are

Table 6. *Factors contributing to the occurrence of complementizer* that *in Quebec City English in frequent main-clause subject-verb collocations*[‡]

Total N:	1118		
Corrected mean:	.049		
		%	N
Main-clause adverbial			
Post-subject	.83	23	22
Pre-subject	.77	17	24
Phrasal	.52	6	102
None	.48	5	970
Range:	35		
Complement-clause subject			
Noun phrase	.68	12	205
Other pronoun	.48	5	463
it/there	.43	4	179
I	.42	4	199
Range:	26		
Adjacency: intervening material			
Present	.72	14	144
Absent	.47	5	974
Range:	25		

Factors not selected as significant: Subject coreferentiality, intervening verbal arguments, complement-clause transitivity

[‡] Excludes main-clause *I don't think* (N = 151)

fixed discourse formulas, the factors conditioning *that* with these forms should differ from that of other subject-verb combinations, which presumably are not used formulaically but rather are instances of (a relatively more) productive grammatical construction. Accordingly, we performed a multivariate analysis of these frequent collocations with the same factors.

Table 6 shows the factors contributing to the occurrence of *that* in frequent subject-verb collocations (*I think*, *I guess*, *I remember*, *I find*, *I'm sure*, *I wish*, and *I hope*). The three factor groups with the greatest magnitude of effect in the earlier analysis of the nonfrequent combinations of main clause subjects and verbs (Table 4) — intervening material, complement-clause subjects, and main-clause adverbials — are also selected as significant here. Crucially, the ordering of factors within each factor group is largely parallel. Within the complement-clause subject factor group, *that* is most favored by NPs and most disfavored by *it/there* and *I*. Similarly, the presence of intervening material favors *that*.

At the same time, however, the ordering of the factor groups by their magnitude of effect (indicated by the range) is not identical. The greatest contribution to *that* in these frequent collocations is the main-clause adverbial (with a range of 35). This is unsurprising, since the presence of a post-subject adverbial, as in (24a)–(24b), detracts from (in fact, nullifies) the formulaic nature of the collocation. Note also that pre-subject adverbials (24c) are almost as favorable to *that*, while phrasal adverbials (24d) have little effect, in contrast with their effects on the less frequent main clauses in Table 4.

(24) a. Well I actually think *that* they were very responsible.
(031.688)
b. I personally think *that* it is well worthwhile.
(027.835)
c. Actually I- I think *that* those were the only two things they said in the entire skit.
(071.737)
d. All of a sudden I think *Ø* I'm getting the heat here.
(057.130)

Thus, the ordering of factor groups by their magnitude of effect provides evidence that highly frequent subject-verb collocations (beyond *I think* and *I guess*) behave as fixed units, since *that* is most likely to occur precisely when the formulaic nature of the collocation is annulled by a co-occurring adverbial. Nevertheless, despite near-categorical absence of *that* (indicated by the low corrected mean, or overall probability, of .049), frequent collocations retain traces of grammatical conditioning: in particular, constraints of intervening material and the complement-clause subject.

6. Conclusion: on persistence

The findings of this study were made possible by adopting the variationist approach to the study of language. Since the variationist method is "pre-theoretical" (Laks 1992), it can be used to test any linguistic theory, provided that theory makes predictions that can be operationalized as factors (cf. Sankoff 1988a: 151, 1988b: 984). Furthermore, multivariate analysis provides a way out of the methodological dilemma of univariate methods, allowing us to extricate and assess the relative contribution of factors proposed by disparate accounts of *that*.

In this study, we tested hypotheses about the function of complementizer *that* and their contribution to speaker use via multivariate analysis. We saw first the skewed distribution of complement-taking predicates by lexical type (Figure 2, Table 1) and identified certain high-frequency collocations of first-person subject and simple present-tense verbs which show near-categorical association with the absence of *that* (Table 2). Multivariate analysis revealed that even beyond the frequent subject-verb collocations, the presence or absence of *that* is associated more with particular main-clause lexical types than with frequency or semantic class (Table 3). At the same time, *that*-variation is conditioned by a number of language-internal factors. Thus, we conclude that while much of *that*-variation is lexically specific, it remains an active, if relatively restricted, area of the grammar.

The choice of *that* is favored by material intervening between the complement-taking predicate and the complement clause, by lexical subjects and the presence of arguments in the complement clause, and by lexical subjects, post-subject adverbials, complex verbal morphology, and arguments in the main clause (Table 4). This conditioning indicates that *that* functions to demarcate two clauses that have lexical content, whereas *that* is absent when there is less semantic content overall in two "clauses" that behave more like a single proposition (cf. Fox and Thompson 1990).

The evidence that frequent subject-verb collocations behave like fixed discourse formulas comes from the strong effect of the co-occurrence of an adverbial after or before *I*: *that* is favored precisely when the presence of other material undermines the collocational association between *I* and the verb (Table 6). However, the two strongest constraints (intervening material and complement subject) are both operative, and with the same direction of effect. Thus, as with the nonfrequent combinations of main-clause subjects and predicates, *that* is used more when there is more lexical content. Notably, even though the rate of *that* is slight, the linguistic conditioning parallels constructions with more robust variation.

The parallelism in linguistic conditioning shows that, despite their high frequency and formulaic status, these fixed units, such as *I think*, *I guess*, *I remember*, are not completely autonomous (in Bybee's [2003: 618] sense) from other instances of the *that*-complement construction. This empirical result, which, we emphasize, emerged only through quantitative multivariate analysis, has broader import in making gradualness explicit in the principle of decategorialization (Hopper 1991). It suggests that the hypothesis of persistence or retention should be extended from semantics to syntax: not only does lexical meaning persist in grammaticizing constructions (Bybee and Pagliuca 1987; Hopper 1991), but grammatical properties also persist in the development of discourse formulas.

In discourse, grammatical persistence is manifested in parallelism in linguistic conditioning. We have shown that fixed formulas are not entirely autonomous from the productive constructions from which they originate, but that constraints on the source construction persist in the formulaic unit. Grammatical retention, like semantic retention, is consonant with the gradualness of change, that is, autonomy (Bybee 2003) applies gradually and, probably, never completely. Thus, properties from the previous stage persist, not only from lexicon to grammar but also from productive grammatical construction to conventionalized formula.

Studies of usage provide evidence that much of "grammar" consists of combinations of prefabricated (Hopper 1987; Bybee 2006) or "reusable" (Thompson 2002) fragments. In usage-based approaches, "grammar" is the conventionalization of co-occurrence patterns in language use (e.g., Haiman 1994; Bybee 2001: 12). In this dynamic view of language, grammar and lexicon are not neatly separated or compartmentalized. Our findings support a fluid, rather than modular, relationship between fixed units (lexicon) and productive constructions (grammar) (Bybee 1998; Hopper 1998). The empirical examination of some of these ideas has yielded new insights into this question, most importantly, on the persisting relationship between formulaic units and the constructions from which they emerge.

Received 29 March 2005
Revised version received
11 January 2006

University of New Mexico
York University

Notes

* The data on which this study is based were taken from the *Quebec English Corpus* (Poplack et al. 2006), housed in the Sociolinguistics Laboratory at the University of Ottawa. Financial support was provided by the Social Sciences and Humanities Research Council of Canada in the form of a standard research grant to Shana Poplack and James Walker and a postdoctoral fellowship to Rena Torres Cacoullos. We thank Professor Poplack for her support and for the infrastructure and resources afforded by the Sociolinguistics Laboratory. Earlier versions of this study were presented at the annual meeting of the Canadian Linguistic Association (Dalhousie University, Halifax, June 2003), the New Ways of Analyzing Variation conference (University of Pennsylvania, Philadelphia, October 2003), and invited talks at the University of New Mexico, York University, and the University of Edinburgh. We thank audiences there for their comments, and Joan Bybee, Sandra Thompson, Gerard Van Herk, and several anonymous reviewers for comments on earlier drafts of this paper. Correspondence address: Rena Torres Cacoullos, Department of Spanish and Portuguese, MSC03-2100, University of New Mexico, Albuquerque, NM 87131-0001, USA. E-mail: rcacoull@umn.edu.

[1] Examples taken from the Quebec City component of the *Quebec English Corpus* (Poplack et al. 2006) are reproduced verbatim from speakers' utterances and are identified by speaker number and line number in the transcription.

[2] Rissanen (1991: 289) suggests that this reversal may also reflect the development in Modern English of nonfinite structures of complementation.

3. The paramount importance of genre or register is supported by the fact that Finegan and Biber (2001: 258–259) find no difference among social groupings in present-day British English.
4. A further complication is the different methods of calculating rates of occurrence. We have attempted to avoid this problem by presenting the results shown in Figure 1 in line with our study: that is, as the proportion of zero out of all complementizers.
5. We thank Gerard Van Herk for researching the grammarians' treatment of *that*. This research was possible thanks to the *Ottawa Grammar Resource on Early Variability in English* (Poplack et al. 2002), a corpus of historical grammars (1577–1898) compiled to track variability and housed in the Sociolinguistics Laboratory at the University of Ottawa.
6. However, on the basis of evidence from conversational English, Thompson (2002) argues that "complement" does not form a unitary category and that complements are neither arguments nor subordinate.
7. "In *It's right they should have the money* the two harmonize and no *that* is needed. In **It's wrong they should have the money* they clash and *that* is required" (Bolinger 1972: 38).
8. To whom we owe the title of Section 4.3.5.
9. Predicate adjectives include *afraid*, *convinced*, *glad*, *happy*, *lucky*, and *positive*; predicate nominals are *bad*, *better*, *fortunate*, *funny*, *good*, *hilarious*, *ironic*, *likely*, *nice*, *obvious*, *rare*, *sad*, and *seldom*. The rate of *that* for predicate adjective is not different from verbal main clause predicates (29%), though it is higher for predicate nominals (74%) and extraposed structures (80%). Given the small number of tokens, we do not pursue this difference further, except to note that the predicate nominals are overwhelmingly "commentative" predicates (see Table 3).
10. Lexical types were considered high frequency if they had more than 200 tokens, medium if they had between 50 and 200 tokens, and low if they had between 10 and 49 tokens. These distinctions were made on the basis of the distribution of tokens in the dataset (see Table 1).
11. Demonstrative *that* complement-clause subjects, as in (i), turn out to disfavor *that* (with a rate of 22% in the 45 tokens), perhaps in accordance with traditional grammarians' censure of the "inelegance" or redundancy of repeating *that*'s.

    (i) I know *Ø* that sounds really funny.
    (Q059.1629)

12. However, Bybee (2003: 603) argues that frequent co-occurrences are automatized as single processing units. Thus, some periphrastic or nonfinite collocations may be just as formulaic or prefabricated as single words.
13. If information status and assertion play a role in conditioning *that*, and if modals are less likely when the content of the complement clause is asserted, the presence of a modal in the complement clause should favor *that* (cf. Underhill 1988). However, it is an empirical question whether (all) English modals have a greater or lesser occurrence than nonmodals in assertive or main clause uses. In any case, we found no particular favoring effect for modals (see also note 21).
14. Also, the *that*-complement structure is more predictable for *think* than for *say* or *know*. In a sample of 100 tokens of *think*, we found 59% occurred in this structure (another 13% were parenthetical uses and 10% were *I think so*), while only 13% of *say* and 7% of *know* tokens did (43% of *say* were quotatives, while 37% of *know* were clauses introduced by *if*, *where*, *what* or another interrogative; *I don't know* made up 33% of *know* tokens [cf. Scheibman 2000]).

15. Note that, in contrast with most quantitative studies, which have been concerned with zero, the following analyses focus on the factors influencing the occurrence of *that*.
16. The probabilities, or "factor weights", in this and subsequent tables show the relative contribution of each factor to the occurrence of *that*. Factor weights above .50 favor *that*, while factor weights below .50 disfavor. The "range", or the difference between the lowest and highest factor weights within each factor group, indicates the strength of each factor group relative to all other factor groups in the same variable-rule analysis. The "corrected mean" is the overall tendency for *that* to occur.
17. The lexical types included in this factor group are the most frequent types once the frequent subject-verb collocations are removed from the analysis.
18. The chi-square value is twice the difference between the log likelihoods of the two variable-rule analyses being compared. The degrees of freedom in the comparison is the difference between the degrees of freedom for each analysis, which itself is the total number of factors in the analysis minus the number of factor groups (cf. Guy 1993: 246–247).
19. Including the lexical-type factor group (see Table 3) in the analysis changes neither the ordering of the factor groups nor the direction of effect within factor groups. However, we exclude this factor group from the remaining analysis because the magnitude of its effect prevents other factor groups (intervening verbal arguments and complement-clause transitivity) from being selected as significant.
20. Thompson and Mulac (1991b: 245), who hypothesize that *that* may be correlated with the greater syntactic complexity of interrogative or negative verb forms in the main clause, also find no statistically significant effect.
21. We find no significant difference between modal and nonmodal periphrastic forms in complement clauses, contrary to what might be predicted if modals signal less "binding", or greater semantic independence of the complement clause, in Givón's (1980) sense.
22. In their conversational data, nongeneric *you* aligned with *I* in the low rate of 9% (6/61), and 82% (55/67) of *you* tokens were in questions (Thompson and Mulac 1991b: 242–243). In the present corpus, 60% (12/20) of questions had a *you* subject but the rate of *that* was 33% (4/12).
23. Findings that pronouns and definite NPs disfavor *that* have been reconciled with Bolinger's (1972) ascription of anaphoricity to *that* by arguing that its presence imbues the complement clause with anaphoric force (Elseness 1984: 531) or that it functions "as a referential marker in relation to the hearer's knowledge" rather than as "a discourse-bound anaphoric marker" (Yaguchi 2001: 1141). Of course, the disclaimer that speakers can present old information as new (Bolinger 1972: 70) renders this idea empirically untestable.
24. *It*-subject clauses may be more likely to be prefabricated units, as may be the case with "it's a joke" in (23a). Bybee (2002) suggests that subordinate clauses are often like constructions and may be processed as unitary chunks.
25. The complement-clause subject effect may not be (solely) a grammatical factor but may also reflect considerations of processing (cf. Roland, Elman and Ferreira 2003) or prosody. For example, different subject types may correspond to different configurations of prosodic structure (cf. Walker 2000). We are exploring this possibility in future work (Walker and Torres Cacoullos in prep.).
26. An anonymous reviewer suggests that this statement is problematic, under the argument that some verbs (e.g., *say*, *tell*) have no less lexical content without *that*. Our point is that neither the subject nor the verb in fixed discourse formulas is truly lexical.

Just as the formula *I think* may not literally refer to thinking, formulaic (i.e., noncompositional) uses with *say* or *tell* may not literally refer to saying or telling.

27. We exclude *I don't think* because of its invariance with respect to *that*.

References

Aijmer, Karin (1997). *I think* — an English modal particle. In *Modality in Germanic Languages: Historical and Comparative Perspectives*, Toril Swan and Olaf Jansen Westvik (eds.), 1–47. Berlin and New York: Mouton de Gruyter.

Ash, John (1979 [1763]). *Grammatical Institutes; or, An Easy Introduction to Dr. Lowth's English Grammar, Designed for the Use of Schools.* Reprint. Ann Arbor, MI: Scholars' Facsimiles and Reprints.

Bayly, Anselm (1969 [1772]). *A Plain and Complete Grammar of the English Language; to which is Prefixed thee English Accidence: with Remarks and Observations on a Short Introduction to English Grammar.* Reprint. Menston: Scolar Press.

Bex, Tony and Watts, Richard J. (eds.) (1999). *Standard English: The Widening Debate*. London: Routledge.

Biber, Douglas (2000). Investigating language use through corpus-based analysis of association patterns. In *Usage-Based Models of Language*, Michael Barlow and Suzanne Kemmer (eds.), 287–313. Stanford: CSLI.

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

Bresnan, Joan (1972). Theory of complementation in English syntax. Unpublished doctoral dissertation, Massachusetts Institute of Technology.

Brittain, Lewis (1788). *Rudiments of English Grammar*. London: L. J. Urban.

Bybee, Joan L. (1988). The emergent lexicon. *Chicago Linguistics Society* 34, 421–435.

Bybee, Joan L. (2001). *Phonology and Language Use*. Cambridge: Cambridge University Press.

Bybee, Joan L. (2002). Main clauses are innovative, subordinate clauses are conservative: consequences for the nature of constructions. In *Complex Sentences in Grammar and Discourse: Essays in Honor of Sandra A. Thompson*, Joan Bybee and Michael Noonan (eds.), 1–17. Amsterdam: John Benjamins.

Bybee, Joan L. (2003). Mechanisms of change in grammaticization: the role of frequency. In *The Handbook of Historical Linguistics*, Richard Janda and Brian Joseph (eds.), 624–647. Oxford: Blackwell.

Bybee, Joan L. (2006). From usage to grammar: the mind's response to repetition. *Language* 82(4), 711–733.

Bybee, Joan L. and Hopper, Paul (eds.) (2001). *Frequency and the Emergence of Linguistic Structure*. Amsterdam: John Benjamins.

Bybee, Joan L. and Pagliuca, William (1987). The evolution of future meaning. In *Papers from the 7th International Conference on Historical Linguistics*, A. G. Ramat, O. Carruba, and G. Bernini (eds.), 109–122. Amsterdam: John Benjamins.

Bybee, Joan L.; Perkins, Revere; and Pagliuca, William (1994). *The Evolution of Grammar: Tense, Aspect, and Modality in the Languages of the World*. Chicago: University of Chicago Press.

Bybee Joan L. and Thompson, Sandra A. (1997). Three frequency effects in syntax. *Berkeley Linguistics Society* 23, 377–388.

Chafe, Wallace (1992). Information flow. In *Oxford International Encyclopedia of Linguistics*, William Bright (ed.), 215–218. Oxford: Oxford University Press.

Company Company, Concepción. (2006). Zero in syntax, ten in pragmatics: subjectification as syntactic cancellation. In *Subjectification: Various Paths to Subjectivity*, Angeliki

Athanasidou, Costas Canakis and Bert Cornillie (eds.), 375–398. Berlin and New York: Mouton de Gruyter.

Diessel, Holger and Tomasello, Michael (2001). The acquisition of finite complement clauses in English: a corpus-based analysis. *Cognitive Linguistics* 12(2), 97–141.

Doherty, Cathal (2000). *Clauses without* that*: The Case for Bare Sentential Complementation in English*. New York: Garland.

Dor, Daniel (2005). Toward a semantic account of *that*-deletion in English. *Linguistics* 43(2), 345–382.

DuBois, John W. (1987). The discourse basis of ergativity. *Language* 63, 805–855.

Ellinger, J. (1933). Substantivsätze mit oder ohne *that* in der neueren englischen Literatur. *Anglia* 57, 78–109.

Elphinson, James (1765). *The Principles of the English Language Digested, or, English Grammar Reduced to Analogy*. Edinburgh: James Bettenham.

Elsness, Johan (1984). *That* or zero? A look at the choice of object clause connective in a corpus of American English. *English Studies* 65, 519–533.

Erman, Britt and Kotsinas, Ulla-Britt (1993). Pragmaticalization: the case of *ba'* and *you know*. *Studier i modern sprakvetenskap* 10, 76–92.

Fanego, Teresa (1990). Finite complement clauses in Shakespeare's English. *Studia Neophilologica* 62, 3–21.

Ferreira, Victor S. and Dell, Gary S. (2000). Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive Psychology* 40, 296–340.

Finegan, Edward and Biber, Douglas (1985). *That* and zero complementizers in Late Modern English: exploring ARCHER from 1650–1990. In *The Verb in Contemporary English*, Bas Aarts and Charles F. Meyer (eds.), 241–257. Cambridge: Cambridge University Press.

Finegan, Edward and Biber, Douglas (2001). Register variation and social dialect variation: the Register Axiom. In *Style and Sociolinguistic Variation*, Penelope Eckert and John R. Rickford (eds.), 235–267. Cambridge: Cambridge University Press.

Fox, Barbara A. and Thompson, Sandra A. (1990). A discourse explanation of the grammar of relative clauses in English conversation. *Language* 66, 297–316.

van Gelderen, Elly (2004). Economy, innovation, and prescriptivism: from spec to head and head to head. *Journal of Comparative Germanic Linguistics* 7, 59–98.

Givón, Talmy (1980). The binding hierarchy and the typology of complements. *Studies in Language* 4, 333–377.

Givón, Talmy (1993). *English Grammar, Volume II*. Amsterdam and Philadelphia: John Benjamins.

Givón, Talmy (1995). Isomorphism in the grammatical code. In *Iconicity in Language*, R. Simone (ed.), 47–76. Amsterdam: John Benjamins.

Gorrell, J. Hendren (1895). Indirect discourse in Anglo-Saxon. *Publications of the Modern Language Association of America* 10, 342–485.

Grimshaw, Jane (1997). Projection, heads and optimality. *Linguistic Inquiry* 28(3), 373–422.

Guy, Gregory R. (1988). Advanced VARBRUL analysis. In *Linguistic Change and Contact*, Kathleen Ferrara, Becky Brown, Keith Walters and John Baugh (eds.), 124–136. Austin: Department of Linguistics, University of Texas at Austin.

Guy, Gregory R. (1993). The quantitative analysis of linguistic variation. In *American Dialect Research*, Dennis Preston (ed.), 223–241. Amsterdam and Philadelphia: John Benjamins.

Haiman, John (1994). Ritualization and the development of language. In *Perspectives on Grammaticalization*, William Pagliuca (ed.), 3–28. Amsterdam and Philadelphia: John Benjamins.

Harris, Alice C. and Campbell, Lyle (1995). *Historical Syntax in Cross-Linguistic Perspective*. Cambridge: Cambridge University Press.

Haugland, Kari E. (1995). *Is't allow'd or ain't it?* On contraction in early grammars and spelling books. *Studia Neophilologica* 67, 165–184.

Hawkins, John A. (2002). Symmetries and asymmetries: their grammar, typology and parsing. *Theoretical Linguistics* 28, 95–149.

Heine, Bernd and Kuteva, Tania (2002). *World Lexicon of Grammaticalization*. Cambridge: Cambridge University Press.

Hodgson, Isaac (1783). *A Practical English Grammar, for the Use of Schools, and Private Gentlemen and Ladies; with Exercises of False Orthography, and Syntax at Large*. London: B. Law.

Hopper, Paul J. (1987). Emergent grammar. *Berkeley Linguistics Society* 13, 139–157.

Hopper, Paul J. (1991). On some principles of grammaticization. In *Approaches to Grammaticalization*, Volume 1, Elizabeth Closs Traugott and Bernd Heine (eds.), 17–35. Amsterdam: John Benjamins.

Hopper, Paul J. (1998). Emergent grammar. In *The New Psychology of Language: Cognitive and Functional Approaches to Language Structure*, Michael Tomasello (ed.), 155–176. Mahwah, NJ: Lawrence Erlbaum.

Hopper, Paul J. and Traugott, Elizabeth Closs. (2003). *Grammaticalization*, 2nd ed. Cambridge: Cambridge University Press.

Horie, Kaoru (ed.) (2000). *Complementation: Cognitive and Functional Perspectives*. Amsterdam: Benjamins.

Horn, Laurence R. (2001). *A Natural History of Negation*. Stanford: CSLI.

Hudson, Richard (1995). Competence without COMP? In *The Verb in Contemporary English: Theory and Description*, Bas Aarts and Charles F. Meyer (eds.), 40–53. Cambridge: Cambridge University Press.

Jespersen, Otto (1967). *Growth and Structure of the English Language*. Oxford: Oxford University Press.

Kayne, Richard (1991). ECP extensions. *Linguistic Inquiry* 12(1), 93–133.

Kirch, Max S. (1959). Scandinavian influence on English syntax. *Publications of the Modern Language Association of America* 74, 503–510.

Kirkby, J. (1971 [1746]). *A New English Grammar*. Reprint. Menston: Scolar Press.

Labov, William (1972). *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

Labov, William (1984). Field methods of the project on linguistic change. In *Language in Use: Readings in Sociolinguistics*, John Baugh and Joel Sherzer (eds.), 28–54. Englewood Cliffs, NJ: Prentice-Hall.

Laks, Bertrand (1992). La linguistique variationniste comme méthode. *Languages* 108, 34–50.

Langacker, Ronald W. (1991). *Foundations of Cognitive Grammar*, Vol. II. Stanford, CA: Stanford University Press.

López Couso, María José (1996). A look at that/zero variation in Restoration English. In *English Historical Linguistics 1994: Papers from the 8th International Conference on English Historical Linguistics*, Derek Britton (ed.), 271–286. Amsterdam and Philadelphia: John Benjamins.

Lowth, Robert (1967 [1762]). *A Short Introduction to English Grammar, with Critical Notes.* Reprint. Menston: Scolar Press.

Milroy, James and Milroy, Lesley (1999). *Authority in Language: Investigating Standard English*, 3rd ed. London: Routledge.

Mitchell, Bruce (1985). *Old English Syntax*. Oxford: Clarendon Press.

Moulton, Keir (2002). Clausal modification in Old English: the case of the correlative. Unpublished Master's thesis, University of Toronto.

Murray, L. (1968 [1795]). *English Grammar*. Reprint. Menston: Scolar Press.

Noonan, Michael (1985). Complementation. In *Language Typology and Syntactic Description*, Volume II, Timothy Shopen (ed.), 42–139. Cambridge: Cambridge University Press.

Ogura, Michiko (1979). *Cweðan* and *secgan* in Old English prose. *Bunken Ronshu* 4, 1–30.

Palander-Collin, Minna (1999). *Grammaticalization and Social Embedding: I THINK and METHINKS in Middle and Early Modern English*. Helsinki: Société Néophilologique.

Paolillo, John C. (2002). *Analyzing Linguistic Variation: Statistical Models and Methods*. Stanford, CA: CSLI.

Pesetsky, David (1998). Some optimality principles of sentence pronunciation. In *Is the Best Good Enough? Optimality and Competition in Syntax*, Pilar Barbosa, Danny Fox, Paul Hagstrom, Martha McGinnis, and David Pesetsky (eds.), 337–383. Cambridge, MA: MIT Press.

Poplack, Shana (1989). The care and handling of a megacorpus: the Ottawa-Hull French project. In *Language Change and Variation*, Ralph Fasold and Deborah Schiffrin (eds.), 411–444. Amsterdam and Philadelphia: John Benjamins.

Poplack, Shana (1992). The inherent variability of the French subjunctive. In *Theoretical Analyses in Romance Linguistics*, Christiane Laeufer and Terrell A. Morgan (eds.), 235–263. Amsterdam and Philadelphia: John Benjamins.

Poplack, Shana and Meechan, Marjory (1998). How languages fit together in codemixing. *International Journal of Bilingualism* 2, 127–138.

Poplack, Shana and Tagliamonte, Sali (1996). Nothing in context: variation, grammaticization and past time marking in Nigerian Pidgin English. In *Changing Meanings, Changing Functions: Papers Relating to Grammaticalization in Contact Languages*, Philip Baker and Anand Syea (eds.), 71–94. London: University of Westminster Press.

Poplack, Shana and Tagliamonte, Sali (1999). The grammaticization of *going to* in (African American) English. *Language Variation and Change* 11, 315–342.

Poplack, Shana and Tagliamonte, Sali (2001). *African American English in the Diaspora*. Oxford: Blackwell.

Poplack, Shana; Van Herk, Gerard; and Harvie, Dawn (2002). Deformed in the dialects: An alternative history of nonstandard English. In *Alternative Histories of English*, Peter Trudgill and Richard Watts (eds.), 87–110. London: Routledge.

Poplack, Shana; Walker, James A.; and Malcolmson, Rebecca (2006). An English "like no other"? Language contact and change in Quebec. *Canadian Journal of Linguistics* 51(2/3), 185–213.

Quirk, Randolph; Greenbaum, Sidney; Leech, Geoffrey; and Svartvik, Jan (1985). *A Comprehensive Grammar of the English language*. London and New York: Longman.

Rand, David and Sankoff, David (1990). *GoldVarb 2.1: A Variable Rule Application for the Macintosh*. Montréal: Centre de recherches mathématiques, Université de Montréal.

Rickford, John R. and McNair-Knox, Faye (1994). Addressee- and topic-influenced style shift: a quantitative sociolinguistic study. In *Sociolinguistic Perspectives on Register*, Douglas Biber and Edward Finegan (eds.), 235–276. Oxford: Oxford University Press.

Rissanen, Matti (1991). On the history of *that*/zero as object clause links in English. In *English Corpus Linguistics: Studies in Honour of Jan Svartik*, Karin Aijmer and Bengt Altenberg (eds.), 272–289. London: Longman.

Rohdenburg, Günter (1996). Cognitive complexity and increased grammatical explicitness in English. *Cognitive Linguistics* 7(2), 149–182.

Roland, Douglas; Elman, Jeffrey L.; and Ferreira, Victor S. (2003). Why 'that'? Paper presented at Architectures and Mechanisms for Language Processing 2003, University of Glasgow.

Rosenbaum, Peter S. (1967). *The Grammar of English Predicate Complement Constructions.* Cambridge, MA: MIT Press.

Rydén, Mats (1979). *An Introduction to the Historical Study of English Syntax*. Stockholm: Almqvist and Wiksell.

Sankoff, David (1988a). Sociolinguistics and syntactic variation. In *Linguistics: The Cambridge Survey*, vol. iv, Frederick J. Newmeyer (ed.), 140–161. Cambridge: Cambridge University Press.

Sankoff, David (1988b). Variable rules. In *Sociolinguistics: An International Handbook of the Science of Language and Society*, Ulrich Ammon, Norbert Dittmar, and Klaus J. Mattheier (eds.), 984–997. Berlin and New York: Walter de Gruyter.

Sankoff, David and Rousseau, Pascale (1989). Statistical evidence for rule ordering. *Language Variation and Change* 1(1), 1–18.

Scheibman, Joanne (2000). *I dunno*: a usage-based account of the phonological reduction of *don't* in American English conversation. *Journal of Pragmatics* 32, 105–124.

Schiffrin, Deborah (1987). *Discourse Markers.* Cambridge: Cambridge University Press.

Schwenter, Scott (1994). The grammaticalization of an anterior in progress: Evidence from a Peninsular Spanish dialect. *Studies in Language* 18(1), 71–111.

Stowell, Tim (1981). *Origins of Phrase Structure*. Unpublished doctoral dissertation, Massachusetts Institute of Technology.

Suárez Gómez, Cristina (2000). *That*/zero variation in private letters and drama (1420–1710): a corpus-based approach. *Miscelánea: A Journal of English and American Studies* 21, 179–204.

Tagliamonte, Sali and Smith, Jennifer (2005). No momentary fancy! The zero 'complementizer' in English dialects. *English Language and Linguistics* 9(2), 289–309.

Temperley, David (2003). Ambiguity avoidance in English relative clauses. *Language* 79, 464–484.

Terrell, Tracey and Hooper, Joan (1974). A semantically based analysis of mood in Spanish. *Hispania* 57, 484–494.

Thompson, Sandra A. (2002). 'Object complements' and conversation. *Studies in Language* 26(1), 125–164.

Thompson, Sandra A. and Hopper, Paul J. (2001). Transitivity, clause structure, and argument structure: evidence from conversation. In *Frequency and the Emergence of Linguistic Structure*, Joan Bybee and Paul Hopper (eds.), 27–60. Amsterdam and Philadelphia: John Benjamins.

Thompson, Sandra A. and Mulac, Anthony (1991a). A quantitative perspective on the grammaticization of epistemic parentheticals in English. In *Approaches to Grammaticalization, Volume II*, Bernd Heine and Elizabeth C. Traugott (eds.), 313–329. Amsterdam and Philadelphia: John Benjamins.

Thompson, Sandra A. and Mulac, Anthony (1991b). The discourse conditions for the use of the complementizer *that* in conversational English. *Journal of Pragmatics* 151, 237–251.

Torres Cacoullos, Rena (2001). From lexical to grammatical to social meaning. *Language in Society* 30(3), 443–478.

Traugott, Elizabeth Closs (1989). On the rise of epistemic meanings in English: an example of subjectification in semantic change. *Language* 65, 31–55.

Traugott, Elizabeth Closs (1995). Subjectification in grammaticalisation. In *Subjectivity and Subjectivisation: Linguistic Perspectives*, Dieter Stein and Susan Wright (eds.), 31–54. Cambridge: Cambridge University Press.

Traugott, Elizabeth Closs (2003). Constructions in grammaticalization. In *The Handbook of Historical Linguistics*, Brian D. Joseph and Richard D. Janda (eds.), 624–647. Oxford: Blackwell.

Underhill, Robert (1988). The discourse condition for *that*-deletion. Unpublished manuscript, San Diego State University.

Verspoor, Marjolijn (2000). Iconicity in English complement constructions. *Complementation: Cognitive and Functional Perspectives*, Kaoru Horie (ed.), 199–225. Amsterdam and Philadelphia: John Benjamins.

Walker, James A. (2000). Rephrasing the copula: contraction and zero in Early African American English. In *The English History of African American English*, Shana Poplack (ed.), 35–72. Oxford: Blackwell.

Walker, James A. and Torres Cacoullos, Rena (in prep.). Prosody and processing in *that*-variation. York University and University of New Mexico.

Warner, Anthony (1982). *Complementation in Middle English and the Methodology of Historical Syntax: A Study of the Wyclifite Sermons*. London and Canberra: Croom Helm.

Wierzbicka, Anna (1988). *The Semantics of Grammar*. Amsterdam and Philadelphia: John Benjamins.

Wright, Laura (ed.) (2000). *The Development of Standard English 1300–1800: Theories, Descriptions, Conflicts*. Cambridge: Cambridge University Press.

Yaguchi, Michiko (2001). The function of the nondeictic *that* in English. *Journal of Pragmatics* 33, 1125–1155.