A multifactorial analysis of that/zero alternation

Shank, Christopher; Plevoets, K.; Van Bogaert, J.

Corpus-based approaches to Construction Grammar

Published: 08/09/2016

Peer reviewed version

To appear in “Corpus-based approaches to Construction Grammar” John Benjamin Publishing Company (CAL series).

A multifactorial analysis of *that*/zero alternation: The diachronic development of the zero complementizer with *think*, *guess* and *understand*.

C. Shank, K. Plevoets and J. Van Bogaert.
Bangor University, University College Ghent and Ghent University.

Email: c.shank@bangor.ac.uk

Corresponding author: Christopher Shank

Running head: Diachronic development of the zero complementizer - think, guess and understand.

Address:
Christopher Shank
School of Linguistics & English Language
Bangor University
Bangor, Gwynedd LL57 2DG
United Kingdom
c.shank@bangor.ac.uk
Phone: +44 01248 38 3590

Koen Plevoets
University College Ghent − Faculty of Applied Language Studies
Groot-Brittanniëlaan 45
B-9000 Ghent
Belgium
koen.plevoets@ugent.be
Tel: +32 9 224 97 30

Julie Van Bogaert
Ghent University
Department of Linguistics
Blandijnberg 2
B-9000 Ghent
Belgium
Julie.VanBogaert@UGent.be
Tel: +32 9 264 37 91

Abstract

This corpus-based study uses a stepwise logistic regression analysis to examine the diachronic development of *that*/zero alternation with three verbs of cognition, viz. *think, guess* and *understand* in both spoken and written corpora from 1560-2012. Eleven structural features which have been claimed in the literature to predict the presence of the zero complementizer form are tested to see if (1) there is indeed a diachronic trend towards more zero use, (2) whether the conditioning factors proposed in the literature indeed predict the zero form, (3) to what extent these factors interact and (4) whether the predictive power of the conditioning factors becomes stronger or weaker over time. The analysis disproves the hypothesis that there has been an overall diachronic development towards more zero use and that the interactions with verb type brings to light differences between verbs in terms of the predictive power of the individual structural features.

1.0 Introduction

This paper is concerned with the alternation between the complementizer *that* and the zero complementizer in constructions with an object clause, as in (1) and (2).

(1) I think that he is a powerful man. (COCA)

(2) I think they're going to blame him. (COCA)

In previous studies, it has been suggested that this complementation construction has been evolving towards more zero use (Rissanen 1991; Thompson & Mulac 1991; Palander-Collin 1999). The present paper seeks to test this hypothesis by means of a stepwise logistic regression analysis of (n= 9759) tokens of *think, guess* and *understand*, three of the most frequently used complement-taking verbs of cognition, spanning the time period from 1560 to 2012.The literature has put forward a number of conditioning factors promoting the zero form. Our regression model will test whether these features indeed predict the zero form, whether they gain or lose predictive power when combined and what happens to their predictive power over time. Determining the interaction of time with each of the structural conditioning factors, this study adds an innovative diachronic perspective to existing research into zero/*that* alternation by testing the effect of each factor over time on the selection of the zero complementizer.

We start off with a review of the literature dealing with *that*/zero alternation in order to characterize the construction under investigation and to review the factors that have previously been said to condition the use of either *that* or zero complementation. In Section 3 our data and methodology are explained. After presenting our results in Section 4, we offer a conclusion in Section 5.

2.0 Review of the Literature

2.1 That*/zero alternation and the emergence of discourse formulas and parentheticals*

In usage-based approaches to the *that*/zero alternation (Thompson and Mulac 1991a, 1991b; Aijmer 1997; Diessel and Tomasello 2001; Thompson 2002), frequently occurring subject–verb combinations, e.g. *I think* and *I guess*, are considered to have developed into conventionalized "epistemic phrases" (Thompson and Mulac 1991a, 1991b) or "discourse formulas" (Torres Cacoullos and Walker 2009). Torres Cacoullos and Walker (2009) argue that such discourse formulas have reached a high degree of autonomy (see Bybee 2003, 2006) from their productive complement-taking source construction. The frequency with which the zero complementizer is used is seen as an indication of this increasing autonomy. Following this rationale, Thompson and Mulac (1991b) argue that the absence of *that* points towards the blurring of the distinction between matrix clause and complement clause, i.e. to a reanalysis of this [MATRIX + COMPLEMENT CLAUSE] construction as a monoclausal utterance in which the complement clause makes the "main assertion" (Kearns 2007a), for which the matrix clause provides an epistemic or evidential "frame" (Thompson 2002).[1] Thompson and Mulac (1991b) show that the subject–verb collocations with the highest frequency of occurrence have the greatest tendency to leave out the complementizer *that.* It is exactly these sequences that "are most frequently found as EPAR [epistemic parenthetical] expressions" (Thompson and Mulac 1991b: 326),[2] which occur in clause-medial or final position with respect to the (erstwhile) complement clause.

[1]Bas Aarts (p.c.) has pointed out that syntactically *I think* can never be a clause; it has no syntactic status as it is not a constituent. Therefore, strictly speaking, in a sentence like (1), the matrix clause is the entire sentence starting with *I* and ending in *man*. In the literature, however, the terms "matrix clause" and "main clause" are commonly used to denote the matrix clause without its complement, i.e. in the case of (1), to refer to *I think.* For the sake of clarity and consistency, this practice will be followed in the current paper.

[2] What Thompson and Mulac mean by this is that the bulk of all the "matrix clauses" in their data are tokens of *think* and *guess* and that these same verbs make up the largest share of all parenthetical uses in

(3) *We have to kind of mix all this together, I think, to send the right message to girls.* (COCA)

These synchronic, frequency-based findings lead Thompson and Mulac (1991b) to propose that *that* complementation (1), zero complementation (2), and parenthetical use (3) embody three degrees or three stages in a process of grammaticalization into epistemic phrases/parentheticals.[3] A study on the use of *I think* in Middle and Early Modern English by Palander-Collin (1999) adds support to the diachronic validity of this grammaticalization path. Her data show an increase in the use of *I think* with the zero complementizer and a concomitant rise in parenthetical use.

Brinton (1996), on the other hand, takes issue with what she calls the "matrix clause hypothesis" and presents an alternative model which posits a paratactic construction with an anaphoric element rather than a complement-taking construction as the historical source construction. Brinton's proposal is consistent with Bolinger (1972: 9), who states that "both constructions, with and without *that*, evolved from a parataxis of independent clauses, but in one of them the demonstrative *that* was added".

(4) Stage I: *They are poisonous*. That I think.

the corpus, i.e. 85%. This does *not* mean that *think* and *guess* have the highest rates of parenthetical use when all instances of each target verb are aggregated and the share of parenthetical use is calculated for each separate verb. When this method is applied to Thompson and Mulac's data, the respective parenthetical rates of *think* and *guess* are 10% and 29%.

[3]For a discussion of the applicability of grammaticalization, pragmaticalization, and lexicalization to this type of construction, see Fischer (2007) and Van Bogaert (2011).

Stage II: *They are poisonous,* {that I think, I think that/it, as/so I think}. = 'which I think'

Stage III: *They are poisonous,* I think. OR

They are poisonous, as I think. = 'as far as I think, probably'

Stage IV: I think, *they are poisonous*. *They are,* I think, *poisonous*.

(Brinton 1996: 252)

Along similar lines, Fischer (2007) posits two source constructions for present-day parentheticals: what Quirk et al. (1985: 1111) have called subordinate clauses of proportion and the seeming zero-complementation patterns that Gorrell (1895: 396–397; cited in Brinton 1996: 140 and Fischer 2007: 103) designates as "simple introductory expressions like the Modern English 'you know'", which stand in a paratactic relationship with the ensuing clause. Fischer (2007: 106) classifies the anaphoric connective element introducing such independent clauses as an adverbial derived from a demonstrative pronoun.

The notion of reanalysis, on which Thompson and Mulac's (1991a, 1991b) account of epistemic parentheticals is based, has been subject to additional criticism. An important point here is the role of zero complementation. Kearns (2007a), for example, does not regard the occurrence of the zero complementizer with epistemic phrases/parentheticals as a diagnostic of the syntactic reanalysis involved in their formation; rather, she accounts for zero complementation in strictly pragmatic terms: it signals a shift in information structure such that the complement clause conveys the main assertion while the matrix clause loses prominence and has a modifier-like use (see also Diessel and Tomasello 2001; Boye and

Harder 2007). These studies allow for a hybrid analysis in which some occurrences with zero complementation are adverbial in terms of function while syntactically retaining their matrix clause status. A further criticism regarding reanalysis concerns the necessity of *that* omission to the use of *I think* (and similar epistemic phrases) as discourse formulas. Both Kearns (2007a) and Dehé and Wichmann (2010) argue that complement-taking predicates followed by *that*, e.g. *I think that*, may also be analyzed as discourse formulas, the whole sequence having become routinized as a whole. In addition to providing prosodic evidence for this position, Dehé and Wichmann (2010: 65) remark that this view is supported by the historical origins of *that* as a demonstrative pronoun (see the discussion of Brinton 1996 and Fischer 2007 above).[4]

In this study, we adopt the matrix clause hypothesis insofar as we aim to test Thompson and Mulac's grammaticalization hypothesis that there is a tendency across time for the zero complementizer to be preferred over the complementizer *that*, i.e. that the verbs under investigation in this study (*think, suppose, believe*) have tended towards higher frequencies of the zero complementizer as conditioned by the factors presented in Section 3. Ascertaining the main effects of these conditioning factors, we determine which ones are good predictors of the zero form. The present study is innovative in approaching the *that*/zero alternation from both a quantitative and a diachronic point of view. While Tagliamonte and Smith (2005) and Torres Cacoullos and Walker (2009) have performed multifactorial analyses of the synchronic conditioning of *that* and zero complementation, the current paper adds a diachronic dimension along with a parallel analysis of diachronic spoken and written data sets, and investigates, by means of a stepwise regression analysis,

[4] For more references on the question whether clause-initial occurrences of "parenthetical verbs" should be considered as matrix clauses or as parentheticals, see Kaltenböck (2007: 5–6).

whether the zero form is on the increase and how time affects the predictive power of the factors. In addition to interactions with time, this study seeks to lay bare any other significant interactions between factors, notably mode (i.e. spoken versus written data), and to identify any resulting similarities and/or differences between the three verbs of cognition.

2.2 *A concise history of the* that*/zero alternation*

There is general agreement on the historical development of the complementizer *that* from an Old English neuter demonstrative pronoun (see, for instance, Mitchell 1985), but the question which of the two complementation patterns, *that* or zero, is older is strictly speaking impossible to answer as both the *that* and the zero complementizer occur in the earliest extant texts (Rissanen 1991).[5] This renders the notion of "*that*-deletion" or "omission" somewhat problematic. On the other hand, it should be observed that in Old English and throughout most of the Middle English period, occurrences of zero are scant. In Warner's (1982) study of the Wycliffe Sermons, for example, *that* is used 98% of the time. It is not until the Late Middle English period that the zero complementizer gradually takes off (Rissanen 1991; Palander-Collin 1999), a trend that continues in Early Modern English. Rissanen (1991) notes a steady increase between the fourteenth and the seventeenth century, but the most dramatic rise in the zero complementizer can be observed in the second half of the sixteenth century and in the early seventeenth century, when its frequency jumps from 40% to 60%. In addition, Rissanen (1991) shows that the zero form

[5]According to Bolinger (1972), there is a semantic difference between constructions with and without *that* due to a trace of the original demonstrative meaning being retained in present-day uses of explicit *that*. For Yaguchi (2001), too, this demonstrative meaning continues to condition the contemporary function of *that.*

is more common in speech-like genres (i.e. trials, comedies, fiction, and sermons) and that its increase is more pronounced with *think* and *know* than with *say* and *tell*. Finegan and Biber (1985), too, find that the zero complementizer is more frequent in the more colloquial genre of the personal letter than in the formal genres of medical writing and sermons.[6] In the eighteenth century, we witness a temporary drop in zero use. Both Rissanen (1991) and Torres Cacoullos and Walker (2009) attribute this change to the prevalence of prescriptivism, which advocated the use of *that* out of a concern with clarity.

2.3 *Conditioning factors in the literature*[7]

Jespersen puts the variability between *that* and zero down to nothing more than "momentary fancy" (1954: 38, cited in Tagliamonte and Smith 2005: 290); as will be seen, this is a claim that several scholars have tried to refute through an examination of a wide range of conditioning factors. Some of these factors are of a language-external nature; many are language-internal.

Many previous studies have tried to account for *that*/zero variability from the point of view of register variation (Quirk et al. 1985: 953; Huddleston and Pullum 2002: 317; see Rohdenburg 1996 for more references); *that* tends to be regarded as the more formal option while zero is associated with informal registers (see Kaltenböck 2006: 373–374 for references. For example Kearns (2007b) observes some significant differences across varieties in newspaper prose and attributes these to different degrees of sensitivity to some of the conditioning factors discussed further down in this section.

[6] This predilection for zero in speech is confirmed in studies of contemporary English (see Tagliamonte and Smith 2005: 291–293).

[7] Although the scope of this article is restricted to *that*/ zero complementizer alternation in so-called object clauses, some of the studies discussed in this section also deal with subject clauses.

There is also a wide range of language-internal factors. One semantic factor is discussed in Dor (2005), who notes that the semantic notion of the "truth claim" is crucial to the *that*/zero alternation, in that *that*-clauses denote "propositions" while zero-clauses denote "asserted propositions". Also, particular semantic classes of verbs, notably "epistemic verbs" (Thompson and Mulac 1991a) or "propositional attitude predicates" (Noonan 1985; Quirk et al. 1985) turn out to have a stronger preference for zero complementation than other complement-taking verbs, such as utterance or knowledge predicates (Thompson and Mulac 1991a; Tagliamonte and Smith 2005; Torres Cacoullos and Walker 2009).

Importantly, various studies have shown certain high-frequency subject-verb collocations to be strongly associated with zero use (among these are "epistemic verbs" mentioned above). Torres Cacoullos and Walker (2009: 32) therefore hypothesize that the conditioning factors for complementizer choice should be different for these highly frequent "discourse formulas" (viz. *I think, I guess, I remember, I find, I'm sure, I wish*, and *I hope*) than for the (relatively more) productive complement-taking construction, and indeed they find a number of differences in terms of significance and effect size.

Finally, a wide array of language-internal, structural factors operating on the selection of zero or *that* have been proposed in previous studies, some of which employ statistical methods, of diverse levels of refinement, to ascertain the import of these factors. In the following three sections, the structural conditioning factors favoring the use of zero will be discussed on the basis of the literature. The factors have been divided into three groups depending on whether they concern matrix clause features, complement clause features, or the relationship between the two. At the end of each section, a table provides a

summary of the factors discussed. For each factor, we indicate whether previous studies have or have not statistically tested the factor's predictive power, and if so, whether it came out as significant or not.

2.3.1 *Matrix clause elements*

The subject of the matrix clause has often been said to play a role in the selection of either *that* or zero. In many studies, it is argued that pronouns, particularly *I* or *you* (5), favor the use of zero (Bolinger 1972; Elsness 1984; Thompson and Mulac 1991a; Tagliamonte and Smith 2005; Torres Cacoullos and Walker 2009).[8] While it is mostly assumed that the pronouns *I* and *you* in particular promote the use of zero, Torres Cacoullos and Walker (2009: 26) demonstrate that the difference in effect size between pronouns (5a) and full NPs (5b) is greater than that between *I* or *you* versus all other subject types, including full NPs. They conclude that the strong effect attributed specifically to *I* and *you* in Thompson and Mulac (1991a: 242) is due to the inclusion of discourse formulas like *I think* and *I guess* in the data, which Torres Cacoullos and Walker consider separately.

(5) a. *but <u>I</u> think a portion of it must have fallen down upon the straw*. (OBC)

b. *<u>Some people</u> think that maybe it was a crazy person that stalked Tara*. (COCA)

[8] In these studies, no distinction is made between declarative and interrogative second person use, although Thompson and Mulac (1991b: 322) indicate that the majority (82%) of their second-person instances of epistemic parentheticals are in the interrogative mood. In the current study, interactions between mood and person as conditioning factors for the selection of *that* or zero are taken into account.

Another matrix clause factor that has received considerable attention is the presence or absence of additional material in the matrix clause. It is believed that matrix clauses containing elements other than a subject and a (simplex) verb are more likely to be followed by *that.* Such elements may be adverbials, negations, or periphrastic forms in the verbal morphology of the matrix clause predicate (Thompson and Mulac 1991a; Torres Cacoullos and Walker 2009).[9] For Tagliamonte and Smith (2005: 302), "additional material" is operationalized as "negation, modals, etc.", including adverbials (Tagliamonte p.c.). In Torres Cacoullos and Walker (2009: 26–27), as far as discourse formulas are concerned, adverbial material in the matrix clause is the conditioning factor making the greatest contribution to the selection of *that*. The authors explain that "this is unsurprising, since the presence of a post-subject adverbial … detracts from (in fact, nullifies) the formulaic nature of the collocation". Distinguishing between single-word (6a) as opposed to phrasal adverbials (6), and pre-subject (6) as opposed to post-subject (6) adverbials in the matrix clause, they find that post-subject adverbials affect both discourse formulas and "productive" constructions while the effect of pre-subject adverbials is restricted to discourse formulas. Phrasal adverbials are different again, promoting the use of *that* only with productive constructions.

(6) a. *I expected <u>maybe</u> that we would be talking about it*.

b. *<u>At the beginning</u>, we told the guy that we were gonna both-each have our own*.

[9]Although periphrastic verb forms in the matrix clause is generally believed to "reduce the likelihood that the main subject and verb are being used as an epistemic phrase" (Thompson and Mulac 1991a: 248), both Kearns (2007a) and Van Bogaert (2010) have argued that such modifying use is not restricted to the prototypical first (or second) person simple present form.

c. *Now I find Ø like, even adults use slang words*.

d. *I totally thought Ø he was a big jerk.*

(Torres Cacoullos and Walker 2009: 15-16)

As for verbal morphology, the presence of auxiliaries in the matrix clause (7) is also believed to be conducive to the use of *that* (Thompson and Mulac 1991a: 246; Torres Cacoullos and Walker 2009: 16). As such, Tagliamonte and Smith (2005) show the simple present to be a significant factor contributing to the use of zero and in Torres Cacoullos and Walker (2009: 27) finite matrix verbs are more favorably disposed towards zero complementation than non-finite forms.[10]

Negation (8), subsumed under "additional material" in Tagliamonte and Smith (2005), is treated as a separate conditioning factor for the use of the complementizer *that* in Thompson and Mulac (1991a: 245), but was found to be not significant. By the same token, the interrogative mood (9) failed to reach significance.

(7) *I would guess that Al Gore will not endorse anyone*. (COCA)

(8) *I don't think they said it was a match*. (COCA)

(9) *Do you think he was talking to the left*? (COCA)

A summary of matrix clause factors is presented in Table 1.

Table 1: Matrix clause factors potentially favoring the zero complementizer

[10] Tagliamonte and Smith (2005: 25) use the term "present", but in fact "simple present" is meant: "present tense, when there are no additional elements in the matrix verb phrase".

Factor	No statistics	Significant	Not significant
subject = pronoun		Torres Cacoullos and Walker (2009)	
subject = *I*		Tagliamonte and Smith (2005)	
subject = *I* or *you*	Elsness (1984)	Thompson and Mulac (1991b)	Kearns (2007a, 2007b)
absence of matrix-internal elements		Tagliamonte and Smith (2005)	
absence of post-subject adverbials		Thompson and Mulac (1991b) Torres Cacoullos and Walker (2009)	
absence of pre-subject adverbials		Torres Cacoullos and Walker (2009)	
absence of phrasal adverbials		Torres Cacoullos and Walker (2009)	
positive polarity	Finegan and Biber (1985)		Thompson and Mulac (1991b)
declarative mood			Thompson and Mulac (1991b)

2.3.2 *Complement clause elements*

Concerning the subject of the complement clause, it has been suggested that pronominal subjects (10) as opposed to full NPs (11) favor the use of zero (Warner 1982; Elsness 1984; Finegan and Biber 1985; Rissanen 1991; Thompson and Mulac 1991a; Rohdenburg 1996, 1998; Tagliamonte and Smith 2005; Torres Cacoullos and Walker 2009).

(10) *Bill, I understand you have a special guest with you*. (COCA)

(11) *Well, I'm not, because I understand that most of his girlfriends have either been, you know, like the hooker or porn star types*. (COCA)

The high discourse topicality of pronouns has been proposed as an explanatory principle (Thompson and Mulac 1991a: 248), as well as Rohdenburg's (1996: 151) complexity principle, which states that "in the case of more or less explicit grammatical options the more explicit one(s) will tend to be favored in cognitively more complex environments". While Elsness (1984) regards *I* and *you* as particularly conducive to zero complementation, Torres Cacoullos and Walker's (2009: 28) multivariate study results in the following ordering of subjects from least to most favorable to *that: it/there* < *I* < other pronoun < NP. Elsness (1984) adds that short NPs and NPs with definite or unique reference are more likely to select the zero variant than longer and indefinite NPs. In Kearns (2007a: 494), first and second person subjects (i.e. *I, you* but also *we*) are compared to third person subjects, but identical rates of zero and *that* are found for both data sets. Kearns (2007a: 493; 2007b: 304) also examines the length of the complement clause subject as a possible factor, operationalizing it in terms of a three-way distinction between pronouns, short NPs

(one or two words) and long NPs (three or more words). The study reveals significant differences, including one between short and long NPs.

As an additional complexity factor, Rodhenburg (1996: 164) mentions the overall length of the complement clause. He suggests that longer complement clauses tend to favor explicit *that* and in this regard he finds that at least with the verbs *think* and *know*, complement clauses introduced by *that* are "on average much longer than those not explicitly subordinated" (Rohdenburg 1996: 164).

A summary of complement clause factors is presented in Table 2.

Table 2: Complement clause factors potentially favoring the zero complementizer

Factor	**No statistics**	**Significant**	**Not significant**
subject = pronoun	Warner (1982) Elsness (1984) Finegan and Biber (1985) Rissanen (1991) Rohdenburg (1996, 1998)	Thompson and Mulac (1991b) Tagliamonte and Smith (2005) Torres Cacoullos and Walker (2009)	
subject = *I* or *you*	Elsness (1984)		
subject = *I, you* or *we*			Kearns (2007a, 2007b)

subject = nominative pronoun			Kearns (2007a, 2007b)
short subject	Elsness (1984)	Kearns (2007a, 2007b)	
definite/unique reference	Elsness (1984)		
referential *it*			Kearns (2007a, 2007b)
long complement clause	Rohdenburg (1996)		
intransitive verb		Torres Cacoullos and Walker (2009)	

2.3.3 *The relationship between matrix and complement clause*

Finally, the presence of intervening material between matrix and complement has been widely discussed as a factor favoring the complementizer *that* (Bolinger 1972; Warner 1982; Finegan and Biber 1985; Rissanen 1991; Rohdenburg 1996; Tagliamonte and Smith 2005; Torres Cacoullos and Walker 2009). Besides potentially leading to ambiguity, which Rohdenburg (1996: 160) regards as a special type of cognitive complexity, the presence of intervening material, as in (12), has been related to a heavier cognitive processing load. In Rohdenburg' (1996: 161) words, "any elements capable of delaying the processing of the object clause and thus the overall sentence structure favor the use of an explicit signal of

subordination". Conversely, adjacency of matrix and complement clause is believed to minimize syntactic and cognitive complexity (Torres Cacoullos and Walker 2009), and thus promote the zero complementizer. In Kearns (2007b), adjacency came out as a key factor responsible for regional differences in zero-complementizer rates, with some varieties being more dependent on adjacency for the licensing of zero than others.

(12) *Well, I'm not, because I understand that most of his girlfriends have either been, you know, I think <u>personally</u> that with time we're going to continue to see positive change*. (COCA)

In Torres Cacoullos and Walker's (2009: 27) study, intervening material – on a par with the complement clause subject – is the factor with the greatest effect on complementizer alternation, at least as regards regular, productive complement-taking verbs; as for high-frequency discourse formulas, the factor with the biggest effect size is the use of matrix clause adverbials (2009: 32–33).

Thompson and Mulac (1991a), Rohdenburg (1996), and Torres Cacoullos and Walker (2009) examine the effect of intervening verbal arguments, as in (13). The factor came out as significant in both Thompson and Mulac (1991a) and Torres Cacoullos and Walker (2009), although in the latter study, the effect is smaller than with other intervening material. As with complement clause subjects, Rohdenburg (1996: 162) points out that pronominal arguments as opposed to full NPs are more amenable to the zero form.

(13) *Within a week, I told him that I'm transgendered, and he was like, you know, what are you talking about?* (COCA)

In Torres Cacoullos and Walker (2009: 7–8), three factors are tested that fall under the explanatory principle of semantic proximity, which predicts the selection of the zero form when the conceptual distance between matrix and complement is minimal.[11] Specifically, subject coreferentiality (14), a factor that was significant in one of Elsness's (1984: 526) text types, cotemporality (15), and harmony of polarity (16), first proposed by Bolinger (1972), are examined, but none of these factors reach significance. Subject coreferentiality is also examined by Kearns (2007a: 493; 2007b: 304), but the factor is not selected as significant.

(14) *I think I nodded several times*. (COCA)

(15) *I parted with my money as I thought it was a very good opening*. (OBC)

(16) *And I think it will rebound on the Democrats*. (COCA)

Table 3 summarizes the factors pertaining to the relationship between matrix and complement clause.

Table 3: Factors pertaining to the relationship between matrix and complement which potentially favor zero

Factor	**No statistics**	**Significant**	**Not significant**

[11]Conceptual distance needs to be interpreted in terms of Givón's (1980) hierarchy of clause-binding or in terms of the iconic separation of the two clauses (Langacker 1991; Givón 1995; Torres Cacoullos and Walker 2009).

absence of intervening material	Bolinger (1972) Warner (1982) Finegan and Biber (1985) Rissanen (1991) Rohdenburg (1996)	Tagliamonte and Smith (2005) Torres Cacoullos and Walker (2009)	
absence of intervening arguments	Rohdenburg (1996)	Thompson and Mulac (1991b) Torres Cacoullos and Walker (2009)	
subject coreferentiality		Elsness (1984)	Kearns (2007a, 2007b) Torres Cacoullos and Walker (2009)
cotemporality			Torres Cacoullos and Walker (2009)
harmony of polarity	Bolinger (1972)		Torres Cacoullos and Walker (2009)

2.3.4 *Non-structural factors*

In this final section on factors conditioning the selection of *that* or zero, one last type of non-structural conditioning will be discussed: prosodic realization.

Dehé and Wichmann (2010) argue that there are rhythmic factors constraining the presence or absence of *that.* They point out that the explicit use of *that* may be motivated by a desire to create a more regular stress pattern in which *that* provides an additional unstressed syllable. In (17), *that* results in a regular, dactylic pattern, while in (18), it is required that *that* be *not* realized in order to obtain such regularity. Similarly, *that* may be inserted as an unstressed "buffer" between two stressed syllables in order to avoid a stress clash (Wichmann p.c.). In view of these rhythmic constraints, Dehé and Wichmann (2010: 66) conclude that "the presence or absence of *that* does not affect the way in which we analyze the function of *I* verb *(that)*". In other words, the absence of *that* is neither a necessary nor a sufficient condition for the use of an *I* verb *(that)* as a discourse formula.[12]

(17)	-		x	-	-		x	-	-		x					
	I		think	*that*	the		problem		of		faith …					
(18)	-		-	x	-		-	x	-		-	x	-		-	
	I		believe		I'm		a	bit	of		a	nightmare			then	

(Dehé and Wichmann 2010: 66, data from the ICE-GB)[13]

3.0 Data and Methods

Our analysis was based on tokens retrieved from the following spoken and written corpora:

Table 4: Spoken Corpora

Sub-period of Spoken English	Time span	Corpus	Number of words

[12] See also the discussion in Section 2.2 on the role played by the zero complementizer in the reanalysis of matrix clauses into adverbials/parentheticals/discourse formulas.
[13] The x's stand for stressed syllables the dashes for unstressed syllables.

Early Modern English (EModE)	1560–1710	*Corpus of English Dialogues (CED)*	980,320
Late Modern English (LModE)	1710 -1913	*Old Bailey Corpus (OBC)*	113,253,011
Present-Day English (PDE)	1920–2012	*The British National Corpus – spoken component. (BYU BNC-S).* *The Corpus of Contemporary American English - spoken component (COCA-S)*	95,341,792

Table 5: Written Corpora

Sub-period of Written English	Time span	Corpus	Number of words
Early Modern English (EModE)	1560–1710	*Innsbruck Corpus of Letters* *CEECS I Corpus (1560 - onward)* *CEECS II Corpus* *Corpus of English Dialogues (CED)*	2,848,314

		Corpus of Early Modern English Texts (CMET) *Lampeter Corpus (Early Modern English portion- up to 1710)*	
Late Modern English (LModE)	1710–1920	*Corpus of Late Modern English texts Extended Version* (CLMETEV) *Lampeter Corpus (Early Modern English portion (1710 - onward)*	15,413,159
Present-Day English (PDE)	1850–2009	*The Time Corpus (Time)* *The Corpus of Contemporary American English - written component (COHA)*	500,000,000

The Wordsmith concordance program was used to first to identify the total number of inflected forms of *think* (i.e. *think, thinks, thinking* and *thought*), and *guess* (i.e. *guess, guesses, guessing* and *guessed*) and *understand* (i.e. *understand, understands, understanding* and *understood*) in both the written and spoken corpora from 1560-2012 per period. Results were broken up in smaller 70-year sub-periods, as shown in Tables 6–14. The sub-periods were modeled after those contained in the CLMET corpora (i.e 1710–1780, 1780–1850, 1850–1920) in order to provide a principled template in which to divide and analyze the other diachronic written and corresponding spoken corpus data utilized in this study. The size, scope, and time periods of the other corpora in this study, especially

those outside of 1710-1920, however, did not always correspond (e.g. the Old Baily Corpus ends in 1913 or the BYU-BNC only covers a period from the 1980s to 1993), so some adjustments were necessary but every effort was taken to remain as close to a 70 year period as possible. In addition, following an initial explorative analysis with just the think data, the decision was made to subdivide the first period of 1560-1639 into 1560-1579 and 1580-1639, in order to provide a reference level for the subsequent regression analysis applied to the three verbs discussed in this paper.

For each sub-period, the relative percentage of each inflected verb form per lemma was calculated. These percentages were then applied to the extracted sets (a minimum of (n=2,000) randomized hits for written data and 1,000 randomized hits for the spoken data) in order to ensure that the extracted sets would be proportionally similar in terms of inflected forms to the larger corpora from which they were taken. This two-step process resulted in the datasets described below for each of the verbs under investigation.

Starting with the verb *think*, we began by extracting (n= 3101) tokens from the spoken English corpora and (n= 6619) tokens from the written English corpora (see Table 5). Randomization was achieved by using the Wordsmith randomization function or by selecting the "randomized sample option" available on the web based corpus resources (i.e COCA, Time, BYU-BNU, etc). The full set (n=9,720) of tokens was divided into those containing either a that-clause or a zero-complementizer clause. Those tokens not containing a that or zero form were then discarded. The resulting distributions of these tokens for both the spoken and written data sets are presented in Tables 6 and Table 7.

Table 6: Distribution of *that*-clauses and zero-complementizer clauses from EModE to PDE in spoken English. (n: absolute frequency, N: normalized frequency per million)

***think* –** spoken corpora				
	***think* - that**		***think* - zero**	
Period	n	N	n	N
1560-1579	(n=8)	92.97	(n=28)	324.78
1580-1639	(n=29)	86.37	(n=116)	345.48
1640-1710	(n=10)	23.75	(n=212)	447.47
1710-1780	(n=22)	45.64	(n=412)	854.10
1780-1850	(n=12)	26.09	(n=439)	938.68
1850-1913	(n=16)	47.50	(n=418)	1305.45
1980-1993	(n=20)	449.18	(n=142)	3152.25
1990-2012	(n=22)	471.64	(n=171)	3139.33
Total	**(n=139)**		**(n=1916)**	

Table 7: Distribution of *that*-clauses and zero-complementizer clauses from EModE to PDE in written corpora. (n: absolute frequency, N: normalized frequency per million)

***think* –** written corpora				
	***think* - that**		***think* - zero**	
Period	n	N	n	N
1560-1579	(n=21)	214.00	(n=17)	173.24

1580-1639	(n=18)	59.23	(n=133)	437.65
1640-1710	(n=65)	174.51	(n=200)	558.27
1710-1780	(n=79)	123.19	(n=290)	535.29
1780-1850	(n=103)	151.66	(n=316)	545.23
1850-1920	(n=101)	175.47	(n=359)	680.69
1920-1989	(n=40)	109.44	(n=204)	561.92
1990-2009	(n=24)	106.20	(n=247)	912.90
Total	**(n=451)**		**(n=1766)**	

A comparison of the diachronic relative frequency patterns of the *that* versus zero forms per million words with the verb *think* indicates that the zero form is clearly the more frequent form from 1560 to 2012, in both spoken and written texts, and this comports with all previous literature on *think* and claims regarding diachronic *that*/zero variation patterns.

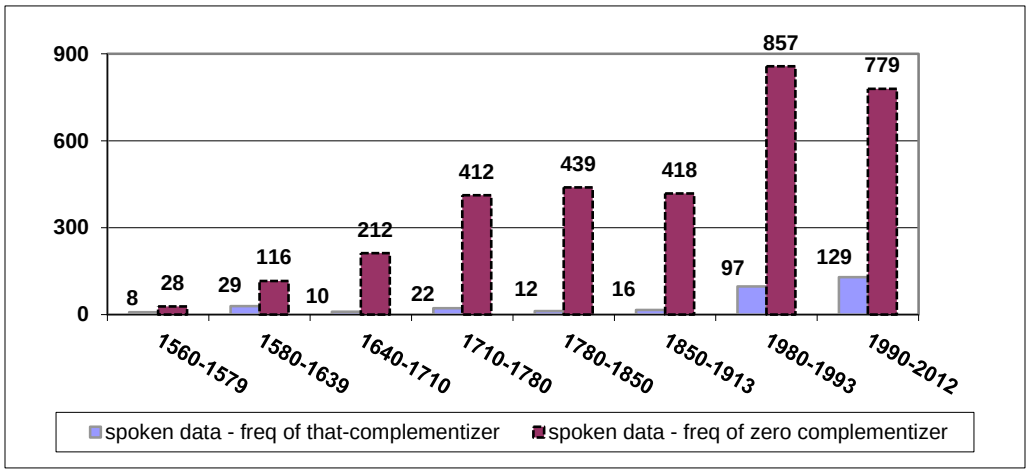

Figure 1.

Think spoken data – *that* versus zero distribution per million words

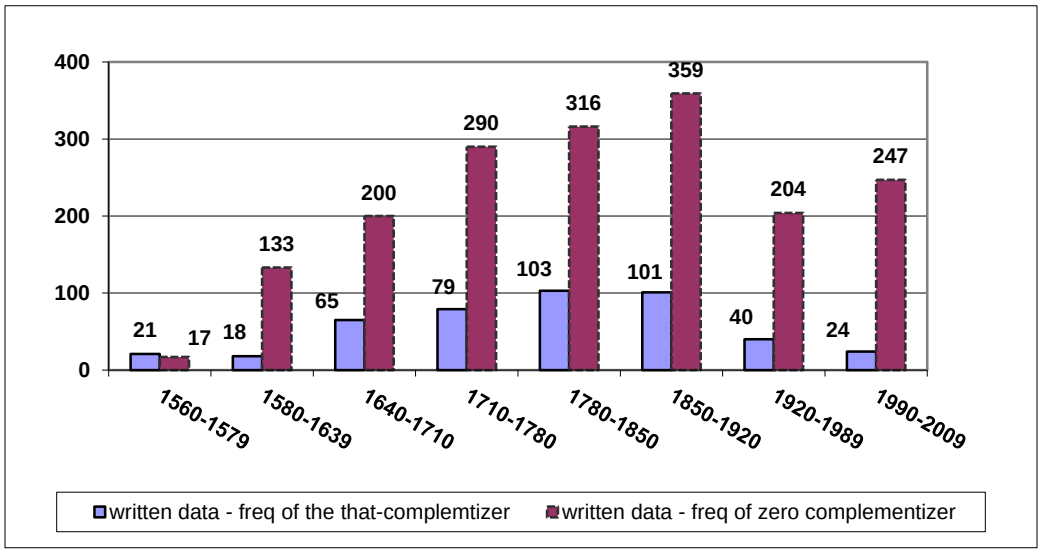

Figure 2. *Think* written data – *that* versus zero distribution per million words

The same extraction process was then performed for the verb *guess*. This yielded (n= 3419) *guess* tokens from the spoken English corpora and (n= 2255) tokens from the written English corpora. The full set (n= 5,674) of tokens was again divided into those containing either a *that*-clause or a zero-complementizer (again, with tokens not containing the *that* or zero form being discarded). The distributions of these tokens for both the spoken and written data sets are presented in Tables 8 and 9.

Table 8: Distribution of *that*-clauses and zero-complementizer clauses from EModE to PDE in spoken English. (n: absolute frequency, N: normalized frequency per million)

***guess* –** spoken corpora				
	***guess* - that**		***guess* - zero**	
Period	n	N	n	N
1640-1780	(n=2)	1.23	(n=17)	1.32
1780-1850	(n=14)	0.30	(n=23)	0.49
1850-1913	(n=14)	0.27	(n=51)	0.97
1960-1993	(n=4)	7.78	(n=348)	677.60
1994-2012	(n=39)	5.84	(n=538)	108.58
Total	**(n=73)**		**(n=977)**	

Table 9: Distribution of *that*-clauses and zero-complementizer clauses from EModE to PDE in written corpora. (n: absolute frequency, N: normalized frequency per million)

***guess* –** written corpora				
	***guess* - that**		***guess* - zero**	
Period	n	N	n	N
1580-1780	(n=15)	1.95	(n=20)	2.75
1780-1850	(n=22)	3.80	(n=38)	6.56
1850-1920	(n=58)	9.25	(n=99)	15.79

1920-1989	(n=93)	10.02	(n=154)	16.37
1990-2009	(n=22)	1.67	(n=312)	41.78
Total	**(n=210)**		**(n=623)**	

When we compare the diachronic relative frequency patterns of the *that* versus zero forms per million words for the verb *guess*, we find that the frequency of the zero form relative to the *that* complementizer is once again more frequent form in both the spoken and written data sets. The distribution pattern for both types of data is presented below in Figures 3 & 4.

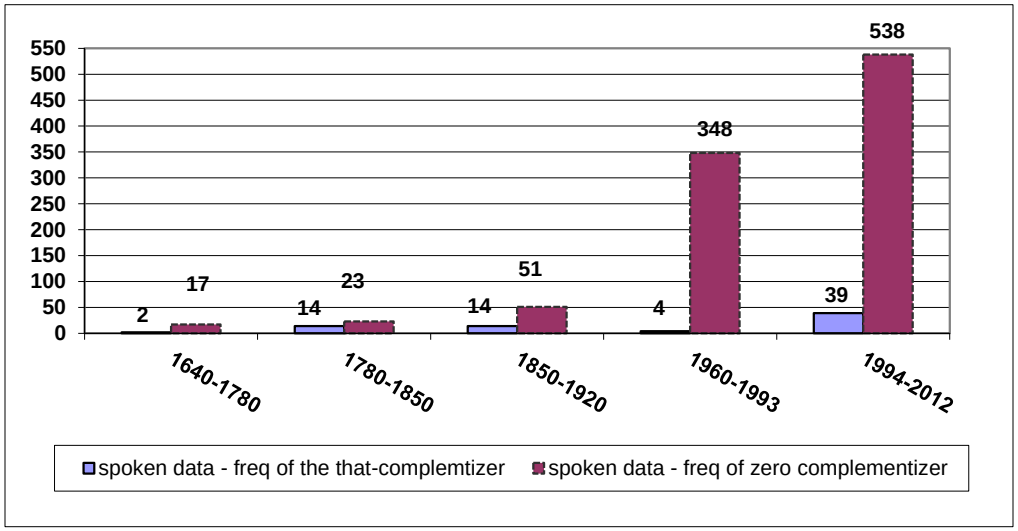

Figure 3. Spoken data – *that* versus zero distribution per million words

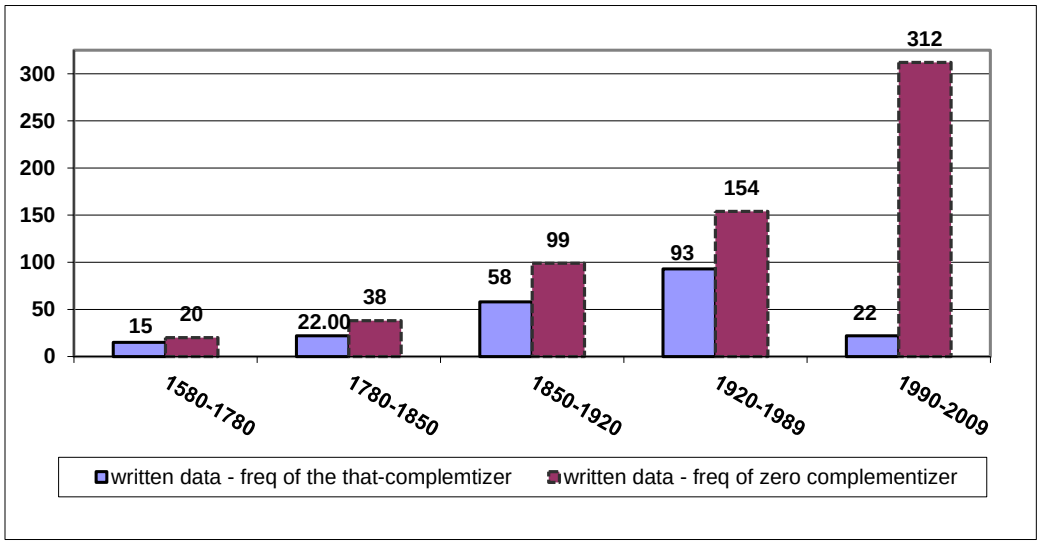

Figure 4. Written data – *that* versus zero distribution per million words

Finally, this process was conducted one last time for the verb *understand*. The extraction yielded (n= 16157) *understand* tokens from spoken English corpora and (n= 6845) tokens from written English corpora. The full set (n= 23,002) of tokens were analyzed and divided into those containing either a *that*-clause or a zero-complementizer. The distributions of these tokens for both the spoken and written data sets are presented in Table 10 and Table 11.

Table 10: Distribution of *that*-clauses and zero-complementizer clauses from EModE to PDE in spoken English. (n: absolute frequency, N: normalized frequency per million)

understand **–** spoken corpora		
	understand **- that**	***understand*** **- zero**

Period	n	N	n	N
1560-1710	(n=15)	12.41	(n=11)	6.31
1710-1780	(n=106)	8.42	(n=200)	15.89
1780-1850	(n=143)	6.48	(n=303)	13.72
1850-1913	(n=613)	33.72	(n=490)	26.96
1960-1993	(n=94)	19.00	(n=68)	10.68
1994-2012	(n=432)	34.83	(n=163)	15.89
Total	**(n=1403)**		**(n=1235)**	

Table 11: Distribution of *that*-clauses and zero-complementizer clauses from EModE to PDE in written corpora. (n: absolute frequency, N: normalized frequency per million)

***understand* –** written corpora				
	***understand* - that**		***understand* - zero**	
Period	n	N	n	N
1580-1710	(n=147)	74.62	(n=61)	24.77
1710-1780	(n=108)	27.45	(n=38)	9.66
1780-1850	(n=143)	24.69	(n=39)	6.73
1850-1920	(n=252)	40.19	(n=31)	4.94
1920-1989	(n=48)	8.76	(n=11)	2.41

1990-2009	(n=63)	25.17	(n=25)	10.33
Total	**(n=761)**		**(n=205)**	

In contrast to the relatively consistent diachronic ratio of the complementizer to zero form patterns observed with the *think* and *guess* data sets the *understand* data presents an unexpectedly different diachronic picture. The trends for that/zero ratio with understand, in both the spoken and written understand data sets, and are presented below in Figures 5 and 6.

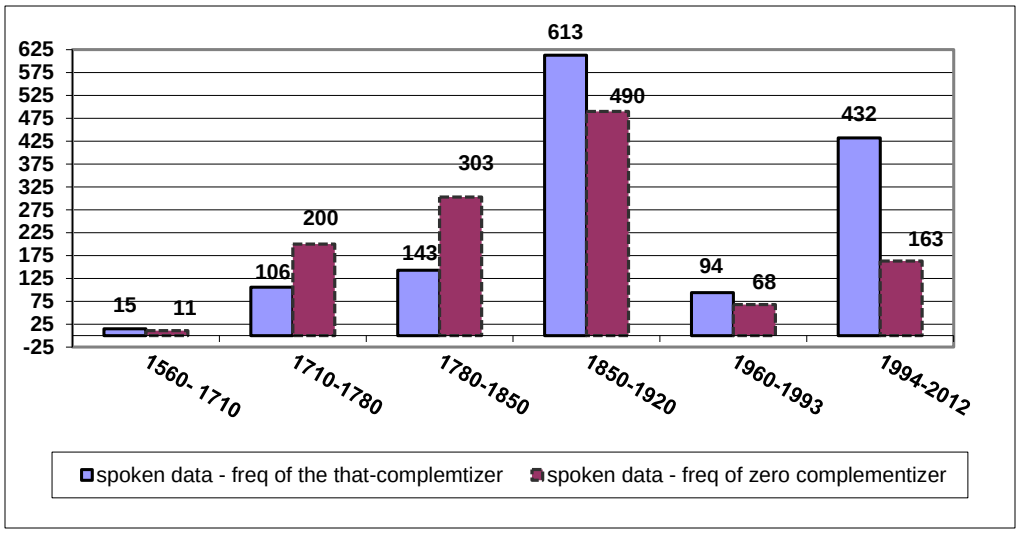

Figure 5. Spoken data *understand – that* versus zero distribution per million words

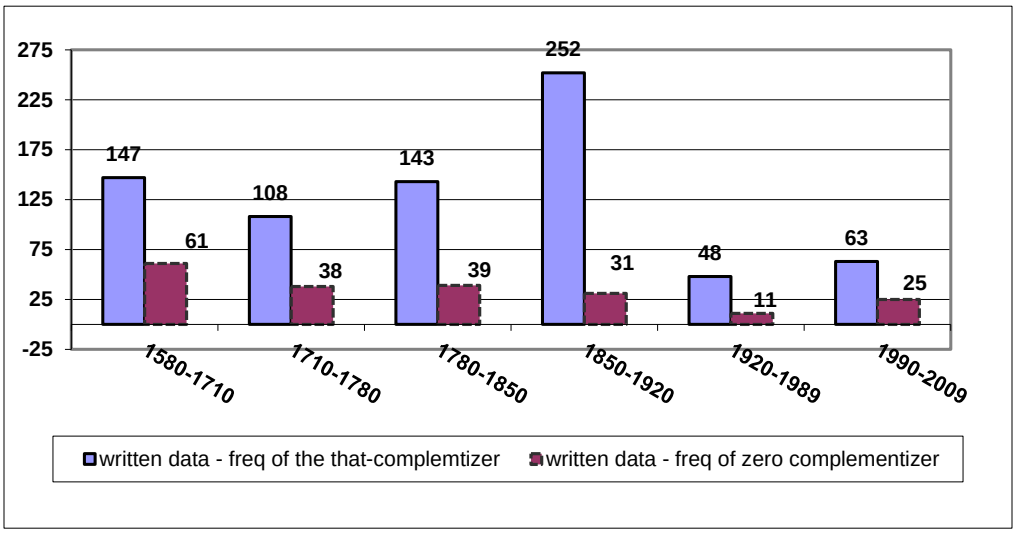

Figure 6. Written data *understand – that* versus zero distribution per million words

The results, presented above for the (n =2638) spoken and (n=966) written *understand* tokens, show that unlike the first two verbs, *understand* is almost always being used more frequently, regardless of the time period, with the *that*-complementizer form. This pattern is reversed between 1710 and 1850 in the spoken data set, and it is reversed again from 1850 to 2012 but this may be simply an idiosyncratic feature of this type of spoken data. The preponderance of *that* is never observed in the parallel written data set; the *that* form remains consistently more frequent for the 400 plus years covered in our corpus resources. This was an unexpected finding and what consequences it has with regard to the factors predicting the presence of the zero form (if any) remains to be seen. Finally, this finding will be integrated into our regression analysis modelling and thus accounted for in Section 4.0.

The (n= 2,055) spoken and (n= 2,217) written *think* tokens, (n= 1050) spoken and (n= 833) written *guess* tokens, and (n= 2638) spoken and (n= 966) written *understand* tokens which contained either a *that* or zero complementizer clause were coded for 26 features, in separate spread sheets, within four categories: corpus information, matrix clause features, complement clause features and the relationship between matrix and complement. The goal of this coding process was to allow for the identification and subsequent statistical analysis, via a regression analysis, upon the following eleven features which in the literature are said to favour the presence of the zero complementizer. A summary of these structural features are presented below in Table 12.

Table 12: Factors which favour the presence zero-complementizer selected for this study (cf. also Tables 1, 2&3)

1) Matrix clause subjects either 'I' or 'You'
2) The absence of extra elements in the matrix clause (viz. auxiliaries, indirect objects, adverbials)
3) The absence of intervening elements between the matrix and complement clause
4) Pronominal subject of the complement clause, co-referential with the matrix clause subject
5) A pronominal subject (versus a NP) in the matrix clause
6) A pronominal subject (versus a NP) in the complement clause.
7) Complement clause subject either 'I' or 'you'
8) The length of the matrix clause subject (pronoun >np-short >np-long)
9) The length of the complement clause subject (pronoun >np-short >np-long)

10) Cotemporality between the matrix and complement clauses

11) Harmony of polarity between the matrix and complement clauses

The corpus information features included information such as the time period of the corpus (e.g. 1710-1780), the inflected form of the token and the full context in which it appeared. The matrix and complement clauses of each extracted token were also coded for the features person, tense, polarity, length of the subject (pronoun / short NP (i.e. 1 to 2 words) / long NP (i.e. 3+ words), and subject coreferentiality. In addition, the presence of additional elements within the matrix clause (elements between the subject and the matrix verb) was also noted along with intervening elements (between the matrix clause and the complement clause) and the location of the intervening elements (either pre-complementizer or post-complementizer and before the complement clause subject).

In addition to the aforementioned categorical coding processes, the data sets for all three verbs were also chronologically reorganized in order to create sufficiently large sample sizes close to or greater than (n= 30) examples per period. This data aggregation procedure was especially important in the early periods (e.g. 1560-1579, 1580-1639 and 1640-1710), where due to the paucity of available data, using every available token and subsequent *that*/zero example still resulted in datasets that fell below the methodologically desirable threshold of (n>30) per period. In such cases data from several periods was combined. For example, with the verb *guess*, this process resulted in an initial period spanning from 1640-1780 in the spoken data set and in the written data 1580 to 1780 and with the verb *understand* it created an initial period spanning 1560 to 1710 in the spoken data and 1580-1701 in the written data sets. The verb *think* was however frequent enough

per period for this step not to be needed. Once the aggregation process was completed, these data sets, per period, were then sufficiently large to function as reference levels for our subsequent diachronic logistic regression analysis. This process was also employed for the PDE spoken data categories from 1980 to 2012 for all three verbs in order to set up a single 20th century period in which to directly compare and contrast with the written data sets which spanned from 1920-2009.

Once these respective processes were completed, the data was loaded into the statistical program *R*, in order to test the effects of the factors represented in Table 12. This was done by means of a *stepwise logistic regression analysis* (using the function stepAIC in the *R* package MASS)[14]. The stepwise selection procedure was both-ways and the minimal model was of course an intercept-only model. The maximal model contained all main effects plus two-way interactions of the factors with period, verb and mode (together with the two-way interactions between period, verb and mode themselves). This necessitated some a priori filtering of the factors. The factor 'I.or.U', for instance, was recoded into two separate factors 'Person' and 'Number', rendering 'I.or.U' itself entirely redundant. Redundancy also applies to the factors 'Mat.Pro.vs. and 'CC.Pro.vs.NP', as the respective factors 'Mat.length' and 'CC.length' contain all the subdivisions of 'it', pronoun, np-short and np-long, and thus capture the important distinctions. The solution was to exclude the redundant factors from the analysis.

The resulting model after stepwise selection contains 11 main effects and 15 interactions (see Table 13), which fits reasonably well: the goodness-of-fit is significant (LLR=5355.511; df=57; p-value=0), the predicted variation (C-score) is 89.3%, but the

[14] The general outline of this methodology was suggested to us by Stefan Th. Gries, for which we wish to express our gratitude.

explained variation (Nagelkerke-R^2) is only 54.2%. This shows that our model still has potential for improvement.

The model diagnostics show a sound model: only 3.5% of the standardized residuals are outside of the range between -2 and 2, and none of the dfbeta's (i.e. the influence of each observation on the coefficients of the effects) fall outside of -1 and 1. In addition, we implemented the procedure in Agresti (2013: 221-224) to dichotomize the fitted probabilities for the *that* zero alternation for comparison with the observed probabilities. This yields a classification accuracy of 84.6%. The significance of this result was finally tested against two baseline models: one that would always predict the most frequent form, and one that would guess an outcome randomly. In both cases, the classification accuracy was highly significant (close to 0). In sum, these diagnostics show that our model is appropriate.

Table 13 gives the ANOVA-table with type III LLR tests. It can be seen that the three strongest predictors are (in decreasing order) the interaction between verb and period, the main effect of length of the complement clause subject ('CC.length'), and the main effect of matrix internal elements ('mat.int'). Only the main effect of cotemporality ('CC.T.co.ref') is not significant, but its interaction with period is border-significant. The interpretation in the next section will discuss all effects. This will be done by means of effect plots (obtained with the *R* package effects).

Table 13: ANOVA-table with type III LLR tests

	Df	**Deviance**	**AIC**	**LRT**	**Pr(Chi)**
<none>		8143.4	8259.4		

Verb	2	8149.7	8261.7	6.284	0.0431921
mat.int	1	8220.1	8334.1	76.701	< 2.2e-16
Person	2	8162.0	8274.0	18.534	9.45e-02
Interv	1	8181.2	8295.2	37.736	8.10e-07
CC.length	3	8228.6	8338.6	85.187	< 2.2e-16
TYPE	1	8184.1	8298.1	40.698	1.78e-07
Tense	3	8158.4	8268.4	14.951	0.0018590
Number	1	8152.3	8266.3	8.857	0.0029196
Mat.length	2	8154.5	8266.5	11.097	0.0038937
Period	1	8167.5	8281.5	24.063	9.33e-04
CC.T.co.ref	1	8144.4	8258.4	0.930	0.3347531
Verb:Person	4	8153.4	8261.4	9.989	0.0406096
Verb:tense	6	8195.2	8299.2	51.779	2.07e-06
interv:TYPE	1	8168.2	8282.2	24.806	6.34e-04
TYPE:tense	3	8151.7	8261.7	8.232	0.0414572
Verb:period	2	8238.6	8350.6	95.191	< 2.2e-16
TYPE:period	1	8185.4	8299.4	41.989	9.18e-08
CC.length:period	3	8169.1	8279.1	25.675	1.12e-02
tense:period	3	8164.3	8274.3	20.879	0.0001115
Verb:TYPE	2	8153.6	8265.6	10.203	0.0060890
interv:period	1	8148.6	8262.6	5.211	0.0224387
Verb:Number	2	8149.5	8261.5	6.031	0.0490240
Person:TYPE	2	8149.1	8261.1	5.713	0.0574741

Verb:mat.int	2	8148.6	8260.6	5.221	0.0734856
Verb:CC.length	6	8157.0	8261.0	13.535	0.0352890
period:CC.T.co.ref	1	8146.5	8260.5	3.032	0.0816159

4.0 Results

In this section, we present the results from the stepwise regression analysis on eleven factors that have been argued in the literature to predict the presence of the zero complementizer form with verbs of cognition such as think, guess and understand (see Section 2.3). Because of the complex structure of our model (with sixteen interactions), this will be done by means of graphical visualization in effect plots that were obtained with the R package effects. The main factors under consideration are the main effects of verb, period, and mode (i.e. spoken versus written), the absence of matrix-internal elements, the absence of intervening elements between the matrix and complement clause, the length of the complement clause subject, matrix clause person, matrix clause number, matrix clause tense, coreferentiality of person between the matrix and complement clause subjects, and harmony of polarity between the matrix and complement clauses.

In 4.1, we discuss the five statistically significant interactions with verb, v viz. interactions with matrix internal elements, length of the complement clause subject, person, number and tense. In 4.2, we show that the following interactions with mode are statistically significant: the absence of intervening elements between the matrix and complement clauses, person and tense. The final set of interactions, presented in 4.3, offers a diachronic account of conditioning factors for zero use. The analysis shows that

there are significant changes across time in the extent to which verb, length of the complement clause subject, person, and harmony of polarity predict the use of zero.

4.1 Verb type

A‘panchronic’ model aggregating all time periods was used to examine the interaction of the factor ‘verb’ with other factors as predictors of the zero complementizer form. This allows us to gauge to what extent the main effects observed above are verb-specific. The significant factors are presented below in figures 7-11.

In Figure 7 we see that the absence of intervening elements in the matrix clause is a strong predictor for the zero form for all three verbs; for each individual verb, the value for zero complementation is significantly higher in the left panel (absence of intervening elements) than in the right panel (intervening material is present). However, due to scarcity of data (there were only 219 occurrences of *guess* with intervening material in comparison to more than 1000 for the other 5 effects) we get a larger confidence interval for *guess* with intervening material. Therefore, some caution is warranted when interpreting this data point. With that proviso, we can say that while the difference in zero use with *think* and *guess* without intervening material is minimal, *guess* has a considerably lower rate of zero than *think* when intervening material is present. Comparing all three verbs, intervening material has the strongest effect on *guess;* the difference in zero use between presence and absence is the greatest for this verb. We also observe that for *understand*; zero rates are much lower overall; there is only a 50% chance of the zero form being used when there is no intervening material compared to values of over 90 and 80 per cent respectively for

think and *guess.* When *understand* occurs with intervening material, its zero rate is lower than 30%.

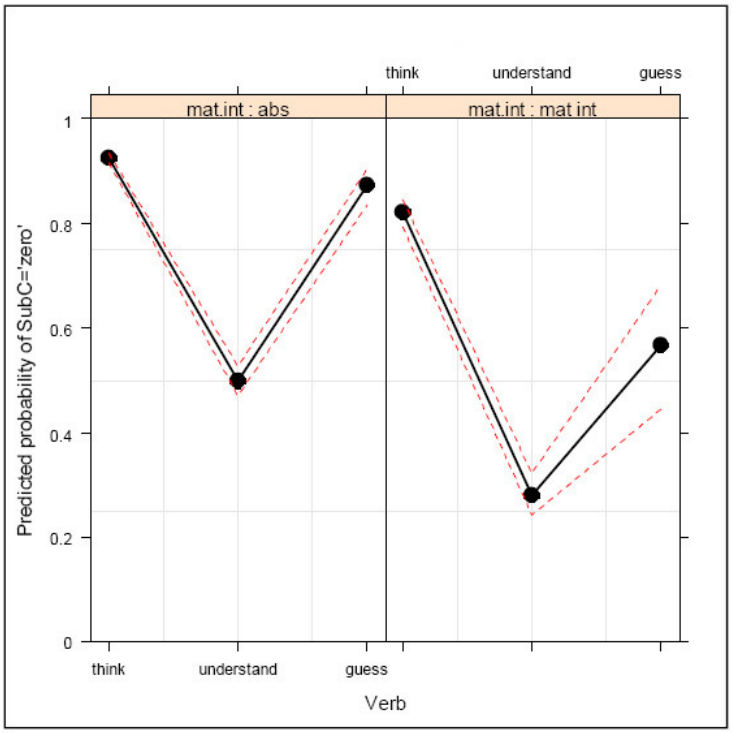

Figure 7. Verb: Matrix-internal elements

In Figure 7 we see that the absence of matrix internal elements is significant for all three verbs, with the proviso that there were so few occurrences of *guess* with matrix-internal elements that the confidence interval for this data point is so large that we cannot make any reliable claims about the effect of the present factor on this verb. Comparing the zero rates of the three verbs in the left panel, i.e. when there are no matrix-internal elements, we observe a strong conditioning effect for *think* and *guess*, but a very weak effect for *understand;* when a matrix clause with *understand* contains additional elements, the chances of getting zero or *that* are split 50-50. The results also reveal that the presence of matrix internal elements is predictive for the *that*-complementizer form for *understand*.

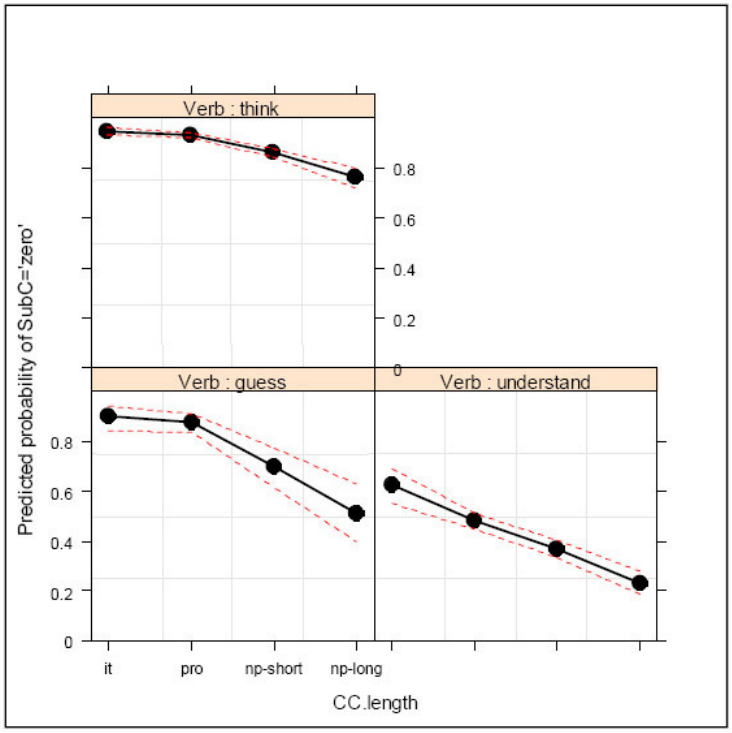

Figure 8. Verb: Complement Clause Length

The analysis of the effect of the length of the complement clause subject reveals additional differences between these three verbs. The results presented in Figure 8 show that *it* and other pronouns as complement clause subjects have largely the same predictive effects for both *think* and *guess*; however, as the weight of the complement clause subject increases *guess* has a lower likelihood of using the zero form relative to *think*. In addition, the analysis reveals that the length of the complement clause subject has a much lower predictive effect overall for the verb *understand*; an np-long complement clause subject with *think* is still more predictive of the zero-form than *it* is for the verb *understand*. In fact, *it* is the only *understand* data point with a +50% value. All other subject types predict the *that* form with *understand.*

This variation across verbs is also seen when we look at the remaining categories of person, number and tense. In Figure 9 below we see the results for the effect of person across all three verbs.

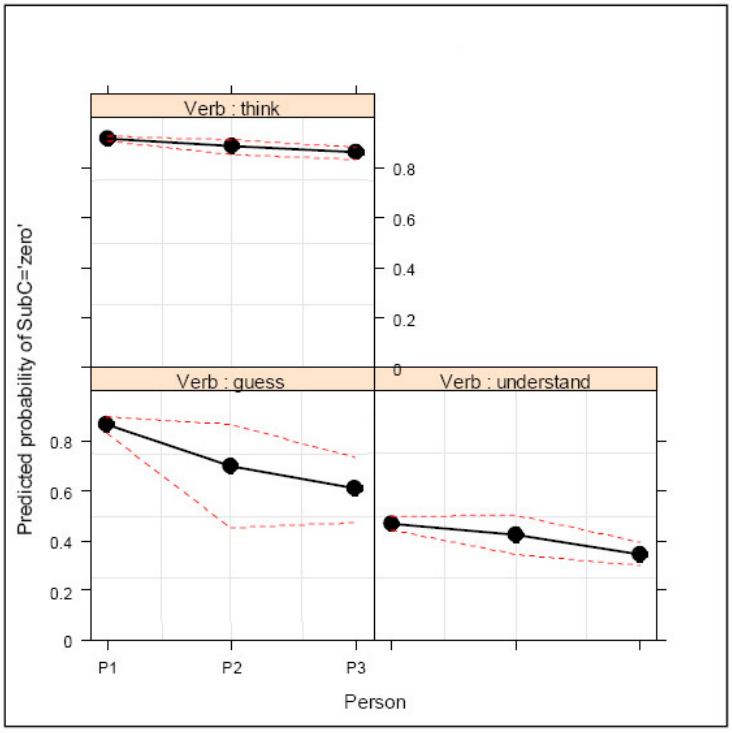

Figure 9. Verb: Person

The results presented Figure 9 indicate that the verbs *think* and *understand* parallel the previous findings regarding person in that a) the overall predictive effect is much stronger for *think* relative to *understand* and b) that P1>P2>P3 in terms for both verbs in terms of serving as a predictor for the presence of the zero form. Furthermore, with the verb *guess* only the result for 1st person is reliable while the 2nd and 3rd person forms are shown to be unreliable predictors due to large confidence intervals.

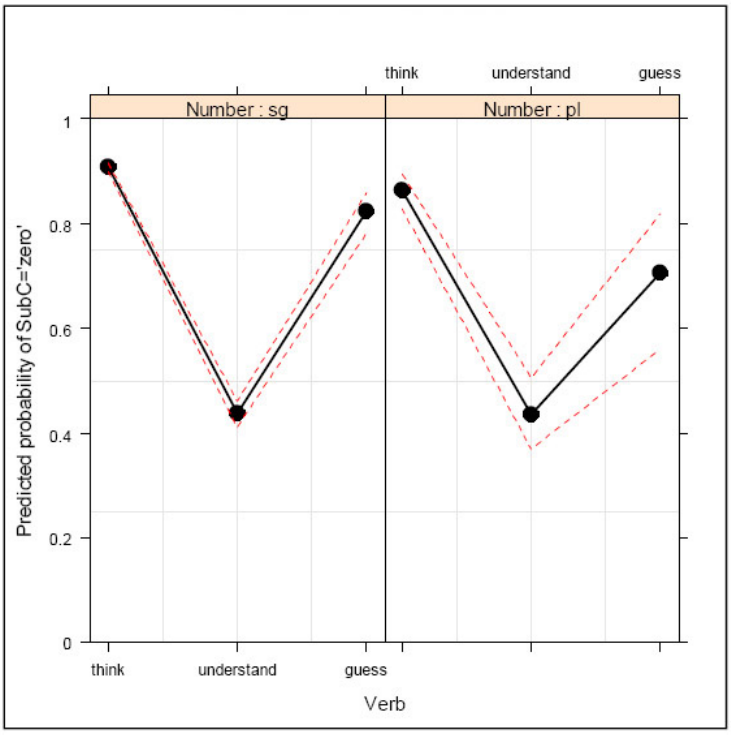

Figure 10. Verb: Number

The analysis of number once again confirms the outsider status of *understand*. The predictive power of singular matrix clause subjects is stronger for *think* and *guess* than it is for *understand*. Since the zero form occurs less often with *understand* than with *think* or *guess* overall (see Figures 5 & 6), zero rates for both singular and plural *understand* are lower than those of the other two verbs. The plot in Figure 10 also indicates that the differences in complementizer use between singular and plural subjects are minimal for all three verbs; the locations of the data points in the two panels do not shift much. In fact, only the difference between singular and plural *think* is significant. In addition, we also see that much like our initial analysis of the main effects, the singular form more strongly predicts the zero form with *think* and to a lesser degree with *guess*, while *understand* is more likely to be used with *that* regardless of number.

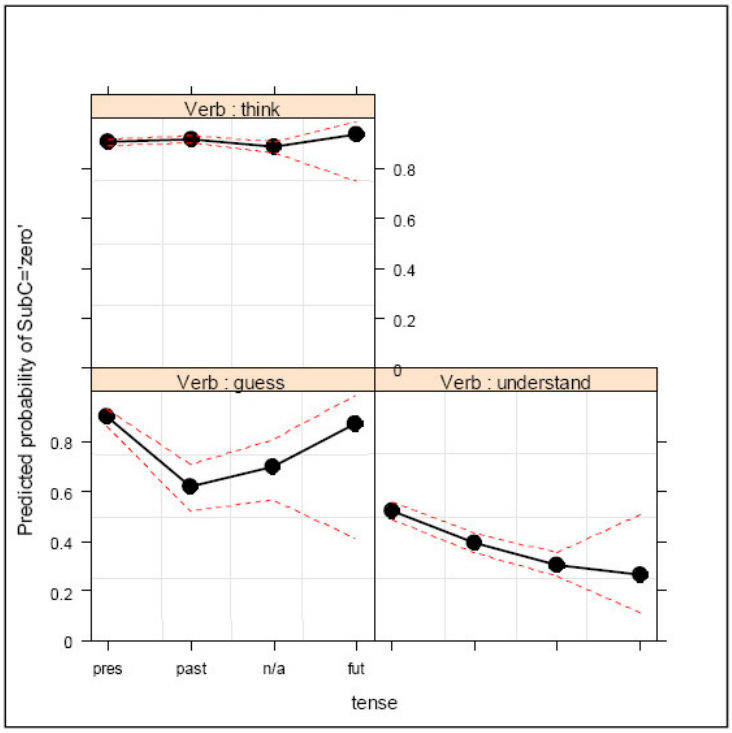

Figure 11. Verb: Tense

The final factor to be discussed in this section is the effect of tense on each of the verbs. Although the main effect for tense was not significant, the interaction of this factor with verb type is. The analysis of tense indicates that the future tense is an unreliable factor across all three verbs and that there are no significant differences in zero rate between past, present and n/a tense forms with *think*. Furthermore, the results for all tense values of *guess* except present are unreliable due to large confidence intervals. Finally, Figure 11 shows that all tense forms of *understand* are predictive of the *that* complementizer.

4.2 Mode (Spoken versus Written data)

Our analysis revealed that the effect of interaction between the factor mode and a number of other factors is significant. We will now compare the extent to which these factors predict the use of the zero form in the spoken and written modes. Once again a stepwise regression procedure was used to examine the effect of the factors presented in Table 15 relative to the mode (i.e. spoken versus written language). This model is also panchronic, i.e. all periods are conflated. Recall that although there was a significant difference between the spoken and written modes in terms of the probability with which zero is used, the main effect of mode on complementizer use was not that strong. In this section, we will see that mode plays a more important role in the zero/*that* alternation than one would expect on the basis of the main effects analysis; the strength of various other factors depends heavily on mode, i.e. some factors may be better predictors of the zero form in one mode as opposed to the other.

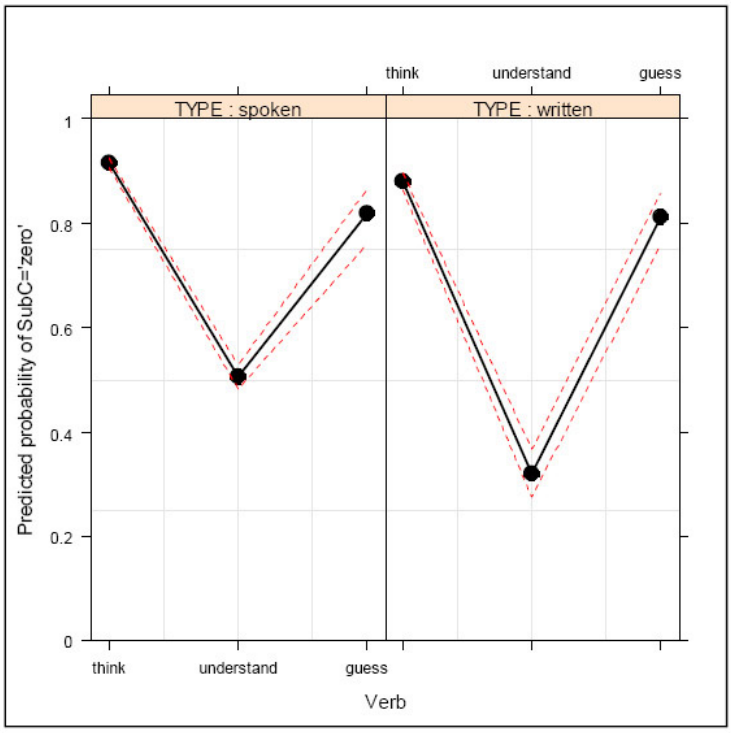

Figure 12. Mode: Verb

In Figure 12 we see that there is little difference with the verbs *think* and *guess* with respect to mode as a predictor for the zero form. The situation is different for *understand*, however. While in the spoken mode, *understand* has a 50% chance of being used with the zero form, there is in fact a greater likelihood of *that* when *understand* is used in the written mode.

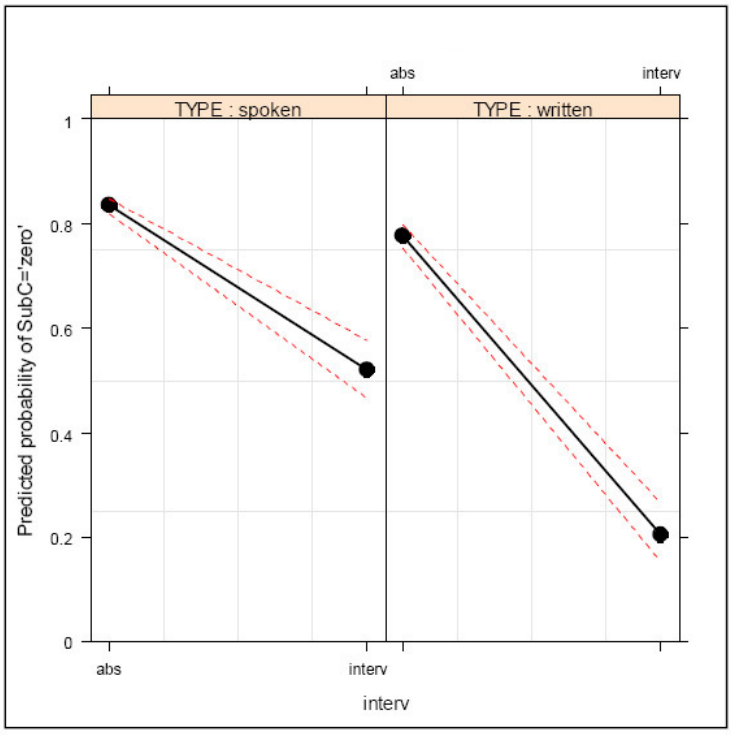

Figure 13. Mode: Absence of Intervening Elements

Figure 13 allows us to compare the conditioning effect of intervening elements between matrix clause and complement clause in the spoken and written modes. Recall that absence of intervening elements was a very good predictor overall. The interaction confirms this earlier finding; in both panels we observe a dramatic difference in complementizer use between presence and absence of intervening material. A notable difference, however, resides in the conditioning effect of the absence of intervening material in the written mode. When there is intervening material in the written mode, we are much less likely to get the zero form than in the spoken mode, so much so that the explicit complementizer *that* in fact becomes more likely; the zero rate drops to 0.2. It may be that writers are more led by

the complexity principle than speakers and feel the need to insert *that* to make clause boundaries clearer when intervening material risks impairing clarity.

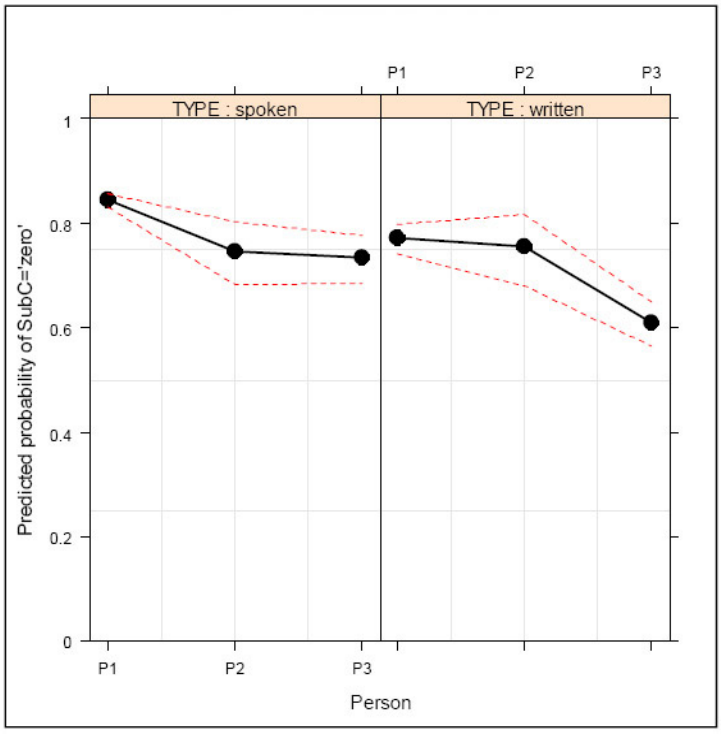

Figure 14. Mode: Person

In Figure 14, we examine the effect of person in the two modes. The plot reveals that in both the spoken and written modes the 1st person subject predicts more zero use; however, in both cases the 2nd person subjects are not significant. We also see that the 1st person subject form is a stronger predictor in the spoken versus the written data and that compared to the spoken mode, 3rd person subjects in written data are less likely to be used with zero complementation.

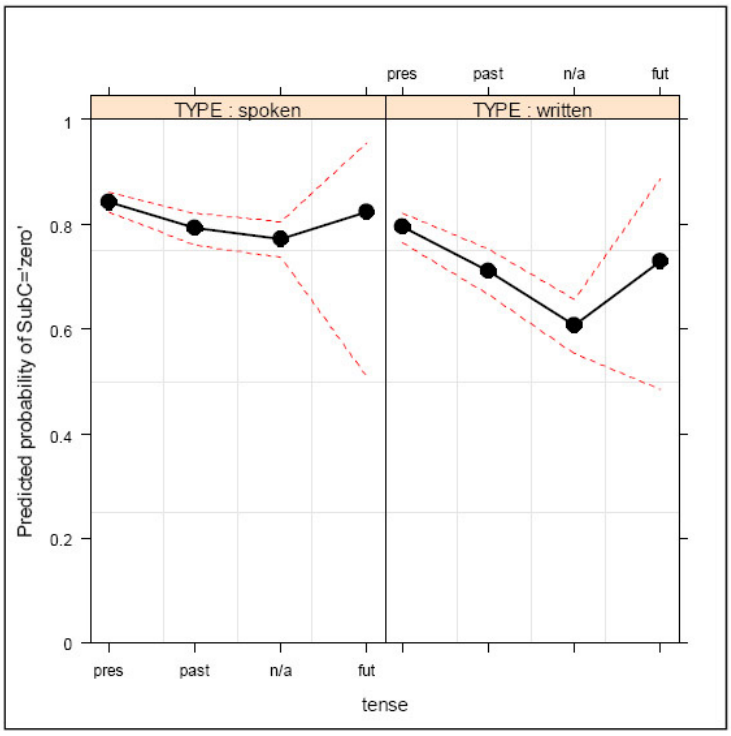

Figure 15. Mode: Tense

The final factor that we will examine in this section is the effect of tense as a predictor of the zero form relative to mode. The analysis of tense, presented above, again follows the pattern established in the preceding discussions of the main effect of tense and its interaction with verb type; in both the spoken and written data, the future form, again due to the sparseness of data, results in large confidence intervals and therefore, we cannot make any claims about the effect of the future on zero use in spoken versus written data.

In addition, Figure 15 reveals that the past, present and n/a forms are not significantly different from one another in the spoken data but they are in the written data. We can thus conceive the following predictive cline for the zero form: n/a>past>pres.

4.2 Period

We now will turn to the final stage of our analysis and look at the effect of the structural factors across the eight time periods. Thus, in the following sections, we adopt a diachronic approach, discussing the interactions with period that came out as significant. The interaction effects with period were significant with the following factors: verb, mode, absence of intervening elements, complement clause length, cotemporality between the matrix and complement clause and tense. This final step in the analysis offers a diachronic perspective; it shows whether the import of a given factor becomes stronger or weaker over time.

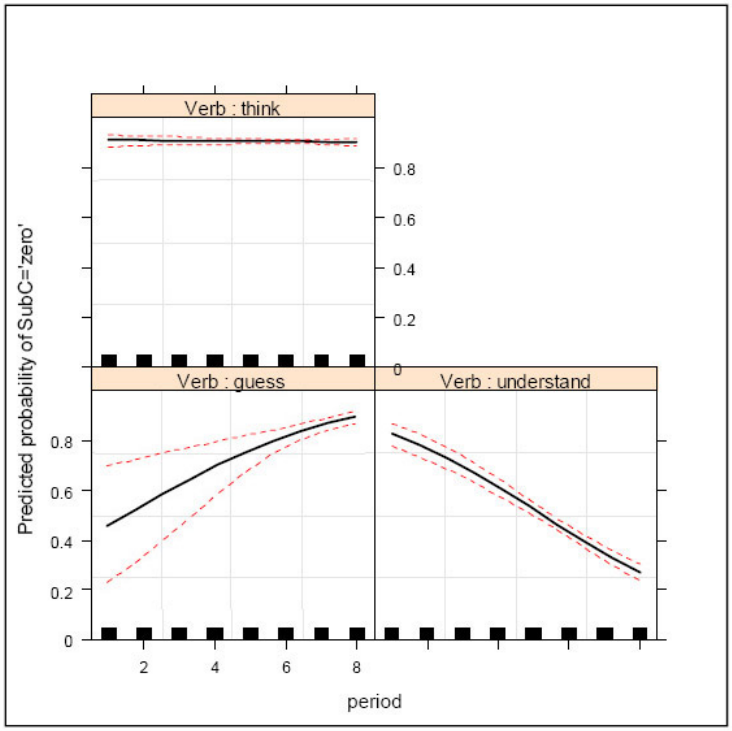

Figure 16. Period: Verb

Figure 16 shows the diachronic development of the zero form for each of the three verbs and it reveals a great deal of variation between them with respect to *that*/zero alternation. The verb *think* starts out with a high rate of zero relative to the *that* form and exhibits a gradual loss of the zero form (relative to *that*) over time. *Guess* on the other hand shows a strong and constant increase in the ratio of the zero from over time, starting out below 0.5 and culminating in value comparable to that seen with *think* in PDE. *Understand,* by contrast, is characterized by a dramatic drop in zero use. It drops below 0.5 in period 6 and in the most recent time period it barely reaches 0.3. Thus, while *understand* used to have a strong preference for the zero form, in more recent years it has come to prefer the explicit complementizer *that.* This shows that there is no homogeneous zero/*that* alternation trend and that interactions with verb type are highly relevant. Also, it opens up perspectives for future research on the basis of a larger number of verb types.

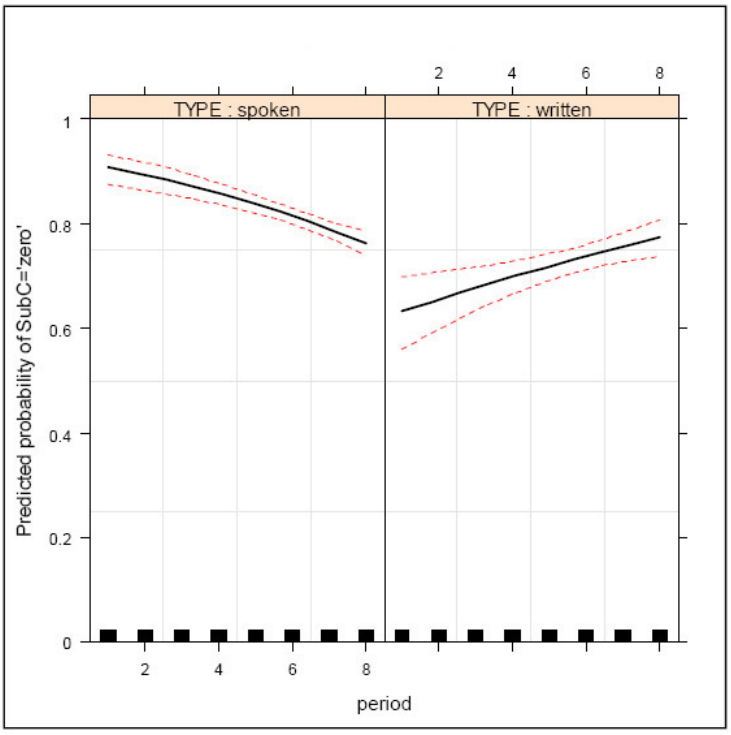

Figure 17. Period: Mode

An analysis of the effect of mode over time shows that in the earliest periods the zero form was far more prevalent in the spoken data relative to the written data but over time, as the zero form has gone down in the spoken mode and increased in the written mode, in PDE the two modes are at the same predictive level. As Figure 17 shows, the endpoints in PDE for both modes are almost identical which suggests that nowadays mode, in and of itself, is no longer a good or a significant predictor of the zero form with these verbs anymore.

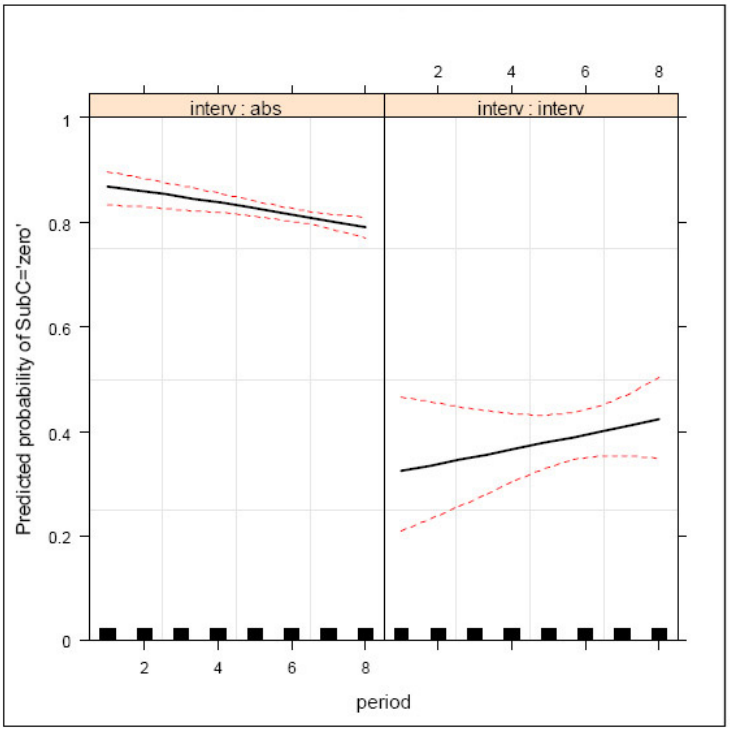

Figure 18. Period: Absence of Intervening Elements

An analysis of the diachronic effect of the absence of intervening elements between the matrix and complement clauses produces a result which confirms what has been argued in the literature on *that*/zero variation, namely that the absence of intervening elements is a strong predictor of the zero form. The results show that this trend is decreasing over time; however, it still remains quite robust relative to the presence of intervening elements. The values in the right panel suggest that intervening elements predict the explicit *that*-complementizer throughout all periods, although the effect gets weaker, but these findings cannot be ascertained due to large confidence intervals.

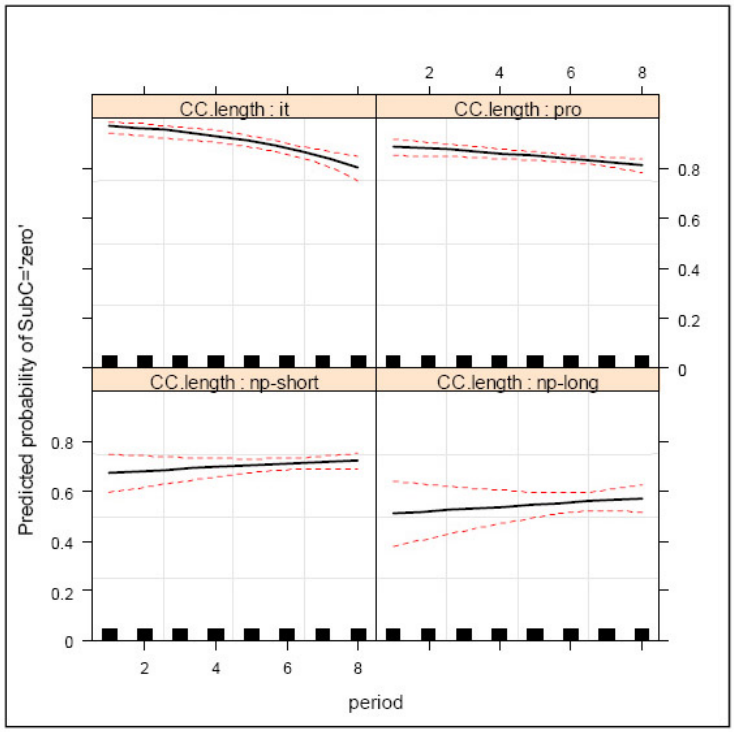

Figure 19. Period: Complement Clause Length

In Figure 19 the analysis of the effect of the length of the complement clause subject over time shows a clear division between *it* and other pronouns versus NPs in that the former two have been and still remain the stronger predictors of the zero form while the latter (i.e. NPs) are actually increasing in their own respective predictive abilities of the zero form but they have yet to reach the level of *it* or other pronouns. Furthermore, an examination of the start and endpoint for *it* and other pronouns shows that they are higher compared to NPs at any stage of their development and that '*it* and other pronouns remain the stronger predictive factors in PDE with the current set of verbs.

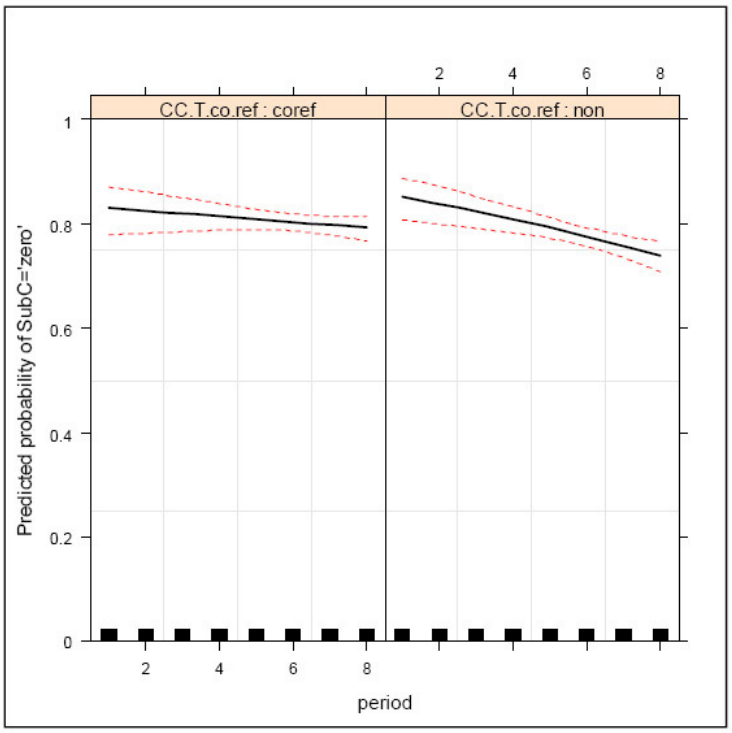

Figure 20. Period: Matrix and Complement Clause Cotemporality

A diachronic analysis of the effect of cotemporality between the matrix and complement clause reveals yet another interesting pattern in that a) the difference between cotemporal and non-cotemporal tense forms is only marginally significant b) both have become less associated with the zero form over time. Figure 20 shows that the predictive power of the non-cotemporal patterns decreases faster than that of the cotemporal patterns. The net result is that the effect of this interaction is significant and that at least in PDE the non-cotemporal pattern, as indicated in the literature, is now a slightly better predictor of the *that* form vis-à-vis its cotemporal counterpart.

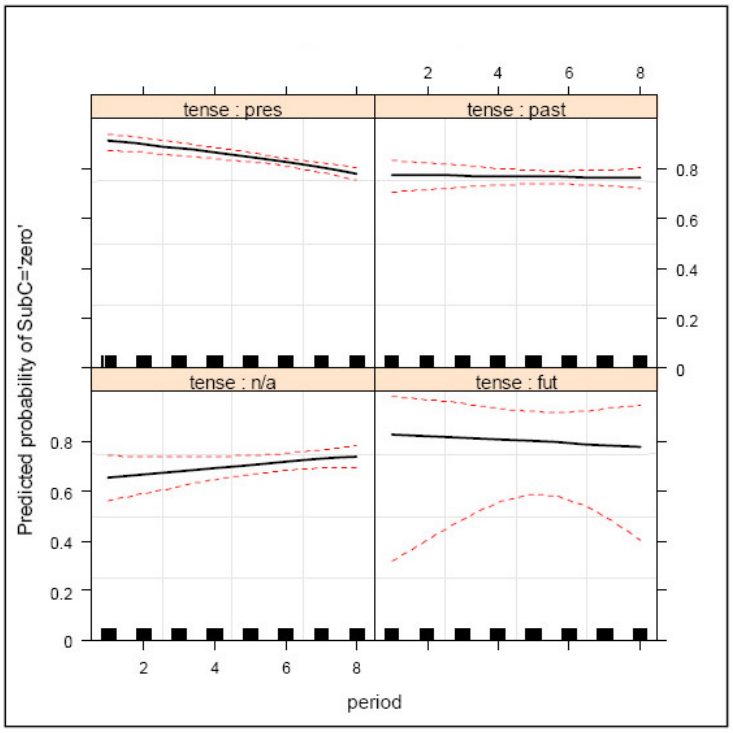

Figure 21. Period: Tense

The final significant effect over time to be discussed in this section is the interaction between tense and period. Figure 21 shows that while present tense is gradually decreasing in predictive effect over time for the zero form its endpoint is largely equal to that of the past tense form regardless of time. This suggests that n there is very little predictive difference between the present and past tense forms. Finally, the future form is also shown to be diachronically problematic and the effects are too uncertain to be of any value for the current discussion.

5.0 Conclusion

This study has shown that, contrary to claims and speculation in the literature to the effect that there has been an overall diachronic tendency towards more zero complementizer use

at the expense of *that*-complementation, the aggregate values for *think*, *guess* and *understand* show a steady *decrease* in zero complementation. In fact, two of the three most frequent complement-taking mental verbs in present-day English, viz. *think* and *understand*, exhibit a diachronic decrease in zero use and a concomitant *increase* in *that* use. *Guess* is the only verb exhibiting a diachronic increase in zero use.

The rigorous methodological approach developed and utilized in this study, and the attention given to ensuring sufficiently large and representative sample sizes when possible from each period has also highlighted the fundamental problems seen in previous work on this topic which have often relied heavily upon descriptive statistical processes. As evidenced by our initial presentation of findings in Section 3.0, reliance of descriptive statistics (often presented in the literature in conjunction with Chi-square analysis) can unintentionally obscure important multicollinear interactions between factors or variable and/or not reveal the stability or robustness of diachronic trend-lines or patterns. From a descriptive perspective it would appear that the zero form for *think* is robust or at least remaining consistent over time and thus one could reasonable infer that the factors which have been proposed to facilitate the zero-form are either equally predictive or also remain significant over time. It is only when a methodology such as the one used in this study is applied that the true significance of the various factors becomes apparent along with diachronic robustness of predicted or expected trends and/or patterns vis-à-vis a dependent variable such as the presence of the zero-complementizer.

In addition to invalidating the long-standing assumption that complement-taking verbs have diachronically developed towards higher levels of zero complementation, this study also highlights the need to differentiate between individual verbs when examining

complementation patterns. It became apparent; firstly, that only one verb examined in this study, viz. *guess*, exhibits the aforementioned diachronic increase in zero use. There is a very slight decrease in zero use with *think* over time and *understand*, though starting out with a preference for the zero form, gradually shifts to being a *that*-favouring verb..

Second, the extent to which the factors mentioned in the literature actually predict zero use may differ from verb to verb, as the interactions with verb type suggest. A striking finding in this regard is the effect of matrix internal elements. A strong predictor overall, lack of matrix internal elements is an especially good conditioning factor with *understand* and *guess*; *understand* actually favours the *that* form when matrix-internal elements are present and *guess* exhibits the largest difference in zero rate as conditioned by this factor..

This study has shown that the effect of conditioning factors is also dependent on mode. Again, intervening material was a case in point. Its predictive power is much stronger in the written mode than in the spoken mode; when intervening material is present in the written mode *that* is favoured. Also, mode is a much more powerful predictor for *understand* than for the other two verbs; in the written mode, the zero rate with *understand* drops to below 30% as compared to over 50% in the spoken mode.

With regard to perspectives for future research, the results of the current study call for a methodologically similar analysis with a larger set of verb types as this may reveal additional differences in the way zero/*that* alternation has evolved with each individual verb and as well as shedding more light on how the effect of a conditioning factor may differ from verb to verb.

An additional avenue for future research consists in looking beyond familiar local conditioning factors that are of a strictly structural nature. Priming effects, as in Jaeger and

Snider's (2008) study of the syntactic persistence of complementation patterns, and prosodic information (cf. Dehé &Wichmann 2010) could be incorporated into the logistic regression model. One drawback to the study of prosody and its effect on zero/*that* use from a diachronic point of view is the absence of audio recordings of older corpus data. This shortcoming could be remedied by reconstructing the natural rhythmic patterns of the data on the basis of current knowledge about prosody.

References

Agresti, Alan (2013). *Categorical Data Analysis*. Hoboken: Wiley.

Aijmer, Karin (1997). *I think* – an English modal particle. In Swan, T. & Westwik, O. J. (eds.), *Modality in Germanic languages: Historical and comparative perspectives.* Berlin: Mouton de Gruyter. 1-47.

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

Boye, Kasper & Peter Harder (2007). Complement-taking predicates: Usage and linguistic structure. *Studies in Language* 31(3): 569–606.

Brinton, Laurel J. (1996). *Pragmatic markers in English: Grammaticalization and discourse functions.* Berlin: Mouton de Gruyter.

Brinton, Laurel J. (2008). *The comment clause in English: Syntactic origins and pragmatic development.* Cambridge: Cambridge University Press.

Bybee, Joan L. (2003). Mechanisms of change in grammaticalization: The role of frequency. In Joseph, B. D. & Janda, R. D. (eds.), *The handbook of historical linguistics*. Oxford: Blackwell. 602-623.

Bybee, Joan L. (2006). From usage to grammar: The mind's response to repetition. *Language* 82(4): 711-734.

Dehé, Nicole & Anne Wichmann (2010). Sentence-initial *I think (that)* and I *believe (that):* Prosodic evidence for uses as main clause, comment clause and discourse marker. *Studies in Language* 34(1): 36-74.

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

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

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

Finegan, Edward & Douglas Biber (1985). *That* and zero complementizers in Late Modern English: Exploring ARCHER from 1650-1990. In Aarts, B. & Meyer, C. F. (eds.), *The verb in contemporary English.* Cambridge: Cambridge University Press. 241-257.

Fischer, Olga (2007). The development of English parentheticals: A case of grammaticalization? In Smit, S. D., Hüttner, J., Kaltenböck, G. & Lutzky, U. (eds.),

Tracing English through time. Explorations in language variation. Vienna: Braumüller. 99-114.

Givón, Talmy (1980). The binding hierarchy and the typology of complements. *Studies in Language* 4(3): 333-377.

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

Gorrell, J.H. (1895). Indirect discourse in Anglo-Saxon. *PMLA* 10: 342-485.

Huddleston, Rodney & Geoffrey K. Pullum (2002). *The Cambridge grammar of the English language.* Cambridge: Cambridge University Press.

Jaeger, Florian T. & Neal Snider (2008). Implicit learning and syntactic persistence: Surprisal and cumulativity. In Love, B. C., McRae, K. & Sloutsky, V. N. (eds.), *Proceedings of the Cognitive Science Society Conference.* Washington, DC. 1061-1066.

Kaltenböck, Gunther (2006). '. . . That is the question': Complementizer omission in extraposed that-clauses. *English Language and Linguistics* 10(2): 371–396.

Kaltenböck, Gunther (2007). Position, prosody and scope: The case of English comment clauses. *Vienna English Working Papers* 16(1): 3-38.

Kearns, Kate (2007a). Epistemic verbs and zero complementizer. *English Language and Linguistics* 11(3): 475-505

Kearns, Kate (2007b). Regional variation in the syntactic distribution of null finite complementizer. *Language Variation and Change* 19: 295–336.

Langacker, Ronald W. (1991). *Foundations of cognitive grammar. Vol II: Descriptive application.* Stanford CA: Stanford University Press.

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

Noonan, Michael (1985). Complementation. In Shopen, T. (ed.), *Language typology and syntactic description. Volume II: Complex constructions.* Cambridge: Cambridge University Press. 42-140.

Palander-Collin, Minna (1999). Grammaticalization and social embedding: I THINK and METHINKS in Middle and Early Modern English. Helsinki: Tome LV.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik ([1985] 1997). *A comprehensive grammar of the English language.* London: Longman.

Rissanen, Matti (1991). On the history of that zero in object clause links in English. In Aijmer, K. & Altenberg, B. (eds.), *English corpus linguistics: Studies in honour of Jan Svartvik.* London: Longman. 272-289.

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

Shank, Christopher, Julie Van Bogaert & Koen Plevoets (under revision). A multifactorial analysis of that/zero alternation: A diachronic study of the grammaticalization of the zero complementizer construction with *think*, *guess* and *understand*. In Jiyoung Yoon & Stefan Th. Gries (eds.), *Construction Grammar beyond English: Current corpus-based approaches.* (Constructional Approaches to Language) Amsterdam: Benjamins.

Storms, G. (1966). *That*-clauses in modern English. *English Studies* 47: 249-270.

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 & Jennifer Smith (2005). *No momentary fancy!* The *zero* 'complementizer' in English dialects. *English Language and Linguistics* 9(2): 289-309.

Thompson, Sandra A. (2002). "Object complements" and conversation: Towards a realistic account. *Studies in Language* 26(1): 125-164.

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

Thompson, Sandra A. & Anthony Mulac (1991b). A quantitative perspective on the grammaticalization of epistemic parentheticals in English. In Traugott, E. C. & Heine, B. (eds.), *Approaches to grammaticalization.* Amsterdam: Benjamins. 313-339.

Torres Cacoullos, Rena & James A. Walker (2009). On the persistence of grammar in discourse formulas: A variationist study of *that*. *Linguistics* 47(1): 1-43.

Underhill, Robert (1988). *The discourse conditions for that-deletion.* Ms., San Diego: San Diego State University.

Van Bogaert, Julie (2010). A constructional taxonomy of *I think* and related expressions: Accounting for the variability of complement-taking mental predicates. *English Language and Linguistics* 14(3): 399-427.

Van Bogaert, Julie (2011). *I think* and other complement-taking mental predicates: A case of and for constructional grammaticalization. *Linguistics* 49(2): 295-332.

Warner, Anthony R. (1982). Complementation in Middle English and the methodology of historical syntax. London: Croom Helm.

Yaguchi, Michiko (2001). The function of the non-deictic *that* in English. *Journal of Pragmatics* 33(7): 1125-1155.