Christopher Shank*, Julie Van Bogaert and Koen Plevoets

The diachronic development of zero complementation: A multifactorial analysis of the *that*/zero alternation with *think*, *suppose*, and *believe*

DOI 10.1515/cllt-2015-0074

Abstract: This corpus-based study examines the diachronic development of the *that*/zero alternation with three verbs of cognition, viz. *think*, *believe*, and *suppose* by means of a stepwise logistic regression analysis. The data comprised a total of ($n = 9{,}720$) *think*, ($n = 4{,}767$) *believe*, and ($n = 4{,}083$) *suppose* tokens from both spoken and written corpora from 1560 to 2012. We test the effect of 11 structural features that have been claimed to predict the presence of the zero complementizer form. Taking our cue from previous research suggesting that there has been a diachronic increase in zero use and applying a rigorous quantitative method to a large set of diachronic data, we examine (i) whether there is indeed a diachronic trend toward more zero use, (ii) whether the conditioning factors proposed in the literature indeed predict the zero form, (iii) to what extent these factors interact, and (iv) whether the predictive power of the conditioning factors becomes stronger or weaker over time. The analysis shows that, contrary to the aforementioned belief that the zero form has been on the increase, there is in fact a steady decrease in zero use. The extent of this decrease is not the same for all verbs. Also, the analysis of interactions with verb type indicates differences between verbs in terms of the predictive power of the conditioning factors. Additional significant interactions emerged, notably with verb, mode (i.e., spoken or written data), and period. The interactions with period show that certain factors that are good predictors of the zero form overall lose predictive power over time.

Keywords: zero complementation, that/zero alternation, multifactorial analysis, logistic regression, verbs of cognition

***Corresponding author: Christopher Shank,** School of Linguistics and English Language, Bangor University, Bangor, Gwynedd LL57 2DG, UK, E-mail: c.shank@bangor.ac.uk
Julie Van Bogaert, Department of Linguistics, Ghent University – FWO Vlaanderen, Blandijnberg 2, B-9000 Ghent, Belgium, E-mail: Julie.VanBogaert@UGent.be
Koen Plevoets, Department of Translation, Interpreting and Communication, Faculty of Arts and Philosophy, Ghent University, Groot-Brittanniëlaan 45, B-9000 Ghent, Belgium, E-mail: koen.plevoets@ugent.be

1 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 [VERB + OBJECT CLAUSE] construction has been evolving toward an increased use of the zero complementizer form (Rissanen 1991; Thompson et al. 1991a, 1991b; Palander-Collin 1999). The present paper seeks to test this hypothesis by means of a stepwise logistic regression analysis of ($n = 9{,}581$) tokens of *think, suppose*, and *believe*, three of the most frequently used complement-taking verbs of cognition, spanning the time period from 1560 to 2012. Previous studies have put forward a number of conditioning factors (structural as well as non-structural) promoting the zero complementizer, or zero form. Our regression model will test whether these structural factors indeed predict the zero form, whether they gain or lose predictive power when combined, and what happens to their predictive power over time. Furthermore, by also testing the effect that time, as a factor, has upon the selection of the zero complementizer, we also show the interaction of time with each of these conditioning factors, thus providing an innovative diachronic perspective to existing research into the *that*/zero alternation.

We start off with a review of the literature dealing with the *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 Background

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

In usage-based approaches to the *that*/zero alternation (Thompson et al. 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 et al. 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 toward 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 et al. 1991b: 326),[2] which occur in clause medial or final position with respect to the (erstwhile) complement clause.

(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*

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 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).

in Middle and Early Modern English (EModE) 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.
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 earlier).[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*, and *believe*) have tended toward 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, 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 vs 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, e.g., 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

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).
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

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 EModE. Rissanen (1991) notes a steady increase between the fourteenth and the seventeenth centuries, 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 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

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.[7] 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

uses of explicit *that*. For Yaguchi (2001), too, this demonstrative meaning continues to condition the contemporary function of *that*.

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.

registers (see Kaltenböck 2006: 373–374 for reference). 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.

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 et al. 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 et al. 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 earlier). 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 et al. 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 (i.e noun phrases) (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 I think a portion of it must have fallen down upon the straw*. (OBC)
b. *Some people think that maybe it was a crazy person that stalked Tara*. (COCA)

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 et al. 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"

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.

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 et al. 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.

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 maybe that we would be talking about it.*
b. *At the beginning, we told the guy that we were gonna both-each have our own.*
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 et al. 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 toward 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.

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

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."

Table 1: Matrix clause factors potentially favoring the zero complementizer.

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)

(Warner 1982; Elsness 1984; Finegan and Biber 1985; Rissanen 1991; Thompson et al. 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 et al. 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 with 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 personally 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
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)

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).

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 Data and methods

Our analysis was based on tokens retrieved from the following spoken and written corpora, each belonging to one of the traditional periods in the history of English (see Tables 4 and 5 below):[14]

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 and the dashes for unstressed syllables.

14 The historical data (CED and OBC) classified as spoken corpora need to be regarded as "speech based" rather than truly spoken (see Culpeper and Kytö 2010: 16–17).

Table 4: Spoken corpora.

Subperiod	Time span	Spoken corpus	Number of words
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)	1980–2012	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.

B	Time span	Written corpus	Number of words
Early Modern English (EModE)	1560–1710	Innsbruck Corpus of English Letters CEECS I Corpus (1560 – onward) CEECS II Corpus Corpus of Early Modern English Texts (CEMET) Lampeter Corpus (Early Modern English portion – up to 1710)	2,848,314
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)	1920–2009	The Time Corpus (Time) The Corpus of Contemporary American English – Written component (COCA-W written component)	500,000,000

First, using the Wordsmith concordance program, *all* instances containing the inflected forms of the verbs *think* (i.e., *think*, *thinks*, *thinking*, *thought*), *suppose* (i.e., *suppose*, *supposes*, *supposed*, *supposing*), and *believe* (i.e., *believe*, *believes*, *believed*, *believing*) were retrieved from the written and the spoken corpora in the time span 1560–2012. Results were broken up in smaller 70-year subperiods, as shown in Tables 6–11. The subperiods were modeled after those contained in the Corpus of Late Modern English texts (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 subperiod, 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 later for each of the verbs under investigation.

Starting with the verb *think*, we began by randomly extracting ($n = 3{,}101$) tokens from the spoken English corpora and ($n = 6{,}619$) 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 7.

As can be seen from Figures 1 and 2, a comparison of the diachronic relative frequency patterns of the *that* versus zero forms per million words with the verb *think* indicates that frequency of the zero form has remained relatively constant vis-à-vis the complementizer *that* from 1560 to 2010 in both the spoken and written genres. The zero form is clearly the more frequent form from 1560 to 2012 and this accords with previous findings on *think* and with claims regarding diachronic *that*/zero variation (see Rissanen 1991; Palander-Collin 1999).

This process was conducted again with the verb *suppose*. The extraction yielded ($n = 2{,}778$) *suppose* tokens from the spoken corpora and ($n = 1{,}305$) tokens from the written corpora. The full set ($n = 4{,}083$) of tokens was again divided into those containing either a *that* clause or a zero complementizer clause (again with tokens not containing the *that* or zero form being discarded).

Table 6: Distribution of *that* clauses and zero complementizer clauses from EModE to PDE in the spoken corpora.

***Think* – spoken data**				
Period	***Think – that***		***Think – zero***	
	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)	

Note: *n*, absolute frequency; *N*, normalized frequency per million.

Table 7: Distribution of *that* clauses and zero complementizer clauses from EModE to PDE in the written corpora.

***Think* – written data**				
Period	***Think – that***		***Think – zero***	
	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)	

Note: *n*, absolute frequency; *N*, normalized frequency per million.

The distributions of these tokens for both the spoken and written data sets are presented in Tables 8 and 9.

When we compare the diachronic relative frequency patterns of the *that* versus zero forms per million words for the verb *suppose*, we find that once again the zero form occurs more frequently than the complementizer

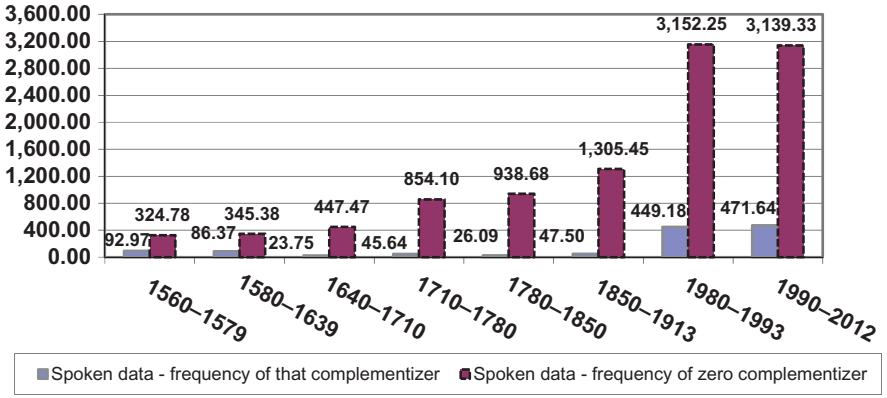

Figure 1: *Think* spoken data – *that* versus zero distribution per million words.

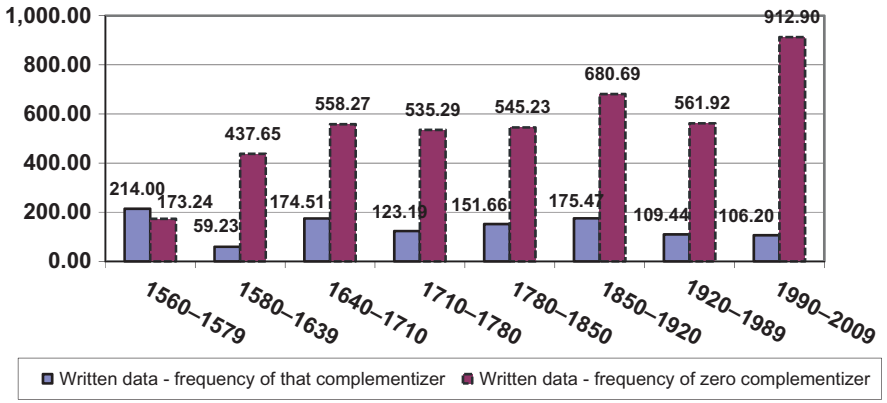

Figure 2: *Think* written data – *that* versus zero distribution per million words.

that in both the spoken and written data sets. These patterns are presented in Figures 3 and 4.

Finally, this extraction process was undertaken with the verb *believe*, yielding ($n = 3{,}061$) *believe* tokens from the spoken English corpora and ($n = 1{,}706$) tokens from the written English corpora. The full set ($n = 4{,}767$) of tokens was again divided into those containing either a *that* clause or a zero complementizer clause. The distributions of these tokens for both the spoken and written data sets are presented in Tables 10 and 11.

In contrast to the relatively consistent *that*/zero complementizer patterns observed with the *think* and *suppose* data sets, the ratio and frequency of the

Table 8: Distribution of *that* clauses and zero complementizer clauses from EModE to PDE in the spoken corpora.

***Suppose* – spoken data**				
Period	***Suppose – that***		***Suppose – zero***	
	n	***N***	***n***	***N***
1560–1579	(*n* = 2)	23.20	(*n* = 2)	23.20
1580–1639	(*n* = 2)	5.96	(*n* = 12)	35.74
1640–1710	(*n* = 2)	1.47	(*n* = 3)	2.20
1710–1780	(*n* = 21)	5.61	(*n* = 451)	124.91
1780–1850	(*n* = 28)	10.97	(*n* = 446)	185.12
1850–1913	(*n* = 32)	9.27	(*n* = 466)	138.68
1980–1993	(*n* = 5)	5.74	(*n* = 144)	165.23
1990–2012	(*n* = 18)	2.70	(*n* = 170)	28.70
Total	(*n* = 110)		(*n* = 1694)	

Note: *n*, absolute frequency; *N*, normalized frequency per million.

Table 9: Distribution of *that* clauses and zero complementizer clauses from EModE to PDE in the written corpora.

***Suppose* – written data**				
Period	***Suppose – that***		***Suppose – zero***	
	n	**N**	**n**	**N**
1560–1579	(*n* = 0)	0.00	(*n* = 0)	0.00
1580–1639	(*n* = 7)	20.83	(*n* = 22)	53.56
1640–1710	(*n* = 24)	28.01	(*n* = 72)	83.54
1710–1780	(*n* = 34)	64.51	(*n* = 72)	136.88
1780–1850	(*n* = 49)	74.78	(*n* = 65)	100.90
1850–1920	(*n* = 44)	44.37	(*n* = 134)	135.49
1920–2009	(*n* = 29)	13.65	(*n* = 162)	81.10
Total	(*n* = 187)		(*n* = 527)	

Note: *n*, absolute frequency; *N*, normalized frequency per million.

zero complementizer form seen in the *believe* data sets present a different diachronic picture. These results are presented in Figures 5 and 6.

The analyses of the relative frequencies of the *that* and zero forms in the spoken data set reveal a greater frequency of the zero form relative to the *that* complementizer up until present-day English (PDE) (1980–2012), when suddenly

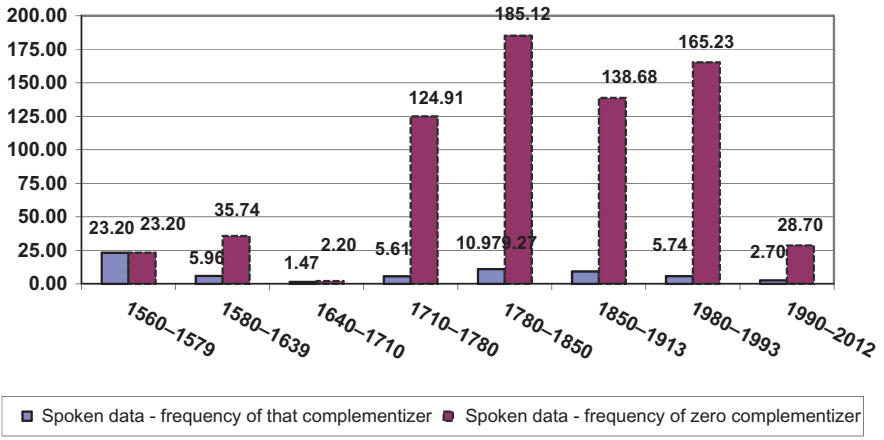

Figure 3: *Suppose* spoken data – *that* versus zero distribution per million words.

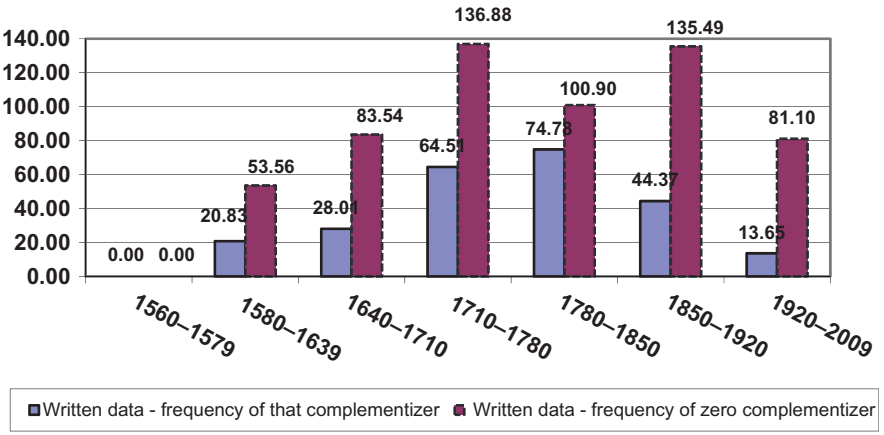

Figure 4: *Suppose* written data – *that* versus zero distribution per million words.

the *that*/zero forms start to appear in almost equal frequencies. An analysis of the parallel written data set reveals a similar pattern: a higher frequency of occurrence for the zero forms relative to *that* up until 1850, whereupon the two forms appear almost equally through 2009. The fact that this shift to *that*/zero parity is seen in the spoken and written data sets and that it occurs within roughly the same period is both unexpected and interesting. What consequences it has with regard to the factors predicting the presence of the zero form (if any)

Table 10: Distribution of *that* clauses and zero complementizer clauses from EModE to PDE in the spoken corpora.

***Believe* – spoken data**				
Period	***Believe – that***		***Believe – zero***	
	n	***N***	***n***	***N***
1560–1579	(*n* = 0)	0.00	(*n* = 0)	0.00
1580–1639	(*n* = 2)	5.96	(*n* = 1)	2.98
1640–1710	(*n* = 25)	18.35	(*n* = 208)	152.69
1710–1780	(*n* = 16)	23.11	(*n* = 482)	695.81
1780–1850	(*n* = 14)	16.79	(*n* = 452)	534.79
1850–1913	(*n* = 41)	43.80	(*n* = 503)	571.84
1980–1993	(*n* = 58)	66.44	(*n* = 67)	76.95
1990–2012	(*n* = 72)	233.64	(*n* = 72)	234.74
Total	(*n* = 228)		(*n* = 1,785)	

Note: *n*, absolute frequency; *N*, normalized frequency per million.

Table 11: Distribution of *that* clauses and zero complementizer clauses from EModE to PDE in the written corpora.

***Believe* – written data**				
Period	***Believe – that***		***Believe – zero***	
	n	***N***	***n***	***N***
1560–1579	(*n* = 0)	0.00	(*n* = 0)	0.00
1580–1639	(*n* = 8)	17.85	(*n* = 17)	47.61
1640–1710	(*n* = 37)	57.36	(*n* = 97)	150.37
1710–1780	(*n* = 38)	63.56	(*n* = 129)	213.25
1780–1850	(*n* = 61)	92.73	(*n* = 82)	123.21
1850–1920	(*n* = 80)	105.59	(*n* = 78)	102.96
1920–2009	(*n* = 73)	79.57	(*n* = 78)	85.01
Total	(*n* = 297)		(*n* = 481)	

Note: *n*, absolute frequency; *N*, normalized frequency per million.

remains, however, to be seen, and this finding will be integrated into and accounted for with our regression modeling in Section 4.

The (*n* = 2,055) spoken and (*n* = 2,217) written *think* sentences, the (*n* = 1,804) spoken and (*n* = 714) written *suppose* sentences, and the (*n* = 2,013) spoken and (*n* = 778) written *believe* sentences which contained either a *that* or zero

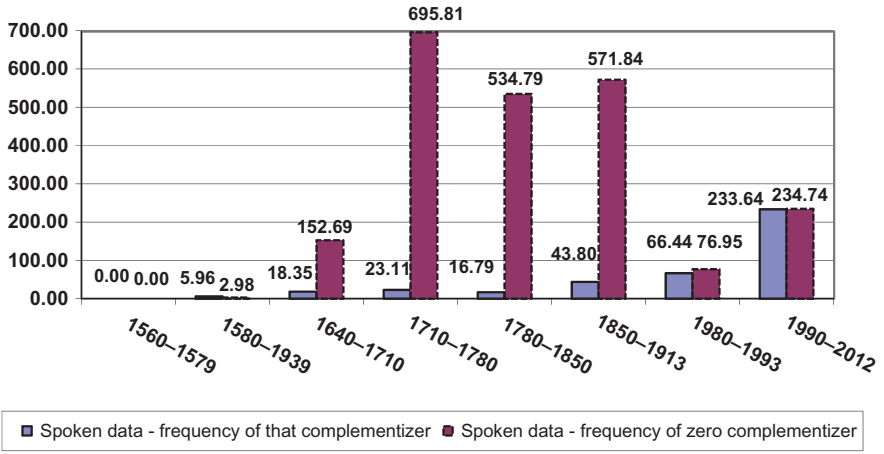

Figure 5: *Believe* spoken data – *that* versus zero distribution per million words.

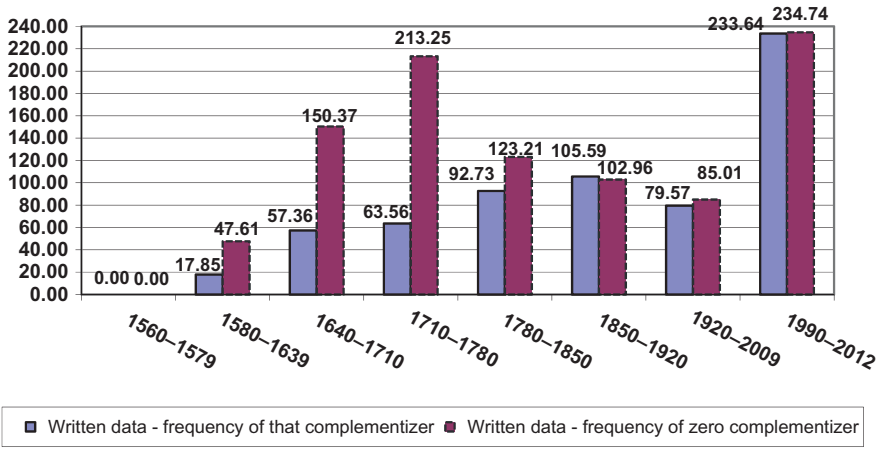

Figure 6: *Believe* written data – *that* versus zero distribution per million words.

complementizer clause were then coded for a number of features. Aside from a token's descriptors "inflected form" and "concordance line," each token was coded for features which, on the basis of the literature described earlier, can be seen as factors potentially favoring or disfavoring zero complementation. The following types of features were distinguished: matrix clause features, complement clause features, features relating to the relationship between matrix and complement, as well as two language-external features (the subperiod to which the token belongs and spoken or written mode).

Matrix features included the verb type (*think*, *believe*, or *suppose*), number, person, and tense[15] of the matrix verb, length of the matrix clause subject (pronoun/NP-short 1–2 words/NP-long 3 + words), and presence (or absence) of additional elements within the matrix clause (elements between the subject and the matrix verb). The complement clause feature was the length of the subject (again expressed in terms of pronoun/NP-short 1–2 words/NP-long 3 + words). Finally, features pertaining to the relationship between matrix and complement comprised coreferentiality of person between the matrix and complement clause subject, harmony of harmony of polarity, intervening elements (between the matrix clause and the complement clause), and cotemporality (i.e., tense agreement across the matrix and complement clauses).

In addition to the aforementioned coding for these variables, 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, we combined data from several periods. For example, with the verb *believe* this process resulted in an initial period spanning 1560–1710, and with the verb *suppose* it created an initial period spanning 1580–1710. The verb *think* was, however, frequent enough per period so that this step was not needed. Once the aggregation process was completed, the periods of the resulting data sets were sufficiently large enough to function as reference levels for our logistic regression analysis. This process was also employed for the PDE spoken data categories from 1980 to 2012, for all three verbs, allowing us to set up a single twentieth-century period with which to directly compare and contrast the written data sets from 1920 to 2009.

Once these processes were completed, the data was loaded into the R statistical software package in order to investigate the effects of the factors[16] via a *stepwise*

15 The coding for tense was divided into four categories: past (which included simple, progressive, perfect, and perfect progressive forms), present (again encompassing simple, progressive, perfect, and perfect progressive forms), future (auxiliary and non-finite future forms), and n/a (forms consisting of an auxiliary or a non-finite form other than future form).

16 Note, we do not specifically consider "I.or.U" (first- or second-person singular pronouns) as an individual factor because of the redundancy vis-à-vis the factors "Person" and "Number" (at the suggestion of Stefan Th. Gries). This methodological decision is also applied to the factors "Matrix subject" (pronoun or NP) and "Complement clause subject" (pronoun or NP) because "Matrix clause subject length" and "Complement clause subject length" contain the levels *it*, pronoun, np-short, and np-long, and thus already capture these important distinctions.

logistic regression analysis (by means of the function stepAIC in the R package MASS) – see Table 12 in Appendix.[17] As indicated earlier, the stepwise selection procedure was bidirectional and the minimal model was an intercept-only model. The maximal model contained all main effects plus two-way interactions of the factors with period, verb, and mode, i.e., spoken versus written mode, together with the two-way interactions between period, verb, and mode themselves. The resulting model after stepwise selection contains 11 main effects (the factors of matrix clause subject length as well as cotemporality were not strong enough to be selected by the stepwise procedure) and 16 interactions. This model performs reasonably well: the goodness of fit is significant (LLR = 3198.121; df = 53; p-value < 0.001), the predicted variation (C-score) is 85.3%, but the explained variation (Nagelkerke R^2) is only 38.1%. This shows that our model may still be improved. For additional validation, we dichotomized the fitted probabilities for our *that*/zero alternation at a cut-off value of 50% in order to compare them with the observed *that*/zero alternation (as outlined by Agresti 2013: 221–224). This yields a classification accuracy (in a confusion matrix) of 87.4%. In other words, 87.4% of all the observations were classified correctly by our regression model as having either the *that* or the zero complementizer. The significance of this result was furthermore tested against two baseline models: one that would always predict the most frequent form and one that would guess randomly. In both cases, our classification accuracy was highly significant (close to 0). All these diagnostics show, in summary, that our model is appropriate.

The next section will discuss all the effects of our regression model. For further statistical details concerning the significance of the factors, the reader is referred to the Appendix where an ANOVA table (Table 12) is given.

4 Results

In this section, we present the results from the stepwise regression analysis on 11 factors that have been argued to predict the presence of the zero complementizer form with verbs of cognition such as *think*, *suppose*, and *believe* (see Section 2.3). Because of the complex structure of our model (with 16 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 vs 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

17 The general outline of this methodology was suggested to us by Stefan Th. Gries, for which we want to express our gratitude.

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 Section 4.1, we discuss the five statistically significant interactions with verb, viz. interactions with person, number, tense, intervening elements between matrix and complement, and harmony of polarity. In Section 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, tense, and the absence of matrix-internal elements, subject coreferentiality, and the length of the complement clause subject. The final set of interactions, presented in Section 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 Interactions with verb

First, we gauge that effects are verb specific (as these effects are aggregated over all time periods, we can call them "panchronic"). The significant factors are presented in Figures 7–11.

In our first interaction plot analyzing the interaction between verb type and matrix clause person, we see that for *think* and *suppose*, the order from best to poorest predictor of the zero form is first, second, and third persons. For *think*, the difference between the three levels is minimal. *Suppose* has similar values to *think* for first and second persons, but a much lower rate of zero with third person. The rate of zero use for both first and second persons is lower overall for *believe* in comparison to *think* and *suppose*. Third-person *suppose* has approximately the same value as third-person *believe*. One possible explanation why third-person *suppose* leads to a much lower probability of *that* may be that in the third person singular simple present, *suppose* is trisyllabic, which gives it more phonetic weight. This extra syllable sets *supposes* apart from all other finite forms of *suppose*, and may lead to more *that*-use. This tentative explanation is consistent with the finding that when additional material is added to the matrix clause, *that* becomes more likely; the third-person ending can be regarded as extra material adding more weight to the matrix clause and making it sound more marked than the default disyllabic verb form.

Although *believe* behaves differently from *think* and *suppose* in having a higher zero rate in the third person than in the second person, this difference is not significant.

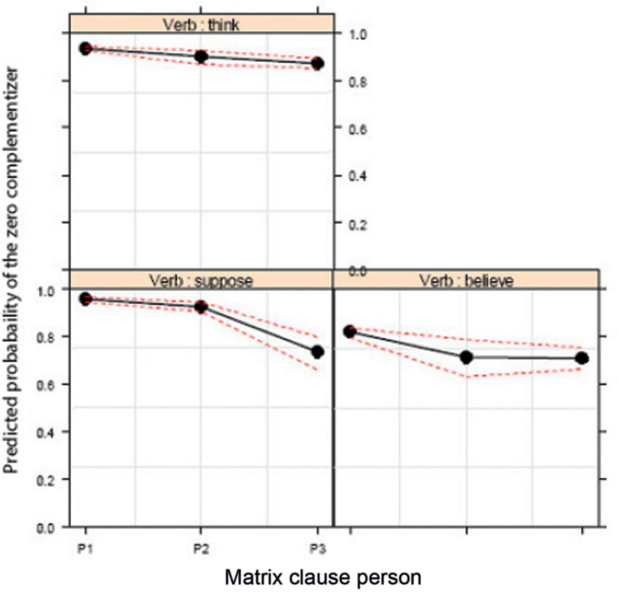

Figure 7: Verb: matrix clause person.

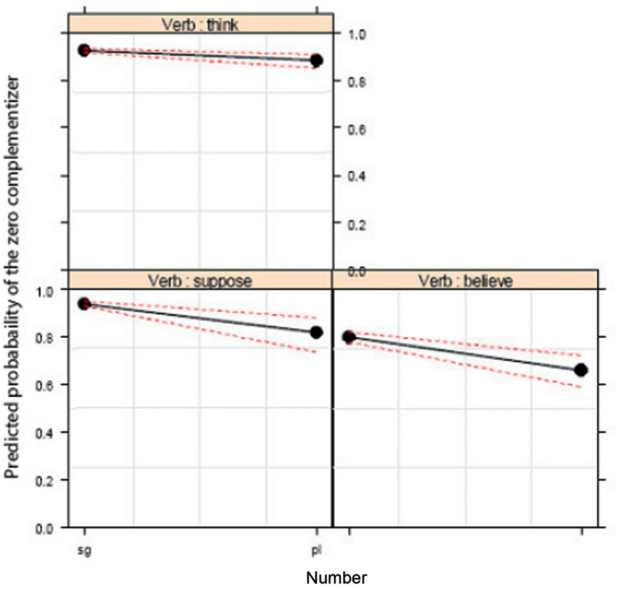

Figure 8: Verb: matrix clause number.

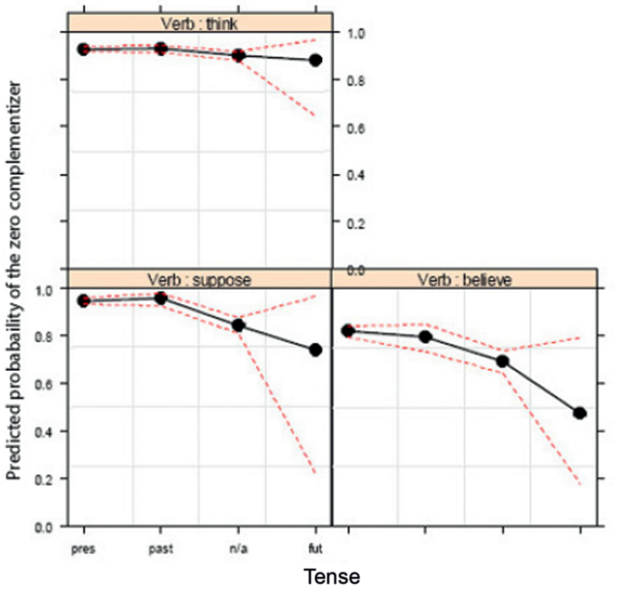

Figure 9: Verb: matrix clause tense.

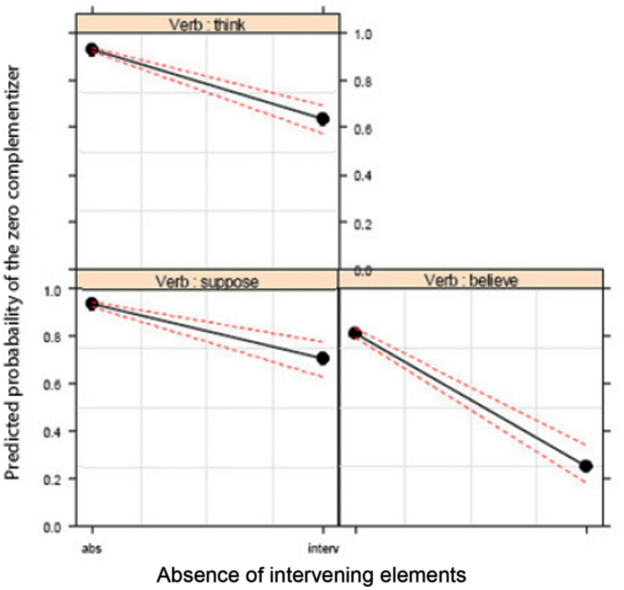

Figure 10: Verb: absence of intervening elements.

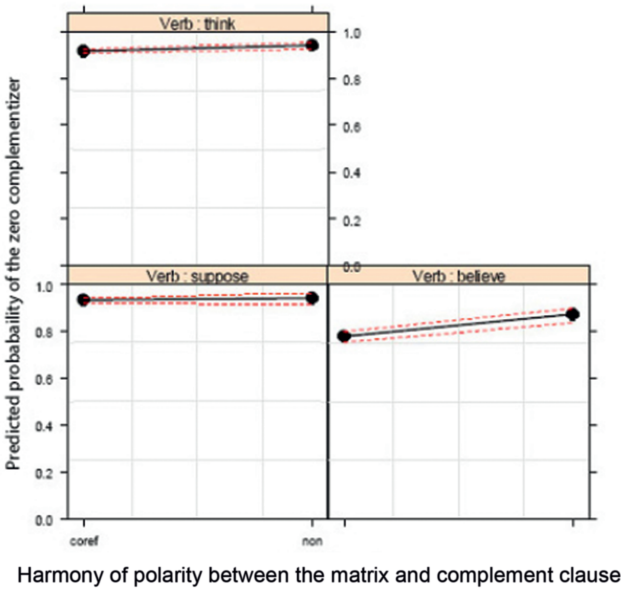

Figure 11: Verb: harmony of polarity between the matrix and complement clause.

In Figure 8, we see that the singular form (sg) more strongly predicts the zero form across all three verbs. The predictive power of singular matrix clause subjects is stronger for *think* and *suppose* than it is for *believe*. The zero rates for both singular and plural (pl) *believe* are lower than those of the other two verbs.

An analysis of tense across all three verbs indicates that past and present tenses do not differ significantly as predictors of the zero form. The effects of verb forms other than present, past, or future, i.e., non-finite forms, or the use of auxiliaries (n/a) are lower for the verb *believe* than for the verbs *think* and *suppose*; in fact, all three verbs can be seen to differ significantly with regard to the n/a form. The future tense, then, is unreliable for all three verbs, as indicated by the large confidence intervals (cf. the dotted lines). Finally, this plot shows that for each of its tense forms, the verb *believe* is less predictive of the zero form than *think* and *suppose*, i.e., the *believe* values are lower than those of the others verbs.

In Figure 10, we see that absence of intervening elements between matrix and complement is a very strong predictor of the zero form, it is significant across all three verbs and that its effect is similar for *think* and *suppose*. However, *believe* is most sensitive to the presence or absence of intervening material. When intervening material is present between a matrix clause with *believe* and the ensuing complement clause, the presence of the explicit *that*

form is more likely than the zero form; the plot shows that the zero rate for *believe* with intervening material is below 0.5.

The fifth and final interaction with verb type is harmony of polarity between the matrix and the complement clause. The results show that *think* and *suppose* slightly favor the zero form in disharmonious patterns and this effect was even more pronounced with *believe*, which has a lower zero rate in harmonious patterns than *think* and *suppose* but comparable values when there is no harmony of polarity.

We will now compare the extent to which these factors predict the use of the zero form in the spoken and written modes.

4.2 Interactions with mode

The interactions of the factor "mode" (i.e., spoken vs written mode) with other factors (see Table 12 in Appendix) are also panchronic, i.e., all periods are conflated. In this section, we will see that mode plays a more important role in the zero/*that* alternation, in that it has an impact on the strength of the other factors: some factors may be better predictors of the zero form in one mode as opposed to the other.

Figure 12 allows us to compare the conditioning effect of intervening elements between matrix clause and complement clause in the spoken and written

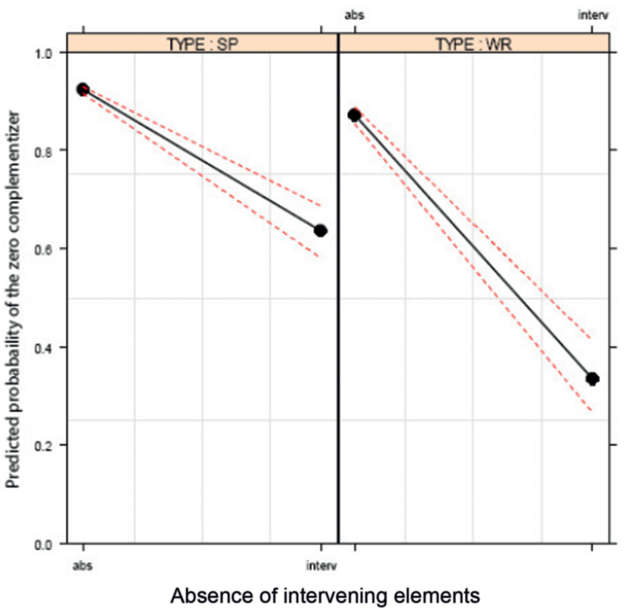

Figure 12: Mode: absence of intervening elements.

modes. In both panels of Figure 12, we observe a dramatic difference in complementizer use between presence and absence of intervening material. A notable difference, however, resides in the extent to which the presence of intervening material in the written mode predicts the zero form. 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 below 0.4. 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.

In Figure 13, we examine the effect of matrix clause person in the two modes. In the spoken mode, first-person subjects (P1) predict more zero use than second- and third-person forms (P2 and P3). In the written mode, the difference between first- and second-person subjects is not significant, but the difference between these values and the much lower zero rate with third-person subjects is significant. Also, compared to the spoken mode, third-person subjects in written data are much less likely to be used with a zero complementizer.

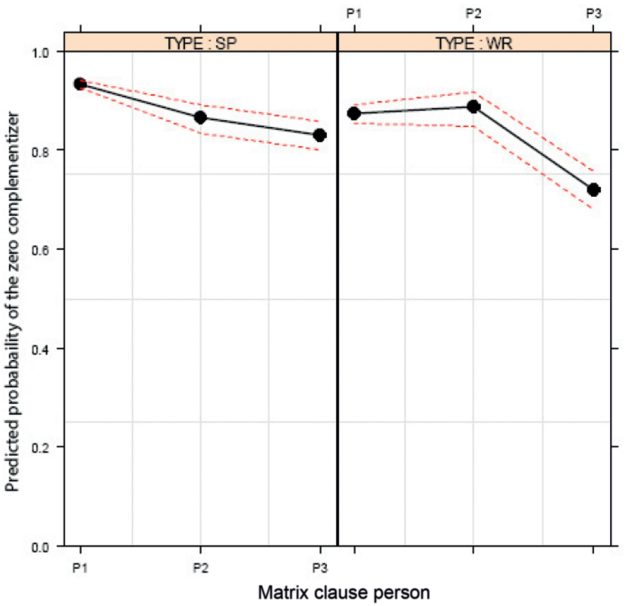

Figure 13: Mode: matrix clause person.

The analysis of tense relative to mode follows the pattern of tense and its interaction with verb; once again, in both the spoken and written data, past and present tense forms are not significantly different from one another. The n/a forms condition for zero use, but much less so in the written mode than in the spoken mode. Due to the sparseness of future forms and resultant large confidence intervals, we cannot make any claims about the effect of the future on zero use in spoken versus written data (Figure 14).

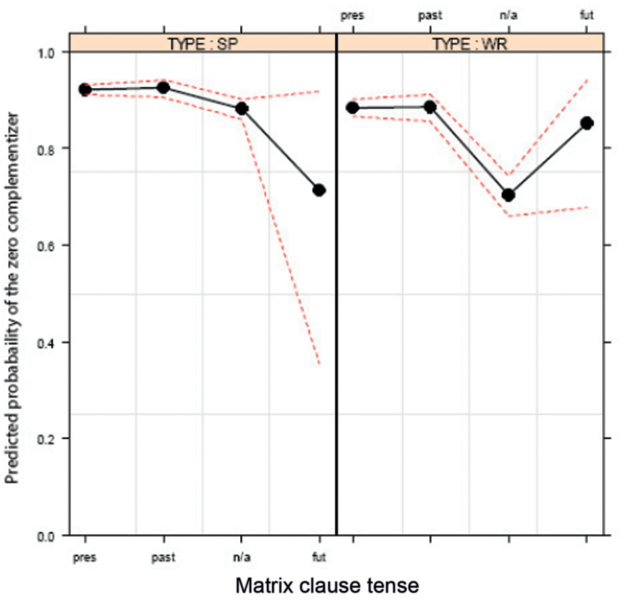

Figure 14: Mode: matrix clause tense.

Figure 15 presents the results for the effect of elements between the matrix clause subject and verb in the spoken and written modes, respectively. Seeing that this interaction only approaches significance, we can say that the absence of matrix-internal elements is a fair predictive factor for both spoken and written data. The steepness of the plot lines shows that the difference in predictive power between presence and absence of matrix internal elements is comparable for both modes, but in the written mode, the zero form is used less overall, as the lower points for both levels indicate.

An analysis of the effect of the coreferentiality of person between the matrix and complement clause subjects reveals an interesting difference with regard to mode. Coreferentiality of person leads to higher levels of zero in the spoken data. In the written data, when subjects are coreferential, zero is slightly less

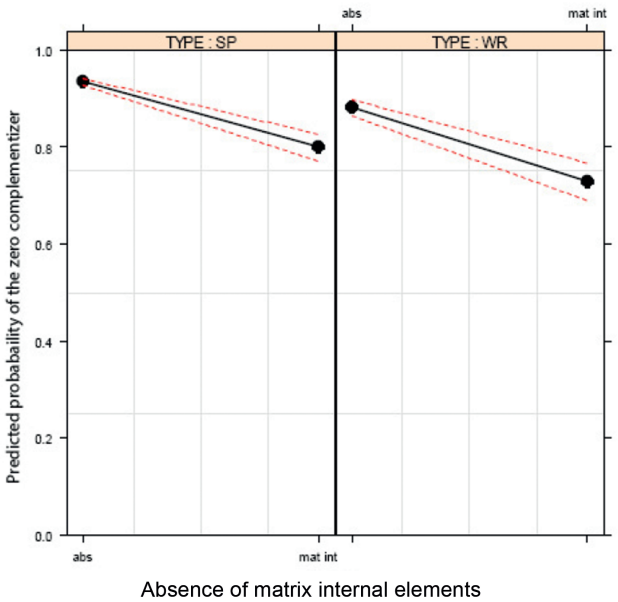

Figure 15: Mode: absence of matrix-internal elements.

likely than when they are not coreferential. However, the difference is minimal and the confidence interval of the non-coreferential written data point is rather large (Figure 16).

The final factor that we will examine in this section is the predictive effect of the length of the complement clause subject on the zero form relative to mode. The plot in Figure 17 shows that within the spoken data the following cline exists: *it* > pro > np-short > np-long (which is in line with the main effect observed for this factor). A comparison between the two modes shows that short and long NPs tend more strongly toward *that* in the written mode than in the spoken mode. Overall, length of the complement clause subject has a stronger effect on written data than on spoken data. Again, the complexity principle, i.e., the need to mark off clause boundaries, may motivate writers' choice of the *that* complementizer as opposed to the zero form. In addition, the concern with clarity fostered by standardization and prescriptivism may also play a role.

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. In the following sections, we will discuss the interactions with period that came out as significant.

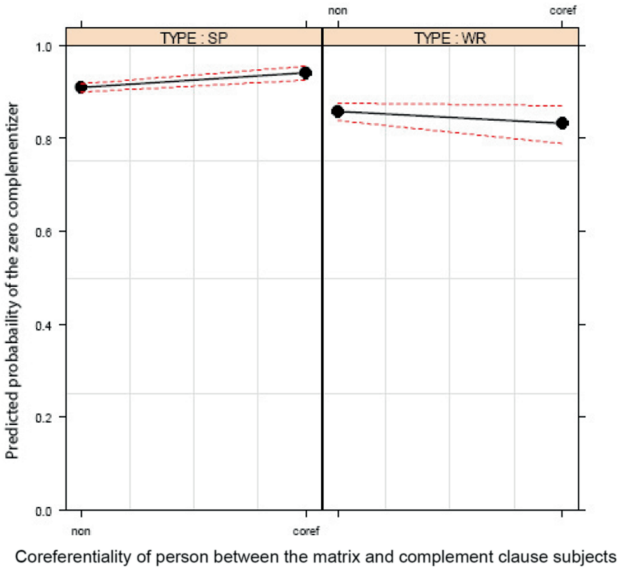

Figure 16: Mode: coreferentiality of person between the matrix and complement clause subjects.

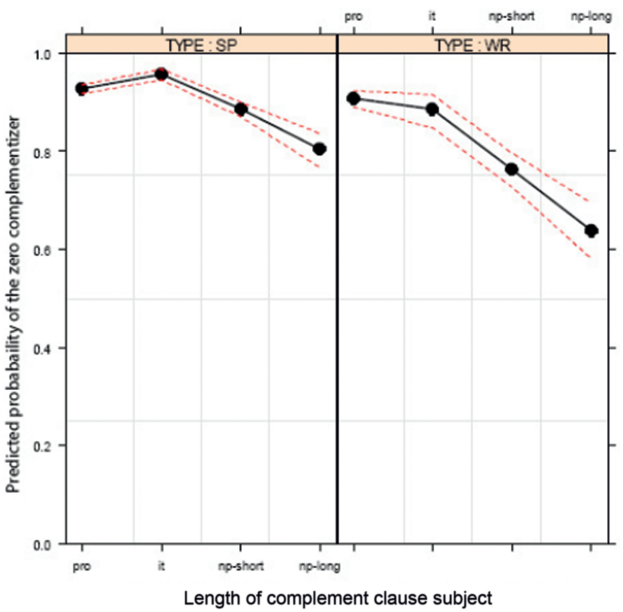

Figure 17: Mode: length of the complement clause subject.

4.3 Interactions with period

The interaction effects with period are significant for the following factors: verb, length of the complement clause subject, matrix clause person, and harmony of polarity. 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 18 shows the diachronic development of the zero form for each of the three verbs. It reveals that *think, suppose*, and *believe* start out at roughly the same zero rate and that all three verbs indeed exhibit a loss of zero. However, there are notable differences as to the extent of this downward trend. The decrease is minimal for *suppose* and even more so for *think*. *Believe*, by contrast, is characterized by a dramatic drop in zero use, especially from 1780 up to the present day. In the most recent data set (period 8), it even plummets below 50%, i.e., in most recent times, *believe* has come to prefer *that* over zero.[18]

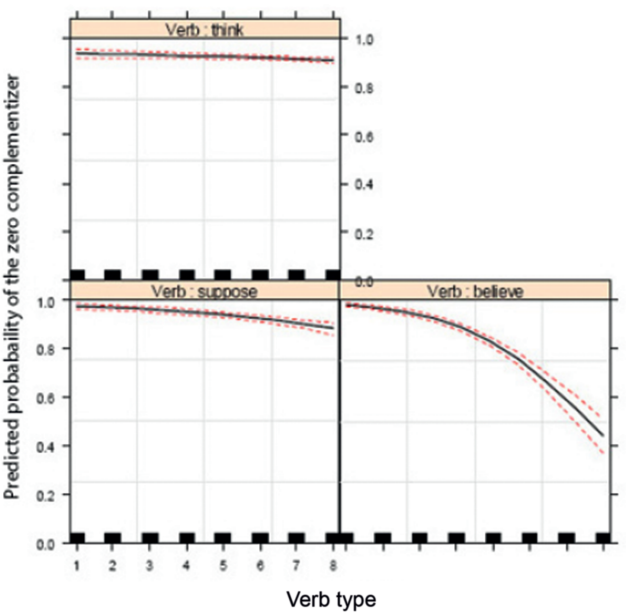

Figure 18: Period: verb.

18 A related study (Shank et al. In preparation), in which we investigate the *that*/zero alternation with *think*, *guess*, and *understand*, reveals that *guess* undergoes an increase in zero

An analysis of the effect of the length of the complement clause subject over time shows that all types of subject used to be better predictors of the zero form than they are now, but that NPs lose more of their zero forms than pronouns (*it* and other pronouns). Thus, over time, short and long NPs tend more strongly toward *that* than *it* and other pronouns (Figure 19).

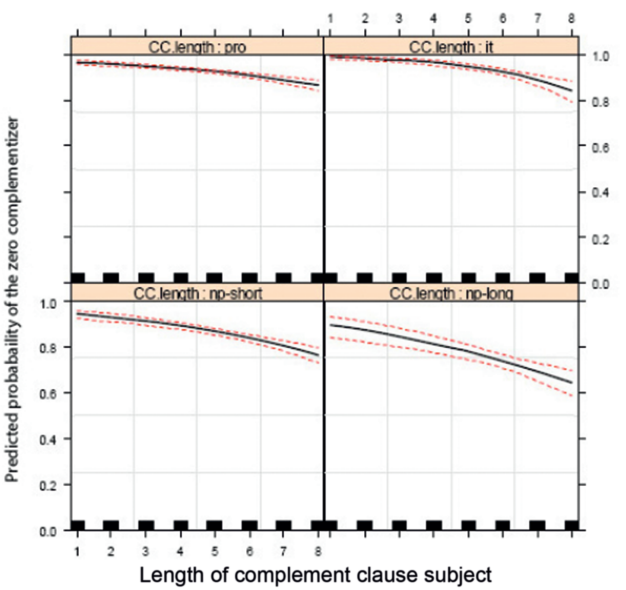

Figure 19: Period: length of the complement clause subject.

The loss of the zero form over time is also seen in Figure 20, which represents the diachronic effect of person. It can be observed that zero rates with both first and third persons decline gradually over time, with third person dropping off more dramatically than first person. By contrast, no such decrease in zero can be observed in the second-person data.

The final significant effect over time is the interaction between harmony of polarity between matrix and complement clause and period. Figure 21 shows that when there is harmony of polarity, there is a distinct tendency toward more *that* over time; in other words, harmony of polarity used to be a stronger predictor of the zero form than it is now. This trend is absent from the non-harmonious data; here, the level of zero use remains more or less stable.

complementation. This confirms that there is no homogeneous *that*/zero alternation trend and that interactions with verb are highly relevant. Also, it opens up perspectives for future research on the basis of a larger number of verbs.

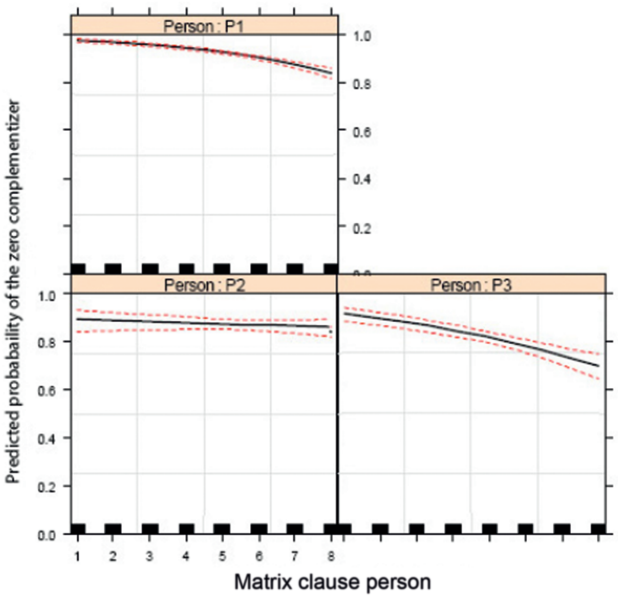

Figure 20: Period: matrix clause person.

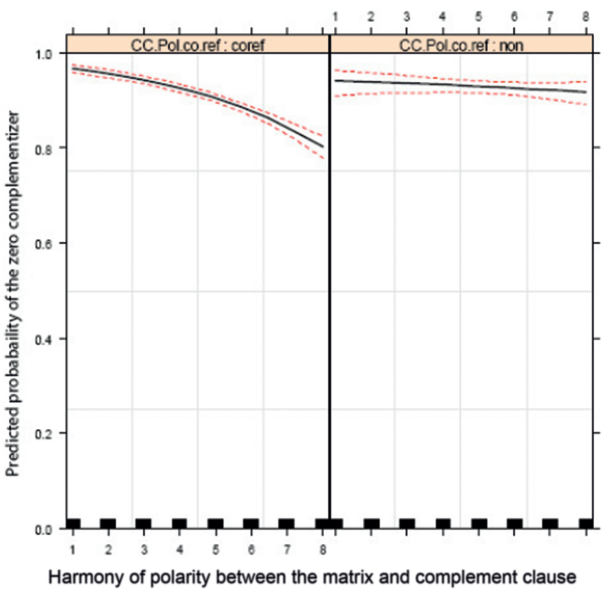

Figure 21: Period: harmony of polarity between the matrix and complement clause.

5 Conclusion

This study has shown that, contrary to claims in the literature that there has been an overall diachronic tendency toward more zero complementizer use at the expense of the *that* complementizer, the three most frequent complement-taking mental verbs in PDE, viz. *think*, *suppose*, and *believe*, in fact exhibit a diachronic *decrease* in the zero complementizer and a concomitant *increase* in *that* use. This trend can be observed for each individual verb, as the interaction between verb type and period shows.

As for the other effects tested in this study, viz. interactions with verb, mode, and period, the absence of intervening material is by far the best predictor, followed by matrix-internal elements for the zero form. The results for complement clause subject length confirm Torres Cacoullos and Walker's (2009) findings: *it* most strongly conditions zero followed by other pronouns, short NPs, and long NPs. Singular matrix clause subjects are more amenable to zero than plural subjects and the conditioning effect of first-person subjects is stronger than that of the second person, which in turn has a stronger effect than the third person. Contrary to expectations in the literature (Bolinger 1972; Torres Cacoullos and Walker 2009), when there is no harmony of polarity, zero is more likely to be selected. Person coreferentiality is a valid predictor, but tense was not significant.

Earlier work on the *that*/zero alternation has often relied on descriptive statistics only; the present study has tried to avoid the shortcomings of this approach by utilizing a rigorous methodology as well as creating sufficiently large and representative sample sizes for each historical period. As evidenced by our initial presentation of findings in Section 3, a reliance on descriptive statistics (sometimes supported by chi-squared tests) can unintentionally obscure interaction effects and/or not reveal the robustness of diachronic trend lines. From a descriptive perspective, it would appear that the zero form for *think* and *suppose* is robust or at least remained consistent over time, and one could thus reasonably infer that the factors proposed as promoting the use of the zero form are either equally predictive or remain significant over time. It is only when a methodology such as the one used in this study is applied that the changing impacts of the various factors become apparent.

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

became apparent, first, that the verbs examined in this study exhibit the aforementioned diachronic increase in the use of *that* to differing degrees. In this regard, *believe* stands out from *think* and *suppose*, exhibiting a more pronounced rise of the complementizer *that*; in the most recent time period, the proportion of *that* is higher than that of the zero form. Second, the extent to which the factors mentioned in the literature actually predict zero use may differ considerably from verb to verb, as was revealed by the analysis of interactions between the factor verb type and other factors. A striking finding in this regard is the effect of intervening material. A strong predictor overall, lack of intervening material is an especially good conditioning factor with *believe*, leading to zero rates below 50% for this verb. Also, *believe* has a greater tendency toward zero in disharmonious polarity patterns than *think* and *suppose*. Third, the effect of many conditioning factors is highly dependent on the mode. As mentioned above, the absence of intervening material between matrix and complement clause strongly conditions each of the individual verbs, but the interaction with mode also reveals that the written mode is especially susceptible to its conditioning effect. The same goes for complement clause length and matrix clause person. Conversely, coreferentiality of person favors zero in the spoken mode only. Fourth, interactions with period show that some factors, notable complement clause subject length and matrix clause person, lose some of their zero forms over time.

This study has further shown that the effect of conditioning factors is also dependent on mode. Again, intervening material was a case in point. Its zero forms are much stronger in the written mode than 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 verbs as this may reveal additional differences in the way *that*/zero alternation has evolved with each individual verb 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 would be to look 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 (Dehé and Wichmann 2010) could be incorporated into the logistic regression model. One drawback to the study of prosody and its effect on *that*/zero 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.

Appendix

The following table presents the so-called type III tests for our 11 main effects and 16 interactions, i.e., the indications of how *poorer* our model would become if the factor in question were *removed*. The first row signifies that no predictors are removed, i.e., the current model. The order of the predictors in the table is determined by the selection of the stepwise procedure and is therefore completely arbitrary. The column "Deviance" gives a measure of *lack of fit* with the actual data; hence, it should ideally be as low as possible. The column "AIC" lists *Akaike's Information Criterion*, which is related to Deviance and has therefore the same meaning: better models have lower AIC scores. The third column

Table 12: Type III LLR tests of 11 main effects and 16 interactions.

	df	Deviance	AIC	LRT	Pr(>Chi)
		7952.5	8060.5		
Absence of matrix-internal elements	1	8087.4	8193.4	134.848	<2.2e-16
Complement clause subject length	3	7983.3	8085.3	30.771	9.500e-07
Intervening elements	1	8059.4	8165.4	106.808	<2.2e-16
Person	2	7985.8	8089.8	33.239	6.057e-08
Verb type	2	8006.1	8110.1	53.506	2.406e-12
Mode	1	8017.9	8123.9	65.352	6.265e-16
Number	1	7962.6	8068.6	10.079	0.0015000
Tense	3	7954.5	8056.5	1.959	0.5809472
Period	1	7974.9	8080.9	22.378	2.239e-06
Harmony of polarity	1	7959.6	8065.6	7.049	0.0079294
Person coreferentiality	1	7963.4	8069.4	10.892	0.0009657
Verb type: matrix clause person	4	7994.1	8094.1	41.529	2.089e-08
Verb type: period	2	8097.3	8201.3	144.714	<2.2e-16
Mode: period	1	8039.8	8145.8	87.212	<2.2e-16
Period: harmony of polarity	1	7971.6	8077.6	19.094	1.245e-05
Verb type: tense	6	7973.9	8069.9	21.375	0.0015706
Person: period	2	7973.9	8077.9	21.366	2.294e-05
Mode: tense	3	7974.9	8076.9	22.369	5.466e-05
Person: mode	2	7968.2	8072.2	15.627	0.0004043
Complement clause subject length: mode	3	7971.7	8073.7	19.182	0.0002507
Absence of intervening elements: mode	1	7962.8	8068.8	10.252	0.0013655
Verb type: number	2	7958.9	8062.9	6.363	0.0415324
Complement clause subject length: period	3	7960.4	8062.4	7.883	0.0484970
Absence of matrix-internal elements: mode	1	7955.6	8061.6	3.069	0.0798110
Absence of intervening elements: verb type	2	7959.6	8063.6	7.083	0.0289655
Mode: person coreferentiality	1	7962.4	8068.4	9.839	0.0017081
Verb type: harmony of polarity	2	7956.6	8060.6	4.004	0.1350579

"LRT" gives the Likelihood Ratio statistic of the predictor removal, which is chi-square distributed. The last column gives the *p*-value, indicating which predictor removals are statistically significant. In other words, significance indicates which predictor removals make the model significantly *worse*. As can be seen from the table, only the main effect of tense and the interaction between verb type and harmony of polarity are not significant. They stay in the model, however, for different reasons. Tense, on the one hand, has interactions with both verb type and mode which are significant. Removal of the interaction between verb type and harmony of polarity, on the other hand, would lead to a higher AIC.

References

Agresti, Alan. 2013. *Categorical data analysis*. Hoboken: Wiley.

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

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. 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 Brian D. Joseph & Richard D. Janda (eds.), *The handbook of historical linguistics*, 602–623. Oxford: Blackwell.

Culpeper, Jonathan & Merja Kytö. 2010. *Early Modern English dialogues: Spoken interaction as writing*. Cambridge: Cambridge University Press.

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. 36–74.

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

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

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.

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

Fischer, Olga. 2007. The development of English parentheticals: A case of grammaticalization? In Stefan Dollinger Smit, Julia Hüttner, Gunther Kaltenböck &

Ursula Lutzky (eds.), *Tracing English through time: Explorations in language variation*, 99–114. Vienna: Braumüller.
Givón, Talmy. 1980. The binding hierarchy and the typology of complements. *Studies in Language* 4. 333–377.
Givón, Talmy. 1995. Isomorphism in the grammatical code. In John Haiman (ed.), *Iconicity in syntax*, 47–76. Amsterdam: John Benjamins.
Gorrell, Joseph Hendren. 1895. Indirect discourse in Anglo-Saxon. *Publications of the Modern Language Association of America* 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 Brad C. Love, Ken McRae & Vladimir N. Sloutsky (eds.), *Proceedings of the Cognitive Science Society Conference*. Washington, DC. 1061–1066.
Jespersen, O. H. 1954. *A modern English grammar on historical principles: Part III: Syntax* (second volume). London: George Allen & Unwin.
Kaltenböck, Gunther. 2006. '... That is the question': Complementizer omission in extraposed that-clauses. *English Language and Linguistics* 10. 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. 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 Timothy Shopen (ed.), *Language typology and syntactic description. Volume II: Complex constructions*, 42–140. Cambridge: Cambridge University Press.
Palander-Collin, Minna. 1999. Grammaticalization and social embedding: I THINK and METHINKS in Middle and Early Modern English. Helsinki: Société Néophilologique.
Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. 1985. *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 Karin Aijmer & Bengt Altenberg (eds.), *English corpus linguistics: Studies in honour of Jan Svartvik*, 272–289. London: Longman.
Rohdenburg, Günter. 1996. Cognitive complexity and increased grammatical explicitness in English. *Cognitive Linguistics* 7(2). 149–182.
Shank, Christopher, Julie Van Bogaert & Koen Plevoets. In preparation. A multifactorial analysis of *that*/zero alternation: A diachronic study of the grammaticalization of the zero complementizer construction with *think*, *guess* and *understand*.
Tagliamonte, Sali & Jennifer Smith. 2005. *No momentary fancy!* The *zero* 'complementizer' in English dialects. *English Language and Linguistics* 9. 289–309.
Thompson, Sandra A. 2002. "Object complements" and conversation: Towards a realistic account. *Studies in Language* 26. 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 Elizabeth Closs Traugott & Bernd Heine (eds.), *Approaches to grammaticalization*, vol. II, 313–329. Amsterdam: John Benjamins.

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

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. Special issue of. *Linguistics* 42(2). 295–332.

Warner, Anthony R. 1982. *Complementation in Middle English and the methodology of historical syntax. A study of the Wycliffite Sermons*. London: Croom Helm.

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

Corpora

A Corpus of English Dialogues 1560–1760. 2006. Kytö, Merja & Jonathan Culpeper.

Corpus of Early Modern English Texts. Department of Linguistics. 2005. De Smet, Hendrik. University of Leuven.

Innsbruck Computer Archive of Machine-Readable English Texts. (CD-ROM version). 1999. Markus, Manfred (ed.). University of Innsbruck.

Parsed Corpus of Early English Correspondence, text version. 2006. Nevalainen, Terttu, Helena Raumolin-Brunberg, Jukka Keränen, Minna Nevala, Arja Nurmi, Minna Palander-Collin & Ann Taylor. University of Helsinki and University of York. Distributed through the Oxford Text Archive.

The British National Corpus, version 3 (BNC XML Edition). 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/ [2013].

The Corpus of Contemporary American English: 450 million words, 1990–present.

Davies, Mark. 2008–. Available online at http://corpus.byu.edu/coca/ [2013].

The Corpus of Historical American English: 400 million words, 1810–2009. Davies, Mark. 2010-. Available online at http://corpus.byu.edu/coha/ [2013].

The Corpus of Late Modern English Texts (Extended Version). 2006. De Smet, Hendrik. Department of Linguistics, University of Leuven.

The Lampeter Corpus of Early Modern English Tracts. 1999. Schmied, Josef, Claudia Claridge & Rainer Siemund. (In: ICAME Collection of English Language Corpora (CD-ROM), Second Edition, eds. Knut Hofland, Anne Lindebjerg, Jørn Thunestvedt, The HIT Centre, University of Bergen, Norway.)

The Old Bailey Corpus. Spoken English in the 18th and 19th centuries, 2012. Huber, Magnus, Magnus Nissel, Patrick Maiwald & Bianca Widlitzki. www.uni-giessen.de/oldbaileycorpus [2013].

TIME Magazine Corpus: 100 million words, 1920s-2000s. Davies, Mark. 2007-. Available online at http://corpus.byu.edu/time/ [2013].