Investigating the impact of structural factors upon that/zero complementizer alternation patterns in verbs of cognition: a diachronic corpus-based multifactorial analysis

Shank, Christopher; Plevoets, Koen

Research in Corpus Linguistics

DOI:
10.32714/ricl.06.07

Published: 07/05/2019

Publisher's PDF, also known as Version of record

RiCL Research in Corpus Linguistics

Investigating the impact of structural factors upon *that*/zero complementizer alternation patterns in verbs of cognition: a diachronic corpus-based multifactorial analysis

Christopher Shank / Koen Plevoets
Bangor University, UK / Ghent University, Belgium

Abstract – This corpus-based study examines the diachronic development of the *that*/zero alternation with nine verbs of cognition, viz. *think*, *believe*, *feel*, *guess*, *imagine*, *know*, *realize*, *suppose* and *understand* by means of a stepwise logistic regression analysis. The data comprised a total of (n=5,812) *think*, (n=3,056) *believe*, (n=1,273) *feel*, (n=1,885) *guess*, (n=2,225) *imagine*, (n=1,805) *know*, (n=1,244) *realize*, (n=2,836) *suppose* and (n=3,395) *understand* tokens from both spoken and written corpora from 1580–2012. Taking our cue from previous research suggesting that there has been a diachronic increase in the use of the zero complementizer form from Late Middle / Early Modern to Present-day English, we use a large set of parallel spoken and written diachronic data and a rigorous quantitative methodology to test this claim with the nine aforementioned verbs. In addition, we also investigate the impact of eleven structural features, which have been claimed to act as predictors for the use or presence of the zero complementizer form for 'panchronic' (i.e. effects are aggregated over all time periods) and diachronic effects. The objectives of this study are to examine the following: (i) whether there is indeed a diachronic trend towards 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, but the extent of this decrease is not the same for all verbs. In addition, 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

1. INTRODUCTION

The focus of this paper is upon *that*/zero complementizer alternation patterns in constructions with an object clause, as seen in examples (1) and (2):

(1) I think that I shall dedicate the book to the Professor. (BNC)
(2) I think I have one of the finest woman in England. (BNC)

Previous studies by Rissanen (1991), Thompson and Mulac (1991a, 1991b) and Palander-Collin (1999) have suggested that this [verb + Object clause] construction has been diachronically moving, from the Early Modern English period forward, towards an increased use of the zero complementizer form. The present paper seeks to test this hypothesis by means of a stepwise logistic regression analysis of (n=23,531) tokens of *think*, *believe*, *feel*, *guess*, *imagine*, *know*, *realize*, *suppose* and *understand*, nine of the most frequently used complement-taking verbs of cognition, spanning the time period from 1580 to 2012, in both spoken and written data sets. 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 specifically focus upon and test whether these proposed structural factors indeed predict the zero form, whether they gain or lose predictive power when combined, the impact of verb type and mode (i.e. spoken versus written data) and what happens to their ability to foretell the presence/use of the zero forms 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 a cutting-edge diachronic perspective to existing research into structural factors acting as predictors for *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 structural 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 and suggestions for future research 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 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.

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

[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 (2) in the main text, the matrix clause is the entire sentence starting with *I* and ending in *England*. 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 (2), 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 (1991b) 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 percent. 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 percent.

These synchronic, frequency-based findings lead Thompson and Mulac (1991b: 323–326) 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.

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".

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; see also Fischer 2007: 103) designates as "simple introductory expressions like the Modern English 'you know'", which stand in a paratactic relationship with the ensuing clause.

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 nine verbs under investigation in this study 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. Furthermore, it investigates, by means of a stepwise regression analysis, whether the zero form is on the increase and how time affects the factors in terms of foretelling the presence/use of the zero forms. 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 nine 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 of 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).[4] 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 Wyclifite Sermons, for example, *that* is used 98 percent 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 percent. In addition, Rissanen (1991) shows that the zero form is more common in speech-like genres (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

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

[4] 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.*

of medical writing and sermons.[5] 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[6]

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).

There is also a wide range of language-internal factors. Some have argued that 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). A number of studies have also have shown certain high-frequency subject-verb collocations to be strongly associated with zero use (among these are the "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" (*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 section, the structural conditioning factors favouring the use of zero will be discussed based on 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 ability to foretell the presence/use of the zero form, 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* (4), favour the use of zero (Bolinger 1972; Elsness 1984; Thompson and Mulac 1991a; Tagliamonte and Smith 2005; Torres Cacoullos and Walker 2009).[7] 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 (4) and full NPs (5) 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.

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

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

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).[8] 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 predictor 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" (2009: 33). Distinguishing between single-word (6a) as opposed to phrasal adverbials (6b), and pre-subject (6c) as opposed to post-subject (6d) 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.

(6a) I expected *maybe* that we would be talking about it.
(6b) *At the beginning*, we told the guy that we were gonna both-each have our own.
(6c) *Now* I find Ø like, even adults use slang words.
(6d) 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 (6d) 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 favourably disposed towards zero complementation than non-finite forms.[9] Negation, in (8), subsumed under "additional material" in Tagliamonte and Smith (2005), is treated as a separate foretelling 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.

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)

Table 1: Matrix clause factors potentially favouring the zero complementizer

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

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) favour 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 favoured 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 favourable 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 favour 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.

Factor	No statistics	Significant	Not significant
subject = pronoun	Warner (1982) Elsness (1984) Finegan and Biber (1985) Rissanen (1991) Rohdenburg (1996)	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)	

Table 2: Complement clause factors potentially favouring the zero complementizer

2.3.3. The relationship between matrix and complement clause

The presence of intervening material between matrix and complement has been widely discussed as a factor favouring 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's (1996: 161) words, "any elements capable of delaying the processing of the object clause and thus the overall sentence structure favour 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 (12). The factor came out as significant in both Thompson and Mulac (1991a) and Torres Cacoullos and Walker (2009).

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.[10] 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.

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)

Table 3: Factors pertaining to the relationship between matrix and complement that potentially favour zero

[10] 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).

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:[11]

Sub-period	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 London-Lund Corpus* *The American National Corpus* – Spoken component *The British National Corpus* – Spoken component. (BYU-BNC-S) *The Corpus of Contemporary American English* – Spoken component (COCA-S)	99,026,000

Table 4: Spoken corpora

Sub-period	Time span	Spoken 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)	500,000,000

Table 5: Written corpora

First, using the Wordsmith Tools concordance program, all instances containing the inflected forms of all nine verbs were retrieved from the written and the spoken corpora in the time spans 1580–2009 and 1580–2012. For example, with the verb *think*, the following four inflected forms were utilized as search terms (*think*, *thinks*, *thinking*, *thought*). This search and extraction process was repeated for all nine verbs. Results were broken up in smaller 70-year sub-periods, as shown in Tables 9–14 in the Appendix. The sub-periods were modelled after those contained in the CLMET corpora (i.e. 1710–1780, 1780–1850 and 1850–1920) in order to provide a principled template in which to divide and analyse 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 Bailey Corpus ends in 1913 or the BYU-BNC spoken component 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 use the first period of 1580–1639 as the reference level for the subsequent regression analysis applied to the nine verbs discussed in this article.

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 (n=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 data sets described below for each of the verbs under investigation. A total of (n=45,028) examples from the spoken corpora and of (n=25,584) from the written corpora were extracted and analysed for this initial stage.

[11] The historical data in Table 4 (CED and OBC) classified here as spoken corpora need to be regarded as "speech-based", rather than as truly spoken (see Culpeper and Kytö 2010: 16–17).

Within this set of spoken and written examples of (n=70,612) only those examples which contained either a *that* clause or a zero complementizer clause, (n=16,036) spoken and (n=8,513) written examples, were retained and subjected to further analysis.[12] The results from this process are presented below in Table 6.

	Spoken data	Written data
think	(n=3,550)	(n=2,251)
believe	(n=2,583)	(n=777)
feel	(n=670)	(n=606)
guess	(n=1,050)	(n=834)
imagine	(n=1,303)	(n=916)
know	(n=831)	(n=975)
realize	(n=782)	(n=460)
suppose	(n=2,629)	(n=714)
understand	(n=2,638)	(n=980)
Total	(n=16,036)	(n=8,513)

Table 6: Total number of *that*-clauses and zero complementizer clauses coded for all nine verbs in the spoken and written corpora

The (n=16,036) spoken and (n=8,513) written examples were then coded for descriptors such as 'inflected form', 'concordance line' and for structural features which, on the basis of the literature described above, can be seen as factors potentially favouring or disfavouring zero complementation. This resulted in the following categorization rubric: matrix clause features, complement clause features, features relating to the relationship between matrix and complement, as well as two language-external features, namely the period to which the token belongs and either 'spoken' or 'written' mode.

The specific matrix features coded included the verb type (*think*, *believe*, *feel*, *guess*, *imagine*, *know*, *realize*, *suppose* or *understand*), number, person and tense[13] 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 features that were coded included the length of the subject (again expressed in terms of the pronoun *it* / any other pronoun[14] / 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 subjects, 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 nine 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. 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 data sets 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 *suppose* it created an initial period spanning 1580 to 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. This process was also employed for the PDE spoken data categories from 1960 to 2012, for all nine verbs, allowing us to set up a single twentieth-century period with which to directly compare and contrast the written data sets from 1920–2009.

Once these processes were completed, the data was loaded into the statistical software R (R Core Team 2018) in order to investigate the effects of the factors.[15] That was done by means of stepwise logistic

[12] The full details for all nine verbs in terms of *that*/zero forms per year and their frequency of occurrence per million words per period is found in the Appendix, Tables 9–15.

[13] 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 a future form).

[14] In Shank et al. (2014) we found that the pronoun form *it* was a significantly strong predictor itself relative to other pronouns as a complement clause subject for the zero form; therefore, *it* is now coded independently from all other pronominal forms.

[15] Note that 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.

regression analysis (with the function *stepAIC* in the R package *MASS*; Venables and Ripley 2002) – see Table 8 in Appendix.[16] Stepwise selection is a search procedure which looks for relevant combinations of predictors, and in our case it moved in-between an intercept-only model (minimal model) and the model with all main effects plus two-way interactions of the factors with period, verb and mode, i.e. spoken vs. written mode, together with the two-way interactions between period, verb and mode themselves (maximal model). The resulting model after stepwise selection contains eleven main effects (the factors of coreferentiality of person as well as coreferentiality of tense were not strong enough to be selected by the stepwise procedure) and twenty interactions. This model performs reasonably well: the goodness-of-fit is significant (G^2=11,593.45; df=134; p-value<0.0001), the predicted variation (C-index) is 0.887 percent and the explained variation (Nagelkerke-R^2) is 52.7 percent. This shows that our model is fairly good. For additional validation, we dichotomized the fitted probabilities for our *that*/zero alternation at a cut-off value of 50 percent 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 82.5 percent. In other words, 82.5 percent 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 (p-value<0.0001). Finally, we checked the (standardized) residuals, and only 2.8 percent of them lie outside of the interval between -2 and 2, which is well below the threshold of 5 percent (Faraway 2015: 84–85). All these diagnostics show, in summary, that our model is appropriate.

The next section discusses 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 of so-called *Type III tests* (Table 8) is given.

4. RESULTS

Due to the complex structure of our model (with sixteen interactions), the discussion of the effects will be done by means of graphical visualization in effect plots that were obtained with the R package effects (Fox 2003). The main factors under consideration are the main effects of verb, period and mode (i.e. spoken versus written), absence of matrix-internal elements, absence of intervening elements between the matrix and complement clauses, matrix clause person, matrix clause number, matrix clause tense, coreferentiality of polarity between matrix and complement clauses, length of the matrix clause subject and length of the complement clause subject.[17]

In Section 4.1 we discuss the seven statistically significant interactions with verb, viz. mode, absence of matrix-internal elements, absence of intervening elements between matrix and complement, person, number, tense and coreferentiality of polarity between the matrix and complement clauses. In 4.2 we show that the following interactions with mode are statistically significant: absence of intervening elements between the matrix and complement clauses, person, tense, coreferentiality of polarity between matrix and complement clauses, length of the matrix clause subject and length of the complement clause subject. The final set of interactions, presented in Section 4.3 and labelled 'period', 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 mode, verb, absence of intervening elements between the matrix and complement clauses, person, coreferentiality of polarity between the matrix and complement clauses, length of the matrix clause subject and length of the complement clause subject predict the use of zero.

4.1. Effects by verb

First, we gauge which effects are verb-specific (as these effects are aggregated over all time sub-periods, we can call them 'panchronic'). The significant factors are presented below in Figures 1–7.

[16] The general outline of this methodology was suggested to us by Stefan Th. Gries.

[17] The results for the individual main effects have not been included because these would be redundant, given the fact that any significant main effect interactions are captured and reported within the VERB, MODE and PERIOD results presented in this section.

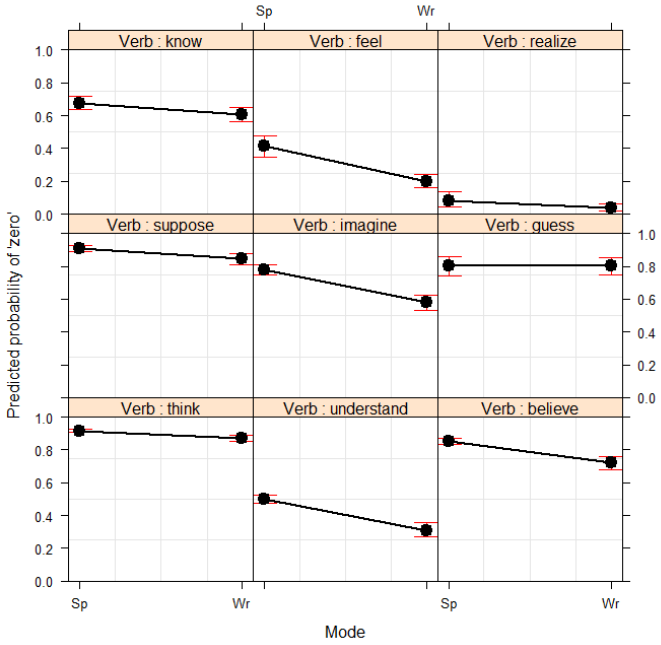

Figure 1: Verb : Mode

In our first effects plot, which presents the interaction between verb type and mode, we see with the verbs *think*, *know*, *suppose*, *imagine* and *believe* that the spoken genre or mode predicts the zero form more often than the written mode. The results for the verb *guess*, however, are not significant. This non-significance can also be seen with the verb *realize*; nevertheless, it should be noted that the results for *realize* are limited by a very low overall probability prediction rate. Finally, while the verbs *feel* and *understand* also revealed an overall probability below 50 percent, both indicated that the spoken mode still predicted the zero form significantly more than the written mode.

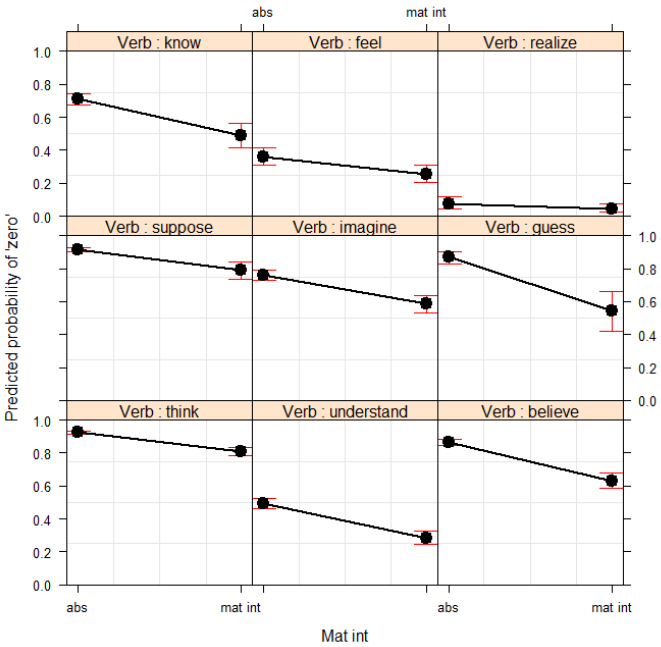

Figure 2: Verb : Absence of matrix-internal elements

In Figure 2, we see that the absence of intervening elements within the matrix clause significantly predicts the zero form for the verbs *think*, *know*, *suppose*, *imagine*, *guess* and *believe*. This factor is also significant with the verbs *feel* and *understand*; however, the results show that with these two verbs the

overall predictive probability fell below 0.5. Finally, this factor does not appear to be a significant predictor at all with the verb *realize*.

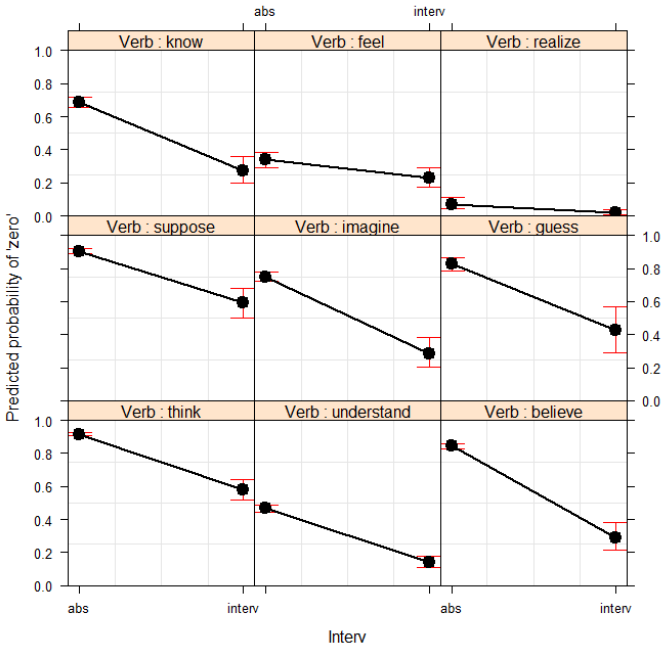

Figure 3: Verb : Absence of intervening elements

Figure 3 suggests that absence of intervening elements between matrix and complement is a very strong predictor of the zero form for the verbs *think*, *know*, *suppose*, *imagine*, *guess* and *believe*. The verbs *understand* and *feel* are also affected by the presence or absence of intervening material. The plot shows that while the zero rates for both verbs are below 0.5, only *understand* has a significant effect, but the effect for *feel* is borderline significance at best. Once again, no significant difference between absence and presence of intervening elements is seen with the verb *realize*.

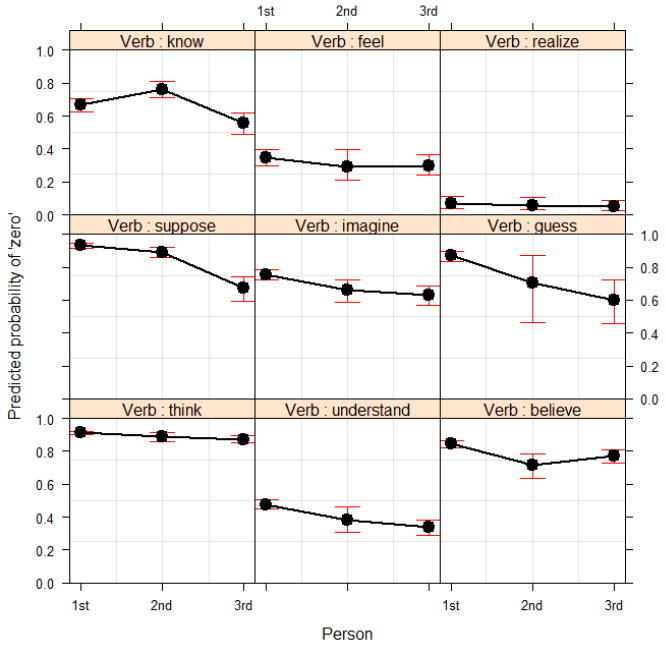

Figure 4: Verb : Person

In our third plot of the interaction between verb type and matrix clause person, we see that the three persons vary greatly in their prediction of the zero form. In addition, for several verbs there was no significance found at all with respect to person. For example with *think*, the differences between the three persons is minimal and not significant. A lack of significance for person was also found with *feel* and

realize. However, in these cases, and unlike *think*, both verbs had an overall frequency below 0.5, and r*ealize* in particular had a very low overall relative frequency across the entire person category. For the remaining verbs, the following patterns emerged: with *know*, and only with *know*, second person is the best predictor of the zero form and, furthermore, first and second person together are significantly different from third person with respect to predicting the zero form. *Believe* presents a different pattern, whereby first person is the best predictor of the zero form relative to second and third person. The emerging picture with this set of mental state predicates becomes more obscured with the inclusion of *imagine* and *understand*. Figure 4 shows that for these verbs first person is significantly different from second person in predicting the zero form, first person is also significantly different from third person and second and third person are not significantly different from each other. However, much like *feel* and *realize*, the predicted probability of the zero form with *understand* still falls below 0.5. Lastly, *suppose* and *guess* show further variation in that with *suppose* first and second person together are significantly different from third person, while with *guess* only first person is significantly different from third person in predicting the zero form.

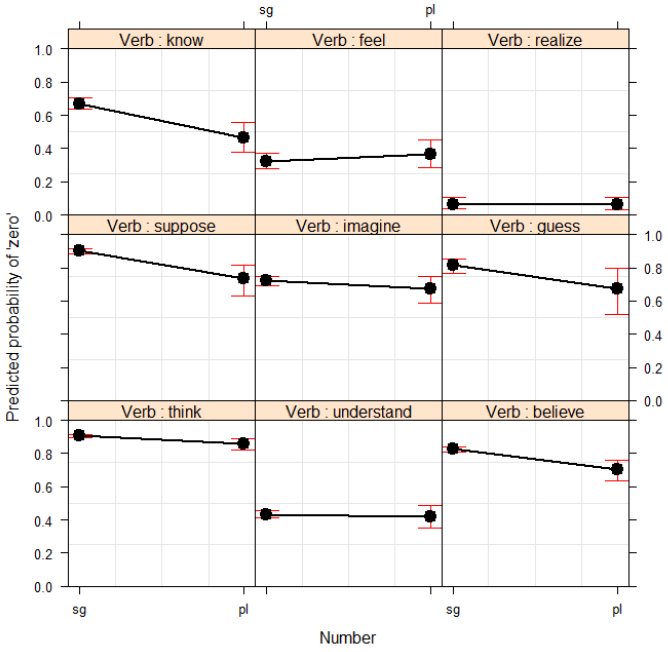

Figure 5: Verb : Number

In Figure 5, we see that the singular form more strongly predicts the zero form for the four verbs *think, know*, *suppose* and *believe.* The predicted probability of the zero form is strongest for *think* and less for *suppose*, *believe* and *know*. The predicted probability, for both singular and plural matrix clause subject forms, for *know* is lower than those of the other three verbs. In addition, we see that there are no significant differences for *feel*, *imagine*, *understand*, *realize* and *guess*.

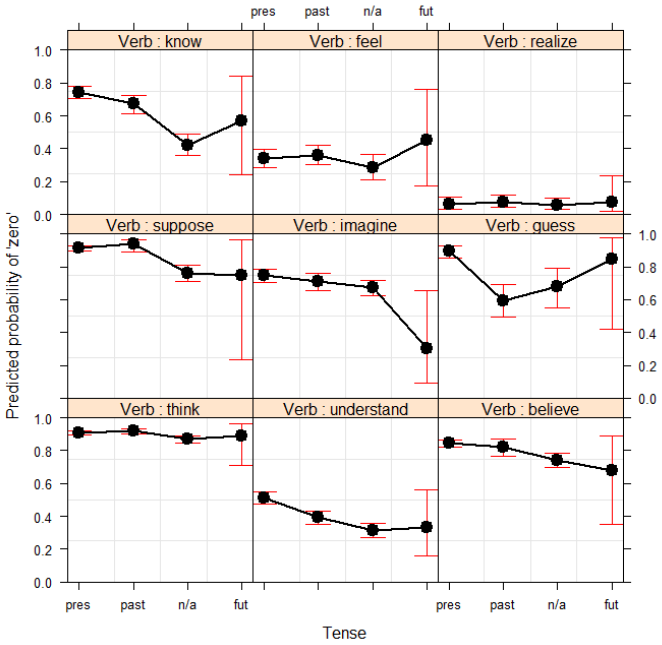

Figure 6: Verb : Tense

An analysis of tense across all nine verbs indicates that past and present tense do not differ significantly in predicting the zero form for eight verbs; *understand* is the only exception. Furthermore, the future tense is an uninformative category for all of the verbs, as indicated by the large confidence intervals. A closer look at the plot reveals the following patterns: while past and present tense are not significantly different from each other with *think*, *know* and *suppose*, they are significant with respect to use of auxiliaries (n/a). The verb *believe* also patterns much like *think* with respect to past and present tense. However, the plot indicates that with *believe* the present tense by itself is significantly different from use of auxiliaries (n/a). Next, the verb *understand* is the only verb where the present tense form is both significantly different from past tense and from use of auxiliaries (n/a). With *guess*, the plot reveals that the present tense is significantly different from both past tense and n/a. Present tense is also, albeit marginally, significantly different from the future tense with *imagine* but, unlike all of the previously mentioned verbs, the plot indicates that present tense, past tense and n/a themselves show no signs of being significantly different vis-à-vis one another. Finally, we see in Figure 7 that there are no significant differences for the verbs *feel* and *realize*.

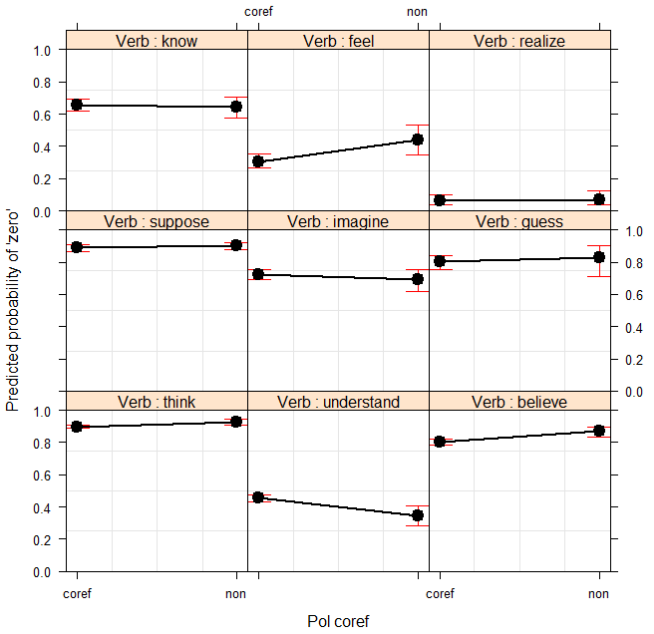

Figure 7: Verb : Coreferentiality polarity between the matrix and complement clauses

The fifth and final interaction effect with verb type is harmony of polarity between the matrix and the complement clauses. The results show that *think* and *believe* actually favour the zero form in disharmonious patterns (the confidence intervals of 'coref' and 'non' for both verbs do not overlap), which is counter to the expectation for this factor: Coreferentiality of polarity (i.e. harmony) between the matrix and complement clauses is supposed to be predicting the zero form according to the literature, but we observe the opposite result. The only verb where there appears to be a significant difference, where harmony of polarity significantly predicts the zero form, is *understand*. No significant differences were found with the remaining verbs *know*, *suppose*, *feel*, *imagine*, *realize* and *guess*.

4.2. Effects by mode

The interactions between mode and other factors (see Table 8 in Appendix) are also panchronic, i.e. all sub-periods are conflated. In this section, we will see that mode plays a more important role in the *that*/zero alternation, since it has an impact on the strength of these factors: some factors may better predict the zero form in one mode as opposed to the other.

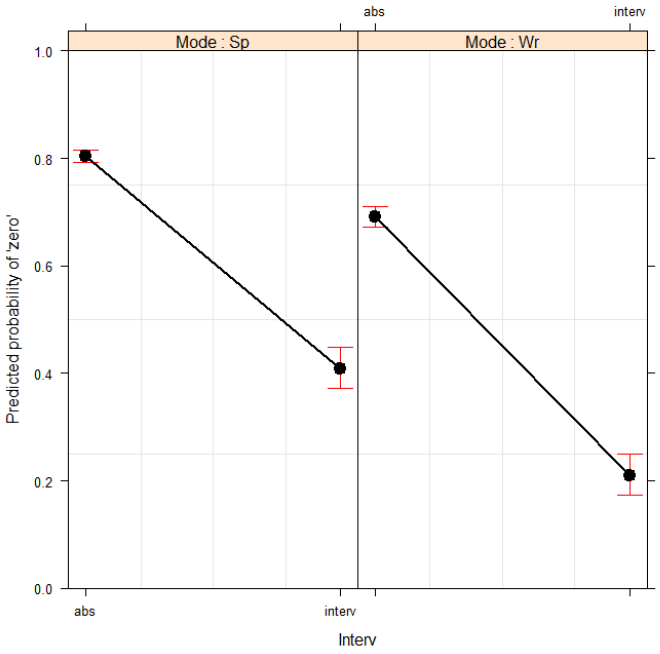

Figure 8: Mode : Absence of intervening elements

Figure 8 shows the predictive effect of (absence of) intervening elements between matrix clause and complement clause in the spoken and written modes. In both modes, we observe a considerable significant difference in complementizer use between presence and absence of intervening elements. However, we can note that the predicted probabilities for both presence and absence of intervening material are significantly higher in the spoken mode than in the written mode. When there is intervening material in the written mode, the predicted probability of the zero form drops below 0.4, so that the explicit complementizer *that* in fact becomes more likely. It may be that writers are guided more by the complexity principle than speakers and therefore feel the need to insert *that* to make clause boundaries clearer when intervening material risks impairing clarity.

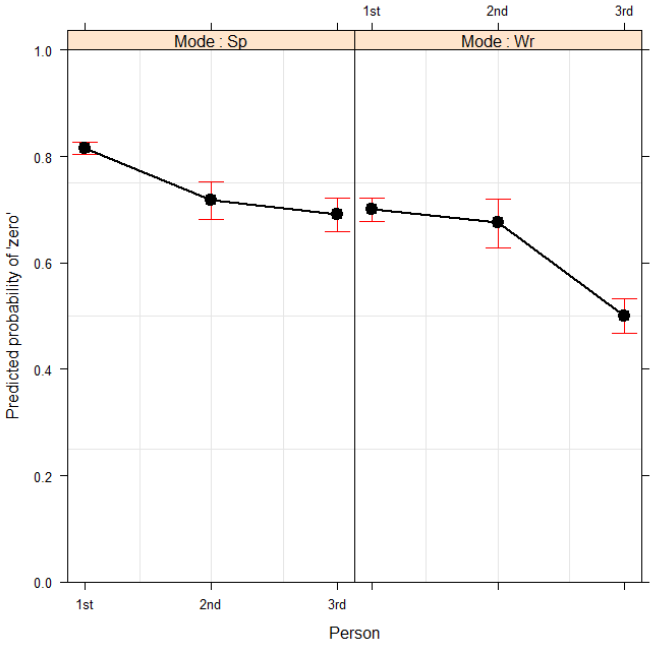

Figure 9: Mode : Matrix clause person

Figure 9 shows the effect of matrix clause person in the two modes. In the spoken mode, first person subjects significantly predict more zero use than second and third person forms. In the written mode, the difference between first and second person subjects is not significant, but the difference between these values and third person is significant. Also, compared to the spoken mode, both first and third person subjects in the written data are much less likely to be used with the zero form.

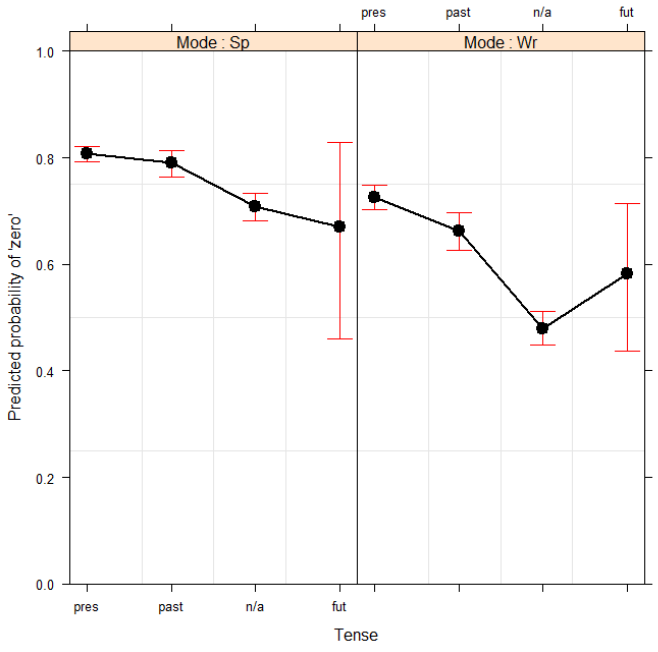

Figure 10: Mode : Matrix clause tense

The effect of tense by mode follows the pattern of tense by verb (see above). In both the spoken and the written data, past and present tense forms are not significantly different from one another. The auxiliary forms (n/a) predict the zero form significantly more in the spoken mode than in the written one; however, in the latter the results show that half of (n/a) forms occur with *that* and half with the zero form. Lastly, due

to the sparseness of future forms and the resultant large confidence intervals, we cannot make any claims about the prediction of the future tense for the zero form in spoken versus written data.

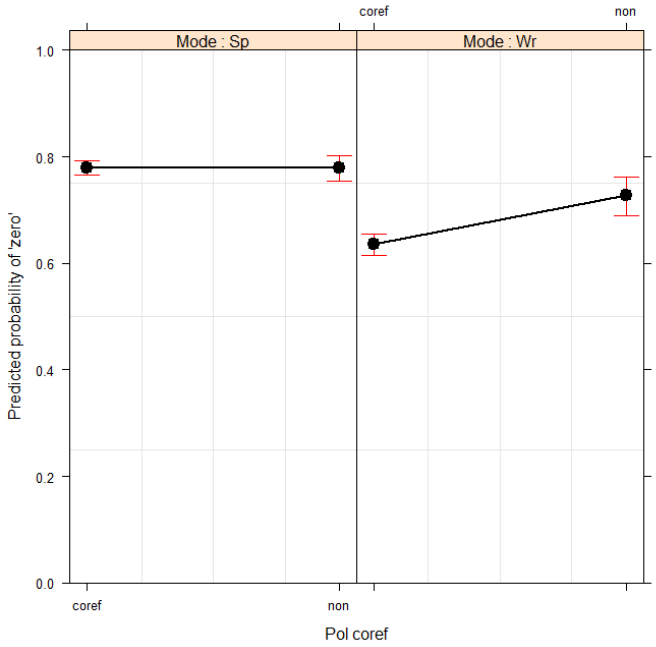

Figure 11: Mode : Coreferentiality of polarity between the matrix and complement clause

In Figure 11, we find evidence of a factor acting in a way opposite to expectations. The plot shows that in the spoken data there is no significant difference between harmony and disharmony of polarity in predicting the zero form. In the written data, however, non-coreferentiality actually predicts the zero form significantly more than coreferentiality (harmony). This finding is in opposition to those reported by and Elsness (1984: 526) and Torres Cacoullos and Walker (2009).

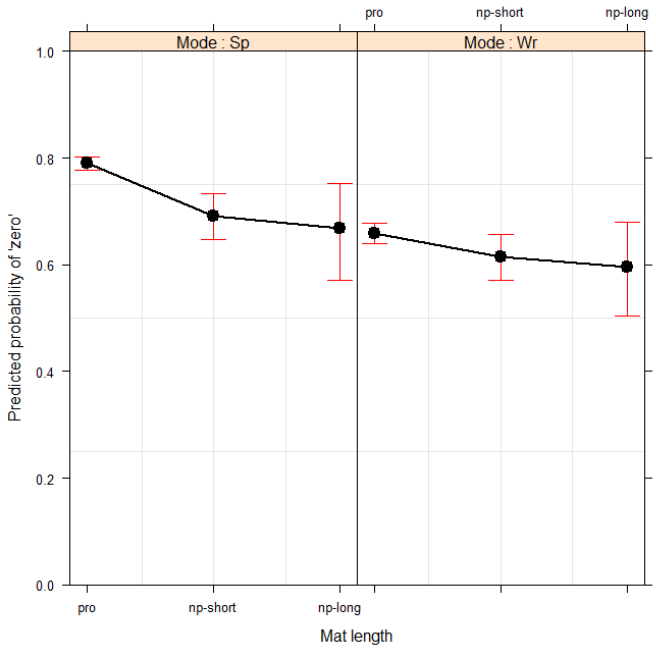

Figure 12: Mode : Length of the matrix clause subject

In Figure 12, we turn our attention to the impact of the length of the subject in the matrix clause in the written and spoken modes. The results show that in the spoken data pronominal subjects (pro) significantly predict the zero form more than NP-short and NP-long matrix clause subjects, which are not significant. In the written data set, there are no significant differences. This is in sharp contrast to the effect below in Figure 13 concerning the length of the complement clause subject.

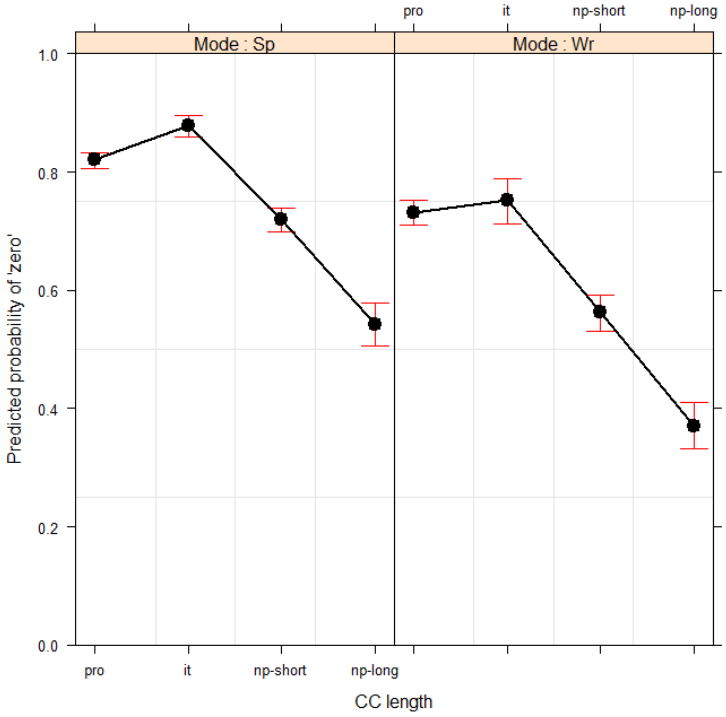

Figure 13: Mode : Length of the complement clause subject

In Figure 13, we see many significant differences in predicting the zero form. In the spoken data, there is a clear cline: *it* > pro > NP-short > NP-long (which is in line with the main effect of this factor, which is not reported due to space limitations). In the written data, however, there is no significant difference between pro and *it*, and both are equally strong. The comparison between the two modes shows that while short NPs still have a high predicted probability of the zero form in the spoken data, this is much lower in the written data. Lastly, long NPs have the lowest predicted probability of favouring the zero form more than that in either mode, and in the written data the probability falls below 0.5. Overall, the predicted probabilities for all four length categories are significantly higher in the spoken data than in the written data, where the *that* form is still more present. 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.

4.3. Effects by period

The interaction effects with period are the following: mode (written versus spoken), verb, absence of intervening elements, person, coreferentiality of polarity between the matrix and complement clause, length of the matrix clause subject and length of the complement clause subject. This final part of the analysis offers a diachronic perspective; it shows whether the impact of a given factor becomes stronger or weaker over time.

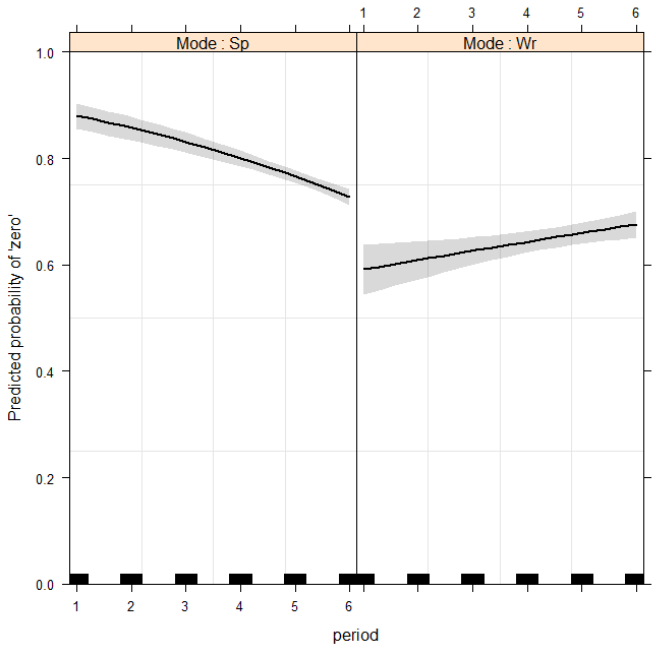

Figure 14: Period : Mode (written versus spoken)

The first effect in a diachronic perspective, presented in Figure 14, is that of mode. The results show that from 1580–2012 the zero form has occurred more frequently in the spoken data than in the written data; however, the predicted probability of the zero form has been steadily decreasing over time from nearly 90 percent to just above 70 percent. The trend in the written data is in the opposite direction, with the predicted probability of the zero form going from just below 60 percent in 1580 to nearly 70 percent by 2009. This means that in PDE (viz. 'Period'), there is still a significant difference in zero form use between the spoken and the written mode, but the predicted probability of the zero form in both modes is similar.

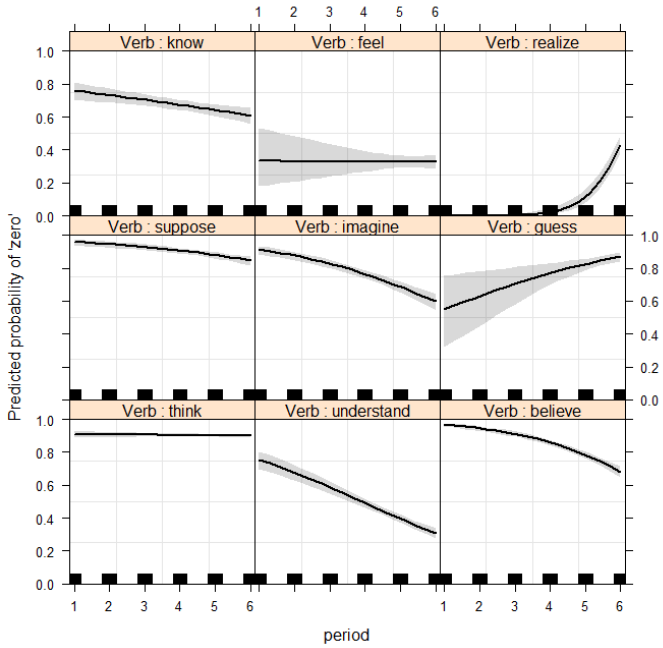

Figure 15: Period : Verb

Figure 15 shows the diachronic development of the zero form for each of the nine verbs. It reveals that *think* appears to have remained largely consistent across time with a high probability of the zero form; however, a closer look at the plot makes it clear that its probability is in fact gradually decreasing over time. The verbs *believe*, *suppose* and *imagine* all start out with roughly the same high probability of the zero form, after which all three verbs exhibit a loss over time. The decrease observed with *suppose*, however, is

minimal; the zero form still remains quite frequent in PDE. However, *believe* and *imagine* show much stronger downward trends ending up in the mid to low 60 percent range. *Know* and *understand* also reveal a consistent decrease of the zero form over time; *know* shows a gentle downward progression to just above 60 percent in Period 6 (i.e. 1920–2012), whereas *understand* drops to almost an inverse ratio from 80/20 percent in Period 1 to 20/80 percent in Period 6. *Guess* and *realize*, however, show consistent increasing trends in the predicted probability of the zero form, with the highest frequencies occurring in Period 6. In spite of this upsurge, the overall frequency of use with *realize* remains low, below the 50 percent threshold, and this phenomenon appears to start in the nineteenth century, although this may be due to the paucity of data in both the spoken and the written data sets for this verb. Finally, as the plot indicates, the use of the zero form with *feel* remains consistently low and steady over time, never breaking above 30 percent.

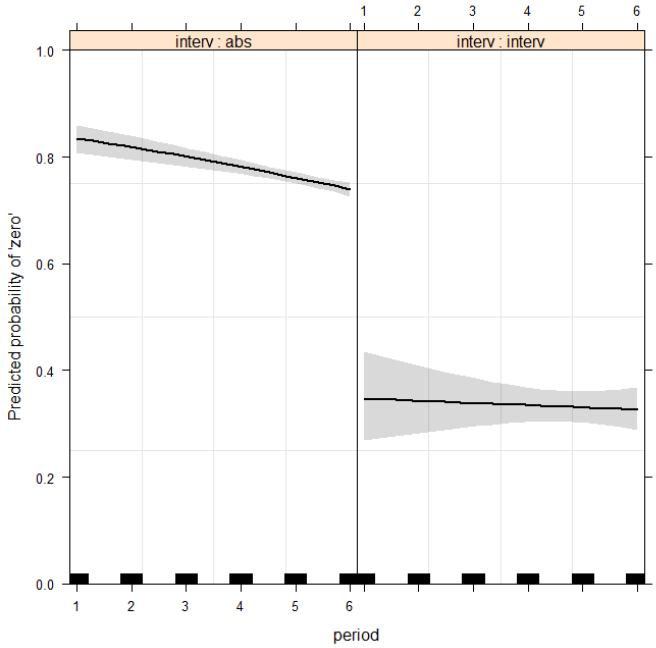

Figure 16: Period : Absence of intervening elements

In Figure 16, we see that the absence of intervening elements shows a decline over time in the zero form, while the evolution for the presence of intervening elements remains at a constant level. Nonetheless, the predicted probability of the zero form in the first case remains much higher than in the second one.

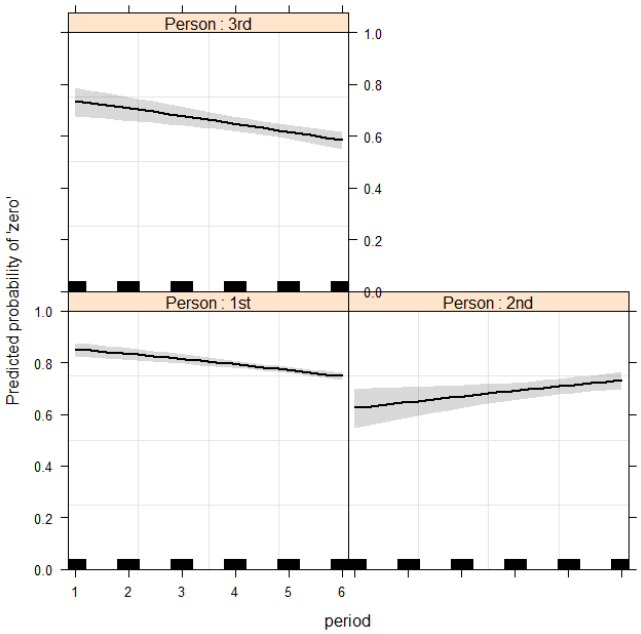

Figure 17: Period : Person

Figure 17 shows the diachronic effect of person. It can be observed that the predicted probability of the zero form declines over time with both first and third person, with third person dropping off more dramatically than first person. By contrast, the second person shows no change over time (the confidence bands demonstrate that the slight increase is not statistically significant). However, it is clear from Figure 17 that in PDE (i.e. Period 6) there is no significant difference anymore between first person and second person. In other words, the predicted probability of the zero form for the first person has converged with the predicted probability of the zero form for the second person in PDE.

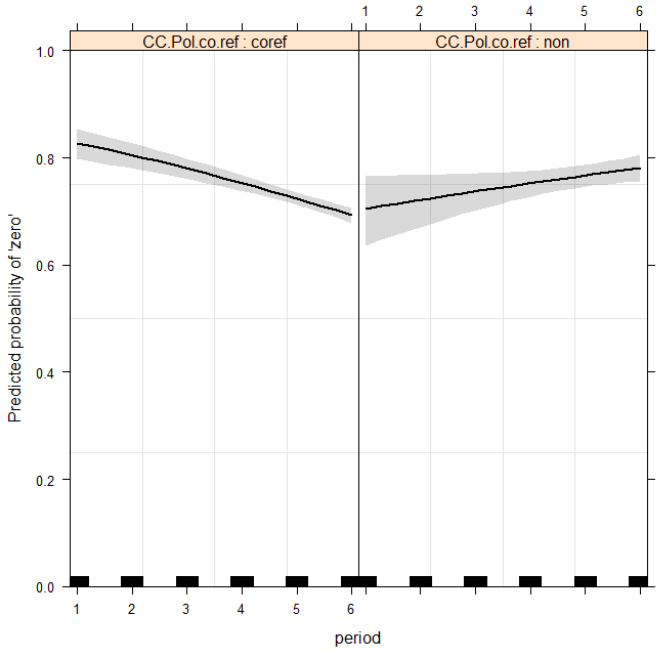

Figure 18: Period : Coreferentiality of polarity

The next effect over time presented in Figure 18 is the interaction of harmony of polarity between matrix and complement clauses and period. The plot shows that in case of harmony of polarity, there is a distinct decrease of the zero form, i.e. a clear tendency towards more *that* over time; however, harmony still remains a predictor and is thus retained within the model. Furthermore, the plot reveals that non-harmonious polarity has an increase of the zero form over time. This results in a situation by Period 6 whereby the non-harmonious constructions actually have a higher predicted probability for the zero form than harmonious constructions – contra expectations (see Section 2.3.3).

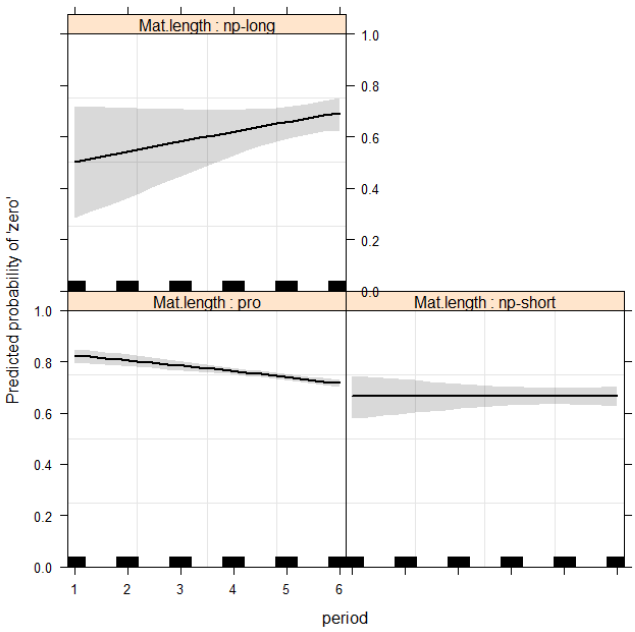

Figure 19: Period : Length of the matrix clause subject

Figure 19 shows the effect of the subject length in the matrix clause from a diachronic perspective. We see that while pronouns initially have the strongest probability of predicting the zero form, they decrease over time to become almost equal in terms of the NP-short forms by Period 6. The NP-short form remains a reasonably strong and consistent factor with little to no change over time. Lastly, we observe a concurrent diachronic increase in the NP-long form to around 70 percent by Period 6 (1920–2012). The results reveal that over time there is a convergence with respect to the length of the matrix clause subject acting as a predictor of the presence of the zero form; however, a pronoun as the subject, thus the shortest form, continues to remain the strongest predictor.

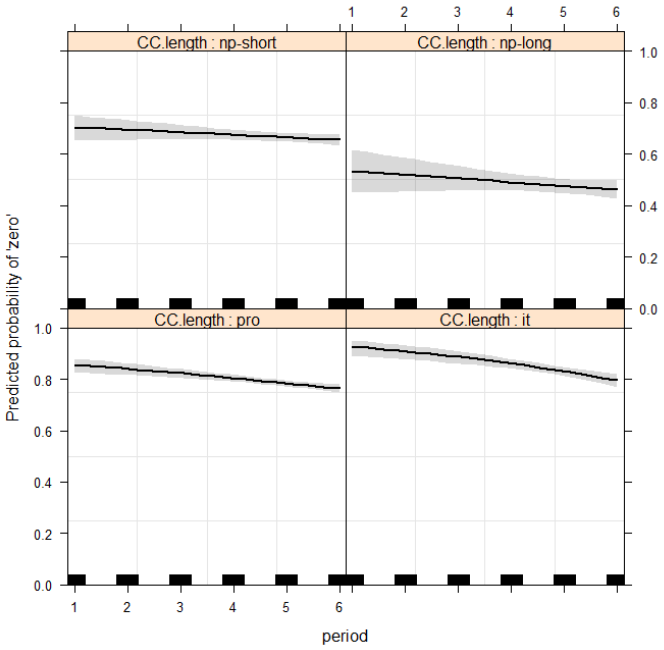

Figure 20: Period : Length of the complement clause subject

The plot for the length of the complement clause subject over time (Figure 20) shows that all four-length categories used to have higher frequencies of the zero form in the past than they do now. In addition, NPs remain consistently below the frequencies of the pronouns (*it* and other pronouns). Pronouns and *it* do show a greater relative decrease over time, but by Period 6 they still have a higher predicted probability than NP-

short and NP-long. Finally, the plot indicates that NP-long has the lowest probability of all four lengths, which remains constant at about 50 percent over time.

5. Conclusion

This study has shown that, contrary to claims in the literature on historical *that*/zero complementizer alternation, there has been an overall diachronic tendency towards more zero complementizer use at the expense of the *that* complementizer. Six of the nine (i.e. 66 percent) of the most frequent complement-taking mental verbs in Present-day English, viz. *think*, *suppose*, *know*, *imagine*, *understand* and *believe*, in fact exhibit a diachronic decrease in the zero complementizer and a concomitant increase in the use of *that*. This trend can be observed for each of the six individual verbs, as the interaction between verb type and period (see Figure 15, Period : Verb) shows. Of the three remaining verbs under investigation, the verb *feel* appears to have been used quite consistently across all six periods with the *that* complementizer, with little to no real increase in the use of the zero form taking place. The verbs *guess* and *realize*, however, show the opposite diachronic pattern and exhibit an increase in the use of the zero form over time; as seen in the Period : Verb plot, *guess* shows a steady increase across all periods, while *realize*, again perhaps due to a paucity of written and spoken data, shows a sharp increase in the use of the zero form from the late nineteenth century onwards.

As for the other effects and interactions tested in this study (see Table 8: Type III LLR tests of 11 main effects and 20 interactions), viz. interactions with verb, mode and period, absence of intervening material is by far the strongest predictor, followed by matrix-internal elements for the zero form. The results for the impact of complement clause subject length confirm Torres Cacoullos and Walker's (2009) findings: *it* most strongly predicts the zero form, followed by other pronouns, short NPs (1 to 2 words) and long NPs (3+). In the spoken data, singular matrix clause subjects are more amenable to zero use than plural subjects and the effect of first person subjects is higher than that of the second person; however, there is no difference between second and third person. In the written data, the length of the matrix clause subjects was not significant for all three forms. In addition, and contrary to expectations drawn from the literature (Bolinger 1972; Torres Cacoullos and Walker 2009), when there is no harmony of polarity, zero is more likely to be selected. Finally, coreferentiality of person is shown to be predicting the zero form, but tense was not significant.

In addition to contradicting 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. Firstly, and as discussed above, of the nine verbs in this study, six exhibit an increase, two a decrease and one shows essentially no meaningful change in the use of *that* over time. Second, this study also showed that the extent to which the factors mentioned in the literature actually predict zero use may differ considerably from verb to verb. One important finding in this regard is the apparent effect of intervening material, or the lack thereof, on predicting the zero form. A strong predictor overall, the lack of intervening material between the matrix and complement clauses is very clear with *know*, *imagine*, *believe*, *suppose*, *think* and *guess*; however, with *understand* and *feel* it clearly is not, as the predicted probability falls below 0.5 and 35 percent, respectively, for these two verbs. This factor has an even smaller predicted probability with *realize*, where it drops to 10 percent.

A second important finding concerns the impact of the mode (spoken versus written data) on the probability of the zero complementizer. The plots show that for the verbs *think* and *suppose*, while the spoken mode has a higher predicted probability than the written mode, both modes are above 80 percent. This pattern is also seen with *believe*, where the spoken mode is above 90 percent and the written mode is above 75 percent. The analysis of *guess*, however, reveals that both modes are not significantly different above 80 percent. The verb *imagine* breaks with this pattern and, as the plots show, the spoken mode is above 80 percent and the written mode falls just below 60 percent as a predictor. The predicted probabilities of both modes are very low with the remaining three verbs, *understand*, *feel* and *realize*, which are all at or below a 50 percent threshold. It should be noted that even in these cases the predicted probability in the spoken mode still remains, even if marginally, higher than in the written mode. The impact of number (singular versus plural with respect to the matrix clause subject) was also informative in that a singular subject has a higher predicted probability for the zero form with the verbs *think*, *believe*, *suppose* and *know*, while the single-plural distinction was not significant with *guess*, *understand*, *imagine*, *feel* and *realize*.

Third, the effect of many factors is also highly dependent on the mode. As mentioned above, the absence of intervening material between matrix and complement clause strongly affects each of the individual verbs, but the interaction with mode also reveals that the written mode is especially susceptible to it. The following

factors are revealed to have higher predicted probabilities of the zero form in the spoken mode: matrix clause person (first person only), matrix clause length (first person only) and complement clause length (both subject pronouns and *it*). Conversely, coreferentiality of polarity favours the *that* form in the written mode only.

Fourth, this study has shown that interactions with period also reveal a number of interesting diachronic trends. First, and foremost, and as we have mentioned in the introduction to this section, was the finding that two thirds of the complement-taking mental state verbs examined in this study – *think, suppose, know, imagine, understand* and *believe* – have shown a diachronic decrease in the zero complementizer and a concomitant increase in the use of *that*. This trend has also been accompanied by evidence that some factors, notably the absence of intervening elements, person and complement clause subject length lose some of their strength as predicting the zero form over time. In addition, Figure 18 (Period : Coreferentiality of polarity) shows that as the predicted probability of coreferential polarity for the zero form decreased over time, this was accompanied by a concomitant increase in the predicted probability of non-coreferential forms – thus negating each other out over time. Finally, as previously discussed with regard to Figure 14 (Period: Mode – written versus spoken), this study has shown that due to changes over time, both the spoken and written modes have the same frequency of the zero form in PDE.

With regard to perspectives for future research, the results of the current study call for a methodologically similar analysis to be carried out in at least three different domains: verb type, genre and register. First, this study only examined mental state verbs. Therefore, by expanding the scope to include verbs from other domains, such as 'locutionary' (*say, tell, ask, answer, mention, remark*), 'cogitation' (*see, get, remember, recognize, learn, notice*), 'appeal' (*hope, wish, pray*) and 'volition' (*accept, admit, agree, assume, doubt*), additional differences should be revealed in the way *that*/zero alternation has evolved with each individual verb, as well as more light should be shed on how the effect of any factor may differ from verb to verb. Secondly, this study examined, in a broad sense, the differences between the spoken and the written language. Future studies should ideally re-examine the potential impact that variables such as formality of context, gender and age of the speaker may play in facilitating *that*/zero alternation patterns. Lastly, the role of register has been examined in the past, but newly available corpus resources and tagging techniques, as discussed in Biber et al. (2015) and Biber and Egbert (2016), could allow greater insight into the roles that registers such as 'instructional', 'how to', 'narrative', 'descriptive', 'informational', 'opinion', 'blog', 'encyclopaedic', 'research focused' play in predicting the use of the zero complementizer form.

References

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

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

Biber, Douglas and Jessie Egbert. 2016. Register variation on the searchable web: a multi-dimensional analysis. *Journal of English Linguistics* 44: 95–137.

Biber, Douglas, Jesse Egbert and Mark Davies. 2015. Exploring the composition of the searchable web: a corpus-based taxonomy of web registers. *Corpora* 10/1: 11–45.

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

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

Bybee, Joan L. 2003. Mechanisms of change in grammaticalization: the role of frequency. In Brian D. Joseph and Richard D. Janda 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.

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

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

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.

Faraway, Julian. 2015. *Linear models with R.* 2nd ed. Boca Raton, FL: Chapman & Hall/CRC.

Finegan, Edward and Douglas Biber. 1985. *That* and zero complementisers in Late Modern English: exploring ARCHER from 1650–1990. In Bas Aarts and Charles F. Meyer 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 Ute Smit, Stefan Dollinger, Julia Hüttner, Gunther Kaltenböck and Ursula Lutzky eds. *Tracing English through time: explorations in language variation*. Vienna: Braumüller, 99–114.

Fox, John. 2003. Effect displays in R for generalized linear models. *Journal of Statistical Software* 32/1: 1–27.

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*. Amsterdam: John Benjamins, 47–76.

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

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

Jespersen, Otto H. 1954. *A modern English grammar on historical principles. Part III: Syntax. Vol. II*. London: George Allen & Unwin.

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

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*. 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: Société Néophilologique.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik. 1985. *A comprehensive grammar of the English language.* London: Longman.

R Core Team. 2018. *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Rissanen, Matti. 1991. On the history of *that*/zero as object clause links in English. In Karin Aijmer and Bengt Altenberg 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: 149–182.

Shank, Christopher, Koen Plevoets and Hubert Cuyckens. 2014. A diachronic corpus-based multivariate analysis of 'I think that' vs. 'I think zero'. In Dylan Glynn and Justyna A. Robinson eds. *Corpus methods for semantics. Quantitative studies in polysemy and synonymy*. Amsterdam: John Benjamins, 279–303.

Tagliamonte, Sali and 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. and 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. and Anthony Mulac. 1991b. A quantitative perspective on the grammaticalization of epistemic parentheticals in English. In Elizabeth Closs Traugott and Bernd Heine eds. *Approaches to grammaticalization. Vol. II*. Amsterdam: John Benjamins, 313–329.

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

Venables, William N. and Brian D. Ripley. 2002. *Modern applied statistics with S*. New York: Springer.

Warner, Anthony R. 1982. *Complementation in Middle English and the methodology of historical syntax. A study of the Wyclifite 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 and Jonathan Culpeper comps.

American National Corpus Project. 2002–2015.

Corpus of Early Modern English Texts. 2005. De Smet, Hendrik comp.
Innsbruck Computer Archive of Machine-Readable English Texts. ICAME. CD-ROM version. 1999. Markus, Manfred comp.
London-Lund Corpus of Spoken English. 1990. Jan Svartvik comp.
Parsed Corpus of Early English Correspondence. Text version. 2006. Nevalainen, Terttu, Helena Raumolin-Brunberg, Jukka Keränen, Minna Nevala, Arja Nurmi, Minna Palander-Collin and Ann Taylor comps.
The British National Corpus, version 3 (BNC XML Edition). 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium.
The Corpus of Contemporary American English. 1990–present. Davies, Mark. (2008–).
The Corpus of Historical American English: 1810–2009. 2010–. Davies, Mark. (2007–).
The Corpus of Late Modern English Texts (extended version). 2006. De Smet, Hendrik comp.
The Lampeter Corpus of Early Modern English Tracts. 1999. Schmied, Josef, Claudia Claridge and Rainer Siemund comps.
The Old Bailey Corpus. Spoken English in the 18th and 19th centuries. 2012. Huber, Magnus, Magnus Nissel, Patrick Maiwald and Bianca Widlitzki comps.
TIME Magazine Corpus. 1920s-2000s. Davies, Mark. (2007–).

Corresponding author
Christopher Shank
School of Languages, Literature and Linguistics
Bangor, Gwynedd LL57 2DG
e-mail: c.shank@bangor.ac.uk

received: July 2018
accepted: December 2018

Appendix

The estimated coefficients of our model together with their standard errors and significance tests are given in Table 7 below:

	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	4.3418	0.2432	17.8520	< 0.0001
Verbunderstand	-0.3589	0.2717	-1.3210	0.1864
Verbbelieve	2.0239	0.2986	6.7780	< 0.0001
Verbsuppose	1.7870	0.3668	4.8720	< 0.0001
Verbimagine	0.7087	0.2973	2.3840	0.0171
Verbguess	-1.4282	0.6321	-2.2590	0.0239
Verbknow	-0.7496	0.2788	-2.6890	0.0072
Verbfeel	-3.2789	0.5832	-5.6220	< 0.0001
Verbrealize	-13.3916	1.3169	-10.1690	< 0.0001
mat.intmat int	-1.0641	0.1140	-9.3320	< 0.0001
CC.lengthit	0.9888	0.2992	3.3040	0.0010
CC.lengthnp-short	-0.9315	0.1711	-5.4440	< 0.0001
CC.lengthnp-long	-1.6460	0.2426	-6.7840	< 0.0001
Person2nd	-1.4780	0.2635	-5.6080	< 0.0001
Person3rd	-0.3511	0.2273	-1.5450	0.1224
intervinterv	-2.3975	0.2838	-8.4480	< 0.0001
ModeWr	-1.5553	0.1966	-7.9130	< 0.0001
tensepast	0.2195	0.1382	1.5880	0.1123
tensen/a	-0.2131	0.1479	-1.4410	0.1496
tensefut	-0.2469	0.7075	-0.3490	0.7271
period	-0.2247	0.0430	-5.2270	< 0.0001
Numberpl	-0.4912	0.1472	-3.3360	0.0008
Mat.lengthnp-short	-1.0867	0.2592	-4.1930	< 0.0001
Mat.lengthnp-long	-1.9497	0.6133	-3.1790	0.0015
CC.Pol.co.refnon	-0.8277	0.2404	-3.4430	0.0006
Verbunderstand:tensepast	-0.6449	0.1628	-3.9610	< 0.0001
Verbbelieve:tensepast	-0.3167	0.2291	-1.3820	0.1669
Verbsuppose:tensepast	0.2153	0.3542	0.6080	0.5433
Verbimagine:tensepast	-0.3372	0.2032	-1.6590	0.0971
Verbguess:tensepast	-1.9016	0.2764	-6.8800	< 0.0001
Verbknow:tensepast	-0.5095	0.1970	-2.5870	0.0097
Verbfeel:tensepast	-0.0628	0.1960	-0.3200	0.7486
Verbrealize:tensepast	0.0430	0.2343	0.1830	0.8544
Verbunderstand:tensen/a	-0.4576	0.1938	-2.3610	0.0182
Verbbelieve:tensen/a	-0.2463	0.2041	-1.2070	0.2275
Verbsuppose:tensen/a	-0.7976	0.2216	-3.5990	0.0003
Verbimagine:tensen/a	0.0324	0.2073	0.1560	0.8758
Verbguess:tensen/a	-0.9906	0.3794	-2.6110	0.0090
Verbknow:tensen/a	-0.9945	0.2126	-4.6770	< 0.0001
Verbfeel:tensen/a	0.1218	0.2540	0.4790	0.6317
Verbrealize:tensen/a	0.3450	0.2419	1.4260	0.1539
Verbunderstand:tensefut	-0.5363	0.7503	-0.7150	0.4748
Verbbelieve:tensefut	-0.7377	0.8731	-0.8450	0.3982

Verbsuppose:tensefut	-1.0447	1.2728	-0.8210	0.4118
Verbimagine:tensefut	-1.6813	0.8793	-1.9120	0.0559
Verbguess:tensefut	-0.2155	1.1707	-0.1840	0.8540
Verbknow:tensefut	-0.5792	0.8825	-0.6560	0.5116
Verbfeel:tensefut	0.6892	0.8983	0.7670	0.4430
Verbrealize:tensefut	0.4492	0.8822	0.5090	0.6106
Verbunderstand:period	-0.3644	0.0486	-7.5030	< 0.0001
Verbbelieve:period	-0.5048	0.0520	-9.7060	< 0.0001
Verbsuppose:period	-0.2667	0.0604	-4.4140	< 0.0001
Verbimagine:period	-0.3686	0.0531	-6.9410	< 0.0001
Verbguess:period	0.3631	0.1066	3.4070	0.0007
Verbknow:period	-0.1194	0.0484	-2.4650	0.0137
Verbfeel:period	0.0172	0.1000	0.1720	0.8634
Verbrealize:period	1.7827	0.2199	8.1080	< 0.0001
ModeWr:period	0.2764	0.0322	8.5820	< 0.0001
Verbunderstand:Person2nd	-0.1345	0.2382	-0.5650	0.5724
Verbbelieve:Person2nd	-0.5002	0.2597	-1.9260	0.0541
Verbsuppose:Person2nd	-0.2307	0.2393	-0.9640	0.3351
Verbimagine:Person2nd	-0.2189	0.2316	-0.9450	0.3445
Verbguess:Person2nd	-0.7596	0.5667	-1.3410	0.1801
Verbknow:Person2nd	0.7329	0.2286	3.2070	0.0013
Verbfeel:Person2nd	0.0134	0.2958	0.0450	0.9640
Verbrealize:Person2nd	0.0546	0.2818	0.1940	0.8465
Verbunderstand:Person3rd	-0.1646	0.1652	-0.9970	0.3189
Verbbelieve:Person3rd	-0.0464	0.1753	-0.2650	0.7912
Verbsuppose:Person3rd	-1.4590	0.2381	-6.1290	< 0.0001
Verbimagine:Person3rd	-0.1832	0.1756	-1.0430	0.2969
Verbguess:Person3rd	-1.0682	0.3029	-3.5260	0.0004
Verbknow:Person3rd	-0.0487	0.1824	-0.2670	0.7895
Verbfeel:Person3rd	0.2190	0.1947	1.1250	0.2605
Verbrealize:Person3rd	0.0909	0.2139	0.4250	0.6707
Verbunderstand:intervinterv	0.4151	0.1989	2.0870	0.0369
Verbbelieve:intervinterv	-0.5168	0.2500	-2.0670	0.0387
Verbsuppose:intervinterv	0.2009	0.2373	0.8470	0.3972
Verbimagine:intervinterv	0.0429	0.2722	0.1570	0.8749
Verbguess:intervinterv	0.1936	0.3236	0.5980	0.5497
Verbknow:intervinterv	0.2948	0.2504	1.1770	0.2391
Verbfeel:intervinterv	1.5147	0.2321	6.5260	< 0.0001
Verbrealize:intervinterv	0.7520	0.3019	2.4910	0.0127
Verbunderstand:ModeWr	-0.3342	0.1536	-2.1760	0.0296
Verbbelieve:ModeWr	-0.3600	0.1618	-2.2260	0.0260
Verbsuppose:ModeWr	-0.1555	0.2017	-0.7710	0.4406
Verbimagine:ModeWr	-0.4619	0.1577	-2.9290	0.0034
Verbguess:ModeWr	0.4571	0.2574	1.7760	0.0758
Verbknow:ModeWr	0.1666	0.1611	1.0340	0.3010
Verbfeel:ModeWr	-0.5782	0.2099	-2.7550	0.0059
Verbrealize:ModeWr	-0.3212	0.2265	-1.4180	0.1561
Verbunderstand:Numberpl	0.4214	0.2081	2.0240	0.0429
Verbbelieve:Numberpl	-0.2041	0.2125	-0.9610	0.3368
Verbsuppose:Numberpl	-0.7095	0.2888	-2.4570	0.0140
Verbimagine:Numberpl	0.2563	0.2365	1.0840	0.2785
Verbguess:Numberpl	-0.2836	0.3492	-0.8120	0.4166
Verbknow:Numberpl	-0.3592	0.2460	-1.4600	0.1442
Verbfeel:Numberpl	0.6801	0.2330	2.9190	0.0035
Verbrealize:Numberpl	0.4548	0.2344	1.9400	0.0524
Verbunderstand:mat.intmat int	0.1658	0.1648	1.0060	0.3144
Verbbelieve:mat.intmat int	-0.2465	0.1774	-1.3890	0.1647
Verbsuppose:mat.intmat int	0.0151	0.2078	0.0730	0.9422
Verbimagine:mat.intmat int	0.2542	0.1786	1.4230	0.1548
Verbguess:mat.intmat int	-0.6615	0.3164	-2.0910	0.0366
Verbknow:mat.intmat int	0.1331	0.2062	0.6460	0.5186
Verbfeel:mat.intmat int	0.5624	0.2026	2.7770	0.0055
Verbrealize:mat.intmat int	0.4940	0.2046	2.4150	0.0157
CC.lengthit:period	-0.1144	0.0548	-2.0870	0.0369
CC.lengthnp-short:period	0.0771	0.0322	2.3930	0.0167
CC.lengthnp-long:period	0.0645	0.0446	1.4460	0.1482
intervinterv:period	0.0980	0.0492	1.9940	0.0461
Verbunderstand:CC.Pol.co.refnon	-0.8437	0.2062	-4.0920	< 0.0001
Verbbelieve:CC.Pol.co.refnon	0.0993	0.2078	0.4780	0.6328
Verbsuppose:CC.Pol.co.refnon	-0.2589	0.2121	-1.2200	0.2223
Verbimagine:CC.Pol.co.refnon	-0.5536	0.2061	-2.6870	0.0072
Verbguess:CC.Pol.co.refnon	-0.2201	0.3614	-0.6090	0.5425
Verbknow:CC.Pol.co.refnon	-0.4514	0.2153	-2.0970	0.0360
Verbfeel:CC.Pol.co.refnon	0.1804	0.2452	0.7360	0.4620
Verbrealize:CC.Pol.co.refnon	-0.2444	0.2346	-1.0420	0.2975
period:CC.Pol.co.refnon	0.2311	0.0412	5.6090	< 0.0001
Person2nd:period	0.2316	0.0432	5.3550	< 0.0001
Person3rd:period	-0.0041	0.0398	-0.1030	0.9178
period:Mat.lengthnp-short	0.1234	0.0469	2.6300	0.0085
period:Mat.lengthnp-long	0.2861	0.1048	2.7290	0.0063
ModeWr:CC.Pol.co.refnon	0.4293	0.1232	3.4860	0.0005
ModeWr:tensepast	-0.1947	0.1141	-1.7070	0.0878
ModeWr:tensen/a	-0.5071	0.1044	-4.8550	< 0.0001
ModeWr:tensefut	0.0828	0.5076	0.1630	0.8704
Person2nd:ModeWr	0.4407	0.1381	3.1920	0.0014
Person3rd:ModeWr	-0.1639	0.1171	-1.3990	0.1617

intervinterv:ModeWr	-0.3626	0.1496	-2.4230	0.0154
CC.lengthit:ModeWr	-0.3521	0.1468	-2.3980	0.0165
CC.lengthnp-short:ModeWr	-0.1790	0.0942	-1.9000	0.0575
CC.lengthnp-long:ModeWr	-0.1895	0.1284	-1.4760	0.1399
ModeWr:Mat.lengthnp-short	0.3239	0.1407	2.3020	0.0214
ModeWr:Mat.lengthnp-long	0.3550	0.2740	1.2960	0.1951

Table 7: Parameter estimates of 11 main effects and 20 interactions

Table 8 below presents the so-called 'Type III tests' for our eleven main effects and twenty 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 '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, the interaction between mode and length of the matrix clause subject is borderline significant. It stays in the model, however, because its removal would lead to a higher AIC-score.

	Df	Deviance	AIC	LRT	Pr(>Chi)
<none>		19045	19315		
Verb	8	19346	19600	301.06	< .0001
mat.int	1	19129	19397	84.205	< .00001
CC.length	3	19133	19397	88.852	< .00001
Person	2	19075	19341	30.304	< .00001
interv	1	19113	19381	68.265	< .00001
Mode	1	19109	19377	64.692	< .00001
tense	3	19052	19316	6.969	0.072901
period	1	19073	19341	28.451	< .00001
Number	1	19055	19323	10.779	0.001026
Mat.length	2	19068	19334	23.866	< .00001
CC.Pol.co.ref	1	19056	19324	11.585	0.000665
Verb:tense	24	19165	19387	119.938	< .00001
Verb:period	8	19313	19567	268.49	< .00001
Mode:period	1	19119	19387	74.894	< .00001
Verb:Person	16	19137	19375	92.591	< .00001
Verb:interv	8	19107	19361	62.59	< .00001
Verb:Mode	8	19078	19332	33.219	< .00001
Verb:Number	8	19085	19339	40.619	< .00001
Verb:mat.int	8	19073	19327	28.563	0.000378
CC.length:period	3	19059	19323	14.762	0.002031
interv:period	1	19049	19317	3.967	0.046394
Verb:CC.Pol.co.ref	8	19080	19334	35.478	< .00001
period:CC.Pol.co.ref	1	19076	19344	31.196	< .00001
Person:period	2	19075	19341	30.846	< .00001
period:Mat.length	2	19057	19323	12.158	0.002291
Mode:CC.Pol.co.ref	1	19057	19325	12.074	0.000511
Mode:tense	3	19069	19333	24.323	< .00001
Person:Mode	2	19060	19326	14.919	0.000576
interv:Mode	1	19051	19319	5.981	0.014457
CC.length:Mode	3	19053	19317	8.006	0.045889
Mode:Mat.length	2	19051	19317	5.977	0.05037

Table 8: Type III LLR tests of 11 main effects and 20 interactions

	believe – Spoken data Total extracted: (n=4,692) With *that*/zero alternation: (n=2,583)				*feel* – Spoken data Total extracted: (n=5,261) With *that*/zero alternation: (n=660)				*guess* – Spoken data Total extracted: (n=3,419) With *that*/zero alternation: (n=1,050)			
	believe – that		*believe* – zero		*feel – that*		*feel* – zero		*guess – that*		*guess* – zero	
	n	N	n	N	n	N	n	N	n	N	n	N
1580–1639	(n=2)	8.93	(n=1)	2.98	(n=0)	0.00	(n=0)	0.00	(n=0)	0.00	(n=0)	0.00
1640–1710	(n=25)	18.35	(n=208)	152.69	(n=0)	0.00	(n=0)	0.00	(n=2)	0.73	(n=4)	2.94
1710–1780	(n=16)	23.11	(n=482)	695.81	(n=3)	0.24	(n=3)	0.24	(n=0)	0.00	(n=13)	1.03
1780–1850	(n=14)	16.79	(n=452)	534.79	(n=27)	4.81	(n=27)	4.84	(n=14)	0.30	(n=23)	0.49
1850–1913	(n=41)	43.80	(n=503)	571.84	(n=55)	8.70	(n=57)	8.78	(n=14)	0.27	(n=51)	0.97
1960–2012	(n=435)	151.85	(n=598)	208.55	(n=314)	160.39	(n=182)	92.74	(n=43)	1.89	(n=886)	396.50
Total	(n=533)		(n=2,050)		(n=401)		(n=269)		(n=73)		(n=977)	

Table 9: Distribution of *that*-clauses and zero complementizer clauses from EModE in the spoken corpora
(n: absolute frequency; N: normalized frequency per million)

	imagine – Spoken data Total extracted: (n=2,624) With *that*/zero alternation: (n=1,309)				*know* – Spoken data Total extracted: (n=5,632) With *that*/zero alternation: (n=831)			
	imagine – that		*imagine* – zero		*know – that*		*know*– zero	
	n	N	n	N	n	N	n	N
1580–1639	(n=9)	26.80	(n=4)	11.91	(n=13)	38.72	(n=14)	41.70
1640–1710	(n=5)	3.67	(n=4)	2.94	(n=57)	73.99	(n=86)	112.93
1710–1780	(n=46)	3.34	(n=536)	40.51	(n=34)	110.44	(n=84)	275.15
1780–1850	(n=45)	1.94	(n=173)	8.32	(n=16)	57.73	(n=90)	313.19
1850–1913	(n=76)	2.80	(n=181)	6.77	(n=77)	277.36	(n=89)	319.76
1960–2012	(n=97)	11.05	(n=127)	14.71	(n=48)	250.19	(n=222)	1104.17
Total	(n=278)		(n=1,025)		(n=245)		(n=586)	

Table 10: Distribution of *that*-clauses and zero complementizer clauses in the spoken corpora (n: absolute frequency; N: normalized frequency per million)

	realize – Spoken data Total extracted: (n=2,696) With *that*/zero alternation: (n=782)				*think* – Spoken data Total extracted: (n=5,525) With *that*/zero alternation: (n=3,550)			
	realize – that		*realize* – zero		*think – that*		*think* – zero	
	n	N	n	n	n	N	n	N
1580–1639	(n=0)	0.00	(n=0)	(n=0)	(n=26)	89.12	(n=121)	394.66
1640–1710	(n=0)	0.00	(n=0)	(n=0)	(n=10)	23.75	(n=212)	447.47
1710–1780	(n=0)	0.00	(n=0)	(n=0)	(n=22)	45.64	(n=412)	854.10
1780–1850	(n=0)	0.00	(n=0)	(n=0)	(n=12)	26.09	(n=439)	938.68
1850–1913	(n=89)	1.67	(n=89)	(n=89)	(n=16)	47.50	(n=418)	1305.45
1960–2012	(n=382)	45.75	(n=382)	(n=382)	(n=226)	389.53	(n=1636)	2868.63
Total	(n=471)		(n=471)	(n=471)	(n=312)		(n=3,238)	

Table 11: Distribution of *that*-clauses and zero complementizer clauses in the spoken corpora (n: absolute frequency; N: normalized frequency per million)

	suppose – Spoken data Total extracted: (n=4,578) With *that*/zero alternation: (n=2,629)				*understand* – Spoken data Total extracted: (n=16,157) With *that*/zero alternation: (n=2,638)			
	suppose – that		*suppose* – zero		*understand – that*		*understand* – zero	
	n	N	n	N	n	N	n	N
1580–1639	(n=4)	11.91	(n=1)	2.98	(n=4)	8.93	(n=1)	2.98
1640–1710	(n=2)	1.47	(n=3)	2.20	(n=11)	8.08	(n=10)	7.34
1710–1780	(n=21)	5.62	(n=451)	125.11	(n=106)	8.42	(n=200)	15.89
1780–1850	(n=28)	10.98	(n=446)	185.05	(n=143)	6.48	(n=303)	13.72
1850–1913	(n=32)	9.27	(n=466)	139.05	(n=613)	33.72	(n=490)	26.96
1960–2012	(n=149)	5.00	(n=1,013)	33.89	(n=524)	71.69	(n=233)	32.70
Total	(n=236)		(n=2,393)		(n=1,401)		(n=1,237)	

Table 12: Distribution of *that*-clauses and zero complementizer clauses in the spoken corpora (n: absolute frequency; N: normalized frequency per million)

	believe – Written data Total extracted: (n=1,706) With *that*/zero alternation: (n=777)				*feel* – Written data Total extracted: (n=4,676) With *that*/zero alternation: (n=606)				*guess* – Written data Total extracted: (n=2,255) With *that*/zero alternation: (n=834)			
	believe – that		*believe* – zero		*feel – that*		*feel* – zero		*guess – that*		*guess*– zero	
	n	N	n	N	n	N	n	N	n	N	n	N
1580–1639	(n=8)	17.85	(n=17)	47.61	(n=0)	0.00	(n=1)	2.98	(n=0)	0.00	(n=0)	0.00
1640–1710	(n=37)	57.36	(n=96)	148.82	(n=0)	0.00	(n=0)	0.85	(n=3)	0.85	(n=16)	4.52
1710–1780	(n=38)	63.56	(n=129)	213.25	(n=24)	6.10	(n=12)	3.05	(n=12)	3.05	(n=5)	1.27
1780–1850	(n=61)	92.73	(n=82)	123.21	(n=258)	87.16	(n=39)	13.23	(n=22)	3.63	(n=38)	6.56
1850–1913	(n=80)	105.58	(n=78)	102.96	(n=156)	103.81	(n=23)	15.31	(n=58)	9.25	(n=99)	15.79
1920–2009	(n=73)	79.57	(n=78)	85.01	(n=52)	80.30	(n=42)	64.63	(n=115)	4.79	(n=466)	33.58
Total	(n=297)		(n=480)		(n=490)		(n=116)		(n=209)		(n=624)	

Table 13: Distribution of *that*-clauses and zero complementizer clauses in the written corpora (n: absolute frequency; N: normalized frequency per million)

	imagine – Written data Total extracted: (n=2,837) With *that*/zero alternation: (n=867)				*know* – Written data Total extracted: (n=5,011) With *that*/zero alternation: (n=975)				*realize* – Written data Total extracted: (n=944) With *that*/zero alternation: (n=460)			
	imagine – that		*imagine* – zero		*know – that*		*know* – zero		*realize – that*		*realize* – zero	
	n	N	n	N	n	N	n	N	n	N	n	N
1580–1639	(n=0)	0.00	(n=1)	2.98	(n=25)	74.39	(n=69)	205.32	(n=0)	0.00	(n=0)	0.00
1640–1710	(n=64)	18.08	(n=75)	21.18	(n=64)	103.87	(n=100)	161.87	(n=0)	0.00	(n=0)	0.00
1710–1780	(n=101)	51.05	(n=96)	48.08	(n=63)	96.72	(n=105)	159.85	(n=0)	0.00	(n=0)	0.00
1780–1850	(n=123)	41.22	(n=57)	19.46	(n=96)	147.53	(n=121)	185.62	(n=0)	0.00	(n=0)	0.00
1850–1913	(n=181)	14.66	(n=113)	9.17	(n=110)	185.61	(n=101)	160.23	(n=298)	59.06	(n=8)	1.57
1920–2009	(n=49)	9.81	(n=56)	11.21	(n=42)	165.59	(n=79)	312.20	(n=105)	72.06	(n=49)	37.91
Total	(n=469)		(n=398)		(n=400)		(n=575)		(n=403)		(n=57)	

Table 14: Distribution of *that*-clauses and zero complementizer clauses in the written corpora (n: absolute frequency; N: normalized frequency per million)

	suppose – Written data Total extracted: (n=1,305) With *that*/zero alternation: (n=714)				*understand* – Written data Total extracted: (n=6,850) With *that*/zero alternation: (n=980)				*think* – Written data Total extracted: (n=6,619) With *that*/zero alternation: (n=2,251)			
	suppose – that		*suppose* – zero		*understand – that*		*understand* – zero		*think – that*		*think* – zero	
	n	N	n	N	n	N	n	N	n	N	n	N
1580–1639	(n=22)	53.56	(n=7)	20.83	(n=43)	127.95	(n=11)	32.73	(n=52)	129.34	(n=248)	616.97
1640–1710	(n=72)	83.54	(n=24)	28.01	(n=109)	30.97	(n=49)	13.28	(n=65)	174.51	(n=200)	554.88
1710–1780	(n=72)	136.88	(n=34	65.51	(n=110)	27.96	(n=38)	9.66	(n=79)	123.19	(n=290)	548.62
1780–1850	(n=65)	100.90	(n=49)	74.78	(n=143)	24.69	(n=40)	6.91	(n=103)	151.66	(n=316)	564.03
1850–1913	(n=134)	135.49	(n=44)	44.57	(n=256)	40.66	(n=34)	5.42	(n=101)	175.47	(n=359)	696.31
1920–2009	(n=162)	81.30	(n=29)	13.65	(n=111)	22.27	(n=36)	7.64	(n=48)	78.93	(n=390)	624.96
Total	(n=527)		(n=187)		(n=772)		(n=208)		(n=448)		(n=1,803)	

Table 15: Distribution of *that*-clauses and zero complementizer clauses in the written corpora (n: absolute frequency; N: normalized frequency per million)