Title: *That*-variation in German and Spanish L2 English

First and corresponding author:

Stefanie Wulff

University of Florida

Linguistics Department

Turlington Hall 4015

Gainesville, FL 32611-5454

swulff@ufl.edu

Second author:

Nicholas Lester

University of California at Santa Barbara

Third author:

Maria T. Martinez-Garcia, University of Kansas

Abstract

In English complement clauses, the complementizer *that* is optional. Previous research has identified various factors that determine when native speakers choose to produce or omit the complementizer, including syntactic weight, clause juncture constraints, and predicate frequency. The present study addresses the question to what extent German and Spanish learners of English as a second language (L2) produce and omit the complementizer under similar conditions. 3,622 instances of English adjectival, object, and subject complement constructions were retrieved from the *International Corpus of English* and the German and Spanish components of the *International Corpus of Learner English*. A logistic regression model suggests that L2 learners' and natives' production is largely governed by the same factors. However, in comparison with native speakers, L2 learners display a lower rate of complementizer omission. They are more impacted by processing-related factors such as complexity and clause juncture, and less sensitive to verb-construction cue validity.

1. Introduction

In English complement clauses, speakers have a choice to either produce the complementizer that, signaling the beginning of the complement clause, or to omit the complementizer. A variety of factors have been argued to impact native speakers' use of *that*, such as frequency, formality, structural complexity, and clause juncture; we review these factors in detail in Section 2. As a consequence, the circumstances driving a native speakers' decision to deploy or omit the complementizer are highly context-specific and complex. At the same time, recent studies suggest that the variable presence of the complementizer appears to be a matter of gradient probabilistic preference rather than discrete grammaticality: omitting or producing *that* never renders an utterance ungrammatical, but depending on the specific context, omitting or producing *that* may render the utterance more or less idiomatic. In consequence, second language learners' non-target-like use of the complementizer becomes a matter of non-idiomatic – much in the sense of Pawley and Syder's (1983) term *nativelike selection* – rather than ungrammatical expression. This may account for the fact that to date, studies of *that*-variation in second language learners are rather scarce, and that this phenomenon receives only limited attention in ESL classrooms: at least from the standpoint of both ESL learners and teachers focusing on grammatically correct L2 production, the cost-benefit ratio of producing native-like *that*-patterns is rather poor.

By the same token, the complex nature of *that*-variation and its absence from most foreign language curricula presents us with a unique opportunity to examine the effects of input, L1 background, and L2 processing undiluted by any potential effects of L2 instruction. Do second language learners, at least at advanced stages of proficiency, use *that* in a native-like fashion, and if they do, what factors shape their performance? Are learners sensitive to the same factors that impact native speakers' choices, and to what extent?

In this study, we examine the factors that determine the optional presence of the complementizer *that* in English subject-, direct object-, and adjectival complements as shown in (1) to (3), and we compare the preferences that native speakers exhibit regarding the variable presence of the complementizer *that* (referred as *that* or zero-*that*[1]) with the preferences of German and Spanish learners of English in the same contexts.[2]

(1) a. The problem is that I don't like it. [subject complement]

b. The problem is ø I don't like it.

(2) a. I thought that you're arriving tomorrow. [direct object complement]

b. I thought ø you're arriving tomorrow.

(3) a. I'm glad that you came. [adjectival complement]

[1] In line with the usage-based construction grammar perspective we adopt here, we use the term zero-*that* to refer to instances in which the *that*-construction is not realized without assuming that there are traces or null elements. However, this theoretical position is inconsequential as far as the present study is concerned.

[2] The complementizer can also be omitted in appositions, relative clauses of *it*-clefts, and with extraposed subjects as shown in (4) to (6):

(4) Your idea, (that) we should buy a present, is good.

(5) It's on Monday (that) my parents arrive.

(6) It's obvious (that) she did it.

For the purpose of the present study, we did not include such instances, but focused on the three most frequent complement constructions in English.

b. I'm glad ø you came.

This study aims to address these questions from a usage-based perspective on language learning in which L2 acquisition is understood as a gradual process towards target-like performance, L2 accuracy at advanced levels of proficiency is best measured as knowledge of conditional probabilities of linguistic features, and language and cognition are viewed as intricately intertwined throughout the acquisition process both in terms of processing and mental representation (Ellis, 2007; Gilquin, 2007; Wulff & Gries, 2011). To this end, we present the results of a quantitative case study of intermediate-advanced Spanish and German ESL learners' use of *that*-variation in written production. German and Spanish differ with regard to complementizer optionality in the three types of complement structures examined here, which renders these two languages an interesting pairing for comparative analysis. In German, the complementizer *dass* can be omitted in subject and direct object complements, but not in adjectival complements; when the complementizer is omitted, the complement clause verb is in post-subject position, whereas it is shifted towards clause-final position when the complementizer is realized. Examples (4) to (6) are translation equivalents of (1) to (3).

(4) a. *Das Problem ist, dass ich es nicht mag.*

the problem COP.3SG.PRS COMP I it NEG like.1SG.PRS

'The problem is that I don't like it'

b. *Das Problem ist, ø ich mag es nicht.*

the problem COP.3SG.PRS COMP I like.1SG.PRS it NEG

'The problem is, I don't like it'

(5) a. *Ich dachte, dass du morgen ankommst.*
I think.1SG.PST COMP you tomorrow arrive.2SG.PRS
‘I thought that you arrive/are arriving tomorrow’

b. *Ich dachte, ø du kommst morgen an.*
I think.1SG.PST COMP you arrive tomorrow PRT
‘I thought you arrive/are arriving tomorrow’

(6) a. *Ich bin froh, dass du da bist.*
I COP.1SG.PRS glad COMP you here COP.2SG.PRS
‘I’m glad that you’re here’

b. *I*ch bin froh, ø du bist da.*
I COP.1SG.PRS glad COMP you COP.2SG.PRS here
‘I’m glad you’re here’

In Spanish, the complementizer *que* is obligatory in all three constructions[3], as illustrated in (7) to (9).

[3] As the Real Academia Espanola (2005: s.v. *que*) notes, the complementizer can in fact be omitted when the subordinate clause is a complement of the direct object of verbs such as *rogar* (‘request’) as in (10a) and *temer* (‘fear’) as in (10b), and even more rarely in direct object complements as defined here with verbs expressing opinions as in (10c). However, according to the Academia, these unusual instances are strongly dispreferred.

(10) a. Le rogué (ø) me permitiera acompañarla hasta la entrada.
’I begged him to allow me to accompany her to the entrance.’

b. Ya me temo (ø) no termine nunca [esta guerra].
’I am afraid this war never ends’

c. El comunicado [...] eriza el cabello y supongo (ø) habrá espantado al ministro Belloch.

(7) a. *El problema es que no me gusta.*
the problem COP.3SG.PRS COMP NEG REFL like.3SG.PRS
'The problem is that I don't like it'

b. **El problema es ø no me gusta.*
the problem COP.3SG.PRS COMP NEG REFL like.3SG.PRS
'The problem is I don't like it'

(8) a. *Pensé que vendríais mañana.*
think.1SG.PST COMP come.2SG.COND tomorrow
'I thought that you'd come tomorrow'

b. **Pensé ø vendríais mañana.*
think.1SG.PST COMP come.2SG.COND tomorrow
'I thought you'd come tomorrow'

(9) a. *Me alegro de que vinieras.*
REFL be.glad.1SG.PRS of COMP come.2SG.SUBJ.IPFV
'I'm glad that you came'

b. **Me alegro de ø vinieras.*
REFL be.glad.1SG.PRS of COMP come.2SG.SUBJ.IPFV
'I'm glad you came'

Accordingly, if we want to assume possible L1 transfer effects, we can expect both German and Spanish learners to be more conservative with regard to omission of the

'The statement ruffles her hair and I guess minister Belloch is scared' complementizer than English native speakers, and Spanish learners to be more conservative in turn than the German learners. To our knowledge, this study is the first to look at L2 learners use of *that* in writing that distinguishes between the three complement constructions and consider the type of complement as a possible predictor.

This paper is structured as follows. First, we briefly summarize previous research in *that*-variation in native and learner language. We then describe how the native speaker and learner data were extracted and coded for statistical analysis. Section 5 presents the results of a logistic regression analysis. We close with a discussion of our main findings and desiderata for future research.

2. Previous studies on *that*-variation

The optionality of *that* in English finite complementation constructions has received much attention from grammarians and linguists alike for at least the last two-and-a-half centuries (cf. Kirkby, 1971[1746], p. 126). The earliest writings on variable *that* were predominantly (if not entirely) prescriptive in orientation and took *that* to be optional and thus subject to preferential constraints based on considerations of, above all, register and genre (see Torres Cacoullos and Walker, 2009 for a review of the early prescriptivist literature). Contemporary efforts have set aside this notion of pure optionality (thanks in large part to Bolinger (1972) – see Thompson and Mulac (1991) on this point), and have instead taken up a more complicated view of what exactly influences a speaker’s choice to include or omit *that*. Though the stylistic influences of register and genre still figure prominently in the literature (e.g. Elsness, 1984), more recent studies have incrementally expanded the discussion to include semantic, discourse/epistemic, and formal complexity/processing factors[4].

Despite the wealth of studies focusing on native English, research on *that*-variation in non-native English is virtually non-existent. A notable exception is Durham (2011), which incorporates several of the above-mentioned factors into a multivariate analysis to compare the distribution of *that* and zero-*that* in native and non-native (French, German, and Italian) English emails.

In the remainder of this section, we survey the major findings of the native and non-native literature, focusing whenever possible on the contributions of quantitative studies. As we shall see, although several of the newly proposed conditioning elements have received empirical support, only a handful of studies (Tagliamonte & Smith, 2005; Torres Cacoullos & Walker, 2009; Jaeger, 2010) have examined them in a multifactorial context. In monofactorial studies, one independent variable (at a time) is investigated without reference to any other (possibly critical) independent variables. Rarely is it the case, however, that one independent variable accounts for *all* variation in a dependent variable. Furthermore, most studies on *that*-variation that have used monofactorial methodologies have examined several variables independently. If more than one turns out to be significant, we should like to know whether all contribute equally, or whether and how they interact, or even whether we are correct in attributing the effect to a certain variable (and not, say, to some other confounding variable). Monofactorial methods cannot provide these answers, and so are susceptible to exaggerated or confounded results. However, we should note that monofactorial tests often do provide useful direction for further

[4] These categories are merely heuristic and do not represent discrete classifications of effect types: in many cases, an effect will cross-cut two or more of them. For instance, pronominal subject NPs serve a discourse function in terms of their high topicality, but also figure into processing accounts in terms of their relative structural simplicity.

investigation. Multifactorial studies, in contrast, examine the relative strength and contribution of each variable vis-a-vis all other possible predictors (that one has decided to include), thereby avoiding the pitfalls of the monofactorial approach while deepening and sharpening the resolution with which we can view the system (Gries, 2003). Thus, although monofactorial results provide important points of departure, when attempting to model a complex system, this type of evidence must be approached with caution.

2.1 Register and stylistic differences

As mentioned earlier, style and register perhaps remain the most convincingly demonstrated predictors of the presence of *that*. In terms of register, the studies that have examined written data (e.g. Elsness, 1984; McDavid, 1964) have reported proportions of *that* from as low as 46% to as high as 87-99%. On the other hand, studies of spoken data (Thompson & Mulac, 1991; Torres Cacoullos & Walker; Tagliamonte & Smith, 2005; Jaeger, 2010) have uniformly found overall usage of *that* below 18% (all but Jaeger [2010] report shares at 10% or below). Clearly, then, *that* use is tied integrally to register, a relationship which is especially marked in the spoken data.

But what about the variability in the distribution of *that* in written language? As it turns out, this internal heterogeneity is at least partially attributable to genre (and by implication, formality; together comprising ‘style’ as used above). Elsness (1984) finds that, as predicted by Storms (1966) inter alia, the more formal scientific genres prefer *that* to the near exclusion of zero, whereas the less formal fiction genres (adventures and westerns) exhibit slightly greater proportions of zero over *that* (58%).

As the above references clearly indicate, there has been a strong bias in the recent literature to examine spoken data, presumably given the well-documented effects of formality and (probably) prescriptivism on *that*-mentioning. We seek to counterbalance the prevailing trend by reexamining written language in light of the wealth of insights to surface since the 'spoken-turn' in *that*-variation research.

2.2 Complement Structure

Even though *that*-variation has received much attention over the years, no study has examined the relative preferences of different complement constructions to occur with *that* or zero. In fact, the majority of studies do not even explicitly characterize what types of constructions were considered besides stating that all allowed for optional *that* (Elsness, 1984; Thompson & Mulac, 1991; Tagliamonte & Smith, 2005, Jaeger, 2010). Torres Cacoullos and Walker (2009) do outline the specific structures included – nominal object clauses, predicate adjective forms, extraposed -subject (as in, *It's sad that…*), and extraposed *it*-clauses (as in, *It seems to me that…*). However, none of these studies has differentiated between what ultimately constitute very different structures.

To help close this gap in the literature, we include a predictor classifying the complementation type into three groups: direct object, predicate-adjective, and subject complements.

2.3 Matrix Clause Properties

Matrix Verb. Early proposals for possible lexical constraints on variable *that* centered on the tendencies of particular matrix verb types to occur with or without the complementizer, resulting in catalogs of *that*-favoring and zero-favoring verbs (Ellinger, 1933; Fowler, 1965; Jespersen, 1954, Poutsma, 1929). While recognition of the importance of matrix verb types is nearly categorical in the more recent literature (e.g., Elsness, 1984; Thompson & Mulac, 1991; Torres Cacoullos & Walker, 2009; Jaeger, 2010), the focus has largely shifted away from the inclusivity of the catalog-style approach, with many studies analyzing only a few privileged high-frequency lemmas (Thompson & Mulac, 1991; Torres Cacoullos & Walker, 2009) or considering absolute type frequency irrespective of the distributions of individual lemmas. This practice finds its source in the Thompson and Mulac's observation that several of the high frequency subject-verb pairings (e.g., *I think, I/you know, I guess*) seem to have grammaticized as fixed epistemic/discourse-functional particles, which no longer operate as embedding predicates and, therefore, appear without the overt complementizer, which would otherwise signal such embedding. Following this line of reasoning, subsequent studies tend to look for purely frequency-based effects (where high frequency predicts zero-*that*; Torres Cacoullos & Walker, 2009) or to contrast those privileged few verb types with the remainder, which get lumped into a rather heterogenous 'other' category (Tagliamonte & Smith, 2005). Jaeger (2010) provides a detailed appendix of all matrix lemmas, including measures of their relative bias towards *that*-mentioning, but does not include these measurements in the final model, opting instead to focus on pure frequency effects.

We attempt to improve and expand these models by including the relative attraction of all matrix verb types to *that*/zero; that is, we account for the frequency *and* distribution of *every* matrix lemma across *that* and zero instances. We measure this association as a function of the probability of occurrence of *that*/zero with a particular lemma and vice versa using the correlation statistic Delta *P* (Ellis, 2007, 2009; Gries, to appear). The primary advantage of this statistic (besides the fact that it provides a relativized image of the effects of frequency) is that it can be applied in both ‘bottom-up’ and ‘top-down’ directions, allowing us to discriminate further between lexically driven associations (in which the matrix verb ‘cues’ the use of *that* or zero) and constructionally driven associations (in which *that* or zero ‘cue’ the use of particular matrix verbs).

Beyond individual verb types, Thompson and Mulac (1991) suggest that entire semantic classes of matrix verbs (more specifically, epistemic verbs) might exhibit preferences to occur with *that* or zero. Dor (2005) similarly argues that truth-claim (matrix) predicates allow zero-*that* on the grounds that zero-marked predicates denote “asserted propositions” (that is, propositions regarding which some epistemic assertion is made regarding their truth validity) made on the part of some “cognitive agent.” These stand in opposition to *that*-marked predicates, which simply signal that a proposition has been made. In the only multivariate analysis to investigate this proposal, however, semantic class was not found to be a significant predictor in either direction (Torres Cacoullos & Walker, 2009), and so it is not included in the present study.

Matrix Verb Complex. Matrix verbs exhibiting greater degrees of internal complexity (as a result of e.g., periphrasis, negation, or modals and other auxiliaries) have been found to correlate with higher shares of *that* (Thompson & Mulac, 1991; Torres Cacoullos & Walker, 2009; Tagliamonte & Smith, 2005, Durham, 2011). Discourse/functional accounts have attributed this finding to the lower likelihood that these more complex structures, especially given the relative variability associated with the complexifying elements (consider auxiliaries), will grammaticize as fixed units. Generally, these variable elements are thought to reinforce the verbs’ status as matrix verbs, and therefore to inhibit such predicates from being reanalyzed as epistemic markers. The presence of the complementizer is then justified as overtly signifying the embeddedness of the following clause (Thomson & Mulac, 1991).

Processing accounts, on the other hand, focus on the complexification resulting from the elaboration of the verb complex itself. Under such analyses, the planning involved in producing a complex matrix verb should increase processing load, thereby increasing the likelihood of *that* as a means of smoothing the overall contour of processing intensity over time (that is, the use of *that* provides one with more time to formulate the upcoming clause in the face of an already taxed processor; Jaeger, 2010).

In response to these claims, we include a continuous measure of verbal complexity based on length in characters of the text spanning from matrix subject head to matrix verbal head. Thus, in keeping with the findings of Jaeger (2010), we do not specify *type* of complexity (auxiliary vs. modal vs. negation, etc.), but rather the *magnitude* of complexity. In this way, we are able to measure not only whether longer (i.e. more complex) textual stretches involving matrix verbs correlate with *that*-mentioning, but also whether incremental increases in length (i.e. complexity; here measured in characters including white space[5]) correlate with incremental variation in the probability of *that*-mentioning. We will also be able to account for the possible contribution of any adverbial material that might intercede between, for instance, modal/auxiliary elements and the verbal head (which would otherwise be ignored in a categorical representation of verbal complexity, e.g., that employed by Thompson and Mulac, 1991). The importance of this last point is underscored by the findings of Torres Cacoullos and Walker (2009), who report a significant effect of adverbial material in the pre-verbal scope of the matrix

[5] The only previous study to measure length continuously (and in a multivariate context), Jaeger (2010), used the word as its basic increment. While this makes sense in the spoken register, it is less obvious how well such a measure fits written data. In order to be as conservative as possible, we selected the minimal written increment, the character, as our measure of length.

clause *only* when it occurs between the matrix subject and verb (though it is unclear whether their ‘post-subject’ measurement included instances such as *The President has recently agreed that his tax cuts are perhaps too shallow*, in which the adverbial appears to the right of the first element of the verb complex). Jaeger (2010) also found a significant effect of pre-verbal material (again, it is unclear what constitutes the beginning of the verb), though this figure included *everything* preceding the matrix verb (i.e. matrix subjects and clause initial adverbs). Thus, our study, in attempting to distinguish the potential loci for complexification as conservatively as possible, greatly improves the resolution on complexity effects in the matrix clause.

Intervening Material (MCverb-CConset). Another variable slot within the matrix predicate which has been proposed to affect *that*-mentioning falls between the verbal head and the onset of the complement clause (either *that* or the first word of the complement subject). Elsness (1984) found greater shares of *that* when adverbials of any length/structure interceded between the matrix verb and complement subject. He attributes this finding to a tendency to avoid ambiguity vis-à-vis the attachment of the adverbial clause, which, in the absence of *that*, could modify either the matrix or complement clause (consider, for example, *He suggested __ on Tuesday __ we should see a movie*, where it is unclear whether the ‘suggesting’ or the ‘seeing’ is associated with *Tuesday*).

Tagliamonte & Smith (2005), Durham (2011), and Jaeger (2010) found similar results, but attribute them to the complexity introduced by the intervening material and its effects on processing efficiency. In the former study, the simple fact that intervening material creates a ‘cognitively more complex environment’ was argued to induce the choice to include the optional complementizer, as predicted by the *Complexity Principle* (Rohdenburg 2000) which states that, given the choice between a more tacit and a more explicit grammatical encoding, speakers will tend to opt for the former in less complex environments and the latter in more complex environments. One shortcoming of this account is that, in seeking to generalize across many more concretely established characterizations of the interactions between grammatical structures and processing constraints, it trades off explanatory precision for broader predictive accuracy. Jaeger (2010), on the other hand, appeals to more explicit models of processing such as *Domain Minimization* (Hawkins 2004; subsumed under the *Complexity Principle*), which predicts that optional elements will be included when they increase the efficiency of processing by minimizing the amount of material necessary for the successful construction of syntactic categories. By this logic, *Domain Minimization* predicts that, in the absence of *that*, any material intervening between the matrix clause and complement clause will obscure the overall complement structure, leading speakers to minimize the potential for parsing difficulty by including *that*.

Again, in order to accommodate the competing analyses put forth in the previous literature, we include a continuous variable to measure the possible effects of intervening material. Using this approach, we will be able to discriminate between general effects of discontinuity (presence vs. absence, operationalized here in the contrast between intervening length =0 and intervening length >0) and more fine-grained incremental effects of increasing complexity.

Matrix Subject. Matrix subjects have held a position of special interest in research on *that*-variation since Thompson & Mulac (1991) observed that first- and second-person-singular matrix subjects correlate strongly (rates of 90% +) with zero-*that.* They suggested that these subjects (along with high-frequency epistemic matrix verbs) are most likely to promote reanalysis as epistemic particles distinct in function from other complement-taking predicates (and thus less likely to occur with *that*). Since then, the majority of studies on *that* have included some measure of the matrix subject, often as part of a collocation involving the matrix verb, with fairly consistent results (Tagliamonte & Smith, 2005; Torres Cacoullos & Walker, 2009; Jaeger, 2010; Durham, 2011). On the whole, *I* and *you* have indeed been found to correlate with greater shares of zero. Studies differ, however, with respect to what other subject types, if any, pattern with or in opposition to *I* and *you*. Tagliamonte & Smith (2005), after excluding tokens of the epistemic pairings identified in Thompson and Mulac (1991), found oppositions between *I* (strong preference for zero), other pronouns (slight preference for *that*), and lexical NPs (strong preference for *that*). Torres Cacoullos and Walker (2009) likewise excluded instances of high-frequency collocations, but only found an opposition between pronominal and full lexical matrix NPs. This leads them to conclude that previous effects for *I* were an artifact of the inclusion of formulaic uses of *I*/*you* + verb, which in fact pattern with the broader trends of *that*-mentioning only inasmuch as *I* and *you* are pronominal (p. 26). Jaeger (2010), without excluding the epistemic pairings, found greater shares of zero with *I* and *you*, as predicted by Thompson and Mulac, but found a more general linear correlation between the various levels of complexity of the matrix subject which seems to operate “beyond the potential effects of grammaticalization” (p. 45).

To develop Jaeger’s findings further, we include a single continuous predictor that measures the complexity of the matrix subject NP as a function of number of characters (as opposed to merely coding for types of NPs). In so doing, we can provide a finer-grained perspective on the previously mentioned linear correlation (Jaeger, 2010).

Other Clausal Material. A handful of studies have also looked at the potential effects of adverbial or other material preceding the matrix clause, with mixed results. Torres Cacoullos and

Walker (2009: 26) found that adverbials, regardless of internal structure, which appeared in clause-initial position correlated with zero as strongly as the absence of adverbials in the matrix clause. They attribute this finding to the fact that initial adverbials are "more likely to have scope over both the matrix and complement clause as a unitary proposition," and thus do not reinforce the status of the matrix predicate as an epistemic marker (as certain other adverbials do, e.g., post-subject adverbials, leading to higher shares of *that*). Jaeger (2010) found that greater numbers of words before the matrix verb correlate with greater shares of *that*, though this figure included matrix subjects and post-subject adverbials as well.

Again, we attempt to account for the conflicting results by including a continuous variable, which in this case measures the length of any clause-initial material up to the onset of the matrix clause subject. By restricting the scope of this predictor, we improve on Jaeger's model by measuring complexity specifically within the clause-initial slot. Moreover, we go beyond Torres Cacoullos and Walker (2009) by allowing for complexity threshold effects that might otherwise be overlooked in their model – a model which classifies single-word initial adverbials (e.g., *Today*) with clause-length or beyond initial adverbials (e.g., *In the morning, after having my coffee and sitting down to the breakfast table, while munching thoughtfully on toast…*).

2.4 Properties of the complement clause

Complement Subject. Within the complement clause, the greatest amount of attention has been paid to the complement subject. Bolinger (1972) observes that complement subjects that do not inflect for case (e.g., *you*, *it*, and lexical NPs) create the potential for ambiguity as they are consistent (if only temporarily) with several conflicting structural interpretations (for example, simple transitive vs. complement structures, as in *I know you* [*like to collect stamps*]. He proposes, therefore, that potentially ambiguous complement subject NPs should correlate with higher shares of *that*. However, Jaeger (2010) did not find any such effect (p>0.2)[6], a result he explains by pointing out that rarely, given the broader context (i.e., beyond the immediate clause), is the 'potential ambiguity' so severe as to impede comprehension.

Others have taken a broader perspective on the role of complement subject type on *that*-mentioning. Elsness (1984) found that pronominal complement subjects in general correlated with higher shares of zero, an effect which was particularly pronounced when the subject was instantiated by first- or second-person-singular pronouns. He accounts for this finding by introducing the intuitively accessible, albeit vaguely defined concept of 'closeness of the clause juncture,' though the exact mechanics of this proposal remain obscure.[7] Similar effects were observed by Thompson & Mulac (1991), who, however, appeal to the high topicality of pronouns (signaling the prominence of the complement vis-à-vis the matrix clause within the broader discourse) as an indicator that the complementizing status of the matrix predicate is questionable (i.e., that the matrix clause is more likely to be an epistemic marker; see also Tagliamonte & Smith, 2005, whose findings mirror those of Thompson and Mulac). Torres Cacoullos and Walker (2009), on the other hand, find that, of the set of personal pronouns, only *I*

[6] See also Elsness (1984).

[7] One might ask, for instance, exactly what "close" means in relation to two propositions. The spatial metaphor notwithstanding, this account also fails to characterize the strength of association between third person pronouns and zero, though Elsness does suggest that syntactically lighter subjects (including *all* pronouns) might be expected in subject position, following general weight-distributional patterns of English (light>heavy). We would therefore not need the overt signification of clause-boundary offered by *that*. Unfortunately, this perhaps more workable hypothesis has received little if any scrutiny in subsequent research.

strongly prefers zero. What is more, *I* exerts a weaker preference for zero than expletive *it* and *there*, which carry no topical import.

Perhaps the most consistent results have come from processing/complexity-driven accounts of the effects of complement clause subjects on *that*-mentioning.[8] Studies in this vein attribute the effects observed for pronouns (mentioned above) not to topicality or propositional 'closeness,' but to the relative simplicity of pronouns vs. lexical NPs. Rohdenburg (1998) argues that *that* should be preferred with complex complement subjects and dispreferred with simple complement subjects (the *Complexity Principle* – see Section 2.3 above). Tagliamonte and Smith (2005) report findings commensurable with either the complexity or topicality account (pronouns favor zero, other NPs *that*), but cautiously abstain from selecting one explanation over the other. Jaeger (2010) similarly found a highly significant difference in *that*-mentioning between pronominal and lexical complement subjects (p<.0001). Beyond this effect, however, Jaeger (2010) found that longer complement clause subjects correlated positively with shares of *that* (p<0.0001), which suggests that complexity may trump topicality as a determinant of *that*-mentioning in this context.

We follow Jaeger (2010) by including a continuous predictor of complement subject complexity (measured, as with all other length-related variables, in terms of number of characters). In so doing, we can further explore the strength of complexity as a determinant of *that*-mentioning.

Subject Coreferentiality. Complement subject types have not only been considered in isolation, but also in coordination with the matrix subject. Accordingly, at least one study has

[8] In fact, this notion extends back at least as far as Elsness (1984), whose data were unfortunately too sparse to make definitive claims about any structural effects of the complement subject.

reported higher shares of zero when the matrix and complement subject refer to the same entity. Thus, Elsness (1984) reports such a finding, though it is unclear what criteria were used for evaluating coreferentiality (other than that the complement clause subjects must be personal pronouns). These results have not been replicated in either of the multivariate analyses to include corefentiality as a predictor, each of which opted for a different definition of what constitutes 'coreference.' Torres Cacoullos and Walker (2009) defined coreferentiality in the broadest possible terms, counting any instances of identity in semantic reference, whereas Jaeger (2010) narrowed the definition to include only string-identical subjects. In neither case did the predictor reach significance, though Jaeger (2010) indicates that the trend was near-significant and in the predicted direction.

Given the lack of consensus regarding subject coreferentiality, and especially provided the fact that the one study to examine subject coreference in written language returned a significant result, we also include a binary predictor measuring the presence or absence of string-identical subjects in the matrix and complement clauses.

Complement Clause. To our knowledge, only two studies have investigated the effects of the complement clause beyond the complement subject. Torres Cacoullos and Walker (2009) found a weak effect of complement verb transitivity on *that*-mentioning. More specifically, they found that, as the number of arguments (i.e. rightward objects or complements) licensed by the complement verb increases, so too do the shares of *that*. It should be mentioned that this variable was not a significant predictor of *that*-mentioning in high-frequency matrix subject-verb pairings. Jaeger (2010) also found an effect of post-verbal material within the complement clause, though he measured absolute length in words of the entire complement predicate without reference to the status of such words as arguments of the verb. He argues that the significant result is due most likely to the speaker having access to a heuristic estimating the general complexity of the upcoming material, which assists in the decision of whether to produce *that* or not (cf. Wasow, 1997).

If Jaeger (2010) is correct, then such an effect should not be restricted to the complement predicate, but encompass everything following the clause juncture (i.e. the clausal pivot representing the optional-*that* slot). For this reason we include two continuous measures of complement complexity, one measuring post-subject material and one measuring the entirety of the complement clause. We again diverge from Jaeger in measuring these lengths by word character.

2.5 That-variation in non-native speaker writing

As mentioned previously, the only study to our knowledge to address the issue of *that*-variation in non-native speaker English is Durham (2011). Durham (2011) compared *that*-use in the English language emails of L1 French, German, and Italian medical students to a sample of comparable native British English emails on several (though by no means all) of the criteria established in the native speaker literature. Despite the fact that two of the three native languages do not permit a zero-variant (French and Italian), Durham (2011) found that, although the learners exhibited much variation across frequency of finite complement use and shares of zero overall (native – 38%, n=328; French – 26%, n=197; German – 46%, n=87; Italian – 34%, n=292), every L1 background showed convergence with the native speaker trends (however, as we can see, the German learners did exhibit the highest overall shares of zero, as predicted by the fact that German has a comparable zero/*dass* distinction in finite complementation patterns). More specifically, all L1 backgrounds except for French showed a proportionally similar sensitivity to clause-juncture constraints to that of native speakers, and all learners produced lower shares of *that* in conjunction with high-frequency first-person-pronoun + epistemic matrix verb pairings such as *I think* and *I hope*.

In a similar fashion, our study includes an L1 background which permits optional complementizers (German) and one that does not (Spanish). However, we expand on Durham (2011) in several key ways: we consider a greater number of variables shown to be relevant to native speaker *that*-mentioning; we distinguish between three finite complementation patterns; and we use a multivariate analysis capable of relativizing the effect of each predictor against all other predictors, while also capturing any interactions among them.

In line with previous research on that-variation in native speakers, and based on Durham's (2011) findings on French, German, and Italian learners of English as a second language, the present study starts out from the hypotheses that

(i) shares of *that* in the learner data should correlate positively with processing cost as measured in the degree of complexity of the constituents involved, as well as the amount of material added before or between constituents;

(ii) learners' use of *that* should be verb-specific such that there is a systematic correlation between specific verbs and the extent to which they license zero-*that*; more specifically, epistemic matrix verbs should display significantly higher shares of zero-*that* than other verbs;

(iii) German learners should approximate target-like distributions of *that* better than Spanish learners because they benefit from positive transfer of possible *that*-omission in the

German translation equivalents of direct object and subject complements where the Spanish corresponding structures in Spanish generally do not license *that*-omission at all.

3. Methods

3.1 Retrieval of corpus data

Data were extracted for native speakers, German learners of English as a second language, and Spanish learners of British English as a second language in Germany and Spain respectively. The native speaker data were obtained from the written sub-section of the British component of the *International Corpus of English* (ICE-GB); the German and Spanish learner data were obtained from the respective sub-samples of the *International Corpus of Learner English* (G-ICLE and SP-ICLE). The British *International Corpus of English* is a balanced corpus of the British English variety; the written sub-section includes printed and non-printed writing such as student essays, correspondence, press news reports, and fiction. The International Corpus of Learner English comprises non-academic argumentative essays by intermediate-advanced learners of English.

Since the ICE-GB is a syntactically annotated corpus, hits for each of the target complementation patterns were retrieved by collecting all complement clauses attested in the written sub-corpus, then checking manually for true hits of direct object-, predicate-adjective-, and subject-complements. Out of an initial sample of 3,090 candidate hits, 1,379 true hits were identified. As far as the learner data are concerned, a different search procedure had to be adopted because the ICLE corpora are not syntactically annotated. In order to ensure maximally exhaustive retrieval from these two corpora, we first searched for all verbs lemmas attested with any of the three complementation patterns in the ICE-GB; this resulted in lists of candidate sentences that comprised 17,622 hits obtained from G-ICLE and 16,969 hits obtained from SP-ICLE. As with the native sample, these candidate lists were then manually checked for true hits of the three complement types. The final data sample included 1,378 tokens for the native speaker data, 1,349 tokens for the German learner data, and 895 tokens for the Spanish learner data. Figure 1 provides a breakdown of the frequencies of *that*/zero by L1 background and complement type.

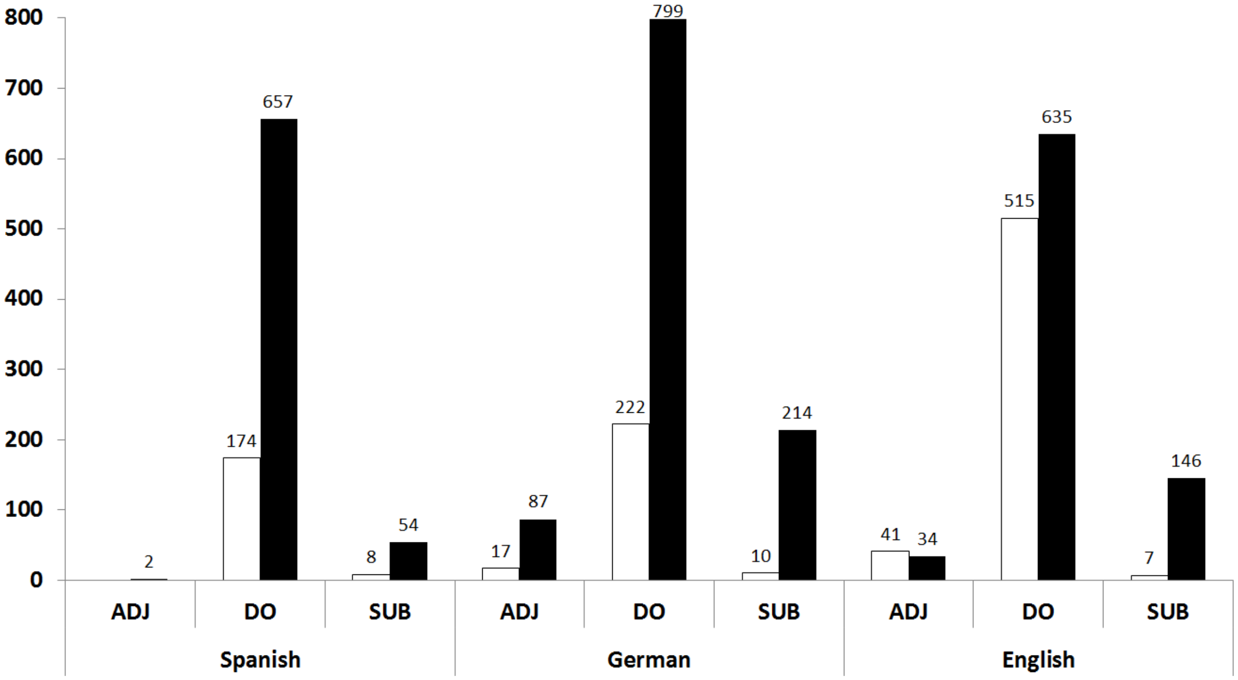

Figure 1. Frequencies of *that*/zero by L1 background and complement type.

When we look at the frequencies of the three complement types in the native speaker data, we can see that direct object complements are by far the most frequent type, with shares of zero being highest as well. In the far less frequent subject complements, zero is extremely rare. In adjectival complements, however, there are more instances of zero than there are instances in which the complementizer is realized (if only by a small margin of 41 instances of zero and 34 instances of *that*). A slightly different picture emerges in the German learner data. While the relative frequencies of direct object, predicate-adjective, and subject complements are about the same as in the native speaker data, we can see here that shares of zero are considerably lower overall, especially in direct object and adjectival complements. In a similar vein, the Spanish learners have a pronounced preference to produce the complementizer across complement types. Moreover, there are only two instances of adjectival complementation in the Spanish data, and also fewer instances of subject complements than in the German learner data.

3.2 Data coding and statistical evaluation

Each attestation was coded with regard to the variable presence of the **Complementizer** (absent/present) as the response variable, as well as for the following predictor variables:

- the **L1** background of the speaker (English/German/Spanish);
- **ComplementType** instaniated (ADJ/DO/SUB);
- the complexity of the matrix subject (**MatrixSubjLength**) as measured in number of word characters[9];
- the complexity of the complement subject (**ComplementSubjLength**) as measured in number of word characters;
- the complexity of the complement clause (**ComplementLength**) as measured in number of word characters;
- the complexity of potential clause-initial material preceding the matrix clause subject (**CIMLength**) as measured in number of word characters;

[9] All length-related predictors were converted to their natural logarithms before statistical analysis.

- the complexity of potential material intervening between the matrix clause subject and the matrix clause verb (**MCSubjMCVerbLength**) as measured in the number of word characters;
- the complexity of potential material intervening between the matrix clause verb and the complement clause (**MCVerbCCLength**) as measured in the number of word characters;
- the cue validity of the matrix clause verb for either *that* or zero (**DeltaPWC**) as measured by a DeltaPWC value;
- the cue validity of *that*/zero for matrix clause verb (**DeltaPCW**) as measured by a DeltaPCW value.

5. Results

The minimal adequate model of a binary logistic regression analysis turned out highly statistically significant (log-likelihood ratio χ^2=1445.19; df=17; *p*=0). [10] Nagelkerke's R^2, an

[10] We computed the logistic regression in *R* using the function stepAIC(model.glm.1, direction=”both”, scope (lower=~1, upper=Complementizer ~ (L1 + ComplementType + ComplementSubjLength + ComplementLength + DeltaPCW + DeltaPWC + CIMLength + MCSubjMCVerbLength + MCVerbCCLength)^2)) for model selection, the function Anova(model.glm, type = “III”, test.statistic = “Wald”) for model comparison, and Stefan Th. Gries' function logregR2s(model.glm.final) for model evaluation.

We tested the final model identified by stepAIC for collinearity issues using the function vif(model.glm.final, and reduced the model further in a stepwise fashion by discarding of predictors/interactions of predictors with variance inflation factors >10, and as long as the classification accuracy of the new model would not be significantly lower than that of the minimal adequate model identified by stepAIC. The final model had slightly higher classification accuracy and nearly identical model quality compared to the minimal adequate model initially identified by stepAIC (classification accuracy minimal adequate model stepAIC=83.11%; Nagelkerke's R^2 minimal adequate model stepAIC=0.503; C=0.879), and does not contain any (interactions of) predictors with variance inflation factors>10.

indicator of general correlational strength, amounts to 0.495. Additionally, the model has good classificatory power (C=0.876). On the basis of the minimal adequate model, 83.29% of all instances can be predicted correctly as either *that* or zero (the random classification accuracy amounts to 72.56%). Table 1 lists all significant predictors of the model (if a predictor was involved in a significant interaction, we list here only the significant interaction; see Appendix A for a complete overview of all predictor levels, standard errors, Wald's *z* scores, and confidence intervals), listing first the main effects and interactions that involved L1 background as a predictor, followed by the main effects and interactions that did not involve L1 background (both groups in descending order of their coefficient values). In the following, we discuss each main effect and interaction in turn in the order provided in Table 1.

Table 1

Significant Predictors of the Minimal Adequate Logistic Regression Model

Predictor	**Coefficient**	***p***
DeltaPWC:L1SP	2.827	0.000
L1SP:ComplementTypeSUB	-1.420	0.017
L1G:ComplementTypeSUB	-1.204	0.038
L1SP:MCVerbCCLength	-0.100	0.011
ComplementSubjLength	0.571	0.000
ComplementLength	0.512	0.000
MCSubjMCVerbLength	0.482	0.000

5.1 Results involving L1 Background

The highest coefficient was yielded by the interaction between L1 background and DeltaPWC. Recall that the higher a given verb's DeltaPWC score (ranging between -1 and 1), the higher the cue validity of that verb for zero-*that*. As Figure 2 illustrates, we see that native speakers (represented by the line interspersed with capital "E"s) behave as predicted such that they are more likely to omit the complementizer when the verb is indeed highly associated with zero-*that*. For the German learners (shown as capital "G"s), we can observe a similar, yet significantly less pronounced trend: they only drop the complementizer with those verbs that are most distinctively associated with zero-*that*. For the Spanish learners (shown as capital "S"s), this "conservative" behavior with regard to zero-that is even more dramatic.

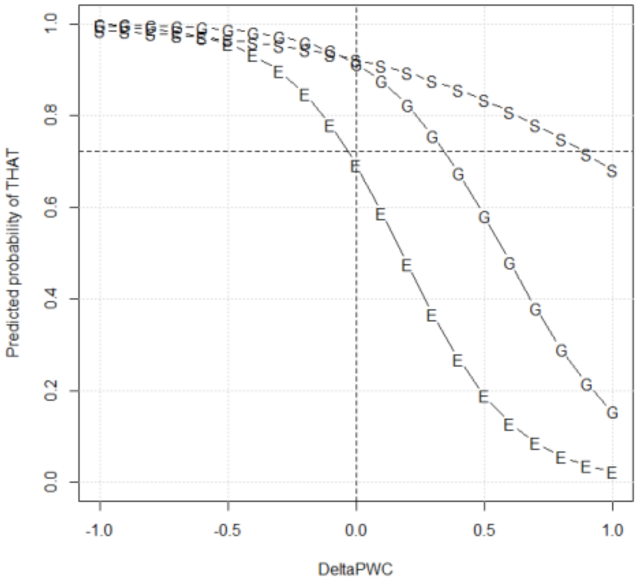

Figure 2. Interaction between DeltaPWC and L1.

A second significant interaction occurred between L1 background and ComplementType. As we can see in Figure 3, all three speaker groups exhibit a tendency to produce the complementizer in subject complements; in direct object complements, however, the learners produce significantly higher shares of that than the native speakers do. With regard to adjectival complements, we see even more diversification by speaker group: native speakers are very likely to omit the complementizer, whereas the German and Spanish learners again adopt a more conservative approach, producing the complementizer with about the same probability as in the overall data. As with the interaction seen in Figure 2, the Spanish learners behave even more conservatively than their German peers.

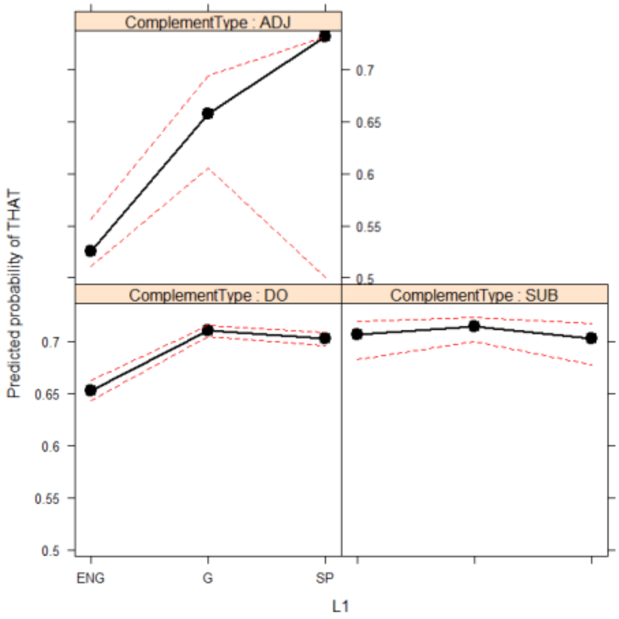

Figure 3. Interaction between L1 and ComplementType.

A third significant interaction was found between L1 background and MCVerbCCLength. As Figure 4 displays, the native speakers exhibit a sensitivity to material intervening between the matrix clause verb and the complement clause such that when there is no material intervening, the predicted probability of that is lower than the overall probability of that (represented by the straight dotted line at y-coordinate 0.72); the probability of that rises as a function of the length of material intervening between verb and complement such that the longer the intervening material, the more likely that is produced. Both learner groups, in contrast, are much more likely to produce that even when no material is present, and so proportions of that do not fluctuate much as a function of MCVerbCCLength.

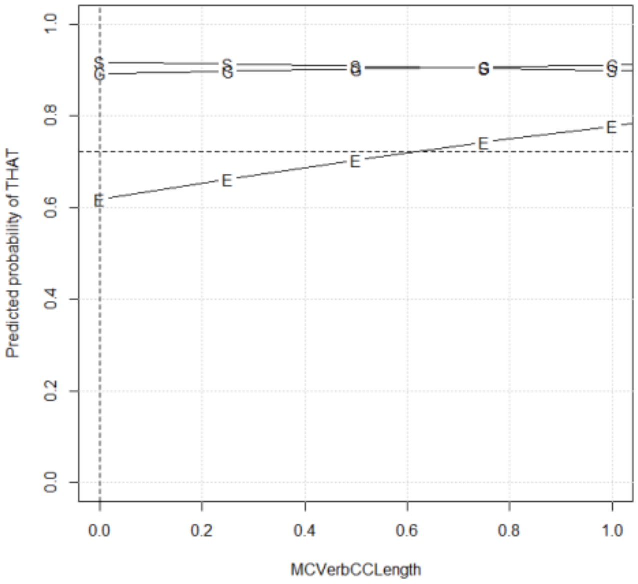

Figure 4. Interaction between L1 and MCVerbCCLength.

5.2 Other Results

Next to the significant interactions involving L1 background, three significant main effects were found to impact that-production across all three speaker groups alike: the longer the complement subjects, the complements at large, and any material intervening between the matrix clause subject and the matric clause verb, the higher the likelihood of *that* being produced. Figures 5 to 7 depict each of these main effects in turn.

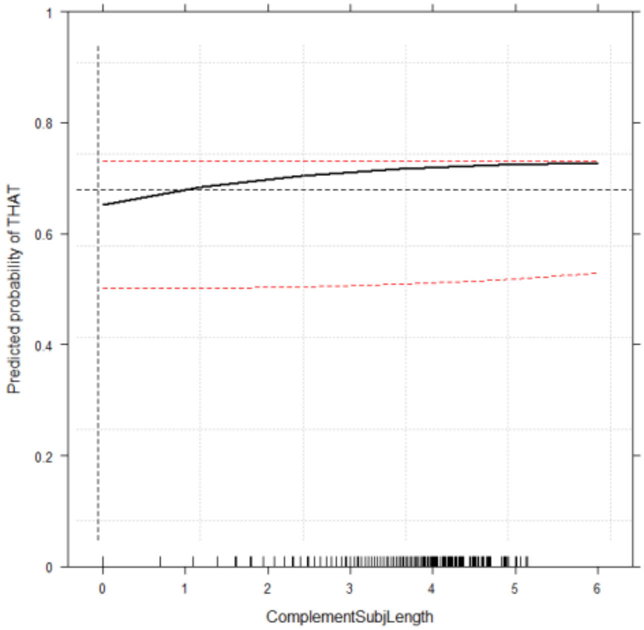

Figure 5. Main effect of ComplementSubjLength.

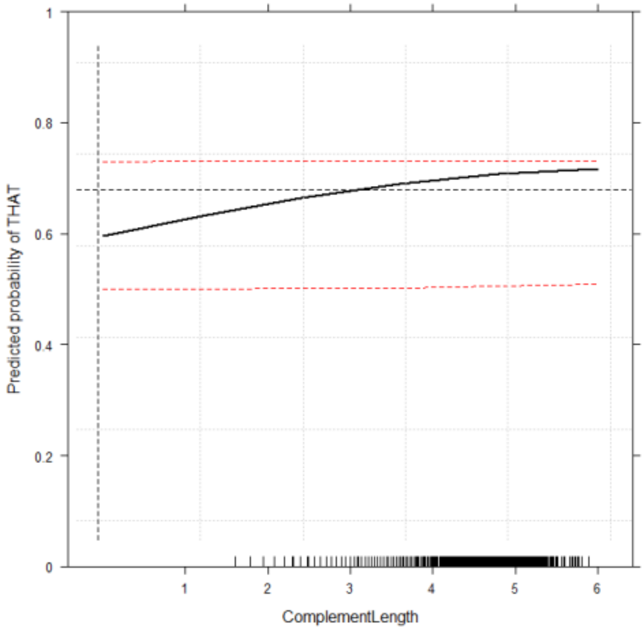

Figure 6. Main effect of ComplementLength.

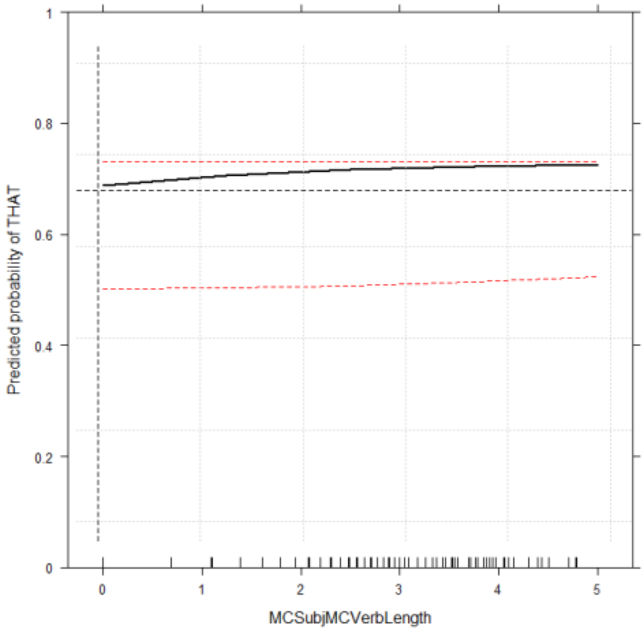

Figure 7. Main effect of MCSubjMCVerbLength.

6. Discussion

In summary, the present study suggests that like native English speakers, English language learners' production of *that* in written discourse is governed by the same array of factors, including complexity, clause juncture, and the cue validity of verb-construction pairings. At the same time, the regression analysis pointed out on which variation parameters learners' approximate target-like behavior more than others. In comparison to native speakers, learners' preferences seem to be more strongly influenced by processing-related factors such as complexity and clause juncture and comparatively less influenced by verb-construction cue validity; overall, learners adopt a more conservative strategy with regard to complementizer omission such that they only drop the complementizer under "safe" circumstances, that is, in contexts that do not entail high processing cost and/or with verbs that are particularly highly associated with zero-*that*.

Returning to the first of the hypotheses laid out in Section 2.5, our results confirm that complex environments correlate positively with higher shares of *that.* From an audience-driven account of language processing, the prevalence of significant complexity/processing-related predictors points to the learners' (perhaps implicit) understanding that omission of *that* has the potential to obscure integration of complex syntactic units (clause-juncture), but crucially not when the complex material is precluded from affecting the relationship between matrix and complement clause (hence the lack of an effect for clause-initial material). This may point not towards a purely syntactic sensitivity, but rather a sensitivity to the preservation of informativeness or communicative efficiency. Alternatively, when adopting a producer-driven account of processing, we could attribute the similarity in patterning across native and non-native writers' use of *that* to the processing difficulties associated with producing embedded structures in increasingly complex environments. From either perspective, our findings provide support for Rohdenburg's Complexity Principle (Section 2.3), in that cognitively more complex environments are attended more frequently by the overt complementizer. Moreover, these effects were observed regardless of L1, suggesting that the learners' emergent grammar is sensitive to the same constraints governing that of the native speakers,

In addition, our results can be seen as an another illustration of the relatively higher cognitive cost associated with deploying one's second (or third or fourth or…) rather than one's native language (Kroll and Dussias, in press; Kroll and Gollan, in press): since cognitive resources are divided, language learners adopt a conservative approach and lean towards producing the complementizer. Future studies could examine if and to what extent *that*-variation in learner language is indeed a function of overall language proficiency, and if native-like *that*-patterns (in the sense of native-like distributions of *that* and zero-*that*) are ever attained by the most advanced language learners.

We also found support for our second hypothesis concerning the role of verb-construction associations. We saw that learners will omit the complementizer only with the verbs most distinctively associated with zero-*that*. On the one hand, this testifies to learners' sensitivity to such associations in the input, which in turn lends credence to recent research from both usage-based and formalist perspectives that argues in favor of a deeper examination of the role of the input (Montrul & Rodríguez Louro, 2006; Rothman, 2009; Rothman & Guijarro-Fuentes, 2010). More specifically, this study supports the hypothesis that a larger than previously assumed number of linguistic properties can be learned from the input not because they are simply frequent, but more adequately because they are particularly frequent *in a specific context*, that is, co-occurring with, or being tied to, the presence of another linguistic property (Saffran, 2003; Ellis, 2012). While we do not wish to imply that all linguistic properties are learnable from the input, and/or that any linguistic property is only learnable if it is distinctively associated with specific contextual parameters, we would like to surmise that *that*-variation is a strong candidate for a linguistic property that is likely to be acquired based on the input. This assumption also gains credibility from the fact that *that*-variation is rarely the subject of explicit language instruction. The more comprehensive college-level TESOL grammar books such as Celce-Murcia and Larsen-Freeman's (1999, pp. 653-654) *The Grammar Book* or Biber et al.'s (2002, pp. 321-322) *Student Grammar of Spoken and Written English*, which both rely on analyses of authentic native speaker corpora in the descriptions of English grammar, provide the reader with some information on the different factors that govern *that*-variation, including lists of specific verb that are trigger zero-*that* like *think* or *know*, the higher share of zero-*that* in informal registers, and increased shares of *that* with co-referential subjects in the main and complement clauses. However, as the results of the present study confirm, the distribution of that is impacted by a complex interplay of these and other factors that is far beyond any one of these rules of thumb. The fact that the intermediate-advanced leaners captured in our data sample indeed approximate this complexly determined distribution of that, therefore, not only supports statistical learning approaches to SLA – what is more, the observed strength of verb-construction interactions in both directions fits nicely with the findings of, e.g., Ellis and Ferreira-Junior regarding the verb-island as a locus of L2 acquisition (Ellis & Ferreira-Junior, 2009a, 2009b).

In a similar vein, while a growing number of studies suggest that explicit (corrective) feedback may indeed be helpful in fostering second language development (Carrol & Swain, 1993; Norris & Ortega, 2000), our findings suggest that explicit feedback is not a necessary condition in the acquisition of all linguistic properties. L2 learners are very unlikely to be corrected on their use of the complementizer in contexts in which it is not required (or even more idiomatic when omitted), and yet the intermediate-advanced learners in our data sample omit *that* in contexts in which native speakers do so as well.

Finally, our findings confirmed our third hypothesis, which predicted that German learners should approximate native English speakers' use of *that* more closely than Spanish learners. The differential performance of Spanish and German speakers provides evidence of negative transfer effects outside of pure grammaticality – transfer, it seems, can be viewed as operative within gradient systems such as preferential mentioning/omission of *that*.

While we were able to elaborate on previous work by Durham (2011), this study presents only a first step towards a more comprehensive understanding of *that*-variation in learner language. Future studies should examine the distribution of *that*-variation across different L1 backgrounds and proficiency levels, as a function of genre and register, and in online experimental settings. So far, our findings support an understanding of L2 acquisition that is based on the input, modulated by L1 background, and constrained by processing demands.

7. References

Diccionario Pnahispánico de Dudas. Real Academia Española (2005). Online at <http://lema.rae.es/dpd/>.

Biber, D., Conrad, S., & Leech, G. (2002). *Longman student grammar of spoken and written English*. London: Longman.

Bolinger, D. (1972). *That's that*. The Hague: Mouton de Gruyter.

Carrol, S., & Swain, M. (1993). Explicit and implicit negative feedback. *Studies in Second Language Acquisition, 15*, 357-386.

Celce-Murcia, M., & Larsen-Freeman, D. (1999). *The grammar book: An ESL/EFL teacher's course*. 2nd ed. Boston, MA: Heinle and Heinle.

Dewaele, J.-M. (2004). Retention or omission of the *ne* in advanced French interlanguage: The variable effect of extralinguistic factors. *Journal of Sociolinguistics*, *8*(3), 433-450.

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

Durham, M. (2011). I think (that) something's missing: Complementizer deletion in nonnative e-mails. *Studies in Second Language Learning and Teaching, 1*(3), 421-445.

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

Ellis, N.C. (2007). The Associative-Cognitive CREED. In B. VanPatten and J. Williams (Eds.), *Theories of second language acquisition: an introduction* (pp. 77-95). Mahwah, NJ: Lawrence Erlbaum Associates.

Ellis, N.C. (2012). What can we count in language, and what counts in language acquisition, cognition, and use? In St.Th. Gries & D.S. Divjak (Eds.), *Frequency effects in language learning and processing* (pp. 7-34). Berlin: de Gruyter Mouton.

Ellis, N.C. & O'Donnell, M. (2012). Statistical construction learning: Does a Zipfian problem space ensure robust language learning? In J. Rebuschat & J. Williams (Eds.), *Statistical learning and language acquisition*. Berlin: Mouton de Gruyter.

Ellis, N.C. & Ferreira-Junior, F. (2009a). Constructions and their acquisition: Islands and the distinctiveness of their occupancy. *Annual Review of Cognitive Linguistics*, *7*, 188-221.

Ellis, N.C. & Ferreira-Junior, F. (2009b). Construction learning as a function of frequency, frequency distribution, and function. *Modern Language Journal*, *93*, 370-385.

Elsness, J. (1984). *That* or zero? A look at the choice of objective clause connective in a corpus of American English. *English Studies*, *65*, 519-33.

Fowler, H. (1965). *A dictionary of modern English usage*. Oxford: Clarendon.

Gilquin, G. (2007). To err is not all: what corpus and elicitation can reveal about thr use of collocation by learners. *Zeitschrift für Anglistik und Amerikanistik*, *55*(30), 273-291.

Gries, St.Th. (2003). *Multifactorial analysis in corpus linguistics: a study of particle placement*. London/New York: Continuum Press.

Hawkins, J.A. (2004). *Efficiency and complexity in grammars*. Oxford: Oxford University Press.

Jaeger, T.F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, *61*, 23-62.

Jespersen, O.H. (1954). *A modern English grammar on historical principles: Part III: Syntax* (second volume). London: George Allen & Unwin.

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

Kroll, J.F., & Dussias, P.E. (in press). The comprehension of words and sentences in two languages. In T. Bhatia & W. Ritchie (Eds.), *The Handbook of Bilingualism and Multilingualism, 2nd Edition*. Malden, MA: Wiley-Blackwell Publishers.

Kroll, J.F., & Gollan, T.H. (in press). Speech planning in two languages: What bilinguals tell us about language production. In V. Ferreira, M. Goldrick, & M. Miozzo (Eds.), *The Oxford handbook of language production*. Oxford: Oxford University Press.

Montrul, S. & Rodríguez Louro, C. (2006). Beyond the syntax of the null subject parameter. A look at the discourse-pragmatic distribution of null and overt subjects by L2 learners of Spanish. In L. Escobar & V. Torrens (Eds.), *The acquisition of syntax in Romance languages* (pp. 401-418). Amsterdam: John Benjamins.

Norris, J.M., & Ortega, L. (2000). Effectiveness of L2 instruction: a research synthesis and quantitative meta-analysis. *Language Learning*, *50*, 417-528.

Pawley, A. & Syder, F. (1983). *Two Puzzles for linguistic theory: nativelike selection and nativelike fluency*. In J. Richards and R. Schmidt (Eds.), *Language and communication* (pp. 191-225). Longman.

Poutsma, H. (1929). *A grammar of late modern English*. Groningen: P. Noordhoff.

Rohdenburg, G. (2000). The complexity principle as a factor determining grammatical variation and change in English. In I. Plag & K.P. Schneider (Eds.), *Language use, language* *acquisition and language history: (Mostly) empirical studies in honour of Rüdiger Zimmermann* (pp. 25-44). Trier: Wissenschaftlicher Verlag.

Rothman, J. (2009). Pragmatic deficits with syntactic consequences: L2 pronominal subjects and the syntax-pragmatics interface. *Journal of Pragmatics*, *41*, 951-973.

Rothman, J., & Guijarro-Fuentes, P. (2010). Input quality matters: Some comments on input type and age-effects in adult SLA. *Applied Linguistics*, *31*(2), 301-306.

Saffran, J.R. (2003). Statistical language learning: mechanisms and constraints. *Current Directions in Psychological Science*, *12*(4), 110-114.

Storms, G. (1966). *That*-clauses in Modern English. *English Studies*, *47*, 249-70.

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

Thompson, S.A., & Mulac, A.J. (1991). The discourse conditions for the use of the complementizer *that* in conversational English. *Journal of Pragmatics*, *15*, 237-51.

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

Wulff, S., & Gries, St.Th. (2011). Corpus-driven methods for assessing accuracy in learner production. In P. Robinson (Ed.), *Second language task complexity: Researching the Cognition Hypothesis of language learning and performance* (pp. 61-88). Amsterdam/Philadelphia: John Benjamins.

Appendix A
Complete output of the minimal adequate logistic regression model (in descending order of absolute coefficient values)

Predictor	**Coeff.**	**S.E.**	**Wald's *z***	***p***
L1SP:ComplementTypeADJ	13.291	228.501	0.058	0.954
DeltaPWC	-4.535	0.300	-15.103	0.000
DeltaPWC:L1SP	2.827	0.447	6.331	0.000
ComplementTypeADJ	-2.693	0.506	-5.322	0.000
L1G	1.707	0.185	9.233	0.000
ComplementTypeSUB	1.432	0.415	3.455	0.001
L1SP:ComplementTypeSUB	-1.420	0.593	-2.393	0.017
L1SP	1.274	0.163	7.794	0.000
L1G:ComplementTypeSUB	-1.204	0.581	-2.072	0.038
L1G:ComplementTypeADJ	1.114	0.719	1.55	0.121
L1SP:MCVerbCCLength	-1.000	0.395	-2.529	0.011
MCVerbCCLength	0.774	0.262	2.952	0.003
ComplementSubjLength	0.571	0.051	11.106	0.000
L1G:MCVerbCCLength	-0.566	0.338	-1.676	0.094
ComplementLength	0.512	0.083	6.14	0.000
DeltaPWC:L1G	0.485	0.487	0.995	0.320
MCSubjMCVerbLength	0.482	0.065	7.413	0.000