Optional *that* in complementation by German and Spanish learners

Stefanie Wulff, University of Florida
Stefan Th. Gries, and Nicholas Lester, University of California, Santa Barbara

Abstract
This study examines the variable presence of the complementizer *that* in English object-, subject-, and adjectival complement constructions as in (1).

(1) a. I thought (that) Nick likes candy.
b. The problem is (that) Nick doesn't like candy.
c. I'm glad (that) Stefan likes candy.

While extensively studied in native speakers, comparatively little is known about the factors governing this alternation in non-native speakers. Apart from determining how similar learners' choices are to those of native speakers, of particular interest here are the questions (i) if and how much learners' choices are subject to individual variation and (ii) the potential role of surprisal for learners' choices. To examine these questions, we revisited a data sample of 9,445 instances from native and intermediate-advanced level learner English (Wulff to appear), including spoken and written corpora (ICE-GB, ICLE, and LINDSEI). The data had been annotated previously for 12 predictors, including the speakers' L1 background; mode; complement type; structural complexity; clause juncture; and the associative bias of the matrix clause verb as either *that*-favoring or zero-favoring (as expressed in a Delta-*P* association measure). For the present study, we expanded the analysis using a regression approach newly-developed for the corpus-based study of learner language or varieties (MuPDAR; Gries & Deshors 2014), and including a mixed-effects modelling approach that also integrates surprisal (Jaeger 2010) and individual variation as predictors.

The results suggest that both learner groups are quite well-aligned with target norms overall, deviating from native speaker behavior more in terms of extent than conditions of *that*-use. For learners, tolerance thresholds for omitting the complementizer appear comparatively lower especially with adjectival and object complements, in complex sentences, when clause juncture is interrupted, and in the written mode.

Key words
That-variation, German L2 English, Spanish L2 English, MuPDAR, mixed effects models

1. Introduction

This study examines the factors that govern the variable presence of the complementizer *that* in English object-, subject-, and adjectival complement constructions as in (1) to (3):[1]

1 The complementizer is optional in other constructions as well, including appositions, relative clauses of *it*-clefts, and with extraposed subjects; instances of these constructions,

(1) a. I thought that Nick likes candy.
b. I thought Ø Nick likes candy.
(2) a. The problem is that Nick doesn't like candy.
b. The problem is Ø Nick doesn't like candy.
(3) a. I'm glad that Stefan likes candy.
b. I'm glad Ø Stefan likes candy.

The conditions under which native speakers (NS) decide to realize or drop the complementizer have been intensively studied (e.g., Thompson & Mulac 1991, Tagliamonte & Smith 2005; Torres Cacoullos & Walker 2009; Jaeger 2010), while few studies have investigated this phenomenon in non-native speakers (NNS) (e.g., Durham 2011; Wulff, Lester, & Martinez-Garcia 2014). In the present study, we therefore address the following research questions:

– What factors govern *that*-variation in intermediate-level German and Spanish L2 learners of English?
– How do these learners' preferences compare to those of native speakers? More specifically, under what conditions, how much, and why do learners deviate from native speaker behavior?

The paper is structured as follows. Section 2 provides a compact overview of the factors suggested to impact *that*-variation; specifically, Section 0 discusses *that*-variation in L1 English whereas Section 2.1 briefly describes the equivalents of *that*-variation in L1 German and L1 Spanish, the native language backgrounds of the L2-learners investigated here. Section 3 gives a brief summary of previous studies on *that*-variation in learner populations. In Section 4, we describe our data sample in detail, explain how the data were annotated for the different variables included in the study, introduce the statistical method employed, MuPDAR, and explain how this method was applied to our data. Section 5 summarizes the results, and Section 6 concludes by recapturing the main findings and their implications, in particular from the perspective of usage-based construction grammar.

2. Factors influencing *that*-variation

That-*variation in native English*

Over the last 25 years, *that*-variation has received a lot of attention. Space does not permit a detailed discussion of this body of research (see Wulff, Lester, & Martinez-Garcia 2014) so here we briefly summarize only those factors which have consistently emerged as relevant:

– *mode* (Bryant 1962; Storms 1966; Biber 1999): the complementizer is omitted more frequently in spoken than in written language; likewise, higher shares of zero-*that* are found in informal registers (both spoken and written).
– *structural complexity* (also referred to as *syntactic weight*; see Elsness 1984; Thompson &

---

which are far less frequent than the three constructions examined here, were not considered in this study.

Mulac 1991; Kaltenböck 2006; Torres Cacoullos & Walker 2009; Jaeger 2010): syntactically light main and/or complement clause subjects as well as light complement clauses are correlated with zero-*that*, and these correlations are strongest with the structurally simple first person pronoun *I* in subject position in the matrix clause.

- *clause juncture* (Thompson & Mulac 1991; Kaltenböck 2006; Torres Cacoullos & Walker 2009; Jaeger 2010): chances of zero-*that* are highest when clause juncture is intact, i.e., when there is no intervening material anywhere. When material intervenes between the matrix clause subject and the verb, the matrix clause verb and the complementizer slot, or the complementizer slot and the ensuing complement clause, this raises the likelihood of *that* being realized. Some studies suggest that material preceding the matrix clause subject may also increase chances of *that* – while clause-initial material does not interrupt clause juncture, it adds to the overall complexity of the message.
- *properties of the matrix clause verb* (Rissanen 1991; Dor 2005; Tagliamonte & Smith 2005; Kaltenböck 2006): several studies point out that zero-*that* is especially likely with (typically highly frequent) matrix clause verbs that denote truth claim predicates (such as *think*, *know*, and *believe*). What is more, Wulff, Lester, & Martinez-Garcia (2014) found that beyond their absolute frequencies, some verbs are zero-favoring while others are *that*-favoring, as can be expressed in the association strength between a given verb and either construction, respectively.
- *surprisal* (Levy 2008, Jaeger 2010): matrix verb lemmas that are biased to occur in the complement clause construction carry enough information about the upcoming clause juncture to make the overt complementizer redundant. This informational boost is quantified using an information-theoretic measure known as surprisal, which Jaeger (2010) shows is positively correlated with rates of *that*-mentioning.
- *individual variation*: just like in many other (psycho)linguistic phenomena, there is individual variation among speakers.

In the next section, we provide a very brief overview of the equivalents of *that*-variation in German and Spanish.

2.1 That-*variation in native German and Spanish*

Regarding complementizer optionality, German is slightly less permissive than English: The complementizer *dass* is optional in subject and direct object complements, but obligatory in adjectival complements. German also differs from English in that the position of the verb in the complement clause is contingent on whether the complementizer is realized or not: When the complementizer is not realized, the verb follows the subject (which is the default word order for main clauses in German); when the complementizer is realized, the verb appears inclause-final position (which is the default word order for subordinate clauses in German). Examples (4) to (6) provide German translations of (1) to (3) respectively.

(4)	a.	*Ich*	*dachte,*	*dass*	*Nick*	*Suesses*	*mag.*
		I	think.3SG.PST	COMP	Nick	candy	like.3SG.PRS
		‘I thought that Nick likes candy’					
	b.	*Ich*	*dachte,*	*Ø*	*Nick*	*mag*	*Suesses.*
		I	think.3SG.PST	COMP	Nick	like.3SG.PRS	candy
		‘I thought Nick likes candy’					

(5) a. *Das Problem ist, dass Nick Suesses nicht mag.*
the problem COP.3SG.PRS COMP Nick candy NEG like.3SG.PRS
'The problem is that Nick doesn't like candy'

b. *Das Problem ist, Ø Nick mag Suesses nicht.*
the problem COP.3SG.PRS COMP Nick like.3SG.PRS NEG candy NEG
'The problem is Nick doesn't like candy'

(6) a. *Ich bin froh, dass Stefan Suesses mag.*
I COP.1SG.PRS glad COMP Stefan candy like.3SG.PRS
'I'm glad that Stefan likes candy'

b. **Ich bin froh, Ø Stefan mag Suesses.*
I COP.1SG.PRS glad COMP Stefan like.3SG.PRS candy
'I'm glad Stefan likes candy'

Spanish, in turn, is even less permissive than German: the complementizer *que* is always obligatory. (7) to (9) are translations of (1) to (3), respectively

(7) a. *Pensé que a Nick le gustaban los dulces.*
think.1SG.PST COMP to Nick CL.DAT. like.3.PL.IMP the candies
'I thought that Nick likes candy'

b. **Pensé Ø a Nick le gustaban los dulces.*
think.1SG.PST COMP to Nick CL.DAT. like.3.PL.IMP the candies
'I thought Nick likes candy'

(8) a. *El problema es que a Nick no le gustan los dulces.*
the problem COP.3SG.PRS COMP to Nick NEG CL.DAT. like.3.PL.IMP the candies
'The problem is that Nick doesn't like candy'

b. **El problema es Ø a Nick no le gustan los dulces.*
the problem COP.3SG.PRS COMP to Nick NEG CL.DAT. like.3.PL.IMP the candies
'The problem is Nick doesn't like candy'

(9) a. *Me alegra que a Stefan le gustan los dulces.*
CL.DAT. makes-happy.3SG.PRS COMP to Stefan

CL.DAT. like.3.PL.PRS the candies
'I'm glad that Stefan likes candy'

b. **Me alegra Ø a Stefan le gustan los dulces.*
CL.DAT. makes-happy.3SG.PRS COMP to Stefan CL.DAT. like.3.PL.PRS the candies
'I'm glad Stefan likes candy'

Given these contrasts between English, German, and Spanish, we can assume that native-like use of *that*-variation should be overall easier to attain for German learners of English than Spanish learners, who should be most reluctant to omit the complementizer. Previous research in fact supports this hypothesis (Wulff, Lester & Martinez-Garcia 2014; Wulff to appear).

3. *That*-variation in L2 production

In contrast to the wealth of studies on native speakers, there are few studies to date that examine *that*-variation in L2 learners. One example is Durham (2011) on native speakers' and French, German, and Italian ESL learners' use of *that*-variation in emails. Durham reports that shares of zero-*that* hover around 35% overall; French and Italian learners are more likely to produce the complementizer than the German learners and native English speakers. Furthermore, Durham confirms that, as in native speakers, combinations of the first person pronoun *I* as the matrix clause subject and verbs like *think* and *hope* trigger the highest shares of zero-*that*. The German and Italian learners display sensitivity also to clause juncture constraints while the French learners do not.

Wulff, Lester and Martinez-Garcia (2014) examine what comprises the written part of the data sample of the present study (i.e., native English speakers, German L2 learners, and Spanish L2 learners). They include all of the factors listed in Section 0 (except for mode, surprisal, and individual variation) in a multifactorial regression analysis. Their findings suggest intermediate-advanced level German and Spanish learners are quite attuned to native-like choices: they appear to be sensitive to the same factors as native speakers, and the directions of the effects for these factors are identical. That said, compared to the native speakers, both learner groups displayed a lower rate of zero-*that*. They also appeared to be more impacted by processing-related factors such as structural complexity and clause juncture as opposed to lexical-semantic properties such as the choice of matrix clause verb.

Wulff (to appear) expands Wulff, Lester, & Martinez-Garcia's (2014) study by adding spoken data to the sample. Her results are mainly in accord with the previous studies, and confirm that, like native speakers, second language learners (at least at an intermediate level of proficiency) are aware of the mode-dependent nature of *that*-variation.

In the present study, we are improving on Wulff's analysis in several important ways. First, the current analysis includes surprisal as a predictor. Second, the statistical analysis presented here is much more sophisticated than the binary logistic regression Wulff (to appear) presents: on the one hand, we are using a two-step regression procedure that has been developed specifically for the analysis of differences between native and non-native language; on the other hand, the regressions we are using involve mixed-effects/multi-level models. This choice of model allows us to take complex hierarchical structures in the data into consideration, such as speaker- and verb-

specific effects. We outline the specifics of this approach in Section 4.3.

4. Methods

4.1 Data

The data for this study were retrieved from different corpora. The NS data were obtained from the British component of the *International Corpus of English* (ICE-GB), a balanced, parsed, 1-million words corpus of British English, which comprises 60% written and 40% spoken data. Using the ICE-CUP software packet that accompanies the corpus, all instances of the three complement constructions that are contained in the corpus were retrieved.

The written NNS data were obtained from the German and the Spanish sub-corpora of the second version of the *International Corpus of Learner English* (G-ICLE and SP-ICLE; see Granger et al. 2009). ICLE comprises 3.7 million words of EFL writing from learners from 16 different L1 backgrounds. The spoken learner data came from the German and Spanish sub-corpora of the LINDSEI corpus (see Gilquin et al. 2010). LINDSEI is a 1-million-word corpus of informal interviews with high intermediate-advanced proficiency EFL learners.

Table 1: Data sample of the present study

L1	Construction	Mode	*that*=absent	*that*=present	Total
English	ADJ	spoken	107	57	164
		written	41	35	76
	OBJ	spoken	2,446	1,235	3,681
		written	528	651	1,179
	SUB	spoken	85	296	381
		written	7	146	153
	Total		3,214	2,420	5,634
German	ADJ	spoken	2	4	6
		written	17	84	101
	OBJ	spoken	643	155	798
		written	224	853	1,077
	SUB	spoken	12	21	33
		written	9	213	222
	Total		907	1330	2,237
Spanish	ADJ	spoken	0	2	2
		written	0	3	3
	OBJ	spoken	437	173	610
		written	176	682	858
	SUB	spoken	4	35	39
		written	8	54	62
	Total		625	949	1,574
Total			4,746	4,699	9,945

Unlike the ICE-GB, neither ICLE nor LINDSEI are syntactically parsed, so in order to retrieve hits from these corpora, the following procedure was adopted: A list of all verb lemmas attested in the ICE-GB across the three constructions was created and used to retrieve all sentences

with these verb lemmas in G-ICLE, SP-ICLE, and LINDSEI. The resulting candidate list was then manually checked for true hits.

Table 1 provides a breakdown of the final data sample of 9,445 hits by L1 background, construction (ADJ vs. OBJ vs. SUB complementation), mode (spoken vs. written), and whether the complementizer was absent or present. Two things stand out immediately when we look at the learner populations: both German and Spanish learners use complementation constructions far less frequently in speaking than in writing (this is especially true for adjectival and subject complementation), which reverses the trend we observe in the native speaker data. Secondly, adjectival complementation is very infrequent in the Spanish learner data.

4.2 Variables and operationalizations

4.2.1 Frequently-used predictors

The 9,445 hits retrieved from the corpora were coded for the factors listed below. In order to understand how each factor was operationalized, let us consider the (fictional) example sentence in (10).

(10) Seriously, I really hope very much that he likes this chocolate.

- **L1 background:** the native language of the speaker: English vs. German vs. Spanish;
- **Mode:** the sub-corpus from which an example came: spoken vs. written;
- **Complementizer**: complementizer presence: absent vs. present;
- **ComplementType**: the type of complement sentence: adjectival vs. object vs. subject;
- **LengthCIM**:[2] the length of any clause-initial material (before the matrix-clause subject) in number of characters;
- **LengthMatrixSubj**: the length of the matrix clause subject;
- **LengthComplementSubj**: the length of the complement clause subject;
- **LengthComplement**: the length of the complement clause;
- **LengthCCRemainder**: the length of any post-verbal material in the complement clause;
- **LengthMCSubjMCVerb**: the amount of material between the matrix clause subject and the matrix clause verb;
- **LengthMCVerbCC**: the amount of material between the matrix clause verb and the complement clause;
- **DeltaPWC/DeltaPCW**: the association of each verb attested in the data sample to *that* or zero-*that* was calculated and vice versa. The specific association measure employed here is a Delta-*P* association measure (using Stefan Th. Gries' *R*-script coll.analysis 3.2; Gries 2007), which involves two different scores: a Delta-P_{WC} value (WC stands for 'word-to-construction') quantifies how predictive the verb is of the absence or presence of *that*, and a Delta-P_{CW} value (CW stands for 'construction-to-word') indicates how predictive the absence or presence of *that* is for the verb in question (see Ellis 2006; Gries 2013). Delta-

2 All length-related predictors were measured as the number of characters. While counting the number of syllables or words might seem more intuitive, for all intents and purposes, length counts in characters or words or phonemes or syllables are so highly correlated (and, thus, come with no conceptual/interpretive disadvantages) that we opted for the ease of operationalizing length with automatically-countable character lengths.

P values range between -1 when the first element strongly repels the second, via 0 (when there is no association), to 1 (when the first element strongly attracts the second).

Consider Table 2 for the annotation of (10):

Table 2: The annotation of example (10)

Complementizer: present	ComplementType: object
LengthCIM: 9 ("Seriously")	LengthMatrixSubj: 1 ("I")
LengthComplementSubj: 2 ("he")	LengthComplement: 20 ("he likes this chocolate")
LengthCCRemainder: 13 ("this chocolate")	
LengthMCSubjMCVerb: 6 ("really")	LengthMCVerbCC: 8 ("very much")
Delta-P_{CW} for *hope*: 0.1148	Delta-P_{WC} for *hope*: 0.167

As previously mentioned, we also included a predictor measuring the surprisal of the material spanning the clause juncture (i.e., the surprisal of moving from *much* to *Nick* in (10)). Given the relative scarcity of such applications in SLA research, we provide a more thorough discussion of this variable in Section 4.2.2. Finally, we added annotation to take into consideration speaker-specific and lexically-specific effects: each example was annotated for the corpus and the file it came from as well as for the verb form and the verb lemma of the main clause.

4.2.2 The information-theoretic notion of surprisal

That-variation has been shown to be affected by various probabilistic relationships between words (and larger units), both within and across the matrix and complement clauses. Jaeger (2010) showed that one particularly important relationship holds between the matrix verb lemma (uninflected stem, e.g., EAT for *eat*, *eats*, *eating*, …) and the syntactic juncture between the matrix and complement clause. When the verb lemma was highly informative about the presence of an upcoming clause juncture, rates of *that* decreased. To measure the expectation of the clause juncture that is projected from the matrix verb lemma (in other words, the *redundancy* of the complementizer), Jaeger used an information-theoretic measure known as *surprisal* or *self-information*. Surprisal measures how uncertain one would be about observing some event – how 'surprising' that event would be – given a known probability distribution of related events. It is calculated by taking the negative binary log of the probability *p* of a given event *x* belonging to probability distribution *P*, as in (11).

(11) $S(x: x \square P) = -\log_2 p(x)$

Because he was interested in the surprisal of the juncture given the matrix verb lemma, Jaeger substituted the conditional probability *p*(juncture | matrix verb lemma) for the simple probability *p*. The generalized form of this substitution, which we shall henceforth refer to as *conditional surprisal* S_c, is

(12) $S_c(y|x: y,x \square P) = -\log_2 p(y|x)$

In the present study, we replace Jaeger's conditional surprisal value with the bi-directional collostructional association measure *P,* and so measure directly the preferences of each matrix verb for the presence or absence of the complementizer (as opposed to the presence or absence of

a complement *clause*). However, the notion of conditional surprisal can be applied at a finer resolution to explore local negotiations of informational load at the clause juncture. For instance, as Jaeger points out, the relative (un)expectedness of the first word following the clause juncture (i.e., the complement clause onset) may influence *that*-mentioning, such that more surprising onsets correlate with greater shares of *that.* Jaeger proposes that, ideally, the surprisal of the onset should be conditioned on the joint probability of the matrix verb occurring in a complement clause construction, that is, S_c(onset | verb, complement construction). However, this measure misses the fact that different verbs are differently associated with rates of the *that*-mentioning apart from their likelihood of occurring within the complement-clause construction (consider the logically possible case of a verb that *only* occurs in the complement-clause construction, but prefers *that*). Moreover, Jaeger's proposal overlooks the possible fluctuations in informational load that can be attributed directly to the words standing at either edge of the clause juncture (the left edge may contain a word other than the matrix verb). The relationships between these words may incrementally or instantaneously overturn (or reinforce) the expectations triggered by the matrix verb. Finally, by taking his measurements at the level of the matrix verb lemma, Jaeger increases the statistical reliability of his estimates, but glosses over the possibility that the different inflected forms of a verb will correlate with different patterns of use.

Therefore, we include among our predictors an additional estimate of conditional surprisal: We take the surprisal of the first word of the complement clause onset conditioned on the last word of the matrix clause prior to the clause juncture, regardless of whether the complementizer separates the words or not. For example, the sequence from (10) *hope* (*that*) *he* would be measured as $S_c(he|hope) = -\log_2 p(he|hope)$, which we operationalize based on data from the complete British National Corpus (World Edition). Thus, we measure how surprising the transition would be if no complementizer had been used, under the assumption that more surprising local transitions will correlate with higher shares of *that*. Importantly, despite the criticisms mentioned above, we do not intend that our measure should be seen as an alternative to the one employed by Jaeger. Rather, we propose that our measure be seen to complement his at a finer granularity.

4.3 Statistical Evaluation: MuPDAR

In order to tease apart how and why the NNS differ from the NS choices of *that*-complementation, we are using an approach called MuPDAR (Multifactorial Prediction and Deviation Analysis using Regressions), which was recently developed in Gries & Deshors (2014) and Gries & Adelman (2014). MuPDAR involves the following three steps:

> fit a regression R_1 that models the choices of speakers of the target language (here, English as operationalized by the ICE-GB) with regard to the phenomenon in question;
> apply the results of R_1 to the other speakers in the data (here, German and Spanish learners of English) to predict for each of their data points what the native speakers of the target language would have done in their situation;
> fit a regression R_2 that explores how the non-native speakers' choices differ from those of the speakers of the target/reference variety.

Crucially, in this study both R_1 and R_2 are mixed-effects models that take into consideration the potential variability that is shared by all examples retrieved from one file and by all examples sharing the same verb (lemma), as will be detailed below.

After preparation of the data (logging several variables and factorizing others, see below),

for R_1, we began with a regression model that predicted *that*-complementation patterns of the NNS on the basis of the following predictors, to which interactions were added as required by likelihood ratio tests: ComplementType, Mode, LengthCIM (factorized into three different levels given the highly skewed distribution of the data), LengthMatrixSubj (factorized into two levels), LengthMCSubjMCVerb (factorized into two levels), LengthMCVerbCC (factorized into two levels), both Delta-*P* values, and (the logged values of) LengthComplementSubj, LengthComplement, and LengthCCRemainder.[3]

We then applied the final version of R_1 to the NNS data and added four columns to them: a column PredictionsNum (the predicted probabilities of a NS using *that* in the situation the NNS is in), PredictionsCat (the dichotomized decision following from PredictionsNum whether a NS would use *that* or not), Correct (whether the NNS made the nativelike choice or not), and, most importantly at present, a column called Deviation. Deviation contains a 0 if the NNS made the nativelike choice, and it contains 0.5-PredictionsNum if the NNS did not make the nativelike choice. That means, Deviation is >0 when the NNS used *that* while the NS wouldn't have, and Deviation is <0 when the NNS did not use *that* while the NS would have.

Finally, we developed a regression model R_2 that tries to predict Deviation, i.e. how nativelike the NNS choices were on the basis of the same predictors as in R_1, but also adding L1 as a predictor that could interact with all others. This last predictor, through interactions, allows us to determine which factors have L1-specific effects. We began with a model involving only main effects, then added interactions of those with L1, then interactions among all predictors (using LR-tests), testing for collinearity at each step and not admitting predictors that would raise variance inflation factors (*VIF*s) to ≥5.1. The final model of R_2 we adopted includes one predictor that was only marginally significant but interesting and was then explored and visualized, as outlined in the next section.

5. Results

5.1 Results of R_1 *on the NS data*

The result of the model selection process for R_1 were encouraging: R_1 featured a variety of highly significant predictors and arrived at a very good classification accuracy: 85.7% of the native speakers' *that* choices were classified correctly, which, according to exact binomial tests, is highly significantly better than either making the more frequent choice all the time ($baseline_1$: 68.5%) or making random choices proportional to the complementation frequencies ($baseline_2$: 56.8%); both p's$<10^{-10}$. The *C*-value for this regression model is 0.91, thus exceeding the typical threshold of

3 While factorizing numeric predictors is typically not recommended given the loss of information it incurs, we nonetheless opted for it here because initial exploratory analyses indicated potentially problematic distributional characteristics for several numeric predictors. For instance, when <10% of all data points of LengthMCSubjMCVerb cover character lengths from 1 to 121, then estimating a regression slope for such a large but sparsely populated range of values is not going to yield reliable results, and a binary factorization of this predictor does not adversely affect the degrees of freedom. Also, note that factorization is a purely methodological choice – it does not reflect particular assumptions of ours regarding the cognitive mechanisms that go into selecting (to omit) a complementizer.

0.8, and the marginal and conditional R^2 are a reassuring 0.48 and 0.59. As for the random-effects structure of the model, we accounted for varying baselines of speakers to use/omit *that* (by including varying intercepts for files in the model) as well as varying preferences of verbs to use/omit *that* (by including varying intercepts for verb forms nested into lemmas in the model).

5.2 Applying R_1 *to the NNS data*

The application of the above regression model to the NNS data also yielded encouraging results: the NS regression model predicted 75.2% of the NNS choices correctly, which again highly significantly (both p's$<10^{-100}$) exceeds both baselines (at 0.5, because the NNS chose to realize *that* nearly half of the time); the *C*-value for this prediction was 0.86.

5.3 Results of R_2 *on the NNS data*

Computing R_2, the model exploring to what degree NNS made nativelike choices, required a few tweaks: because of their high intercorrelations, the two Delta-*P* values as well as LengthCCRemainder and ComplementLength were each combined into a single variable (using principal component scores); the principal component for the Delta-*P*s, however, did not survive the model selection process. As above, we included a simple random-effects structure for files and verbs (forms nested into lemmas). R_2 returned a variety of significant predictors, both main effects and interactions (some pointing to L1-specific effects of the learners, some applying to both learner groups). The overall model R^2-values are less high than those of R_1: marginal and conditional R^2 are 0.13 and 0.3 respectively. Table 3 gives a brief overview of the highest-level predictors in the final model of R_2.

Table 3: Summary results of R_2

Fixed effects predictor	Likelihood ratio test	*p*
LengthCIM	40.103 (*df*=2)	<0.0001
Surprisal	10.434 (*df*=1)	0.0012
ComplementType : LengthComplementSubject	23.902 (*df*=2)	<0.0001
Mode : LengthComplementSubject	18.792 (*df*=1)	<0.0001
Mode : LengthMatrixSubj	19.7 (*df*=2)	<0.0001
ComplementLength/LengthCCRemainder : LengthMatrixSubj	7.531 (*df*=2)	0.0232
L1 : LengthMCSubjMcVerb	8.282	0.004
L1 : LengthComplementSubject	2.896	0.0089 ms

For reasons of space, we can unfortunately not discuss all effects in much detail; here, we will leave out the predictors involving the matrix subject. In our discussion, we will first turn to the main effects (Section 5.3.1), then we will turn to interactions, first those that apply to both learner groups (Section **Error! Reference source not found.**), then the ones that reveal differences between the German and Spanish learners (Section 5.3.3).

5.3.1 Main effects in R_2

Figure 1 shows the main effect of LengthCIM on Deviation: The more material precedes the main clause, the more the NNS make nativelike choices. What are the NS choices? The more material precedes the main clause, the more the NS use *that*, from 29.5% (for none) over 43.6% (for some) to 59.4% (for much). Our results show that the NNS exhibit the same tendency, but with higher proportions of *that*-use throughout: 44.6% over 67.5% to 77.4%. One possible explanation for this pattern is that, as the amount of material before the main clauses grows, both NS and NNS benefit

more from inserting *that* as a structural marker between main clause and complement clause.

Figure 2 shows that, as the first word of the complement clause becomes more surprising given the last word of the main clause, NNS make significantly more nativelike choices. Both NS and NNS increase their complementizer use with higher rates of surprisal, and as before, the NNS just do this with a higher overall baseline of *that*-use. This difference reflects the fact that even what is expected by NS remains rather unexpected to NNS, a likely consequence of their lesser experience with naturalistic English use. Nevertheless, under conditions of high uncertainty, both groups appear to use *that* to smooth spikes in informational load (as reported for NS by Jaeger, 2010).

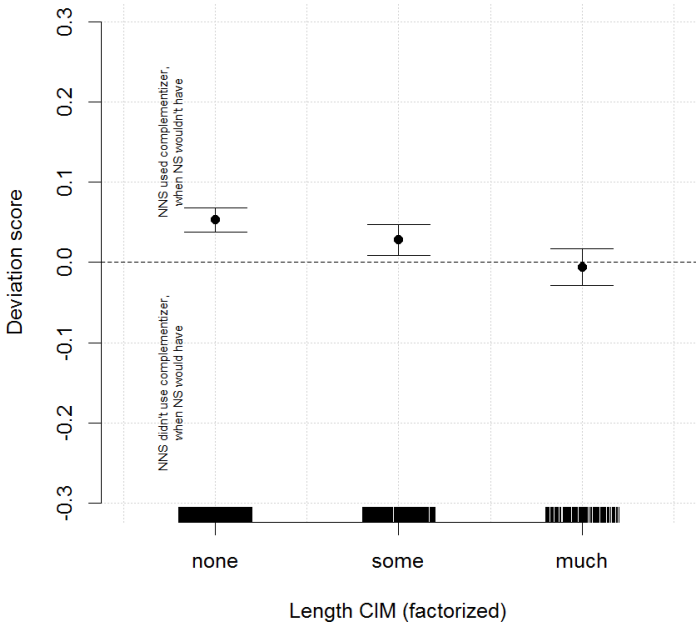

Figure 1: The effect of LengthCIM in R_2

In sum, both overall main effects are compatible with the interpretation that NS and NNS are subject to similar processing pressures and react to them in similar ways even though NNS have a much higher baseline of *that*-use.

5.3.2 Interactions in R_2 that do not involve L1

Figure 3 shows the interaction ComplementType : LengthComplementSubj; the former predictor is represented by three regression lines with the initial letters of the complement types, the latter is represented on the *x*-axis. While the sample size in particular for ComplementType: Adjective is very small, as reflected in the wider confidence band, the corresponding effect in the NS data is that, with increasing length of the subject of the complement clause, speakers use *that* more. The NNS exhibit a similar trend: As the length of the subject of the complement clause increases, they also use *that* more, just like the NS. However, when the subjects of the complement clauses are short, the NNS overuse *that* in adjectival and object complement clauses and are fairly close to NS all the time in subject complement clauses. It is very plausible that this is due to transfer: In Spanish, the complementizer is obligatory in object and adjectival complement clauses, and in German, it is obligatory in adjectival complement clauses. The fact that both NNS L1s require the complementizer in at least one complement construction suggests that functionally specific

transfer could be responsible for the overuse of *that* by our sample of NNS.

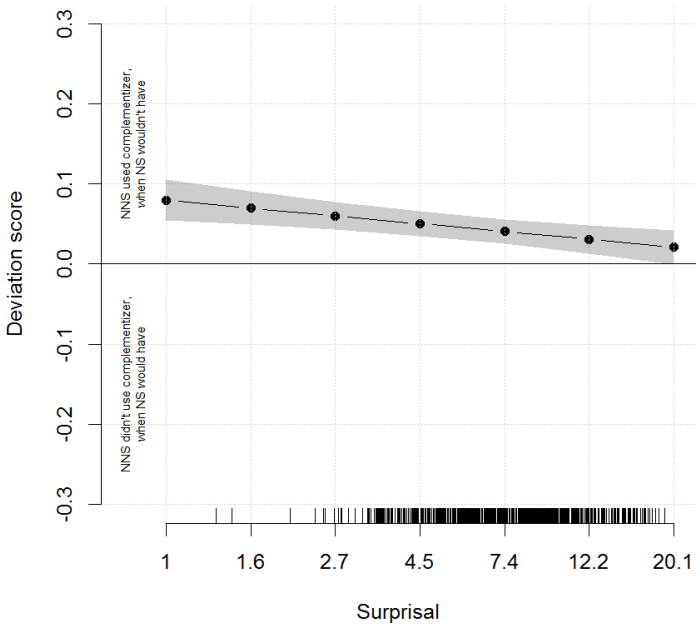

Figure 2: The effect of surprisal in R_2

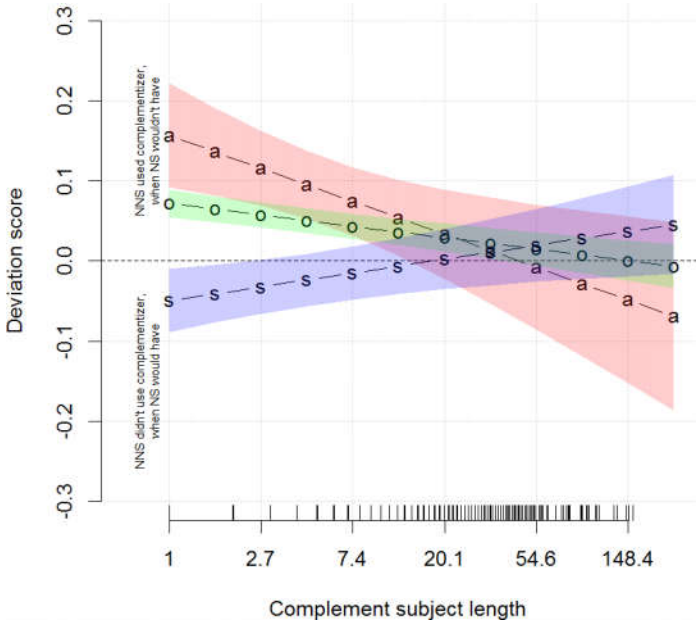

Figure 3: The effect of ComplementType : LengthComplementSubj in R_2

Figure 4 reflects a clear-cut effect. NS use *that* more in writing and less in speaking while the NNS are fairly close to the NS in speaking but still overuse the complementizer regardless of the length of the complement subject. In writing, on the other hand, the NNS are more nativelike with longer subjects, but overuse *that* with short subjects (in particular *I*).

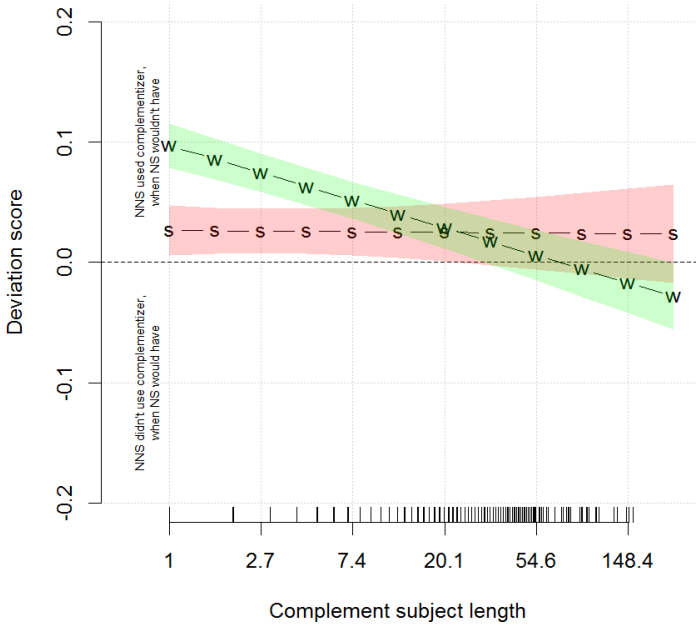

Figure 4: The effect of Mode : LengthComplementSubj in R_2

Both effects show that the length of the complement clause subject is important for all speaker groups and that the learners 'get' the overall preference; however, due to transfer from complementizer use in their L1s and exaggerating the difference between modes, intermediate learners still need to fine-tune their preferences.

5.3.3 Interactions in R_2 that involve L1

Let us finally turn to two interactions that reveal differences between German and Spanish learners. Figure 5 shows how the two learner groups (represented with separate regression lines) react differently to the length of the subject of the complement clause. As discussed above, all speakers – NS and NNS – are more likely to use *that* with longer complement clause subjects. However, the Germans are marginally significantly more similar to the NS with short complement subjects than the Spanish learners, who with short subject overuse *that* more than the Germans.

Finally, Figure 6 shows that, if there is material intervening between the subject and the verb of the main clause, then both German and Spanish speakers behave nativelike and use *that*, but when there is none, then both learner groups overuse *that*, and the Spanish speakers particularly much.

In sum, the German learners produce more nativelike rates of *that*-mentioning than the Spanish learners when it comes to the length effects studied in this section.

Space only permits a brief comment regarding the random-effects structure of the final model of R_2. The largest amount of the variance of the random effects by far was accounted for by the file names, i.e. our proxy for different speakers, namely 12.5%. The second most useful random effect was the verb forms (nested into the verb lemmas), which accounted for an additional 3.5%; verb lemmas contributed an additional 3.1%. While these numbers may not seem high, they point to the need for including such effects for more accurate results than are usually provided in SLA research, and it needs to be borne in mind that our random-effects structure was restricted to varying intercepts only (given data sparsity) – more complex structures might well explain (much) more variability.

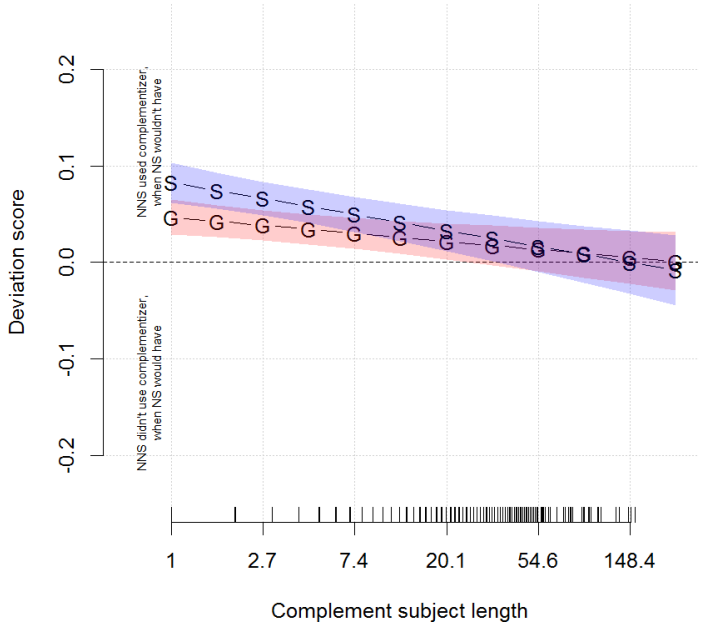

Figure 5: The effect of L1 : LengthComplementSubj in R_2

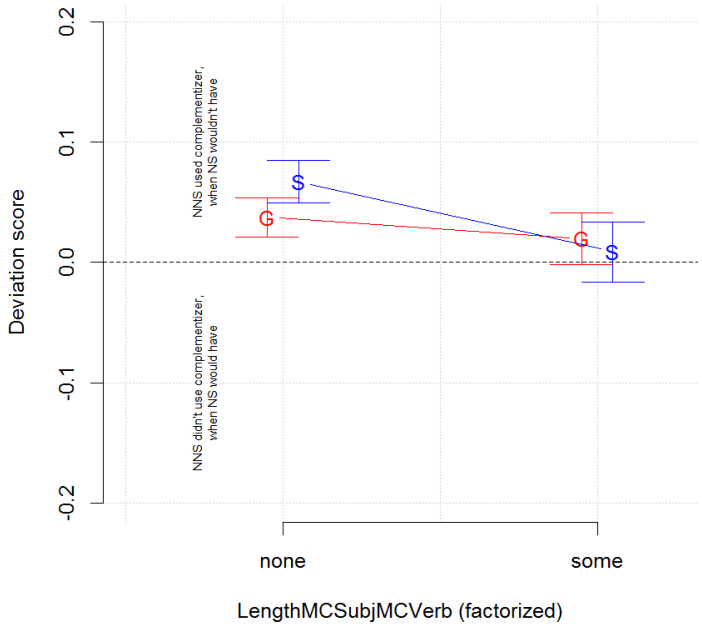

Figure 6: The effect of L1 : LengthMCSubjMCVerb in R_2

6. Discussion

The results of the MuPDAR analysis suggest that the intermediate-advanced German and Spanish learners are quite well aligned with NS norms overall. Minor (yet significant) differences were identified in the second regression: both learners groups employ comparatively higher shares of *that* as the processing demands increase, be it in the form of more material occurring at the onset

of the clause or with longer complement subjects. More pronounced differences between NS and learners become visible when we consider construction-specific uses of *that* – learners overuse the complementizer in adjectival and object constructions – and register-specific uses of *that*: both learners groups overuse the complementizer especially in writing when the main clause subject is *I*. Finally, a few L1-specific differences emerge: the Spanish learners overuse the complementizer more frequently than their German peers do in contexts with sort complement clause subjects and when clause juncture is interrupted.

These findings suggest that the intermediate-advanced learners examined here rely on the same basic mechanisms governing *that*-variation as native speakers, but at the same time display a comparatively more conservative behavior than the native speakers: learners do not realize the complementizer only in what we may call "ideal contexts" associated with low shares of *that* also in NS use, namely in speaking, with short subject and complement clause subjects, and with little or no increased processing costs imposed by optional additional and/or intervening material. When the context is less than ideal, the learners – and the Spanish learners more so than their German peers, arguably reflecting transfer from the L1 – resort to the "safe" strategy of realizing the complementizer as this choice is never, strictly speaking, ungrammatical, if only, at times, non-idiomatic.

Generally speaking, the learner behavior is not fundamentally different from NS behavior; rather, the thresholds for producing the complementizer are significantly lower compared to NS speakers, and they are reactive to the factors mentioned above. This stands in accord with usage-based models of L2 learning such as N. Ellis' Associative-Cognitive CREED model (Ellis & Wulff 2015) or Goldberg's (2006) usage-based Construction Grammar, to name but two examples; these models share the assumption that L2 learning is best characterized as the gradual approximation towards native-like representations. As one reviewer pointed out, questions regarding which specific mechanisms underlie the factors included here – cognitive load, learning as a result of usage, transfer effects, and/or instructional effects --, and how *exactly* each these mechanisms operate in the individual learner – even something as seemingly straightforward as cognitive load can be manifested on different levels of linguistic analysis and can interact with general intelligence, working memory, age, etc. – are beyond the scope of the present analysis, and possibly beyond a purely corpus-based approach. In the following, we can only speculate about the relationship between these factors and the cognitive mechanisms they potentially tap into

Firstly, it is with regard to processing-related factors such as clause complexity and juncture that we see learners in need to further improve their alignments to the target norm.This reminds us of psycholinguistic accounts such as that of Judith Kroll and her colleagues, who argue in favor of a tight link between bilingualism and cognitive cost: according to Kroll & Dussias (2013), speaking a second language entails a higher cognitive load because the speaker constantly has to juggle between the two (or more) languages (Kroll & Dussias 2013). From that perspective, it makes sense that our learners display lower tolerance thresholds for factors that themselves are directly related to cognitive cost, such as complexity or clause juncture: compared to native speakers, the learners have fewer cognitive resources to allot in the first place. As a result, they produce the complementizer more frequently.

In addition, we found that NS and NNS both responded in the expected fashion to spikes in uncertainty (based on Jaeger, 2010) as captured by the conditional surprisal of the first word of the complement clause given the last word of the matrix clause. Both groups were more likely to produce *that* at high-uncertainty transitions. However, NNS also tended to overproduce *that* at lower surprisal junctures, suggesting again a conservative strategy. This effect, like that discussed

above, is amenable to explanation in terms of cognitive cost, with NNS experiencing greater difficulty with transitions that are otherwise unproblematic for native speakers, but converging on native performance when the transitions reach a certain threshold of uncertainty.

As far as the implications of the present study for language teaching are concerned, one may conclude that overall, *that*-variation does not constitute an insurmountable challenge to learners: in spite of the fact that proper complementizer use is hardly if ever a topic of explicit classroom instruction, the intermediate-advanced learners investigated here seem to be well on their way to nearly native-like behavior. *That*-variation may be taken as a powerful example of how much learners can pick up by implicitly scrutinizing the distributional patterns of their input even though the random effects also showcase considerable individual variation. That said, the results, of course, point to room for improvement. For one, instruction could focus more on complementizer variability by comparing the L1 with the L2; especially the Spanish learners may benefit from their attention being directed at the optionality of *that* in adjectival and object complements in particular. Similarly, increasing awareness for mode-dependent differences may be useful for both learner groups examined here.

References

Biber, Douglas. 1999. A register perspective on grammar and discourse: variability in the form and use of English complement clauses. *Discourse Studies* 1. 131-50.

Bryant, Margaret M. 1962. *Current American usage*. New York: Funk & Wagnalls.

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

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

Ellis, Nick C. 2006. Language acquisition as rational contingency learning. *Applied Linguistics 27*. 1-24.

Ellis, Nick C. & Stefanie Wulff. 2015. Usage-based approaches in second language acquisition. In Bill VanPatten & Jessica Williams (eds.), *Theories in second language acquisition: an introduction*, 75-93. London & New York: Routledge.

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

Gilquin, Gaëtanelle, Sylvie de Cock & Sylviane Granger. 2010. *Louvain International Database of Spoken English Interlanguage.* Louvain-la-Neuve: Presses universitaires de Louvain.

Goldberg, Adele E. 2006. *Constructions at work: the nature of generalization in language*. Oxford: Oxford University Press.

Granger, Sylviane, Estelle Dagneaux, Fanny Meunier & Magali Paquot. 2009. *International Corpus of Learner English v2.* Louvain-la-Neuve: Presses Universitaires de Louvain.

Gries, Stefan Th. 2007. *Coll.analysis 3.2*. A program for *R* for Windows.

Gries, Stefan Th. 2013. 50-something years of work on collocations: What is or should be next … *International Journal of Corpus Linguistics* 18(1). 137-165.

Gries, Stefan Th. & Allison S. Adelman. 2014. Subject realization in Japanese conversation by native and non-native speakers: exemplifying a new paradigm for learner corpus research. In Jesús Romero-Trillo (ed.), *Yearbook of corpus linguistics and pragmatics 2014: new empirical and theoretical paradigms*, 35-54. Cham: Springer.

Gries, Stefan Th. & Sandra C. Deshors. 2014. Using regressions to explore deviations between corpus data and a standard/target: two suggestions. *Corpora* 9(1). 109-136.
Jaeger, T. Florian. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology* 61. 23-62.
Kaltenböck, Gunther. 2006. '… *That* is the question': Complementizer omission in extraposed *that*-clauses. *English Language and Linguistics* 10(2). 371-96.
Kroll, Judith F. & Paola E. Dussias. 2013. The comprehension of words and sentences in two languages. In Tej K. Bhatia & William C. Ritchie (eds.), *The handbook of bilingualism and multilingualism*, 216-243. Malden, MA: Wiley-Blackwell Publishers.
Levy, Roger. 2008. Expectation-based syntactic comprehension. *Cognition* 106(3). 1126-1177.
Rissanen, Matti. 1991. On the history of *that*/zero as object clause links in English. In Karin Aijmer & Bengt Altenberg (eds.), *English corpus linguistics*, 272-289. London: Longman.
Storms, G. 1966. *That*-clauses in Modern English. *English Studies* 47. 249-70.
Tagliamonte, Sali A. & Jennifer Smith. 2005. *No momentary fancy!* The *zero* 'complementizer' in English dialects. *English Language and Linguistics* 9(2). 289-309.
Thompson, Sandra A. & Anthony Mulac. 1991. The discourse conditions for the use of the complementizer that in conversational English. *Journal of Pragmatics* 15. 237-51.
Torres Cacoullos, Rena & James A. Walker. 2009. On the persistence of grammar in discourse formulas: a variationist study of *that*. *Linguistics* 47. 1-43.
Wulff, Stefanie. to appear. A friendly conspiracy of input, L1, and processing demands: *that*–variation in German and Spanish learner language. In Andrea Tyler, Lourdes Ortega & Mariko Uno (eds.), *The usage-based study of language learning and multilingualism* (Proceedings of GURT 2014). Georgetown: Georgetown University Press.
Wulff, Stefanie, Nicholas A. Lester & Maria M. Martinez-Garcia. 2014. *That*-variation in German and Spanish L2 English. *Language and Cognition* 6. 271-299.