Language and Cognition (2026), 18, e58, 1–24
doi:10.1017/langcog.2026.10106

ARTICLE

An intermodal approach to the complementizer *that* in constrained and unconstrained language

Qiurong Zhao and Siqi Lin

School of Foreign Studies, University of Science and Technology Beijing, China
Corresponding author: Siqi Lin; Email: siqilin@xs.ustb.edu.cn

(Received 07 October 2025 UTC; Revised 13 July 2026 UTC; Accepted 23 July 2026 UTC)

Abstract

This study investigates the factors conditioning the *that*/zero alternation in four varieties of political discourse: translated and interpreted English as constrained varieties, and original written and spoken English as unconstrained varieties. Distinctive collexeme analysis and multivariate analysis were employed. The results show that constrained varieties use the complementizer *that* significantly more often than unconstrained varieties, with a significantly higher rate in translated than interpreted English. Distinctive collexeme analysis reveals subtle yet systematic differences in the matrix verbs associated with *that*-complement clauses, suggesting that distinct semantic and constructional profiles contribute to variation in the use of *that*. Multivariate analysis shows that all four varieties are largely conditioned by the same conventionality- and complexity-related factors, although production mode interacts with specific linguistic constraints. Translated English favors *that* even in contexts where omission is favored, a pattern consistent with pragmatic risk avoidance. Its broader preference for *that* may also reflect a tendency toward analyticity associated with cognitively demanding bilingual activation. In interpreted English, however, explicit marking is more plausibly attributable to analyticity than to heightened risk avoidance. Taken together, these findings provide intermodal empirical evidence of increased explicitness in constrained language based on Chinese–English language pair, extending previous findings derived primarily from European language pairs.

Keywords: cognitive load; constrained language; explicitation; intermodal; optional complementizer *that*

1. Introduction

Explicitation was introduced as a core concept in translation studies by Vinay and Darbelnet (1995, p. 342). They originally defined it as 'a stylistic translation technique which consists of making explicit in the target language what remains implicit in the source language because it is apparent from either the context or the situation'. Building on this notion, Blum-Kulka (1986, p. 19) subsequently formulated the Explicitation Hypothesis, stating that translation typically exhibits greater clarity

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 17:56:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/langcog.2026.10106

in the use of cohesion and connective mechanisms compared to the source texts. Baker (1993, p. 243) further integrated explicitation into the framework of translation universals, highlighting its empirical observability and theoretical significance in corpus-based studies. Building on these foundations, empirical research generally adheres to a core principle: such an increase in explicitness should be optional in nature; in other words, 'to prove that there was explicitation, there must have been the possibility of a correct but less explicit or less precise version' (Séguinot, 1988, p. 108).

Research on explicitation has mainly followed three perspectives: linguistic, sociocultural and cognitive. The earliest approach, the linguistic perspective, focuses on differences in language use and typology, attributing explicitation phenomena to factors such as typological differences (Molés-Cases, 2019), grammatical metaphor (Steiner, 2001) and language contact (Kruger & Van Rooy, 2012). The sociocultural perspective centers on translation acceptability, arguing that explicitation stems from translators' risk-aversion strategies to comply with target-language norms and readers' expectations (Becher, 2010; Pym, 2005, 2015). From a cognitive perspective, explicitation is interpreted through the lenses of cognitive linguistics and cognitive psychology, with a focus on translators' cognitive load and working memory. Representative theories include the Complexity Principle (Rohdenburg, 1996) and theories of processing efficiency (Hawkins, 2014). This perspective holds that explicitation is an unconscious choice made by translators (Kruger & Van Rooy, 2016b; Olohan & Baker, 2000).

In terms of specific linguistic forms, optional *that* is a key grammatical marker in English for introducing complement clauses. According to the Cambridge International Dictionary of English, *that* not only signals the syntactic dependency between main and subordinate clauses but also exhibits a 'property of encoding' (Hansen-Schirra et al., 2007, p. 246), serving as an important linguistic device to connect subordinate structures (Procter, 1995, p. 1507). The grammatical optionality of *that* is considered an effective indicator of translators' linguistic preference, sociocultural conditions and cognitive processing, making it one of the crucial linguistic units for testing the explicitation hypothesis.

Olohan and Baker (2000) were the first to conduct a corpus-based, contrastive study comparing translated English from the Translational English Corpus (TEC) with original English from the British National Corpus (BNC). They found that the optional complementizer *that* is used far more frequently after the reporting verbs *say* and *tell* in translated English than in original English. This reveals evidence for subliminal processes of explicitation, opening up an empirical research path for investigating the complementizer *that* in translation studies. Subsequent research extended this line of inquiry across different language pairs and text types (Kajzer-Wietrzny, 2012; Kruger & Van Rooy, 2012; Morselli, 2014), consistently finding that translated English exhibits a significantly higher frequency of the complementizer *that* compared to original English, thus further corroborating the hypothesis that translated language is more explicit. De Sutter and Vermeire (2019) further advanced this line of research by applying a multifactorial corpus-based approach, namely MuPDAR, to investigate the *that*/zero alternation across 120 lemmas. Taken together, previous research suggests that the increased use of explicit *that* is associated with several factors, including risk-aversion strategies (De Sutter & Vermeire, 2019; Kruger, 2019), source-language transfer (De Sutter & Vermeire, 2019; Kruger, 2019), cognitive processing load (Kruger, 2019), language proficiency (De Sutter &

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 17:56:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/langcog.2026.10106

Lefer, 2020) and task expertise (De Sutter & Lefer, 2020). These studies reveal the multiple drivers behind the explicit use of *that* in translated English.

Extending the scope to oral contexts, Kajzer-Wietrzny (2018) observed a higher frequency of the complementizer *that* in both L2 and interpreted English compared to original English, suggesting that this phenomenon may not be specific to written discourse but may also characterize spoken modes of language use. In terms of intermodal variation, Hu and Tao's (2013) analysis of the Chinese-English Conference Interpreting Corpus (CECIC) revealed marked differences across production modes: the frequency of the complementizer *that* in interpreted English was 13% higher than in original spoken English and three times that observed in translated English. These findings suggest a pronounced tendency toward explicitation within Chinese–English conference interpreting.

Despite significant advances in exploring the explicit use of *that* in translated and interpreted English, insufficient attention has been paid to the differences in the frequency of *that* across modalities. To the best of our knowledge, Hu and Tao's (2013) study represents the only attempt to adopt such an intermodal perspective. However, several limitations in their study warrant further investigation. First, a control corpus of original written English in a similar register was absent from their study. Second, because the study reports only the frequency of the complementizer *that* and does not account for zero realization, it is unclear whether the higher rate of *that* reflects a genuine preference for lexicogrammatical explicitness or simply a greater incidence of complement clauses. Finally, the unexpectedly low frequency of the complementizer *that* in translated English in Hu and Tao (2013) stands in stark contrast to findings from studies restricted to written translation (e.g., Kruger, 2019; Olohan & Baker, 2000), where translated texts typically exhibit increased explicitness. This discrepancy underscores the need for more comparable, multimodal corpus analyses that treat complementizer choice as a probabilistic alternation.

To address these gaps, this study employs the constrained communication framework (Kotze, 2022), utilizing the Chinese-English Political Interpreting Corpus (CEPIC) for interpreted English and original spoken English, the Parallel Corpus of *Xi Jinping: The Governance of China* (PCGC) for translated English and the Corpus of the *U.S. Annual Report to Congress on Military and Security Developments Involving the People's Republic of China* (CMSC) for original written English. Through distinctive collexeme and multifactorial analysis, we systematically investigate *that*-explicitation across constrained and unconstrained English, addressing two core questions:

(1) What differences exist in the frequency and distribution of optional complementizer *that* across four varieties? How do these differences reflect distinctive register-related or mediation-specific features?
(2) To what extent is the choice of the complementizer *that* in interpreting and in translation driven by linguistic, cognitive and sociocultural factors, and how do these factors interact in shaping modality-specific patterns?

This study aims to identify the factors underlying the use of *that* across the four varieties and to provide intermodal empirical evidence of the increased explicitness in constrained language.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 17:56:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/langcog.2026.10106

2. Literature review

2.1. Constrained language and the constrained communication framework

In recent years, 'constrained language' has gradually gained academic attention as an independent research topic (Kotze, 2019, pp. 344–348). It refers to language production that is shaped by multiple constraints in specific communicative contexts (Kruger & Van Rooy, 2016a, p. 27). Among the theoretical models proposed, Kotze's (2019, 2022) constrained communication framework is one of the most systematic attempts to describe and compare constrained varieties. The framework identifies five core dimensions that jointly shape constrained production, namely, Language Activation, Modality and Register, Text Production, Proficiency and Task Expertise, and highlights the complex interactions among them in shaping the structural features of translationese and second-language varieties. By integrating linguistic, sociocultural and cognitive perspectives, the framework provides a novel lens for explaining language phenomena across registers.

Research on constrained language has so far focused predominantly on written registers, with a particular emphasis on the dimensions of language activation and text production. Existing studies have explored the characteristics of written constrained language through specific lexical and syntactic features, such as collocations (Ferraresi, 2019; Ferraresi & Bernardini, 2023), the optional complementizer *that* (Partridge, 2024) and embedded inversion (Gilquin & Meriläinen, 2024). Beyond these feature-specific analyses, some studies have sought to develop a broader linguistic profile of constrained language by drawing on a more extensive set of linguistic features, as exemplified by 67 lexicogrammatical features (Kruger & Van Rooy, 2016a), POS dependency bigrams (Ivaska et al., 2022) and dependency distance and dependency direction (Ma et al., 2024). Collectively, these studies indicate that written constrained language shares common trends, including increased explicitness, heightened simplification and reduced complexity, primarily associated with cognitive and social constraints arising from bilingual activation.

In research on constrained language that has predominantly focused on written registers, translation is regarded as a highly constrained form of bilingual language production, imposing substantial cognitive pressure (Kajzer-Wietrzny, 2022). Compared with translation, interpreting is subject to stricter temporal constraints, rendering its bilingual activation and language processing more immediate and intensive, and potentially more cognitively demanding. Therefore, expanding the research perspective from written to spoken mediated language and comparing the two constrained modes from an intermodal perspective holds significant theoretical and empirical value.

Against this backdrop, Kajzer-Wietrzny and her collaborators conducted a series of representative intermodal studies on constrained language based on the European Parliament Translation and Interpreting Corpus (EPTIC). At the lexical level, Kajzer-Wietrzny and Ivaska (2020) compared six language varieties, focusing on lexical diversity and its distributional patterns. The study revealed that interpreting exhibited significantly lower lexical diversity than translation, indicating that spoken mediated language tends to adopt more restricted lexical selection strategies. At the information processing level, Kajzer-Wietrzny et al. (2021) took numerals as a starting point to compare simultaneous interpreting and written translation when processing the same source text. The results showed that numeral conversion error rates were significantly higher in interpreting than in translation, with this difference

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 17:56:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/langcog.2026.10106

influenced by multiple factors such as source numeral type, in-sentence numeral density, speech rate and translation directionality. At the level of discourse cohesion, Kajzer-Wietrzny (2022) compared patterns in the use of cohesive devices and found that, despite substantial differences in modality and processing conditions, interpreting and translation displayed a high degree of consistency in their overall frequency. This suggests that mediated texts may exhibit similar explicitation tendencies at the discourse cohesion. Together, Kajzer-Wietrzny's work demonstrates that modality exerts a significant influence on constrained production.

However, existing intermodal research has predominantly focused on European language pairs. Given that Chinese and English differ fundamentally in their typological structures (Lian, 2010), it is reasonable to expect additional challenges for Chinese–English translators and interpreters (Ma, 2021). In this context, the present study adopts an intermodal perspective to investigate the frequency of the complementizer *that* across translated, interpreted, original spoken and original written English. By doing so, we aim to extend the empirical scope and cross-linguistic applicability of the constrained communication framework.

2.2. Multifactorial studies of the factors that condition that/zero alternation

The conditioning factors of *that*/zero alternation, well established in L1 and L2 English research (e.g., Jaeger, 2010; Wulff et al., 2014, 2018), are increasingly being scrutinized within the context of mediated varieties (e.g., Kotze, 2022; Kruger, 2019).

Drawing on Kotze (2022, pp. 83–84), the variables influencing *that*/zero alternation can be divided into complexity-based factors, which stem from cognitive processing constraints and conventionality-based factors, where the preference for either variant is linked by convention to particular contexts.

In terms of complexity-related predictors, the complexity of the matrix subject has been shown to condition the *that*/zero alternation (Jaeger, 2010, pp. 24–25; Torres Cacoullos & Walker, 2009, pp. 32–33). Generally, *I* and *you* correlate with an increased likelihood of zero-complementizer, particularly with simple, high-frequency verbs. Other measures of syntactic complexity in the matrix clause that are reported to influence the *that*/zero alternation include the presence of intervening material between the matrix subject and the matrix verb and the presence of intervening material between the matrix verb and the onset of the complement clause (Jaeger, 2010; Kruger, 2019; Partridge, 2024; Tagliamonte & Smith, 2005; Torres Cacoullos & Walker, 2009).

The *that*/zero alternation is also sensitive to the complexity of the complement clause. Factors such as the complexity of the complement-clause subject (Jaeger, 2010, p. 16; Tagliamonte & Smith, 2005, p. 304) and the presence of intervening material between the onset of the complement clause and the subject of the complement clause (Jaeger, 2010, p. 18; Partridge, 2024, pp. 108–109) correlate positively with the presence of *that*. However, though they have been extensively investigated within L1 and L2 varieties, these factors remain largely under-explored in the field of translation and interpreting studies.

In terms of conventionality-related predictors, the semantics of the matrix verb is said to condition the *that*/zero alternation. 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-complementizer. Dor (2005) similarly

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 17:56:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/langcog.2026.10106

argues that truth-claim (matrix) predicates allow zero-complementizer on the grounds that zero-preferring predicates denote 'asserted propositions' (i.e., 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*-preferring predicates, which simply signal that a proposition has been made.

The frequency of the matrix verb also plays an important role, with high-frequency verbs more likely to cue complementizer omission (Tagliamonte & Smith, 2005; Torres Cacoullos & Walker, 2009; Wulff et al., 2014). However, it is pointed out that the frequency operationalization in many existing studies is problematic, since some studies focus only on high-frequency verbs (e.g., Torres Cacoullos & Walker, 2009), while others consider pure lemma frequency for the matrix verb as a criterion (e.g., Jaeger, 2010). The measure Delta P used by Wulff et al. (2014) and Wulff et al. (2018) is derived from a collostructional analysis, which provides an indication of the attraction of a particular lemma to the construction with or without the complementizer. The robustness of this measure is further corroborated by Partridge (2024), who applied conditional inference trees to investigate the alternation in South African English. Her findings indicate that the Delta P_{WC} variable serves as the highest-level predictor in the conditional inference tree.

Register has repeatedly been found to be one of the variables influencing the *that*/zero alternation and is considered a socio-cognitive conditioner since it describes the environment in which the text or utterance is situated. The main contrast is between spoken and written language with spoken registers displaying a higher rate of zero-complementation than written registers (Jaeger, 2010; Torres Cacoullos & Walker, 2009; Wulff et al., 2014). Written language is often further differentiated on the grounds of formality. Formal registers tend to retain the complementizer *that*, while registers that are more informal are said to display a higher rate of zero-complementizer (Kruger, 2019; Van Rooy & Kruger, 2016).

Taken together, existing research indicates that the realization of the optional complementizer is jointly shaped by cognitive processing constraints and conventionalized associations with particular socio-communicative contexts. A comprehensive account of the constraints governing the *that*/zero alternation therefore requires an integrated approach that considers both sets of factors.

3. Corpus and methods

3.1. Corpus composition

This study focuses exclusively on political discourse and incorporates four English varieties: interpreted English, translated English, original spoken and original written English (henceforth interpreted, translated, original spoken and original written, respectively). Data used for interpreted and original spoken English were extracted from the CEPIC,[1] an open-access corpus consisting of political speeches and their interpreting in both Chinese and English. Speeches covered by the corpus were delivered by top government officials on official occasions. The main locales of the speeches were Hong Kong SAR, Beijing, London, and Washington, DC. According to Pan (2019), the corpus included 16 sub-datasets, totaling over 6 million tokens. The

[1]See https://digital.lib.hkbu.edu.hk/cepic/.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 17:56:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/langcog.2026.10106

Table 1. Corpus composition

	Translated	Interpreted	Original spoken	Original written
Corpus	PCGC	CEPIC	CEPIC	CMSC
Corpus size	520,925	2,578,911	3,815,083	550,251
Time period	2012–2020	1997–2017	1997–2017	2000–2020

Table 2. Constraint matrix for translated, interpreted, original spoken and original written English

Constraint dimension	Translated	Interpreted	Original spoken	Original written
Language activation	Bilingual (L1-L2)	Bilingual (L1-L2)	Monolingual (L1)	Monolingual (L1)
Language proficiency	Proficient	Proficient	Proficient	Proficient
Modality and register	Written	Spoken	Spoken	Written
Text production	Mediated	Mediated	Non-mediated	Non-mediated
Task expertise	Expertise	Expertise	Expertise	Expertise

data used for translated English were sourced from the Multilingual PCGC,[2] a parallel corpus developed by Shanghai International Studies University. The corpus encompasses Volume I, II and III of *Xi Jinping: The Governance of China* in 28 languages, including reports, speeches, talks, addresses, instructions, directives and congratulatory messages delivered by General Secretary Xi Jinping. These texts were translated by a highly experienced team from China International Communications Group specializing in the translation of official Chinese political texts. In total, the corpus contains 270 texts with a total word count of 520,925. Data for the original written English were retrieved from a self-constructed Corpus of the CMSC.[3] Mandated to the U.S. Department of Defense, this report documents the development of China's military capabilities and strategic frameworks and is submitted annually to meet legislative requirements. The self-constructed corpus for the present study comprises 20 texts, with a total word count of 550,251.

CMSC covers the period 2000–2020, PCGC spans 2012–2020 and CEPIC covers 1997–2017. To ensure maximum comparability across the four varieties and to mitigate the potential influence of diachronic linguistic shifts, the present study restricts its sampling period to a synchronized six-year window from 2012 to 2017. A summary of the corpora is presented in Table 1.

Drawing on Kotze's (2019, 2022) constrained language model, this study compares the selected corpora across the five constraint dimensions summarized in Table 2. The table shows that, relative to unconstrained language, the constrained varieties are subject to stronger constraints in language activation and text production. It also highlights a further distinction between interpreted English and translated English in modality and register.

[2]See http://imate.cascorpus.com.

[3]Retrieved from the U.S. Department of Defense (DOD) website https://search.usa.gov/search?query=china%20military%20power%20report&affiliate=defensegov&utf8=%26%23x2713%3B.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 17:56:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/langcog.2026.10106

3.2. Data extraction and coding

Building on prior work (De Sutter & Vermeire, 2019; Olohan & Baker, 2000), we constructed the dataset as follows. To identify *that*/zero constructions, we retrieved all instances in which a finite complement clause with either a *that*- or zero-complementizer occurred after one of the 157 matrix verb lemmas classified by Quirk et al. (1985) as public,[4] private[5] and suasive[6] verbs. This process produced 903 instances from interpreted English, 2,239 from original spoken English, 439 from translated English and 253 from original written English. All hits were manually verified to exclude non-target contexts, following the procedure of Olohan and Baker (2000).

The instances were then manually annotated. The dependent variable is the presence or absence of *that*. Predictors comprise (i) a four-level factor, production mode (ProdMode), distinguishing four varieties (translated, interpreted, original spoken and original written), and (ii) two sets of linguistic predictors: complexity-related predictors and conventionality-related predictors based on the literature discussed in Section 2.2. Following Kotze (2022) and Partridge (2024), these predictors and the levels of each predictor are summarized below.

3.2.1. Conventionality-related predictors

(1) Semantic class of the matrix verb (MCVerbSemantics): A categorical variable with the following values: private, public, suasive.
(2) An association measure that quantifies how likely the matrix verb lemma is to take a *that*/zero-complementizer (DELTAPW2C): A continuous variable ranging from −1 through 0 to 1.[7]
(3) An association measure that quantifies how likely the *that*/zero-complementizer is to occur with the matrix verb lemma in question (DELTAPC2W): A continuous variable ranging from −1 through 0 to 1.[8]

[4]**Public verbs**: *acknowledge, add, admit, affirm, agree, allege, announce, argue, assert, bet, boast, certify, claim, comment, complain, concede, confess, confide, confirm, contend, convey, declare, deny, disclose, exclaim, explain, forecast, foretell, guarantee, hint, insist, maintain, mention, object, predict, proclaim, pronounce, prophesy, promise, protest, remark, repeat, reply, report, retort, say, state, submit, suggest, swear, testify, vow, warn, write*.

[5]**Private verbs**: *accept, anticipate, ascertain, assume, believe, calculate, check, conclude, conjecture, consider, decide, deduce, deem, demonstrate, determine, discern, discover, doubt, dream, ensure, establish, estimate, expect, fancy, fear, feel, find, foresee, forget, gather, guess, hear, hold, hope, imagine, imply, indicate, infer, insure, judge, know, learn, mean, note, notice, observe, perceive, presume, presuppose, pretend, prove, realize, reason, recall, reckon, recognize, reflect, remember, reveal, see, sense, show, signify, suppose, suspect, think, understand*.

[6]**Suasive verbs**: *agree, allow, arrange, ask, beg, command, concede, decide, decree, demand, desire, determine, enjoin, ensure, entreat, grant, insist, instruct, intend, move, ordain, order, pledge, pray, prefer, pronounce, propose, recommend, request, require, resolve, rule, stipulate, suggest, urge, vote*.

[7]The closer the value is to 1, the more likely it is that the matrix verb will attract the *that*-complementizer. The closer the value is to −1, the more likely it is that the matrix verb will attract the zero-complementizer. A value of 0 indicates that there is no association. This association measure is obtained with R script 'coll. analysis 4.1' (Gries, 2024).

[8]The closer the value is to 1, the more likely it is that the *that*-complementizer will attract the verb lemma in question. The closer the value is to −1, the more likely it is that the zero-complementizer will attract the verb lemma in question. A value of 0 indicates that there is no association. This association measure is obtained with R script 'coll.analysis 4.1' (Gries, 2024).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 17:56:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/langcog.2026.10106

3.2.2. Complexity-related predictors

(4) Tense and modality of the matrix clause (MCVerbTenseModality): A categorical variable with the following values: present, past, modal, semi-modal, ing-participle.
(5) Aspect of the matrix clause (MCVerbAspect): A categorical variable with the following values: simple, perfect, progressive.
(6) Polarity of the matrix clause (MCPolarity): A categorical variable with the following values: positive, negative.
(7) The distance between the subject in the matrix clause and the verb in the matrix clause (DistMCSubjToMCVerb): A continuous variable of the number measured in characters (space excluded).
(8) Number of the subject in the matrix clause (MCSubjNum): A categorical variable with the following values: singular, plural, no subject.
(9) Syntactic structure of the subject of the matrix clause (MCSubjForm): A categorical variable with the following values: pronoun, clause, relative pronoun, gap (no overt subject), NP.
(10) The distance between the matrix-clause verb and the onset of the complement clause (DistMCVerbToComp): A continuous variable of the number measured in characters (space excluded).
(11) The distance between the start of the complement clause and the subject in the complement clause (DistCompToCCSubj): A continuous variable of the number measured in characters (space excluded).
(12) The number of the subject of the complement clause (CCSubjNum): A categorical variable with the following values: singular, plural, no subject.
(13) The syntactic structure of the subject in the complement clause (CCSubjForm): A categorical variable with the following values: pronoun, clause, relative pronoun, gap (no overt subject), NP.

The linguistic annotation in this study was conducted by two researchers with specialized backgrounds in linguistics. To ensure the reliability and consistency of the data, an inter-annotator agreement analysis was performed, yielding a Cohen's kappa coefficient of 0.88. Discrepancies between the annotators were infrequent. Any controversial instances were subsequently resolved through comprehensive discussion until a consensus was reached.

3.3. Data analysis

This section first presents descriptive statistics on the use of *that* in four varieties. Four pairwise chi-square tests were conducted on the raw counts using 2 × 2 contingency tables. This step aims to provide a preliminary understanding of the distribution of *that*/zero alternation.

Next, to identify those collexemes that distinguish best among the four varieties with respect to the use of *that*, we conducted a multiple distinctive collexeme analysis with Coll.analysis 4.1 (Gries, 2024), an R script designed for different types of collostructional analysis, running in R version 4.4.1. This analytical approach is capable of measuring not only the dissimilarities between semantically similar constructions but also those of the same construction across different language varieties (Gries & Stefanowitsch, 2004; Wulff et al., 2007, 2014).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 17:56:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/langcog.2026.10106

Finally, to explore the influence of various predictor variables on the use of *that*, we conducted a conditional inference tree as implemented in the 'ctree' function in the R package 'partykit' (Hothorn & Zeileis, 2015), which is suitable for 'messy' data with empty cells and non-normal distribution (Kotze, 2022, p. 81), and is more intuitively interpretable than the coefficients table usually reported for regression models (Baayen et al., 2013, p. 265; Gries, 2020, pp. 618–619). The results of the conditional inference tree analysis are compared with the results of random forest analysis as implemented in the R package 'randomForest' (Liaw & Wiener, 2002). Random forests provide the importance measure for each variable averaged over many conditional inference trees. It adds an additional layer of randomness to the analysis and protects against overfitting (Gries, 2020, pp. 618–619).

4. Findings

This section first provides an overview of the frequency and distribution of the complementizer *that* across the constrained and unconstrained varieties (RQ 1), and then examines the factors that condition the use of *that* across the four varieties (RQ 2).

4.1. Overall frequencies of the complementizer that across four varieties

Table 3 presents the frequencies and proportions of *that* and zero in four varieties, and Table 4 reports four pairwise chi-square tests calculated on raw counts using 2 × 2 tables.

As shown in Table 3, there are clear differences in the use of *that* across the four varieties: translated English exhibits the highest proportion of *that* (86.33%), followed by interpreted English (77.96%), original written English (70.36%) and original spoken English (64.18%). These figures are notably higher than those reported by Kruger (2019), who reported frequencies of 66% and 74% for original written and translated English, respectively, in a multi-register context. Such high frequency likely reflects the formal nature of the political register examined here (see Section 5.3 for a detailed discussion).

Table 3. Occurrences of *that*/zero-complementizer across four varieties

	Translated		Interpreted		Original written		Original spoken	
Complementizer	*that*	zero	*that*	zero	*that*	zero	*that*	zero
Raw frequency	379	60	704	199	178	75	1437	802
%	86.33	13.67	77.96	22.04	70.36	29.64	64.18	35.82

Table 4. Differences between the investigated varieties with respect to occurrences of *that*: results of chi-square tests calculated on raw counts in 2 × 2 contingency tables

	Original written versus translated	Original spoken versus interpreted	Translated versus interpreted	Original spoken versus original written
χ^2	26.091	56.301	13.288	3.801
p-value	<0.001	<0.001	<0.001	0.051

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 17:56:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/langcog.2026.10106

To determine whether the differences are significant, we conducted four chi-square tests calculated on raw counts in 2 × 2 contingency tables. Table 4 presents the results of the four pairwise comparisons between the varieties.

As shown in Table 4, all comparisons involving constrained language (translated or interpreted) versus corresponding unconstrained varieties (written or spoken) yielded significant differences ($p < 0.001$). Conversely, the difference between the two unconstrained varieties (spoken versus written) did not show statistical significance ($p = 0.051$). This pattern indicates that mediation exerts a more profound influence on the use of *that* than the mode of production (spoken versus written) itself. Based on these findings, we conclude that the explicitation hypothesis is supported in both interpreted and translated English.

4.2. *Distinctive collexeme analyses of matrix verbs in* that*-complement constructions*

This section employs Coll.analysis 4.1 (Gries, 2024) to identify those matrix verbs that distinguish among the four varieties, in the sense that they are more typical of one variety than of the others. For each matrix verb, the analysis compares its observed frequency in a given variety with the frequency expected on the basis of its distribution across all four varieties (Gries, 2024; Wulff et al., 2014). The strength and direction of association were measured by Pearson residuals. To facilitate the interpretation of the systematic differences, the distinctive collexemes were subsequently classified into semantic categories following Quirk et al. (1985). Table 5 provides an overview of all distinctive collexemes for the four varieties, ranked according to their Pearson residuals,[9] Table 6 shows the distribution of all distinctive collexemes according to semantic class and variety, and Table 7 provides the corresponding summed Pearson residuals per class and variety.

The semantic classification of the distinctive collexemes reveals systematic cross-variety differences in the matrix verbs occurring in *that*-complement constructions. The clearest contrast concerns public verbs. Original written English has ten distinctive public verbs, with a cumulative Pearson residual of 73.947, while original spoken English has four, with a substantially lower cumulative residual sum of 15.248. Interpreted and translated English each have only one distinctive public verb, with cumulative residuals of 4.087 and 2.293, respectively.

Private verbs constitute the largest semantic class, comprising 23 distinctive collexemes. They are most numerous in translated English (10), followed by interpreted (6), original written (5) and original spoken English (2). Although translated English has the highest residual sum (31.473), interpreted English follows closely (30.654) despite having considerably fewer private verbs, suggesting that individual private verbs are, on average, more strongly associated with interpreted English. Original written English occupies an intermediate position (23.895), whereas original spoken English relies exclusively on *think* and *know*, which have a combined residual of 13.465.

[9]A positive Pearson residual indicates that a matrix verb occurs more frequently than expected with the *that*-complementizer in a particular variety; the larger the residual, the more distinctive the verb is of that variety. The table includes only matrix verbs with positive Pearson residuals greater than 1.96 (Friendly & Meyer, 2015).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 17:56:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/langcog.2026.10106

Table 5. Distinctive matrix verb collexemes in *that*-complement constructions across the four varieties

Interpreted		Original spoken		Translated		Original written	
Matrix verb	Distinctiveness	Matrix verb	Distinctiveness	Matrix verb	Distinctiveness	Matrix verb	Distinctiveness
Ensure	10.220	think	6.873	require	6.434	state	12.312
see	7.949	know	6.592	prove	5.437	suggest	11.436
estimate	7.521	confirm	5.680	propose	4.721	claim	9.713
hope	5.535	say	4.796	stipulate	4.528	maintain	8.840
forecast	4.087	announce	2.609	remember	4.279	indicate	8.192
expect	3.787	report	2.163	demand	3.766	assert	7.318
move	3.235			believe	3.584	rule	6.298
consider	3.117			observe	3.216	note	5.718
notice	2.745			realize	2.876	argue	5.263
				mean	2.713	declare	5.143
				demonstrate	2.432	affirm	5.143
				learn	2.350	contend	5.143
				feel	2.293	judge	4.527
				guarantee	2.293	acknowledge	3.636
				dream	2.293	hold	2.908
						request	2.718
						determine	2.550

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 17:56:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/langcog.2026.10106

Table 6. Distribution of distinctive collexemes according to semantic classes

Semantic class	Interpreted	Original spoken	Translated	Original written	Row totals
Public	1	4	1	10	16
Private	6	2	10	5	23
Suasive	2	0	4	2	8
Column totals	9	6	15	17	47

Table 7. Distribution of summed Pearson Residuals of distinctive collexemes according to semantic classes

Semantic class	Interpreted	Original spoken	Translated	Original written	Row totals
Public	4.087	15.248	2.293	73.947	95.575
Private	30.654	13.465	31.473	23.895	99.487
Suasive	13.455	0.000	19.449	9.016	41.920
Column totals	48.196	28.713	53.215	106.858	236.982

Suasive verbs form the smallest semantic class, comprising eight distinctive collexemes overall. They are most numerous in translated English (4), followed by interpreted and original written English (2 each), while none attains distinctive status in original spoken English. The cumulative residual shows the same general pattern: translated English exhibits the strongest overall association (19.449), followed by interpreted English (13.455) and original written English (9.016). Although interpreted English has only two distinctive suasive verbs, its relatively high cumulative residual again suggests strong associations at the level of individual verbs.

Overall, the distinctive collexemes reveal clear semantic variation across the four varieties. Translated English is characterized by a broad combination of private and suasive verbs, whereas interpreted English exhibits a narrower but individually stronger concentration of private verbs. Original written English is dominated by public verbs, with private verbs forming a secondary component. Original spoken English presents a more restricted profile, comprising public and private verbs but no distinctive suasive verbs.

These semantic and constructional differences may help account for cross-variety variation in *that* retention. In particular, original spoken English, which has the lowest rate of the complementizer *that*, is distinctively associated with the private verbs *think* and *know*. Such verbs typically express intellectual states or acts that are not directly observable by others (Quirk et al., 1985, p. 1181) and often occur with first-person pronominal subjects, especially *I* and *we*. Highly conventionalized constructions, such as *I think* and *I know*, combine syntactically simple subjects with cognitively entrenched patterns, both of which favor zero realization, as discussed in Section 2.2. The prominence of *think* and *know* in original spoken English may therefore contribute, from a constructional perspective, to its comparatively low rate of overt *that*.

4.3. Multivariate analysis of factors conditioning the that*/zero alternation across the four varieties*

To investigate the effects of the predictor variables on the likelihood of the complementizer *that*, the predictor variables were used to grow a conditional inference tree, as explained in Section 3.2. The model formula for the tree is:

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 17:56:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/langcog.2026.10106

$$\begin{aligned} that/zero \sim\ & DELTAPC2W + DELTAPW2C + DistMCSubjToMCVerb \\ & + DistMCVerbToComp + DistCompToCCSubj + ProdMode \\ & + MCVerbSemantics + MCVerbTenseModality + MCVerbAspect \\ & + MCPolarity + MCSubjNum + MCSubjForm + CCSubjNum + CCSubjForm, \end{aligned}$$

and yields the tree[10] presented in Figure 1.

The conditional inference tree presented in Figure 1 visualizes the complex interactions among the predictor variables conditioning the *that*/zero alternation. At first glance, it is clear that there is a large degree of similarity across the four varieties, as evidenced by the absence of ProdMode-based splits at the higher nodes of the tree. However, at lower levels of the tree, there are factors that do affect the choice between *that*/zero in the four varieties in different ways.

The first split in the tree is based on DELTAPW2C (node 1), with a cut-off value of −0.086. The left branch, representing values at or below −0.086, shows a modest overall preference for zero-complementizer (55.0%). By contrast, the right branch, representing values above −0.086, displays a strong preference for *that-complementizer* (83.5%).

Within the left branch of node 1, the next split is based on the complexity of the subject in the complement clause (node 2). When the complement-clause subject is a pronoun, DELTAPW2C produces a further split at −0.186 (node 3). For values at or below this threshold, production mode (ProdMode) conditions the choice between the *that/zero alternation* (node 4).

When the complement-clause subject is a pronoun and DELTAPW2C falls between −0.186 and − 0.086, the distance between the matrix verb and the onset of the complement clause becomes relevant (node 7). If there is no intervening material between the matrix verb and the complement clause, zero-complementizer remains the preferred option (node 8, n = 282; 58.5% zero), as illustrated in Example 1. If there is any intervening material between the matrix verb and the complement clause, the choice is categorically in favor of *that*-complementizer (node 9, n = 16; 100% *that*), as illustrated in Example 2.

(1) The PLA believes $_{[\Delta Pwc\ =\ -0.151]}$ Ø $_{[DistMCVerbToComp\ =\ 0]}$ electronic warfare (EW) is one method to reduce or eliminate U.S. technological advantage. (CMSC: written)

(2) You have said $_{[\Delta Pwc\ =\ -0.086]}$ on many occasions $_{[DistMCVerbToComp\ >\ 0]}$ *that* you must work hard to promote social equity and justice. (CEPIC: spoken)

The other sub-branch of node 2 comprises complement clauses whose subjects are realized as clauses, relative pronouns, gaps, or noun phrases. Within this branch, DELTAPW2C generates a further split at −0.366 (node 10). When DELTAPW2C is at or below this threshold, indicating a strong lexical association with the zero-complementizer, production mode becomes decisive (node 11). Translated English nevertheless favors overt *that* (node 12, n = 48; 75.0% *that*), as illustrated in

[10]The conditional inference tree model achieved an overall classification accuracy of 79.37%, indicating a moderate level of predictive performance beyond chance.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 17:56:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/langcog.2026.10106

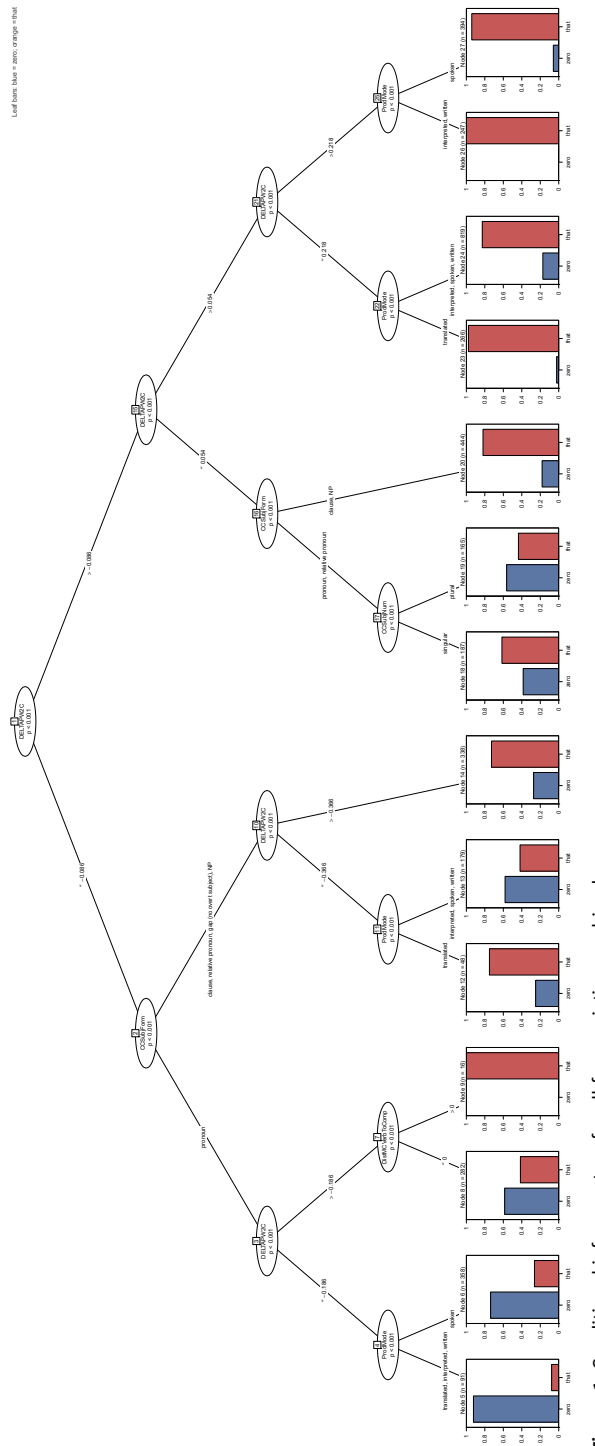


Figure 1. Conditional inference tree for all four varieties combined.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 17:56:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/langcog.2026.10106

Example 3, whereas interpreted, original spoken and original written English collectively show a moderate preference for zero (node 13, *n* = 179; 58.1% zero), as illustrated in Example 4. When DELTAPW2C exceeds −0.366, *that* predominates and no further statistically significant split by production mode emerges (node 14, *n* = 338; 72.8% *that*).

(3) We hope $_{[\Delta Pwc = -0.402]}$ *that* both sides $_{[NP]}$ can work together on the understanding that we are one family. (PCGC: translated)

(4) I do not think $_{[\Delta Pwc = -0.394]}$ Ø House Republicans $_{[NP]}$ would suggest we are there yet. (CEPIC: spoken)

The right branch of node 1, in which DELTAPW2C exceeds −0.086, displays a stronger overall preference for *that*. This branch is divided again at a DELTAPW2C value of 0.054 (node 15). For values between −0.086 and 0.054, complement-clause subject form conditions the subsequent choice (node 16). When the complement-clause subject is a pronoun or relative pronoun, complement-clause subject number produces a further split (node 17). Singular subjects favor *that* (node 18, *n* = 187; 61.5% *that*), whereas plural subjects show a modest preference for zero (node 19, n = 165; 56.4% zero). By comparison, complement clauses with clausal, gaps, or noun-phrase subjects strongly favor *that* (node 20, *n* = 444; 82.0% *that*).

When DELTAPW2C exceeds 0.054, the preference for *that* becomes particularly pronounced (90.2% of cases = *that*). Values between 0.054 and 0.218 are further differentiated by production mode (node 22). Translated language shows a near-categorical preference for *that* (node 23, *n* = 266; 97.7% *that*), while interpreted, original spoken and original written language also favor *that*, though less categorically (node 24, *n* = 819; 82.8% *that*).

At DELTAPW2C values above 0.218, a distinction emerges between interpreted, original written English and spoken English; the former two varieties show near-categorical *that*-usage (node 26, *n* = 247, 100% *that*), as exemplified in example (5), whereas original spoken English retains a slight margin for omission (node 27, *n* = 394; 94.2% *that*), as exemplified in example (6).

(5) This shows $_{[\Delta Pwc > 0.218]}$ *that* we Chinese people have the courage. (CEPIC: interpreted)

(6) Surveys show $_{[\Delta Pwc > 0.218]}$ Ø our standing around the world is higher than when I was elected to this office. (CEPIC: spoken)

Overall, the constrained varieties and their non-constrained counterparts are governed by similar conventionality- and complexity-related constraints, as reflected in the higher-level splits of the tree. Production mode does not operate as the principal global determinant of complementizer choice; instead, it modulates the effects of these constraints under particular linguistic conditions.

To add a layer of randomness to the analysis, the same predictor variables used to grow the conditional inference tree were used to grow a random forest, as explained in Section 3.3. The results of the random forest analysis are shown in the variable

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 17:56:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/langcog.2026.10106

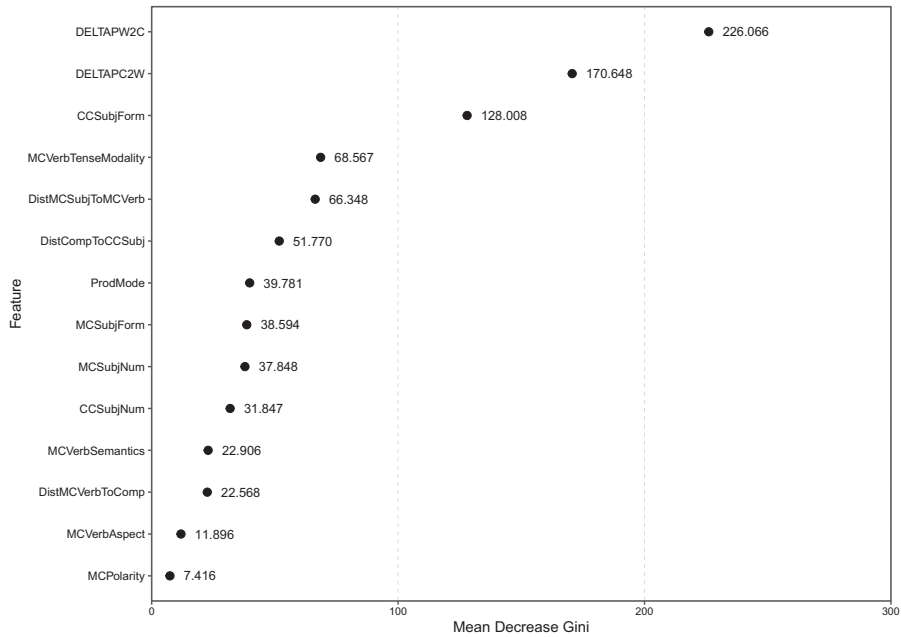


Figure 2. Random forest variable importance for all four varieties combined.

importance plot[11] in Figure 2 and provide information on the importance of the predictor variables.

Figure 2 reveals that two conventionality-related variables, DELTAPW2C and DELTAPC2W, are the two strongest predictors of the *that*/zero alternation, with an importance score of 226.066 and 170.648, respectively. This aligns with the conditional inference tree, where DELTAPW2C served as the highest-level splitting variable. Following these, complexity-related factors, including CCSubjForm, MCVerbTenseModality, DistMCSubjToMCVerb and DistCompToCCSubj, show relatively high importance. Then comes ProdMode, which has a greater impact than the rest of complexity-related factors, namely, MCSubjForm, MCSubjNum and CCSubjNum. Another conventionality-related factor, MCVerbSemantics, ranks relatively lower in the importance score, which aligns with the phenomenon that it is absent from the conditional inference tree.

The relatively low importance of ProdMode in the random forest analysis largely correlates with the findings from the conditional inference tree. This suggests that the four varieties are generally governed by a shared set of constraints when choosing between *that*- and zero-complementizer. However, the nuanced differences among the four varieties revealed by conditional inference tree remain opaque in the random forest. This can be attributed to the difference between the conditional inference trees and random forests. The latter successfully ranks the relative importance of factors, but it cannot illustrate how these variables functionally intersect with specific varieties to condition *that*-omission (Gries, 2020).

[11]A cross-tabulation of the predicted probabilities with observed outcomes demonstrates that the model correctly classifies 80.62% of the total of 3,834 instances.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 17:56:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/langcog.2026.10106

5. Discussion

The results show that the frequency of *that* in constrained language is significantly higher than in unconstrained language. Significant differences also exist between constrained varieties, with translated English using *that* more frequently than interpreted English. This finding is consistent with the findings of Olohan and Baker (2000), Kruger and De Sutter (2018), Kruger (2019), De Sutter and Lefer (2020) and De Sutter and Vermeire (2019), which show that translated English uses *that* more frequently than original English. It also aligns with Kajzer-Wietrzny's (2018) finding that explicit use of *that* also occurs in interpreted English. However, it contradicts Hu and Tao's (2013) finding that the frequency of *that* in interpreted English is three times that of translated English. Their study focused primarily on the frequency of overt *that* rather than the *that*/zero alternation (Hu & Tao, 2013, p. 636). Multivariate analysis shows that constrained and unconstrained varieties are shaped by similar conventionality- and complexity-related constraints, but they exhibit different patterns under specific linguistic conditions. This pattern is explained by the constrained communication framework proposed by Kotze (2019, 2022).

5.1. Language activation dimension

The language activation dimension reflects the degree of multilingual activation during language use and its cognitive impact on language production (Kotze & Van Rooy, 2024, p. 11). Unlike original English, which involves the activation of a single language, translated and interpreted English are produced under the simultaneous activation of both languages. Research has shown that even when only a single language is used during interlanguage production, the other language remains cognitively active (Shook & Marian, 2012). This bilingual activation state requires translators and interpreters to activate highly complex cognitive mechanisms, including the inhibition of the non-target language, the dynamic retrieval of language knowledge and the regulation of activation levels in both languages. This cognitive regulation places greater demands on working memory (Bialystok et al., 2008, p. 860). The use of explicit structures is a strategy for reducing cognitive load and optimizing the allocation of memory resources. However, the cognitive mechanisms of language activation always occur within specific communicative contexts and serve specific communicative purposes (Halverson & Kotze, 2021). Therefore, it is necessary to approach language activation from both cognitive and social perspectives.

From a cognitive perspective, the Complexity Principle holds that 'language users tend to use more explicit grammatical forms in cognitively complex environments' (Rohdenburg, 1996, p. 149). One response to such processing demands is analyticity, that is, the use of more transparent and regular mappings between form and function, often realized through explicit lexicogrammatical encoding (Partridge, 2024). Mesthrie (2012, p. 99) similarly argues that increased analyticity can facilitate L2 processing because greater overt surface marking reduces cognitive complexity by making grammatical relations more transparent. This account is particularly relevant to the present study, in which the translators and interpreters produced English as an L2.

From a social perspective, communicative risks have been argued to increase in bilingual settings (Lanstyák & Heltai, 2012), potentially encouraging translators to minimize semantic ambiguity by selecting communicatively and normatively safer

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 17:56:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/langcog.2026.10106

forms (Pym, 2015, p. 70). The split at node 11 in Section 4.3 lends support to this account. Within a lexical environment that strongly favors complementizer omission, translated English instead favors the complementizer *that*, whereas the other three varieties collectively favor zero. This distribution closely matches the pattern used by Kruger (2019, p. 1) to test the pragmatic risk-aversion hypothesis: translators appear to favor the communicatively and normatively 'safer' option of retaining the complementizer in contexts in which non-translated English would typically favor omission. The contrast at node 25, however, calls for a more cautious interpretation. Here, interpreted English patterns with original written English, with both showing categorical retention of *that*, whereas original spoken English exhibits only a low rate of omission. Since this lexical environment already strongly favors *that*, and interpreted English does not differ from original written English, the result provides at best limited evidence for pragmatic risk aversion in interpreting. It may instead reflect the greater analyticity of interpreted language and the pressures imposed by online processing.

5.2. Text production dimension

The text production dimension emphasizes whether language production depends on prior texts and how this dependence constrains language choices (Kotze & Van Rooy, 2024, p. 12). In both translation and interpreting contexts, the source language is one of the most important factors directly influencing language production.

From a typological perspective, Chinese relies heavily on implicit coherence, making little use of the explicit connectives that are pervasive in English (Lian, 2010, pp. 77–78). Logically, the frequency of *that* in translated and interpreted English should be influenced by the *source-language shining-through effect* (Teich, 2003, pp. 145–146), resulting in a lower frequency of *that* than in original English. However, empirical results show the opposite trend: the frequency of *that* in translated and interpreted English is significantly higher than in the original English. This apparent anomaly may be partly explained by the typological and morphosyntactic distance between Chinese and English. Chinese, as a Sinitic language, is highly analytic and tends to leave interclausal relations implicit, whereas English, a Germanic language, relies more heavily on overt grammatical marking and explicit clausal linkage. Translating or interpreting between such typologically distant languages may therefore impose additional cognitive effort, which in turn weakens the expected *source-language shining-through effect*. Within the constrained communication framework, the language activation dimension may become the dominant constraint in such contexts, overriding source-language influence on target-language grammar (Kotze & Van Rooy, 2024). More specifically, under conditions of intense bilingual activation, typological distance and cognitive load, translators and interpreters may opt for the explicit *that* to reduce processing demands and enhance clarity and stability, thereby producing a higher frequency of *that*.

Furthermore, the translator's target-language proficiency and genre sensitivity also influence the choice of grammatical structures in the translation. Highly proficient translators often generalize and normalize expressions in the target language, which can diminish the influence of source-language structures (Venuti, 2018). The translations in the corpus selected for this article are all from professional translators with high levels of linguistic proficiency in the target language and familiarity with target language usage norms. This phenomenon is consistent with Hansen-Schirra

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 17:56:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/langcog.2026.10106

and Steiner's (2012, p. 272) concept of the norm continuum, which posits that *source-language shining-through* and *translation normalization* (Baker, 1996, p. 183) are not binary opposites but rather points along a dynamic continuum. Translators navigate this continuum by balancing the two tendencies in response to task demands, contextual pressures and their own abilities, thereby directly shaping the linguistic features manifested in the translated texts.

5.3. Modality and register dimension

Modality and register are fundamental dimensions of language use, as communicative goals and participant constellations in a given context profoundly shape linguistic behavior. As demonstrated in Section 4.1, all varieties analyzed in this study exhibit a notably high frequency of *that* compared to prior research (Kruger, 2019).

This higher frequency may partly reflect, first, the specific register under investigation: political discourse. Schäffner (1997, p. 3) emphasizes that politicians often speak as representatives of their parties, which limits what they can do and how they can express themselves. In this sensitive context, translation and interpreting are not merely auxiliary but are integral parts of political activity (Schäffner & Bassnett, 2010, p. 13). Consequently, the demands placed on translators and interpreters in this domain may be intensified by the high-stakes nature of political discourse and by their potential accountability for perceived errors (Buri, 2015). Although the present study does not directly compare registers, these pressures offer one possible explanation for the widespread use of the complementizer *that* across the four varieties.

Second, the sociopragmatic dimensions of interaction and intersubjective alignment manifest differently in spoken and written modalities, which may account for more pronounced risk avoidance in translation than in interpreting. Real-time speech imposes processing constraints but simultaneously permits adjustment, correction and interaction on the basis of continuous feedback, enabling interpreters to adapt to interlocutors and audiences to secure communicative success. Written discourse, by contrast, does not face the challenges of immediacy but lacks real-time feedback and interactivity. Consequently, translators working with written texts must rely on more careful and explicit expression to anticipate how readers will best interpret the discourse, since writing lacks the shared spatiotemporal context of spoken interaction for co-constructing meaning. This may help explain why written translators are more likely to adopt risk-averse strategies in order to align with readers' expectations and the normative conventions of the target language (Kotze & Van Rooy, 2024). Moreover, in contrast to the typically more fragmented input of interpreting, the source texts in written translation are generally characterized by higher syntactic complexity. This suggests that the divergence in the rate of *that* use may stem from intermodal differences in the density of the source material.

6. Conclusion

This study investigates the distribution of *that*- and zero-complementizer across translated, interpreted, original written and original spoken English from an intermodal perspective, using distinctive collexeme analysis and multivariate analysis. The findings can be summarized as follows. At the frequency level, both constrained varieties show significantly higher rates of that than their unconstrained counterparts,

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 17:56:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/langcog.2026.10106

and translated English shows a significantly higher rate than interpreted English. These patterns are consistent with a general tendency toward explicitation. Distinctive collexeme analysis reveals different semantic – constructional profiles across the four varieties. Translated English is characterized by a broad combination of private and suasive verbs, interpreted English by a narrower but individually stronger concentration of private verbs, original written English predominantly by public verbs, and original spoken English by a restricted set of public and private verbs. Multifactorial analysis indicates that the four varieties appear to share a broadly similar constraint system, while production mode introduces conditional differences in specific environments. Taken together, this study provides intermodal empirical evidence from the Chinese–English language pair for increased explicitness in constrained language. It should be noted that the syntactic complexity of the source text may also play a role, as the observed effects on the complementizer *that* might stem from inherent complexity differences in the source language. Future studies should incorporate source-language syntactic complexity as a potential predictor to further disentangle the multifaceted drivers of *that*/zero alternation in mediated language.

Data availability statement. The data, R scripts and outputs are available on the OSF: https://osf.io/h83fq/.

Competing interests. The authors declare none.

References

Baayen, R. H., Janda, L. A., Nesset, T., Endresen, A., & Makarova, A. (2013). Making choices in Russian: Pros and cons of statistical methods for rival forms. *Russian Linguistics*, 37(3), 253–291. https://doi.org/10.1007/s11185-013-9118-6

Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (pp. 233–250). John Benjamins. https://doi.org/10.1075/z.64.15bak

Baker, M. (1996). Corpus-based translation studies: The challenges that lie ahead. In H. Somers (Eds.), *Terminology, LSP and translation: Studies in language engineering in honour of Juan C. Sager* (pp. 175–188). John Benjamins. https://doi.org/10.1075/btl.18.17bak

Becher, V. (2010). Abandoning the notion of "translation-inherent" explicitation: Against a dogma of translation studies. *Across Languages and Cultures*, 11(1), 1–28. https://doi.org/10.1556/acr.11.2010.1.1

Bialystok, E., Craik, F. I. M., & Luk, G. (2008). Cognitive control and lexical access in younger and older bilinguals. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(4), 859–873. https://doi.org/10.1037/0278-7393.34.4.859

Blum-Kulka, S. (1986). Shifts of cohesion and coherence in translation. In J. House & S. Blum-Kulka (Eds.), *Interlingual and intercultural communication: Discourse and cognition in translation and second language acquisition studies* (pp. 17–36). Gunter Narr.

Buri, M. R. (2015). *Interpreting in diplomatic settings*. AIIC. https://www.mariarosariaburi.it/en/interpreting-in-diplomatic-settings-2/

De Sutter, G., & Lefer, M. A. (2020). On the need for a new research agenda for corpus-based translation studies: A multi-methodological, multifactorial and interdisciplinary approach. *Perspectives*, 28(1), 1–23. https://doi.org/10.1080/0907676x.2019.1611891

De Sutter, G., & Vermeire, E. (2019). Grammatical optionality in translations: A multifactorial corpus analysis of *that*/zero alternation in English using the MuPDAR approach. In L. Vandevoorde, J. Daems, & B. Defrancq (Eds.), *New empirical perspectives on translation and interpreting* (pp. 13–37). Routledge. https://doi.org/10.4324/9780429030376-2

Dor, D. (2005). Toward a semantic account of *that*-deletion in English. *Linguistics*, 43(2), 345–382. https://doi.org/10.1515/ling.2005.43.2.345

Ferraresi, A. (2019). Collocations in contact: Exploring constrained varieties of English through corpora. *Textus: English Studies in Italy*, 32(1), 203–222. https://doi.org/10.7370/93190

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 17:56:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/langcog.2026.10106

Ferraresi, A., & Bernardini, S. (2023). Comparing collocations in translated and learner language: In search of a method. *International Journal of Learner Corpus Research*, 9(1), 125–153. https://doi.org/10.1075/ijlcr.22012.fer

Friendly, M., & Meyer, D. (2015). *Discrete data analysis with R: Visualization and modeling techniques for categorical and count data*. CRC Press.

Gilquin, G., & Meriläinen, L. (2024). Constrained communication in EFL and ESL. *English World-Wide*, 45(2), 196–223. https://doi.org/10.1075/eww.23001.gil

Gries, S. T. (2020). On classification trees and random forests in corpus linguistics: Some words of caution and suggestions for improvement. *Corpus Linguistics and Linguistic Theory*, 16(3), 617–647. https://doi.org/10.1515/cllt-2018-0078

Gries, S. T. (2024). *Coll.analysis 4.1: A script for performing collostructional analyses* [Computer software]. https://www.stgries.info/teaching/groningen/index.html

Gries, S. T., & Stefanowitsch, A. (2004). Extending collostructional analysis: A corpus-based perspective on "alternations. *International Journal of Corpus Linguistics*, 9(1), 97–129. https://doi.org/10.1075/ijcl.9.1.06gri

Halverson, S. L., & Kotze, H. (2021). Sociocognitive constructs in translation and interpreting studies (TIS): Do we really need concepts like norms and risk when we have a comprehensive usage-based theory of language? In S. L. Halverson & Á. M.García (Eds.), *Contesting epistemologies in cognitive translation and interpreting studies* (pp. 51–79). Routledge. https://doi.org/10.4324/9781003125792-5

Hansen-Schirra, S., Neumann, S., & Steiner, E. (2007). Cohesive explicitness and explicitation in an English-German translation corpus. *Languages in Contrast*, 7(2), 241–265. https://doi.org/10.1075/lic.7.2.09han

Hansen-Schirra, S., & Steiner, E. (2012). 14 towards a typology of translation properties. In S. Hansen-Schirra, S. Neumann, & E. Steiner (Eds.), *Cross-linguistic corpora for the study of translations: Insights from the language pair English-German* (pp. 255–280). De Gruyter Mouton. https://doi.org/10.1515/9783110260328.255

Hawkins, J. (2014). *Cross-linguistic variation and efficiency*. Oxford University Press.

Hothorn, T., & Zeileis, A. (2015). Partykit: A modular toolkit for recursive partytioning in R. *Journal of Machine Learning Research*, 16(118), 3905–3909.

Hu, K., & Tao, Q. (2013). The Chinese-English conference interpreting corpus: Uses and limitations. *Meta: Translators' Journal*, 58(3), 626–642. https://doi.org/10.7202/1025055ar

Ivaska, I., Ferraresi, A., & Bernardini, S. (2022). Syntactic properties of constrained English: A corpus-driven approach. In S. Granger & M.-A. Lefer (Eds.), *Extending the scope of corpus-based translation studies* (pp. 133–157). Bloomsbury. https://doi.org/10.5040/9781350143289.0013

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, 61(1), 23–62. https://doi.org/10.1016/j.cogpsych.2010.02.002

Kajzer-Wietrzny, M. (2012). Interpreting universals and interpreting style *[Doctoral dissertation]*. Adam Mickiewicz University.

Kajzer-Wietrzny, M. (2018). Interpretese vs. non-native language use: The case of optional *that*. In M. Russo, C. Bendazzoli, & B. Defrancq (Eds.), *Making way in corpus-based interpreting studies* (pp. 97–113). Springer Singapore. https://doi.org/10.1007/978-981-10-6199-8_6

Kajzer-Wietrzny, M. (2022). An intermodal approach to cohesion in constrained and unconstrained language. *Target*, 34(1), 130–162. https://doi.org/10.1075/target.19186.kaj

Kajzer-Wietrzny, M., & Ivaska, I. (2020). A multivariate approach to lexical diversity in constrained language. *Across Languages and Cultures*, 21(2), 169–194. https://doi.org/10.1556/084.2020.00011

Kajzer-Wietrzny, M., Ivaska, I., & Ferraresi, A. (2021). 'Lost' in interpreting and 'found' in translation: Using an intermodal, multidirectional parallel corpus to investigate the rendition of numbers. *Perspectives*, 29(4), 469–488. https://doi.org/10.1080/0907676x.2020.1860097

Kotze, H. (2019). Converging what and how to find out why. In L. Vandevoorde, J. Daems, & B. Defrancq (Eds.), *New empirical perspectives on translation and interpreting* (pp. 333–371). Routledge. https://doi.org/10.4324/9780429030376-14

Kotze, H. (2022). Translation as constrained communication. In S. Granger & M.-A. Lefer (Eds.), *Extending the scope of corpus-based translation studies* (pp. 67–97). Bloomsbury. https://doi.org/10.5040/9781350143289.0010

Kotze, H., & Van Rooy, B. (2024). Introduction: The constrained communication framework for studying contact-influenced varieties. In B. Van Rooy & H. Kotze (Eds.), *Constraints on language variation and change in complex multilingual contact settings* (pp. 1–28). John Benjamins. https://doi.org/10.1075/coll.60.01kot

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 17:56:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/langcog.2026.10106

Kruger, H. (2019). *That* again: A multivariate analysis of the factors conditioning syntactic explicitness in translated English. *Across Languages and Cultures*, 20(1), 1–33. https://doi.org/10.1556/084.001

Kruger, H., & De Sutter, G. (2018). Alternations in contact and non-contact varieties: Reconceptualising *that*-omission in translated and non-translated English using the MuPDAR approach. *Translation, Cognition & Behavior*, 1(2), 251–290. https://doi.org/10.1075/tcb.00011.kru

Kruger, H., & Van Rooy, B. (2012). Register and the features of translated language. *Across Languages and Cultures*, 13(1), 33–65. https://doi.org/10.1556/acr.13.2012.1.3

Kruger, H., & Van Rooy, B. (2016a). Constrained language: A multidimensional analysis of translated English and a non-native indigenised variety of English. *English World-Wide*, 37(1), 26–57. https://doi.org/10.1075/eww.37.1.02kru

Kruger, H., & Van Rooy, B. (2016b). Syntactic and pragmatic transfer effects in reported-speech constructions in three contact varieties of English influenced by Afrikaans. *Language Sciences*, 56, 118–131. https://doi.org/10.1016/j.langsci.2016.04.003

Lanstyák, I., & Heltai, P. (2012). Universals in language contact and translation. *Across Languages and Cultures*, 13(1), 99–121. https://doi.org/10.1556/acr.13.2012.1.6

Lian, S. (2010). *Yinghan duibi yanjiu* 英汉对比研究 *[Contrastive studies of English and Chinese]* (Rev. ed.) Higher Education Press.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, 2(3), 18–22.

Ma, R., Jiang, Y., & Qian, J. (2024). Syntactic and typological properties of constrained language: A study based on dependency treebanks of translated English and non-native English. *Across Languages and Cultures*, 25(1), 91–111. https://doi.org/10.1556/084.2023.00363

Ma, X. (2021). Coping with syntactic complexity in English-Chinese sight translation by translation and interpreting students: An eye-tracking investigation. *Across Languages and Cultures*, 22(2), 192–213. https://doi.org/10.1556/084.2021.00014

Mesthrie, R. (2012). Deletions, antideletions and complexity theory, with special reference to black south African and Singaporean Englishes. In B. Kortmann & B. Szmrecsanyi (Eds.), *Linguistic complexity: Second language acquisition, indigenization, contact* (pp. 90–100). De Gruyter Mouton. https://doi.org/10.1515/9783110229226.90

Molés-Cases, T. (2019). Why typology matters: A corpus-based study of explicitation and implicitation of manner-of-motion in narrative texts. *Perspectives*, 27(6), 890–907. https://doi.org/10.1080/0907676x.2019.1580754

Morselli, N. (2014). *Gli universali dell'interpretazione: Studio sull'esplicitezza nel corpus intermodale EPTIC* [Doctoral dissertation]. University of Bologna.

Olohan, M., & Baker, M. (2000). Reporting *that* in translated English: Evidence for subconscious processes of explicitation? *Across Languages and Cultures*, 1(2), 141–158. https://doi.org/10.1556/acr.1.2000.2.1

Pan, J. (2019). *The Chinese/English political interpreting corpus (CEPIC)*. Hong Kong Baptist University Library. https://digital.lib.hkbu.edu.hk/cepic/

Partridge, M. (2024). Investigating the complementiserthatin the verb complementation of black south African English. In B. Van Rooy & H. Kotze (Eds.), *Constraints on language variation and change in complex multilingual contact settings* (pp. 87–119). John Benjamins. https://doi.org/10.1075/coll.60.04par

Procter, P. (Eds.). (1995). *Cambridge international dictionary of English*. Cambridge University Press.

Pym, A. (2005). Explaining explicitation. In K. Károly & Á. Fóris (Eds.), *New trends in translation studies: In honour of Kinga Klaudy* (pp. 29–34). Akadémiai Kiadó.

Pym, A. (2015). Translating as risk management. *Journal of Pragmatics*, 85, 67–80. https://doi.org/10.1016/j.pragma.2015.06.010

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. Longman.

Rohdenburg, G. (1996). Cognitive complexity and increased grammatical explicitness in English. *Cognitive Linguistics*, 7(2), 149–182. https://doi.org/10.1515/cogl.1996.7.2.149

Schäffner, C. (1997). Editorial: Political speeches and discourse analysis. In C. Schäffner (Eds.), *Analysing political speeches* (pp. 1–4). Multilingual Matters.

Schäffner, C., & Bassnett, S. (2010). Politics, media and translation: Exploring synergies. In C. Schäffner & S. Bassnett (Eds.), *Political discourse, media and translation* (pp. 1–31). Cambridge Scholars Publishing.

Séguinot, C. (1988). Pragmatics and the explicitation hypothesis. *TTR: Traduction, Terminologie, Rédaction*, 1(2), 106–113. https://doi.org/10.7202/037024ar

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 17:56:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/langcog.2026.10106

Shook, A., & Marian, V. (2012). Bimodal bilinguals co-activate both languages during spoken comprehension. *Cognition*, 124(3), 314–324. https://doi.org/10.1016/j.cognition.2012.05.014

Steiner, E. (2001). Intralingual and interlingual versions of a text—How specific is the notion of "translation"? In E. Steiner & C. Yallop (Eds.), *Exploring translation and multilingual text production: Beyond content* (pp. 161–190). De Gruyter Mouton. https://doi.org/10.1515/9783110866193.161

Tagliamonte, S. A., & Smith, J. (2005). No momentary fancy! The zero 'complementizer' in English dialects. *English Language and Linguistics*, 9(2), 289–309. https://doi.org/10.1017/s1360674305001644

Teich, E. (2003). *Cross-linguistic variation in system and text: A methodology for the investigation of translations and comparable texts*. De Gruyter Mouton. https://doi.org/10.1515/9783110896541

Thompson, S. A., & Mulac, A. (1991). The discourse conditions for the use of the complementizer *that* in conversational English. *Journal of Pragmatics*, 15(3), 237–251. https://doi.org/10.1016/0378-2166(91)90012-m

Torres Cacoullos, R., & Walker, J. A. (2009). On the persistence of grammar in discourse formulas: A variationist study of *that*. *Linguistics*, 47(1), 1–43. https://doi.org/10.1515/ling.2009.001

Van Rooy, B., & Kruger, H. (2016). Faktore wat die weglating van die Afrikaanse onderskikker *dat* bepaal. *Tydskrif vir Geesteswetenskappe*, 56(1), 102–116. https://doi.org/10.17159/2224-7912/2016/v56n1a7

Venuti, L. (2018). *The translator's invisibility: A history of translation* (Rev. ed.) Routledge.

Vinay, J.-P., & Darbelnet, J. (1995). *Comparative stylistics of French and English: A methodology for translation*. John Benjamins. https://doi.org/10.1075/btl.11

Wulff, S., Gries, S. T., & Lester, N. (2018). Optional *that* in complementation by German and Spanish learners. In A. Tyler, L. Huang, & H. Jan (Eds.), *What is applied cognitive linguistics? Answers from current SLA research* (pp. 99–120). De Gruyter Mouton. https://doi.org/10.1515/9783110572186-004

Wulff, S., Lester, N., & Martinez-Garcia, M. (2014). *That*-variation in German and Spanish L2 English. *Language and Cognition*, 6(2), 271–299. https://doi.org/10.1017/langcog.2014.5

Wulff, S., Stefanowitsch, A., & Gries, S. T. (2007). Brutal Brits and persuasive Americans. In G. Radden, T. Berg, K. Köpcke, & P. Siemund (Eds.), *Aspects of meaning construction* (pp. 265–281). John Benjamins. https://doi.org/10.1075/z.136.17wul

Cite this article: Zhao, Q., & Lin, S. (2026). An intermodal approach to the complementizer *that* in constrained and unconstrained language, *Language and Cognition*, 18, e58, 1–24. https://doi.org/10.1017/langcog.2026.10106

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 17:56:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/langcog.2026.10106