CrossMark

English Language and Linguistics 20.3: 533–558. © Cambridge University Press 2016
doi:10.1017/S1360674316000344

Variation across two dimensions: testing the Complexity Principle and the Uniform Information Density Principle on adjectival data[1]

HENRIK KAATARI
Uppsala University
(Received 2 October 2015; revised 20 June 2016)

This study tests the applicability of the Complexity Principle (Rohdenburg 1996) and the Uniform Information Density Principle (Jaeger 2010) on adjectival data as regards the variation between retaining and omitting the complementizer *that* in English adjectival complementation constructions. More specifically, the study tests the effect of different factors of potential importance on this variation across extraposed (e.g. *It was inevitable (that) he should be nicknamed 'the Ferret'*) and post-predicate clauses (e.g. *I'm happy (that) we are married*). While both the factors concerned with the Complexity Principle and the Uniform Information Density Principle are found to have an effect on post-predicate clauses, less clear effects are found concerning extraposed clauses. I attribute these findings to the difference between the two constructions in terms of their frequency of co-occurrence with different matrix subject types and with different adjectives.

1 Introduction

This study is concerned with the variation between retaining and omitting the complementizer *that* in English adjectival complementation constructions (e.g. *I'm happy (that) we are married*). The topic is explored with regard to which factors prove to be most important in explaining the choice between these two variants. The factors included in the analysis are those which in previous studies have been shown to have an effect on this particular variation. However, most of these studies have been concerned with *verbal* complement-taking predicates, thereby neglecting to examine the effect of these factors on adjectival data (see Thompson & Mulac 1991b; Rohdenburg 1995, 1996; Jaeger 2010). Furthermore, the choice between these two variants has not been examined from the point of view of comparing the effect of different factors across extraposed (e.g. *It was inevitable (that) he should be nicknamed 'the Ferret'*) and post-predicate clauses (*I'm happy (that) we are married*) (terminology from Biber *et al.* 1999; see Kaltenböck 2006 for a study on extraposed clauses only). As a consequence, the aim of this study is twofold: (i) to test the effect of different factors on adjectival data as regards the variation between *that*-omission and *that*-retention, and (ii) to test whether there are any differences between extraposed and post-predicate clauses where the effect of those factors is concerned.

[1] I would like to express my gratitude to the members of the Uppsala University English linguistics seminar and two anonymous reviewers for their comments and suggestions on earlier versions of this article.

https://doi.org/10.1017/S1360674316000344 Published online by Cambridge University Press

The article starts by examining previous research on the variation between retaining and omitting *that* in section 2, with emphasis placed on the Complexity Principle (Rohdenburg 1996) and the Uniform Information Density Principle (Jaeger 2010). The data and the design of the study are described in section 3. In section 4, the results of the analyses are presented, with the factors presented in sections 4.1–4.5. Section 4 ends with the presentation of multivariate analyses, establishing the effect of the different factors on the choice between retaining and omitting *that*. Finally, a discussion of the results and a summary are given in section 5.

2 Introducing two principles

Regarding the variation between *that*-omission and *that*-retention, there are two principles that merit a short discussion. Due to the bias towards verbal complementation in previous studies, both these principles are concerned with verbs complemented by *that*-clauses, but they can also be applied to adjectives, as will be shown in sections 3 and 4. The first of these two is the Complexity Principle, which states that '[i]n the case of more or less explicit grammatical options the more explicit one(s) will tend to be favored in cognitively more complex environments' (Rohdenburg 1996: 151).[2] In testing the Complexity Principle, Rohdenburg (1995, 1996) is mainly concerned with variable uses of alternating constructions. As *that*-retention is a more explicit grammatical option than *that*-omission, Rohdenburg (1996) expects *that*-retention to be favored in cognitively more complex environments. In analysing the variable use of *that* to introduce complement clauses, Rohdenburg focuses on two different complexity factors, illustrated in (1) and (2).[3]

(1) He told me **yesterday** that **the young lady** had gone away.
(2) He told me **she** had gone away.

The first of these two complexity factors is concerned with the separation between the matrix clause and the complement clause, as illustrated in (1) by *yesterday*.[4] According to the Complexity Principle, *that*-retention, as in (1), should be favoured since 'any elements capable of delaying the processing of the object clause and thus of the overall sentence structure favor the use of an explicit signal of subordination' (Rohdenburg 1996: 161). Rohdenburg (1996) points out that, traditionally, this factor has been attributed to structural ambiguity, since the adverbial phrase could in theory be seen as

[2] An earlier definition of the Complexity Principle, also referred to as the Transparency Principle, states that '[t]he less directly the dependent clause is linked to its superordinate clause, or the more complex the dependent clause turns out to be, the greater is the need to make its sentential status more explicit' (Rohdenburg 1995: 368).

[3] The examples in (1)–(2) are adopted from Rohdenburg (1996: 160), but changed slightly for illustrative purposes.

[4] Rohdenburg (1996) is not the first to note the influence of intervening adverbs on *that*-retention (see Elsness 1984; Thompson & Mulac 1991b for similar observations), but his formulation of the Complexity Principle, and the way it encompasses several different variation phenomena, are of special importance in the present context.

https://doi.org/10.1017/S1360674316000344 Published online by Cambridge University Press

part of the matrix clause or the complement clause. However, the crucial point made by Rohdenburg (1996) is that any element in this position, regardless of whether it causes structural ambiguity, can be seen as a complexity factor as it increases the likelihood of retention of *that*. The second complexity factor is concerned with the realization of the subject of the complement clause. The Complexity Principle predicts that the more complex the subject of the complement clause is, the more likely the retention of *that* becomes. This is manifested by a striking contrast between NPs headed by pronouns and NPs headed by nouns, in that NPs headed by nouns, as in (1), favour *that*-retention whereas NPs headed by pronouns, as in (2), favour *that*-omission.[5]

The second principle of importance for the choice between retaining and omitting *that* is referred to as the Uniform Information Density Principle (UID) (Jaeger 2010). UID is concerned with the influence of lexical items on the choice between grammatical options. UID predicts that 'the more probable (expected) a word is in its context, the less information it carries (the more redundant it is) in that context' (Jaeger 2010: 24).[6] The idea behind UID is that not all words provide the same amount of information, and that speakers aim to distribute information uniformly in production. The way in which this works in practice is illustrated by the difference between (3) and (4) (from Jaeger 2010: 127).

(3) My boss **confirmed** that we were absolutely crazy.
(4) My boss **thinks** I am absolutely crazy.

Retaining *that*, as in (3), distributes the information at the onset of the complement clause over one extra word (i.e. *that*), thereby lowering information density. Omitting *that*, on the other hand, as in (4), increases information density, as the information is distributed over one less word (see Jaeger 2010: 27 for an illustrative figure explaining the difference in information density between (3) and (4)). UID takes the information at the complement clause onset as conditioned by the matrix predicate (i.e. CONFIRM and THINK in (3) and (4)). UID predicts that the lower the information carried by the matrix predicate, the more likely speakers are to omit *that* and vice versa. The information carried by matrix predicates is based on how expected a predicate is in a given context. This is calculated by contrasting the frequency of co-occurrence of a given predicate with *that*-clauses with the frequency of the same predicate in the entire corpus. This means that the less information a given predicate carries, the more expected it is to be followed by a *that*-clause. This is the situation we have in (3) and (4), as CONFIRM carries more information than THINK, and is thus less expected to be followed by a *that*-clause than THINK. Following this, verbs found to be frequently complemented by *that*-clauses, such as THINK, are predicted to be more likely to occur with *that*-omission than verbs found less frequently complemented by *that*-clauses, such as CONFIRM.

[5] See footnote 4 (and also Elsness 1984; Thompson & Mulac 1991b).
[6] Note that the UID could be subsumed under the more general Complexity Principle, since frequency and entrenchment are negatively correlated with cognitive complexity. Frequent or entrenched structures are easier to process than rare ones (other things being equal).

https://doi.org/10.1017/S1360674316000344 Published online by Cambridge University Press

In addition to these two principles and the factors they focus on, there are several other factors that have been found to influence the choice between retaining and omitting *that*. Perhaps the most widely discussed factor is the formality of different genres. Several studies have found that *that*-omission is more frequent in informal genres (see Elsness 1984; Biber 1999; Kaltenböck 2006; Mindt 2011). Other factors discussed in previous studies include the type of matrix clause subject and its interaction with different matrix verbs. In studies on grammaticalization, certain combinations of matrix clause subjects and verbs (e.g. *I think*) are found to frequently co-occur with *that*-omission, and are treated as reanalysed epistemic phrases (see Thompson & Mulac 1991a; Van Bogaert 2011). According to Thompson & Mulac (1991a), the pronominal matrix subjects *I* and *you* are the two pronouns most likely to occur with *that*-omission (see also Torres Cacoullos & Walker 2009; Jaeger 2010).

3 Data and design of the study

The data used in this study consist of a 3-million-word subcorpus of the BNC (*British National Corpus*; see Burnard 2007). The subcorpus has been compiled to facilitate comparisons across genres, with five genres represented: academic prose, popular science, news, fiction and conversation. The subcorpus has been carefully sampled to include as many different texts as possible. Rather than sampling whole texts, samples of 10,000 words were drawn.[7] Each genre contains 600,000 words and 60 different samples, making sure that all genres and subgenres are represented by an equal number of words as well as by equal amounts of text-files (see Kaatari 2012 for more detailed information about the subcorpus). Furthermore, an equal number of samples were drawn from the beginning, the middle and the end of text files to minimize any skewing due to textual colligation and clustering (see Hoey 2005; Evert 2006).

The constructions included in the analysis are extraposed and post-predicate clauses. In both these constructions, omitting *that* is optional, as seen in (5)–(8).[8,9]

(5) Nevertheless, it is *noteworthy* from the annotation of their books that they use translations wherever possible […]. (A1A 199) (Extraposed: +*that*)

[7] The s-unit (the proxy for sentences in the BNC; see Burnard 2007: section 2.4) was taken as the smallest unit sampled in order to ensure, as far as possible, that the corpus would include full sentences. Note also that text files smaller than 10,000 words were excluded.

[8] Examples in this article followed by a six-character code (indicating the file name and the position of the token) are taken from the BNC. Note also that the adjectival matrix predicate is shown italicized throughout the article.

[9] Extraposed object clauses (e.g. *I made it clear (that) she would go*) are not included in the study on the grounds that they do not include non-referring *it* in matrix subject position. Similarly, it is also possible to distinguish a class of post-predicate object clauses (e.g. *He made her aware (that) they were coming soon*). Tokens such as these have also been excluded due to reasons of comparability with extraposed clauses. Finally, a related class of tokens, represented by a closed set of verb + adjective combinations (almost exclusively restricted to the verb MAKE, as in *They made sure (that) there's nothing left*) have also been excluded. Following Huddleston & Pullum (2002: 289), these are treated as 'verbal idioms', i.e. as complex verbs, and are as such not considered instances of adjectival complementation (see also Biber *et al.* 1999: 931–2 and Mindt 2011: 77ff. for a comprehensive survey of the literature on these verb + adjective combinations).

https://doi.org/10.1017/S1360674316000344 Published online by Cambridge University Press

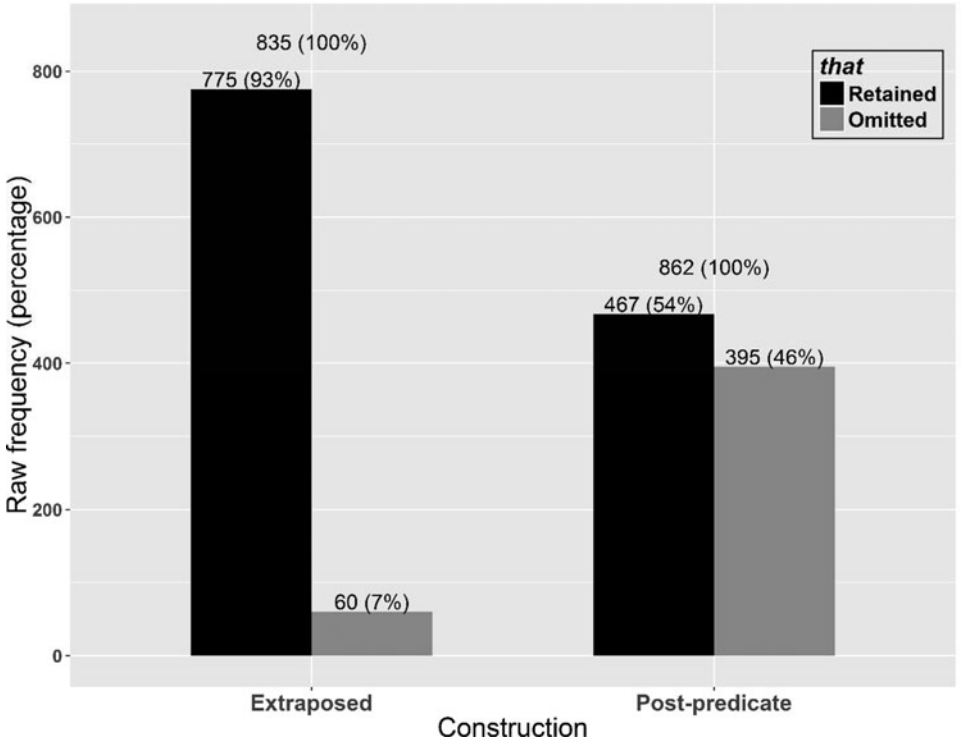


Figure 1. Overview of the data

(6) It was *certain* the display would be repeated. (B09 1162) (Extraposed: *–that*)
(7) I was *glad* then that he did not know the subject of the talk I had just given at the conference […]. (H78 1007) (Post-predicate: *+that*)
(8) 'I'm *sure* no one would have noticed.' (GUU 3674) (Post-predicate: *–that*)

The data used in this study were extracted via a customized program called SVEP (see Garretson & Kaatari 2014). A guiding principle was to include as many instances of adjectives complemented by *that*-clauses as possible, thus avoiding lexical lists as a starting point for the data collection process (see Mindt 2011; Van linden 2012). Accordingly, the dataset has no bias towards certain lexemes, and it also includes instances in which the adjective and the *that*-clause are separated (see Garretson & Kaatari 2014: 68 for detailed information on the patterns used for data extraction).

Figure 1 provides an overview of the data. Extraposed and post-predicate clauses are more or less evenly distributed as regards frequency (835 vs 862), but *that*-omission is far more frequent with post-predicate clauses (46 per cent) than with extraposed clauses (7 per cent).

The analysis in this study is based on five different factors, listed in table 1, and further illustrated in figure 2. Factors (i)–(v) are numbered according to their position in the sentence (from left to right), whereas factor (vi) is extralinguistic. Factor (ii) tests the underlying assumption of the Uniform Information Density Principle, although

https://doi.org/10.1017/S1360674316000344 Published online by Cambridge University Press

Table 1. *Factors included in the analysis*

Factor	Values
(i) MATRIX SUBJECT TYPE	*I*
	you
	Other Pronoun
	Other NP
(ii) COLLOSTRUCTION STRENGTH (UID)	Numeric variable
(iii) SEPARATION	Separation
	No separation
(iv) COMP. CLAUSE SUBJECT TYPE	*I*
	you
	Other Pronoun
	Other NP
(v) COMP. CLAUSE SUBJECT LENGTH	Numeric variable
(vi) GENRE	Academic Prose
	Popular Science
	News
	Novels
	Conversation

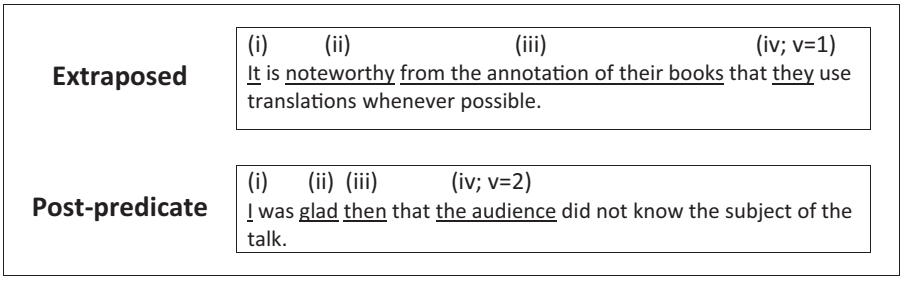


Figure 2. Illustration of constructions and factors included in the analysis

here operationalized as Collostruction Strength (see footnote 13, section 4.2 for a justification), and factors (iii), (iv) and (v) test the Complexity Principle. Factors (i)–(v) are further illustrated in figure 2 together with the constructions included in the study.

Note that factor (i) is only applicable to post-predicate clauses as extraposed clauses do not allow variation of the matrix subject type (see footnote 9). Note also that construction type (i.e. extraposed vs post-predicate) is not included as a factor. The reason behind this is that the influence of the factors listed in table 1 will be tested on extraposed and post-predicate clauses separately in order to determine to what extent the two construction types differ as regards the variation between retaining and omitting *that*. Extraposed and post-predicate clauses are thus treated as two distinct constructions in this study.

https://doi.org/10.1017/S1360674316000344 Published online by Cambridge University Press

4 Results

In this section, the results of each factor will be illustrated (sections 4.1–4.5) and section 4.6 presents a multivariate analysis, assessing the combined explanatory power of all factors.

4.1 Factor (i): MATRIX SUBJECT TYPE

MATRIX SUBJECT TYPE is implemented in this analysis as a variable with four levels: *I*, *you*, 'Other Pronoun', and 'Other NP'. The reason why the pronouns *I* and *you* are distinguished from other pronouns is that they have been found to be particularly salient in triggering *that*-omission (see Thompson & Mulac 1991a; Jaeger 2010). Matrix subject type 'Other NP' is represented by NPs headed by nouns and proper nouns.[10] The four types are illustrated in (9)–(12), with the matrix subject type underlined.

(9) I'm *sorry* I spoilt your sleep. (Type: *I*) (CML 1568)
(10) Are you *sure* you don't mean seven? (Type: *you*) (KNY 89)
(11) In a way he was *glad* that she was 'in trouble'. (Type: Other Pronoun) (CR6 886)
(12) The employers are *adamant* that an agreement on working hours must be part of the deal. (Type: Other NP) (A9W 245)

The influence of the matrix subject type on the choice between retaining and omitting *that* is shown in figure 3.

The two pronominal categories *I* and *you* are characterized by a much higher proportion of *that*-omission than the other categories. Furthermore, not only is *I* the subject type with the highest proportion of *that*-omission, but *I* is also more than twice as frequent as the second most frequent matrix subject type ('Other Pronoun'), and accounts for 50 per cent of the total number of tokens (431/862). It thus appears that there is a divide between the categories *I* and *you* on the one hand, and the two remaining categories on the other hand in terms of the effect of matrix subject type on the choice between retaining and omitting *that*. These findings are in line with findings from studies on verbal predicates in which *I* and *you* have been found to be particularly strong triggers of *that*-omission (see e.g. Thompson & Mulac 1991a; Jaeger 2010).

4.2 Factor (ii): COLLOSTRUCTION STRENGTH

Factor (ii) tests the Uniform Information Density Principle (UID) (Jaeger 2010) and the influence of particular adjectives on the choice between retaining and omitting *that*. UID is operationalized here as COLLOSTRUCTION STRENGTH (Stefanowitsch & Gries 2003), since it closely models the notion of how expected a given word is in a given context. Collostruction strength is a measure of the attraction/repulsion of different lexemes to one particular slot in a given construction, in this case the degree to

[10] Matrix clauses in which the matrix subject is not overtly expressed but can be inferred from the context (e.g. *'Glad you came?' he asked* (JXW 4039)) have been classified according to the class of the inferred subject.

https://doi.org/10.1017/S1360674316000344 Published online by Cambridge University Press

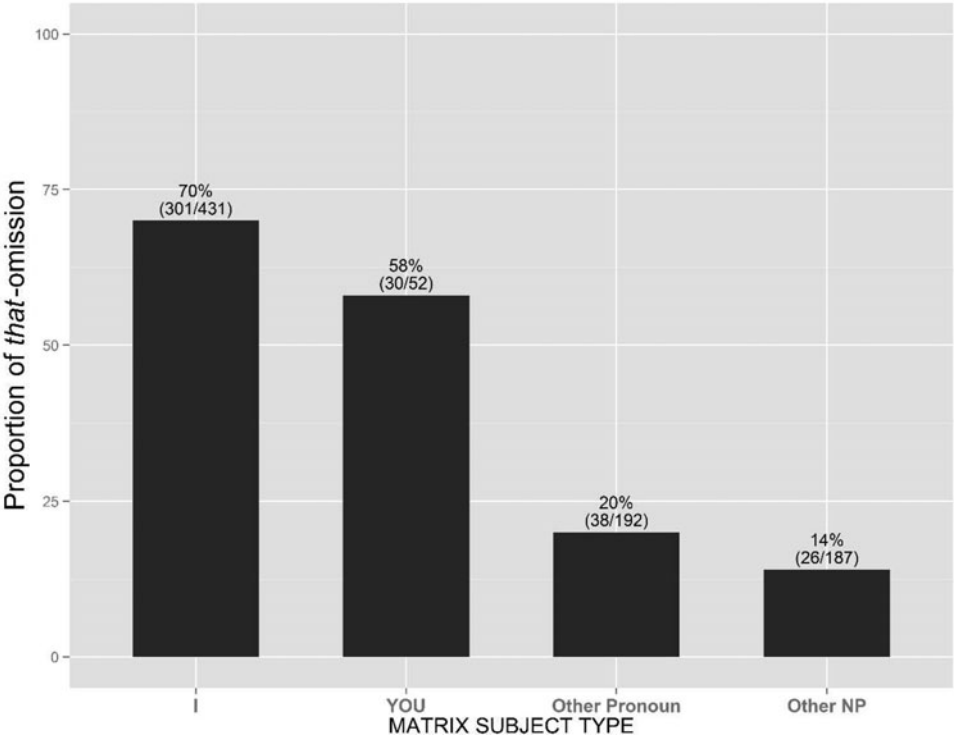


Figure 3. Proportion (raw frequency) of *that*-omission across matrix subject types

which individual adjectives are attracted to extraposed and post-predicate *that*-clauses respectively.[11] The appeal of using collostruction strength is that it can handle low frequency items, weighting them appropriately so as to avoid overestimating their importance. The operationalization of UID in Jaeger (2010), on the other hand, takes into account the frequency of a given word followed by a *that*-clause, divided by the frequency of the same word in the entire corpus, using the following formula:

$$\text{UID} = -\log p(\text{adj.}+\textit{that} \mid \text{adj. total})$$

This formula does not in itself address the problem of low-frequency items, as they receive very low information density scores (remember that the lower the UID score, the more expected a word is in its given context).[12]

[11] Collostruction strength is a measure derived from Collostructional Analysis, itself a cover term for three different analyses. The analysis performed here is called Collexeme Analysis (see Stefanowitsch & Gries 2003). See Gries & Stefanowitsch (2004) for more on Distinctive Collexeme Analysis and Stefanowitsch & Gries (2005) for more on Co-Varying Collexeme Analysis.

[12] The issue of correctly weighting low-frequency items is addressed, but not explained in detail, by Jaeger (2010). For the purposes of the current study, it was considered important to weigh infrequent adjectives in a systematic way, and collostruction strength was considered the most economical solution.

https://doi.org/10.1017/S1360674316000344 Published online by Cambridge University Press

The calculation of collostruction strength, on the other hand, takes into account the frequency of a given adjective followed by a *that*-clause, the frequency of that same adjective in all other constructions, the frequency of the construction (i.e. *that*-clauses) with other adjectives and the frequency of all other constructions with adjectives other than the given adjective (i.e. the total frequency of adjectives in the entire corpus) are also taken into account. These four frequencies are then entered into a 4-by-4 table, and submitted to a statistical test from which the collostruction strength value is derived (see Stefanowitsch & Gries 2003: 218–19; see also Hilpert 2010 for a more practical discussion).

To sum up the discussion thus far, UID is operationalized here slightly differently than in Jaeger (2010), using collostruction strength as a measure of how expected different adjectives are in a given context. Collostruction strength is a numeric variable, assigning higher values to adjectives which are strongly attracted to a particular construction, in this case *that*-clauses. In this study, the Collexeme Analysis (see Stefanowitsch & Gries 2003), from which the collostruction strength values are derived, was performed using the statistical software R (R Development Core Team 2008) by implementing a script written by Gries (2007). The statistical measure used for calculating collostruction strength is log-likelihood.[13] According to the prediction of the Uniform Information Density Principle, we would expect more *that*-omission where adjectives have a high collostruction strength.

Before examining the mean collostruction strengths of those cases where *that* was omitted versus those where it was retained, we first inspect the collostruction strength of different adjectives across extraposed and post-predicate clauses. Table 2 lists the top fifteen adjectives according to collostruction strength (COLL.) between the adjective and extraposed and post-predicate clauses respectively. The raw frequency of each adjective in the entire corpus (WFreq.) and the observed frequency of each adjective complemented by an extraposed or post-predicate clause (OFreq) are also given.

Looking only at the top fifteen adjectives, collostruction strength ranges between 2248.1 and 48.7 for post-predicate clauses, whereas the range among extraposed clauses is only 665.7 to 42.0.[14] In essence this means that *sure* is more strongly expected to be followed by a post-predicate clause than *clear* is expected to be followed by an extraposed clause. Looking at the (log-transformed) mean collostruction strength of extraposed and post-predicate clauses, we also find notable differences between the two constructions: extraposed 4.69; post-predicate 5.72.

Turning to the influence of collostruction strength on the choice between omitting and retaining *that*, we find further differences between the two constructions. The difference in mean collostruction strength between retaining and omitting *that* is

[13] Log-likelihood is used here instead of the default measure of the Fisher-Yates exact test. The reason for this is that the Fisher-Yates exact test assigns an infinite collostruction strength to strongly attracted collexemes, thus preventing collostruction strength to be used as a numeric variable. Log-likelihood is considered the preferred option when faced with this particular problem (Stefan Th. Gries p.c.).

[14] The full range of collostruction strengths of post-predicate clauses is 2248.1 to 0.4 and the corresponding range for extraposed clauses is 665.7 to 0.03.

https://doi.org/10.1017/S1360674316000344 Published online by Cambridge University Press

Table 2. *Top fifteen collexemes across extraposed and post-predicate constructions*

Extraposed				Post-predicate			
Adjective	WFreq	OFreq	COLL.	Adjective	WFreq	OFreq	COLL.
clear	634	110	665.7	*sure*	751	280	2248.1
likely	754	86	441.8	*aware*	327	82	562.6
possible	998	74	314.5	*convinced*	75	53	502.2
true	630	64	311.9	*glad*	120	57	472.2
unlikely	178	42	279.1	*afraid*	159	51	371.6
surprising	117	37	270.7	*confident*	98	26	177.1
obvious	302	42	230.9	*sorry*	425	31	126.8
probable	48	20	159.5	*concerned*	457	31	122.4
important	985	46	152.8	*surprised*	133	21	118.9
inevitable	97	20	126.4	*certain*	782	33	100.4
evident	89	18	112.9	*satisfied*	90	16	94.6
apparent	184	16	72.0	*disappointed*	69	14	86.8
essential	212	16	67.5	*pleased*	129	14	68.21
conceivable	12	7	62.2	*conscious*	114	11	50.9
significant	350	14	42.0	*worried*	126	11	48.7

Table 3. *Mean collostruction strength across constructions (log-transformed)*

Construction	+*that*	–*that*
Extraposed	4.70	4.48
Post-predicate	5.02	6.55

much smaller with extraposed clauses than it is with post-predicate clauses, as seen in table 3.

It is worth pointing out that a single adjective is potentially responsible for much of the mean difference in collostruction strength between the two post-predicate groups. The adjective *sure* has the highest collostruction strength and it is also the most frequent adjective, occurring 280 times, and thus accounts for 32 per cent of all post-predicate clauses, 73 per cent of which (203/280) are found with *that*-omission.

To sum up the results concerning the influence of collostruction strength on the choice between retaining and omitting *that*, we can say that this variable appears to play a larger role when it comes to post-predicate clauses. Furthermore, the results point in the direction predicted by UID, i.e. adjectives with higher collostruction strength are found to be more likely to occur with *that*-omission.

4.3 Factor (iii): SEPARATION

The first factor explicitly linked to the predictions of the Complexity Principle is concerned with separation between the adjectival matrix predicate and the complement

https://doi.org/10.1017/S1360674316000344 Published online by Cambridge University Press

Table 4a. *Raw frequency and percentage of separation (extraposed)*

	+that		*–that*		Total	
Separation	105	14%	0	0%	105	13%
No separation	670	86%	60	100%	730	87%
Total	775	100%	60	100%	835	100%

Table 4b. *Raw frequency and percentage of separation (post-predicate)*

	+that		*–that*		Total	
Separation	30	6%	0	0%	30	3%
No separation	437	94%	395	100%	832	97%
Total	467	100%	395	100%	862	100%

clause. Separation of this type is illustrated by the underlined material in (13) and (14).[15]

(13) It has become *clear* from all that we have said that the composition of the atmosphere is not fixed. (AMS 1554)
(14) He was *certain* too, she recalled crossly, that she and Travis were bedfellows. (JY1 439)

SEPARATION is here defined as intervening elements delaying the processing of the complement clause, regardless of whether these elements cause structural ambiguity or not (see section 2). In other words, it is immaterial to which clause, i.e. the matrix clause or the complement clause, the intervening material belongs. The prediction based on the Complexity Principle (see Rohdenburg 1996) is that *that*-retention should be favoured in separation contexts, as they constitute cognitively more complex environments. An examination of the data reveals that this prediction seems to hold extremely well. As seen in tables 4a–b, there are in fact no instances in which separation co-occurs with *that*-omission, regardless of whether the clause is extraposed or post-predicate.

There is, however, a difference between extraposed and post-predicate clauses in the total share of attested separation (13 vs 3 per cent). We also find a difference in mean separation length (measured in words) between extraposed and post-predicate clauses. Extraposed clauses have a longer mean separation length (0.48) than post-predicate

[15] As pointed out by an anonymous reviewer, it is unclear how much variation is actually permitted in cases such as *become clear (that) the*. There are several factors apart from the separation context in (13) that may cause *that* to be retained in examples such as *it has become clear (that) the*. Taking together the fact that omission is quite rare in extraposed clauses to begin with, and that the complement clause subject type is 'Other NP', an environment in which omission is very infrequent, seems to go a long way in explaining this.

https://doi.org/10.1017/S1360674316000344 Published online by Cambridge University Press

clauses (0.24) when including all data points (i.e. including data points with a value of zero). On the other hand, including only data points with separation (i.e. > 0) post-predicate clauses have a slightly longer mean separation length (3.73) than extraposed clauses (3.60).

It is safe to conclude that the prediction of the Complexity Principle holds, in that SEPARATION is exclusively associated with *that*-retention. SEPARATION is thus treated as a knockout factor, i.e. a factor in which the response (+*that* vs –*that*) is invariant (see also section 4.6).

4.4 *Factors (iv) and (v):* COMP. CLAUSE SUBJECT TYPE *and* LENGTH

The second complexity factor included in this analysis concerns the type and length of the complement clause subject. The prediction based on the Complexity Principle (Rohdenburg 1996) is that pronominal subjects favour *that*-omission whereas NPs headed by nouns favour *that*-retention. COMP. CLAUSE SUBJECT TYPE is implemented as a variable with four levels: *I*, *you*, 'Other Pronoun' and 'Other NP'. All four types are illustrated in (15)–(18). The pronouns *I* and *you* are included as separate categories in the analysis as they have been reported in previous research to play a special role in the choice between retaining and omitting *that* (see Thompson & Mulac 1991a, 1991b; Jaeger 2010). Complement clause subject type 'Other NP' includes NPs headed by nouns and proper nouns.

(15) 'I'm *sorry* I was a bit short with you this afternoon,' Nick said when he rang Kelly later that night. (Type: *I*) (BP7 2835)
(16) I'm *sure* you'll manage to come up with something plausible. (Type: *you*) (JXV 1513)
(17) We are *hopeful* that it will be a true festival of football. (Type: Other Pronoun) (A9N 135)
(18) It is *true* that Franciscan spirituality put new stress on the necessity for integrating action and contemplation. (Type: Other NP) (HY6 989)

The distribution of complement clause subject types is presented in figure 4. First of all, we can see that the prediction of the Complexity Principle seems to hold; there is a clear difference in terms of the proportion of *that*-omission between the three pronominal categories and 'Other NP', supporting a division between NPs headed by pronouns and NPs headed by nouns.

Furthermore, we see that there are large differences between the pronominal categories; *you* is the subject type with the highest proportion of *that*-omission in both constructions. Complement clause subject type *I* takes an intermediate position, and the complement clause subject type 'Other Pronoun' is the pronominal category with the smallest proportion of *that*-omission. It is also clear that there are large differences between the two constructions. We can see that the post-predicate construction consistently has a higher proportion of *that*-omission across all levels of COMP. CLAUSE SUBJECT TYPE. This difference is most pronounced for complement clause subject type *I*, but is also pronounced for complement clause subject type 'Other Pronoun'. With

https://doi.org/10.1017/S1360674316000344 Published online by Cambridge University Press

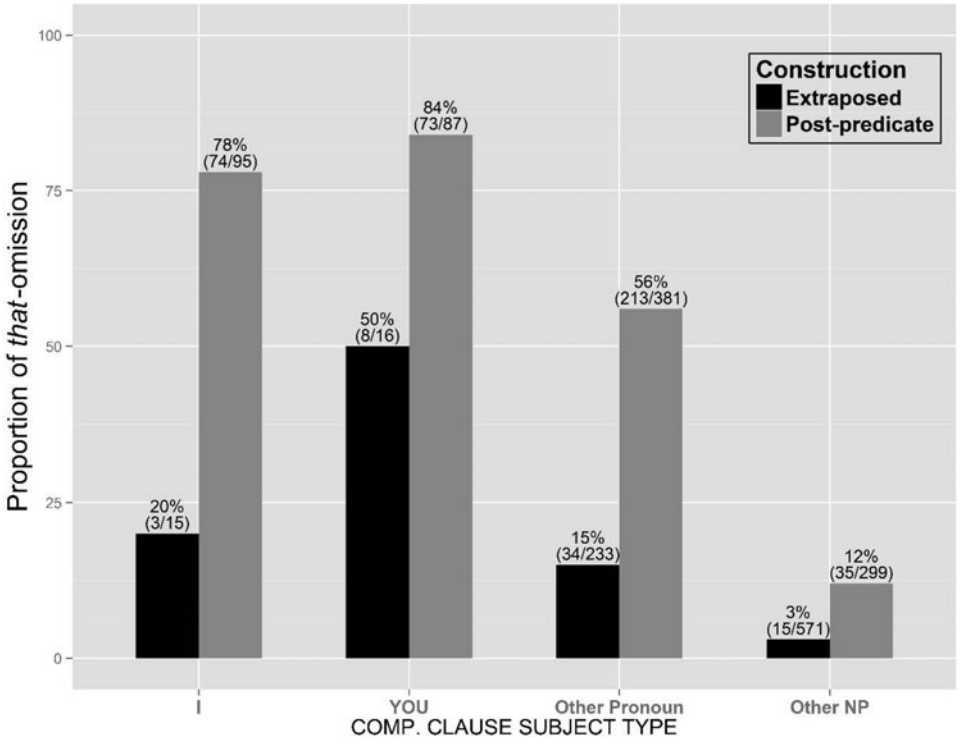


Figure 4. Proportion of *that*-omission across subject types and constructions

extraposed clauses, *you* stands out as the complement clause subject type with the highest proportion of *that*-omission.

Turning our attention to factor (v), COMP. CLAUSE SUBJECT LENGTH (measured in words), we now restrict the focus to complement clause subject type 'Other NP' only, as pronominal subjects do not exhibit any variation in length. As seen in figure 5, longer subjects favor *that*-retention across both constructions (range/standard deviation: Extraposed +*that* 1:21/3.22; Extraposed –*that* 1:5/0.75; Post-predicate +*that* 1:25/3.04; Post-predicate –*that* 1:4/0.45).

This finding suggests that COMP. CLAUSE SUBJECT LENGTH constitutes an important factor. In other words, when the subject of the complement clause is not pronominal, shorter subjects are still found to favour *that*-omission.[16]

[16] An additional length measurement, complement clause length, has been reported in previous studies to influence the choice between retaining and omitting *that* (see e.g. Jaeger 2010). Complement clause length is not included as a factor in this analysis as it proved too time-consuming to manually count the number of words in each complement clause.

https://doi.org/10.1017/S1360674316000344 Published online by Cambridge University Press

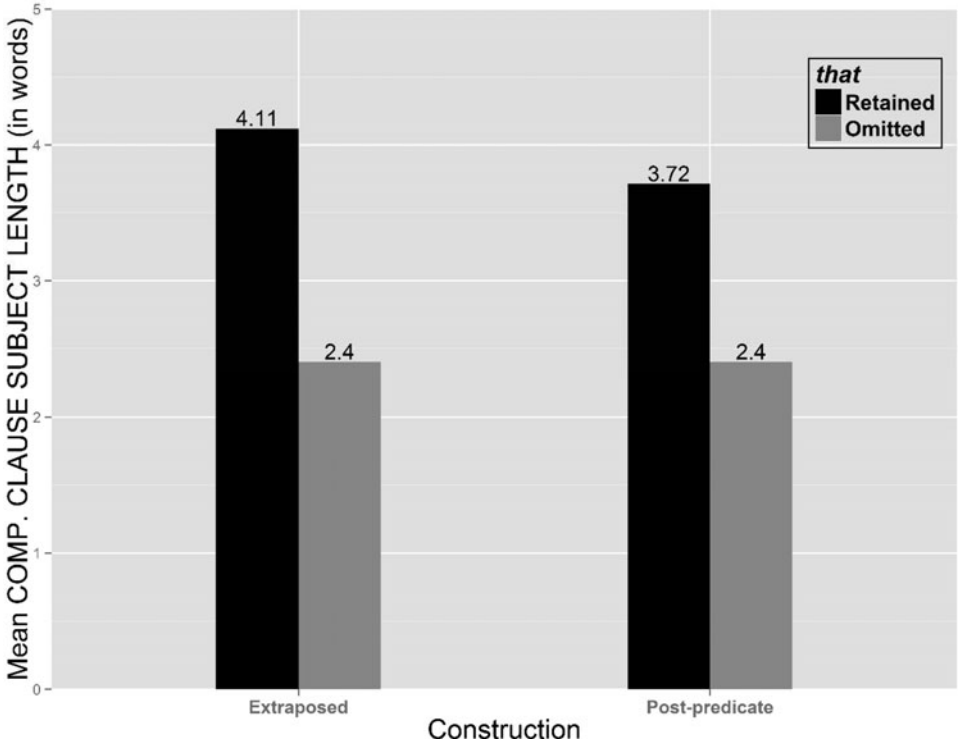


Figure 5. Mean complement clause subject length (in words) of 'Other NP' across constructions

4.5 *Factor (vi):* GENRE

The sixth and final factor is concerned with the role played by GENRE on the choice between retaining and omitting *that*. It has been noted in previous studies that *that*-retention is favoured in more formal genres, whereas *that*-omission shows the reverse genre distribution (see Elsness 1984: 521; Biber 1999: 145; Kaltenböck 2006: 378; Mindt 2011: 162 ff.). Biber *et al.* (1999: 680–1) attribute these differences to the difference in production circumstances and communicative purposes of different genres:

> Conversations are spoken and produced online; they typically have involved, interpersonal purposes; and they generally favor the reduction or omission of constituents that can be easily reconstructed. These characteristics are associated with omission rather than retention of *that* as the norm. Academic prose has the opposite characteristics: written, careful production circumstances; an expository, informational purpose; and a general preference for explicit, elaborated structure. Correspondingly, *that* retention is the norm in academic prose.

The corpus used in this study comprises five genres (see section 3), with each genre represented by 600,000 words. Four of these represent written language and one

https://doi.org/10.1017/S1360674316000344 Published online by Cambridge University Press

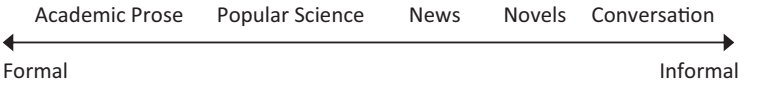


Figure 6. Continuum of formality

represents spoken language.[17] These five genres can be placed along a continuum, indicating their respective degree of formality. As seen in figure 6, academic prose is positioned at the most formal end of the continuum, and conversation at the most informal end (see Biber *et al.* 1999: 680).

Conversation, as the sole representative of the medium of speech, differs from the written genres by being interactive. Furthermore, the main function of conversation is personal communication, a function not shared by the written genres. Of the written genres, fiction is closest to conversation, as it includes written dialogue. The remaining three written genres all share an informational focus but differ on other grounds. News has a wide public audience, whereas academic prose targets a specialist audience. Popular science is situated between news and academic prose, as it targets both a specialist audience and a wider audience, albeit not as wide as news (see Biber *et al.* 1999: 16 for an overview of the situational characteristics of different genres).

The distribution of extraposed and post-predicate clauses across genres is shown in figure 7. First of all, we can see that the prediction that *that*-retention is more frequent in formal genres largely holds; the proportion of *that*-retention generally decreases as we move down the continuum of formality.

There are, however, large differences between extraposed and post-predicate clauses. As shown, extraposed clauses have a strong preference for *that*-retention across all genres; the highest proportion of *that*-omission reaching only 21 per cent (novels). Post-predicate clauses, in contrast, have higher proportions of *that*-omission in general, and *that*-omission, in fact, reaches more than 50 per cent in the two least formal genres. Note also that novels, not conversation, exhibit the highest proportion of *that*-omission for both extraposed and post-predicate clauses.

In summary, the prediction of previous studies (see Elsness 1984; Biber 1999; Kaltenböck 2006; Mindt 2011) is confirmed in that the proportion of *that*-omission increases as we move from formal to informal genres along the continuum of formality, although we find notable differences between extraposed and post-predicate clauses.

4.6 Modelling variation

Now that we have seen the influence of the factors in isolation, it is time to assess the effect of the factors' combined effect on the choice between retaining and omitting

[17] The fact that the corpus used in this study prioritizes written genres might have a bearing on the results obtained for collostruction strength, since the prediction of UID refers to spontaneous speech only (see Jaeger 2010). However, since previous studies have identified genre as an important factor, it was considered useful to include a wide variety of genres with different production circumstances and communicative purposes.

https://doi.org/10.1017/S1360674316000344 Published online by Cambridge University Press

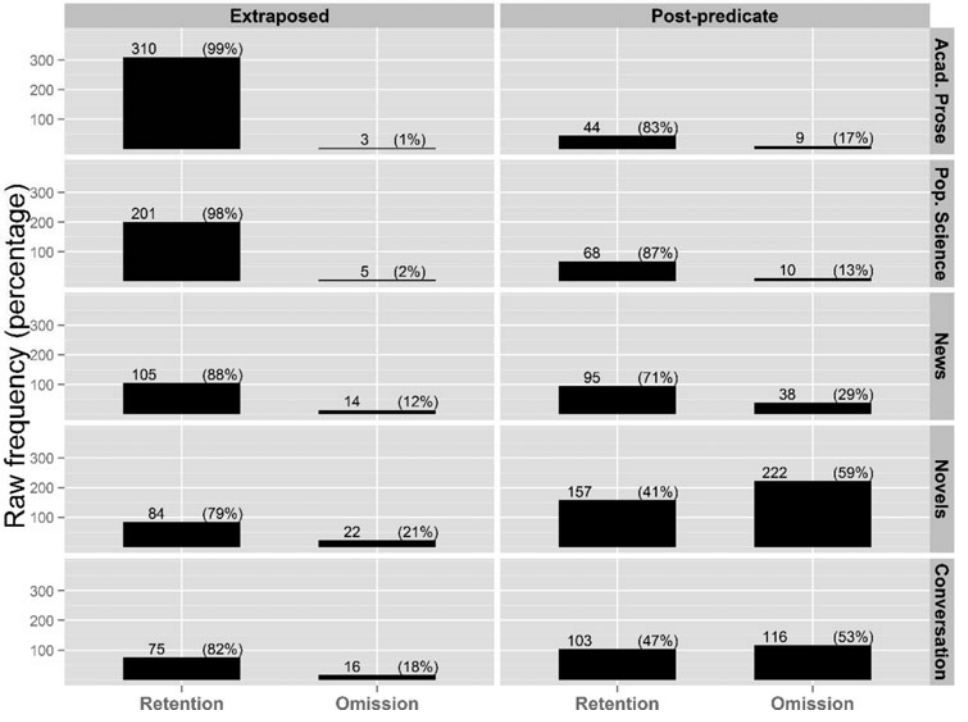


Figure 7. Raw frequency and percentage by genre across constructions

that. This section is thus devoted to multivariate analyses of the factors included in the study. Extraposed and post-predicate clauses will be treated as two different constructions and tested separately in different logistic regression models. All models are binomial models using the *glm*() function (see Crawley 2007: 511ff.) in the statistical software R (R Development Core Team 2008). Factor (iii), SEPARATION, is not included in these analyses as it was identified as a knockout factor, i.e. a factor in which the response (+*that* vs –*that*) is invariant (see section 4.3).

Starting with the models concerned with extraposed clauses in tables 5a–b, we should first note that MATRIX SUBJECT TYPE is not included as this factor is not applicable to extraposed clauses (see section 3). Second, in order to avoid obvious multicollinearity, i.e. that two or more factors are correlated with each other, COMP. CLAUSE SUBJECT LENGTH was removed from the main model as it correlates with COMP. CLAUSE SUBJECT TYPE. The factor was used in a second model which was fitted exclusively to data points with COMP. CLAUSE SUBJECT TYPE 'Other NP' (see table 6).

In the minimal adequate model in table 5b, COLLOSTRUCTION STRENGTH, the only non-significant factor from the saturated model in table 5a has been removed.[18] We see that two factors (factor (iv), COMP. CLAUSE SUBJECT TYPE, and factor (vi),

[18] Removing COLLOSTRUCTION STRENGTH is justified as there is no significant difference between the two models ($p = 0.23$).

https://doi.org/10.1017/S1360674316000344 Published online by Cambridge University Press

Table 5a. *Saturated logistic regression model: extraposed clauses. Predicted outcome: omission of* that

Predictor	Estimate	SE	*z*	Odds	*p*
(Intercept)	− 5.57	0.74	− 7.53	0.004	<0.001
(ii) log(COLLOSTRUCTION STRENGTH)	0.09	0.08	1.17	1.10	non-sig.
(iv) COMP. SUBJECT TYPE = Other NP					
vs Other Pronoun	1.48	0.35	4.27	4.38	<0.001
vs *you*	3.01	0.63	4.79	20.32	<0.001
vs *I*	1.34	0.74	1.80	3.81	<0.1
(vi) GENRE = Academic Prose					
vs Popular Science	0.67	0.75	0.90	1.96	non-sig.
vs News	2.57	0.65	3.94	13.10	<0.001
vs Novels	2.59	0.66	3.93	13.30	<0.001
vs Conversation	2.54	0.66	3.84	12.72	<0.001

$R^2 = 0.29$
$C = 0.85$

Table 5b. *Minimal adequate logistic regression model: extraposed clauses. Predicted outcome: omission of* that

Predictor	Estimate	SE	*z*	Odds	*p*
(Intercept)	− 5.09	0.60	− 8.43	0.006	<0.001
(iv) COMP. SUBJECT TYPE = Other NP					
vs Other Pronoun	1.47	0.35	4.26	4.36	<0.001
vs *you*	2.90	0.62	4.67	18.20	<0.001
vs *I*	1.36	0.74	1.84	3.91	<0.1
(vi) GENRE = Academic Prose					
vs Popular Science	0.66	0.74	0.89	1.93	non-sig.
vs News	2.55	0.65	3.91	12.79	<0.001
vs Novels	2.50	0.65	3.82	12.15	<0.001
vs Conversation	2.45	0.66	3.73	11.64	<0.001

$R^2 = 0.27$
$C = 0.84$

GENRE) have a significant effect on the choice between retention and omission of *that*. COMP. CLAUSE SUBJECT TYPE 'Other Pronoun' and *you* are both found to trigger significantly more *that*-omission when compared to 'Other NP'. We might also note that COMP. CLAUSE SUBJECT TYPE *you* is the factor with the strongest effect size (odds ratio (Odds) = 18.20). COMP. CLAUSE SUBJECT TYPE *I*, on the other hand, is not found to be significantly different from 'Other NP'. This might be explained by the small number of *I*-tokens in the data (3/15 found with *that*-omission) as compared to 'Other NP' (15/571), even though the proportion of *that*-omission with

https://doi.org/10.1017/S1360674316000344 Published online by Cambridge University Press

Table 6. *Logistic regression model of the effect of* COMP. CLAUSE SUBJECT LENGTH: *extraposed clauses. Predicted outcome: omission of* that

Predictor	Estimate	SE	*z*	Odds	*p*
(Intercept)	− 6.96	1.85	− 3.77	0.0009	<0.001
(ii) log(COLLOSTRUCTION STRENGTH)	0.57	0.29	1.97	1.77	<0.05
(iv) COMP. CLAUSE SUBJECT LENGTH	− 0.29	0.17	− 1.66	0.75	non-sig.
(vi) GENRE = Academic Prose					
vs Popular Science	1.00	0.92	1.08	2.72	non-sig.
vs News	2.52	0.80	3.16	12.48	<0.01
vs Novels + Conversation	0.54	1.24	0.44	1.72	non-sig.
$R^2 = 0.22$					
$C = 0.83$					

I is much higher (20 per cent) than with 'Other NP' (3 per cent) (see figure 4, section 4.4). The effect of GENRE is such that Academic Prose and Popular Science contrast with the other genres, the latter triggering far more *that*-omission (as indicated by their positive *z* scores; see also figure 7, section 4.5).

With only two factors, the model is still quite good: Nagelkerke's $R^2 = 0.29$ and the model has good classificatory power $C = 0.85$. R^2 has been described as a measure of 'the proportion of variation of the dependent variable which can be explained by the predictor variables of a given regression model' (Mittlböck & Schemper 1999: 17). *C* indicates how well the model accommodates the data it is based on and is thus concerned with the overall classificatory power of the model (according to Gries 2009: 297, a *C* value of approximately 0.8 is considered good). The good classificatory power of the model could thus be interpreted as a reflection of the fact that the variation between retaining or omitting *that* is quite restricted with extraposed clauses (remember that only 7 per cent of extraposed clauses feature *that*-omission; see figure 1, section 3).

The effect of COMP. CLAUSE SUBJECT LENGTH is tested in table 6 by fitting a model which excludes all pronominal complement clause subjects. In this model, we are thus only concerned with the 'Other NP' category of COMP. CLAUSE SUBJECT TYPE (n = 571; see figure 4, section 4.4). Due to the low-frequency of *that*-omission in certain genres, the two genres farthest towards the informal end of the continuum of formality (novels and conversation; see figure 6, section 4.5) have been conflated in this model for more reliable results.[19]

As seen in table 6, COMP. CLAUSE SUBJECT LENGTH does not prove to be a significant factor. In other words, when the subject of the complement clause is not pronominal, longer subjects are not found to significantly favour *that*-retention.

[19] Note, however, that 'Other NP' as the complement clause subject almost constitutes a knockout factor since only 3 per cent of these tokens are found with *that*-omission (see figure 4, section 4.4).

https://doi.org/10.1017/S1360674316000344 Published online by Cambridge University Press

Table 7. *Logistic regression model: post-predicate clauses. Predicted outcome: omission of* that

Predictor	Estimate	SE	*z*	Odds	*p*
(Intercept)	− 5.04	0.57	− 8.80	0.006	<0.001
(i) MATRIX SUBJECT TYPE=Other NP					
vs Other Pronoun	0.03	0.32	0.08	1.03	non-sig.
vs *you*	0.91	0.45	2.03	2.48	<0.05
vs *I*	1.75	0.29	5.92	5.73	<0.001
(ii) log(COLLOSTRUCTION STRENGTH)	0.37	0.06	6.44	1.45	<0.001
(iv) COMP. SUBJECT TYPE=Other NP					
vs Other Pronoun	2.04	0.24	8.61	7.72	<0.001
vs *you*	2.75	0.40	6.88	15.70	<0.001
vs *I*	2.66	0.36	7.46	14.32	<0.001
(vi) GENRE=Academic Prose					
vs Popular Science	− 1.61	0.59	− 2.71	0.20	<0.01
vs News	0.51	0.50	1.03	1.67	non-sig.
vs Novels	0.48	0.45	1.06	1.62	non-sig.
vs Conversation	− 0.26	0.48	− 0.54	0.77	non-sig.

$R^2 = 0.56$
$C = 0.89$

Shifting focus from extraposed to post-predicate clauses, the model in table 7 does not include the factor COMP. CLAUSE SUBJECT LENGTH as this factor is tested on non-pronominal complement clause subjects only (see table 9), in order to avoid multicollinearity (see tables 5a–b above).

As seen in table 7, all factors included prove to be significant: MATRIX SUBJECT TYPE *I* and *you* stand out as the most salient in triggering *that*-omission (see figure 3, section 4.1); COLLOSTRUCTION STRENGTH is found to be highly significant in that higher COLLOSTRUCTION STRENGTH favours *that*-omission (see section 4.2); the three pronominal levels of COMP. CLAUSE SUBJECT TYPE are all significantly different from 'Other NP' (see figure 4, section 4.4). However, looking at the effect of GENRE in the model in table 7, we see that the results differ from what we would expect, given the distribution in figure 7, section 4.5. Only popular science stands out as significantly different from academic prose, which we would not expect since novels and conversation seem more different from academic prose, given that they include far more instances of *that*-omission. These puzzling results for GENRE suggest that GENRE is correlated with one or more other factors. Upon closer inspection, certain values of factors such as MATRIX SUBJECT TYPE and COMP. CLAUSE SUBJECT TYPE are not represented in academic prose or in popular science. The results for GENRE in table 7 could thus be attributed to correlations between GENRE and other factors, rendering an intuitively insignificant effect the only comparison with any classificatory power left in the combined model. To remedy this problem, a model was fitted which conflates

https://doi.org/10.1017/S1360674316000344 Published online by Cambridge University Press

Table 8. *Updated logistic regression model: post-predicate clauses. Predicted outcome: omission of* that

Predictor	Estimate	SE	*z*	Odds	*p*
(Intercept)	− 5.93	0.52	− 11.40	0.002	<0.001
(i) MATRIX SUBJECT TYPE=Other NP					
vs Other Pronoun	0.002	0.32	0.06	1.00	non-sig.
vs *you*	0.87	0.45	1.94	2.38	<0.1
vs *I*	1.66	0.29	5.69	5.26	<0.001
(ii) log(COLL. STRENGTH)	0.37	0.06	6.43	1.45	<0.001
(iv) COMP. SUBJECT TYPE=Other NP					
vs Other Pronoun	2.00	0.24	8.50	7.40	<0.001
vs *you*	2.72	0.40	6.84	15.21	<0.001
vs *I*	2.56	0.35	7.34	13.00	<0.001
(vi) GENRE=Ac. Prose + Pop. Sci.					
vs News	1.47	0.40	3.70	4.38	<0.001
vs Novels	1.47	0.33	4.46	4.35	<0.001
vs Conversation	0.78	0.35	2.17	2.13	<0.05

$R^2 = 0.56$
$C = 0.89$

academic prose and popular science, the two genres farthest towards the formal end of the continuum of formality (see figure 6, section 4.5), into one genre, as seen in table 8.[20]

The new model in table 8, with academic prose and popular science conflated, has identical R^2 (0.56) and C (0.89) values with the model in table 7, suggesting that this simplification is warranted. More importantly, the new model handles the genre comparisons better in that the three remaining genres are found to be significantly different from academic prose and popular science combined.

Turning to the effect of COMP. CLAUSE SUBJECT LENGTH on post-predicate clauses, we once again restrict our focus to non-pronominal complement clause subjects (i.e. COMP. CLAUSE SUBJECT TYPE = Other NP; n = 299, see figure 4, section 4.4). The results of the model are presented in table 9. Note that for reasons discussed in the previous paragraph, academic prose and popular science are conflated into one genre. Furthermore, due to scarcity of data in certain cells, MATRIX SUBJECT TYPES *I* and *you* have also been conflated.

From table 9 we see that COMP. CLAUSE SUBJECT LENGTH does not have a significant effect on the choice between omitting and retaining *that*.

[20] While conflating Academic Prose and Popular Science solves the problem of empty cells and thus improves the model, there is still reason to believe that there is collinearity left in the model. Most importantly, GENRE and MATRIX SUBJECT TYPE seem to be correlated in that the pronominal subject *I* and *you* are predominantly found in novels and conversation. The fact that GENRE seems to be correlated with other variables further suggests that the explanatory power of GENRE is limited.

https://doi.org/10.1017/S1360674316000344 Published online by Cambridge University Press

Table 9. *Logistic regression model of the effect of* COMP. CLAUSE SUBJECT LENGTH: *post-predicate clauses. Predicted outcome: omission of* that

Predictor	Estimate	SE	*z*	Odds	*p*
(Intercept)	− 8.69	1.58	− 5.50	0.0002	<0.001
(i) MATRIX SUBJECT TYPE=Other NP					
vs Other Pronoun	1.21	0.64	1.88	3.35	<0.1
vs *I* + *you*	2.01	0.67	2.99	7.46	<0.01
(ii) log(COLLOSTRUCTION STRENGTH)	0.62	0.16	3.89	1.85	<0.001
(v) COMP. SUBJECT LENGTH	− 0.23	0.13	− 1.76	7.97	<0.1
(vi) GENRE=Ac. Prose + Pop. Sci.					
vs News	3.46	1.11	3.13	3.19	<0.01
vs Novels	2.88	1.07	2.68	1.78	<0.01
vs Conversation	1.51	1.16	1.31	4.54	non-sig.
$R^2 = 0.38$					
$C = 0.88$					

Finally, a model was also fitted to determine whether the effect of the factors observed for post-predicate clauses is largely due to the frequent adjective *sure* or not.[21] In this model, all tokens of *sure* are excluded (n=280). Furthermore, the COLLOSTRUCTION STRENGTH values of the remaining adjectives have been re-calculated on the basis of *sure* being removed from the dataset (see section 4.2). The model is presented in table 10.

We see that the same factors are still found to be significant, thus indicating that the previously observed effects are not just licensed by *sure*. However, we might note that MATRIX SUBJECT TYPE *you* is no longer significant, indicating that much of the effect of this particular MATRIX SUBJECT TYPE is licensed by *sure*. This is also illustrated by the fact that 56 per cent of all MATRIX SUBJECT TYPE *you* tokens are found with *sure*, 76 per cent of which are found with *that*-omission.

5 Discussion and summary

The aim of this study has been twofold: (i) to test the effect of different factors on adjectival data as regards the variation between *that*-omission and *that*-retention, and (ii) to test whether there are any differences between extraposed and post-predicate clauses where the effect of those factors is concerned. The analysis of the choice between retaining and omitting *that* has identified several factors as significant in explaining this variation. However, looking at the difference between extraposed and post-predicate clauses with respect to this variation, we have seen that not all of these factors have a significant effect on extraposed clauses. We might conclude that

[21] As pointed out by an anonymous reviewer, a more elegant way to address this question would involve fitting a mixed-effect model with adjective lemma as random effect.

https://doi.org/10.1017/S1360674316000344 Published online by Cambridge University Press

Table 10. *Logistic regression model: post-predicate clauses (sure removed). Predicted outcome: omission of* that

Predictor	Estimate	SE	*z*	Odds	*p*
(Intercept)	− 6.06	0.69	− 8.78	0.002	<0.001
(i) MATRIX SUBJECT TYPE = Other NP					
vs Other Pronoun	− 0.05	0.35	− 0.13	0.95	non-sig.
vs *you*	− 0.03	0.64	− 0.05	0.97	non-sig.
vs *I*	1.66	0.33	5.09	5.26	<0.001
(ii) log(COLLOSTRUCTION STRENGTH)	0.34	0.08	3.95	1.41	<0.001
(iv) COMP. SUBJECT TYPE = Other NP					
vs Other Pronoun	2.23	0.31	7.11	9.27	<0.001
vs *you*	3.47	0.57	6.06	32.24	<0.001
vs *I*	2.65	0.42	6.28	14.01	<0.001
(vi) GENRE = Ac. Prose + Pop. Sci.					
vs News	1.34	0.48	2.79	3.82	<0.01
vs Novels	1.47	0.41	3.57	4.36	<0.001
vs Conversation	0.96	0.44	2.19	2.62	<0.05

$R^2 = 0.53$
$C = 0.88$

the majority of the factors identified in previous studies on verbal predicates are primarily applicable to post-predicate clauses. This is not surprising, since adjectival post-predicate clauses closely mirror verbal post-predicate (or object) clauses. In what follows I will discuss the differences between extraposed and post-predicate clauses regarding the variation between *that*-omission and *that*-retention, with a special focus on the factors related to the Uniform Information Density Principle and the Complexity Principle.

The prediction derived from the Uniform Information Density Principle (Jaeger 2010) that adjectives with a higher collostruction strength will favour *that*-omission was only found to hold for post-predicate clauses. The main reason for this may be the fact that adjectives complemented by post-predicate clauses appear to be represented in more entrenched formulaic expressions, such as *I'm sure*, *I'm convinced* (remember that *I* was found to be the matrix subject most strongly triggering omission of *that*), thus promoting *that*-omission, whereas this does not appear to be the case among extraposed clauses. In other words, this is not only a matter of the expectedness of certain adjectives to be followed by a post-predicate *that*-clause, but also a matter of recurring subject–adjective combinations forming a highly recognizable unit more susceptible to *that*-omission. In connection with this, the numerous studies on the grammaticalization of the subject–verb combination *I think* and its preponderance for *that*-omission might serve as a good analogy (see Thompson & Mulac 1991a; Van Bogaert 2010, 2011; Kaltenböck 2011, 2013). In fact, one particular subject–adjective combination can be singled out as especially frequent and a strong trigger for *that*-omission. It was observed in section 4.2 that *sure* was the most frequent adjective complemented by post-predicate clauses. Taking a closer look at the preference of the

https://doi.org/10.1017/S1360674316000344 Published online by Cambridge University Press

Table 11. *Preference for that-omission with* I ('m/am) sure *vs other combinations (post-predicate clauses)*

	+that		*–that*		Total	
I + sure	43	21%	164	79%	207	100%
you + sure	7	24%	22	76%	29	100%
Other *sure*	27	61%	17	39%	44	100%
Other post-predicate	390	67%	192	33%	582	100%
Total	467	54%	395	46%	862	100%

subject–adjective combination *I* ('*m*/*am*) *sure* (including modifiers and negation such as *I'm pretty sure* and *I'm not sure*) for *that*-omission, we find that omitting *that* is quite prevalent (79 per cent, as shown in table 11).

We have already observed that the effect of MATRIX SUBJECT TYPE *you* was largely licensed by *sure* (see table 10, section 4.6), but comparing *I* ('*m*/*am*) *sure* to other combinations including *sure*, the proportion of *that*-omission drops dramatically (39 per cent). It drops even further when all other post-predicate tokens are taken into account (33 per cent). This suggests that the subject–adjective combination *I* ('*m*/*am*) *sure* is a major trigger for *that*-omission, comparable to the role of *I think* reported in previous studies on verbal predicates (see Thompson & Mulac 1991a, 1991b; Jaeger 2010; Torres Cacoullos & Walker 2009; Van Bogaert 2011; Kaltenböck 2013).

Regarding the predictions of the Complexity Principle, SEPARATION and COMP. CLAUSE SUBJECT TYPE were found to be significant factors for both extraposed and post-predicate clauses, whereas COMP. CLAUSE SUBJECT LENGTH was not. A preliminary hierarchy of complexity factors most influential in *that*-omission would thus have SEPARATION at the top (remember that separation contexts invariably feature retention of *that*) followed by COMP. CLAUSE SUBJECT TYPE. It should also be pointed out that COMP. CLAUSE SUBJECT TYPE is the variable with the highest effect size across both extraposed and post-predicate clauses (see table 5b and table 8, section 4.6).

In relation to the complement clause subjects, Elsness (1984) has pointed out that the preference of NPs headed by pronouns for occurring with *that*-omission (as compared to NPs headed by nouns) is largely due to coreferentiality between the matrix and complement clause subjects (here we might want to recall that *I* and *you* were found to be the subjects most consistent in triggering *that*-omission as both matrix subject and complement clause subject in post-predicate clauses). In extraposed clauses, such coreferentiality between the matrix and complement clause subjects is excluded. This fact highlights an important difference between extraposed and post-predicate clauses and might serve as an important clue as to differences in the proportion of *that*-omission between them (see also Kaltenböck 2006: 387). Even though coreferentiality between matrix and complement clause subjects might serve as a powerful factor for triggering *that*-omission, I would like to argue that it is the matrix subject type that does most of the work. Consider table 12, and the apparent triggering of *that*-omission when the matrix subject type is *I* and the complement clause subject type is either *I*

https://doi.org/10.1017/S1360674316000344 Published online by Cambridge University Press

Table 12. *Raw frequency and percentage of selected* MATRIX + COMP. CLAUSE SUBJECT TYPE *combinations*

	+*that*		–*that*		Total	
I + *I*	13	15%	72	85%	85	100%
I + *you*	9	14%	57	86%	66	100%
I + other subject	108	39%	172	61%	280	100%
All other combinations	337	78%	94	22%	431	100%
Total	467	54%	395	46%	862	100%

(i.e. coreferential) (85 per cent) or *you* (86 per cent). Corresponding figures for *you* + *you* (74 per cent) and *you* + *I* (67 per cent) further indicate that matrix subject type *I* combined with COMP. CLAUSE SUBJECT TYPE *I* or *you* could be seen as more important than coreferentiality in triggering *that*-omission, as indicated by the sudden drop in the proportion of *that*-omission for the other combinations, as shown in table 12 (see also Torres-Cacoullos & Walker 2009: 25 for similar observations on the effect of coreferentiality on *that*-omission).

Furthermore, the results suggest that the division between NPs headed by pronouns and NPs headed by nouns, as predicted by the Complexity Principle (Rohdenburg 1996), should be further subcategorized when it comes to the complement clause subject; the results suggest that the division should be between the pronouns *I* and *you,* on the one hand, and 'Other Pronouns' and 'Other NP' on the other hand (see figure 4, section 4.4).

Another thing to note about the complexity factors is that extraposed clauses generally allow much more complexity than post-predicate clauses; they are found to have more attested examples with separation, they are also found more frequently with NPs headed by nouns as the complement clause subject, and thus their complement clause subject is found to be longer than that of post-predicate clauses. Even so, GENRE proved to be a highly significant factor in explaining the variation between retaining and omitting *that* (see table 5b, section 4.6). This might indicate that the frequent occurrence of complexity factors found for extraposed clauses should be seen as an inherent feature of the genres they most frequently occur in rather than as features triggering *that*-omission.

Author's address:

Department of English
Uppsala University
Box 527
751 20 Uppsala
Sweden
henrik.kaatari@engelska.uu.se

https://doi.org/10.1017/S1360674316000344 Published online by Cambridge University Press

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.
Biber, Douglas, Stig Johansson, Susan Conrad, Geoffrey Leech & Edward Finegan. 1999. *Longman grammar of spoken and written English*. Harlow: Longman.
Burnard, Lou. 2007. *Reference guide for the British National Corpus (XML edition)*. www.natcorp.ox.ac.uk/docs/URG/ (22 October 2014).
Crawley, Michael J. 2007. *The R book*. Chichester: Wiley.
Elsness, J. 1984. That or zero? A look at the choice of object clause connective in a corpus of American English. *English Studies* 65, 519–33.
Evert, Stefan. 2006. How random is a corpus? The library metaphor. *Zeitschrift für Anglistik und Amerikanistik* 54(2), 177–90.
Garretson, Gregory & Henrik Kaatari. 2014. The computer as research assistant: A new approach to variable patterns in corpus data. In Lieven Vandelanotte, Kristin Davidse, Caroline Gentens & Ditte Kimps (eds.), *Recent advances in corpus linguistics: Developing and exploiting corpora*, 55–80. Amsterdam: Rodopi.
Gries, Stefan Th. 2007. *Coll.analysis 3.2a.* A program for R for Windows 2.x.
Gries, Stefan Th. 2009. *Statistics for linguists with R: A practical introduction*. Berlin: Mouton de Gruyter.
Gries, Stefan Th. & Anatol Stefanowitsch. 2004. Extending collostructional analysis: A corpus-based perspective on alternations. *International Journal of Corpus Linguistics* 9(1), 97–129.
Hilpert, Martin. 2010. *Collostructional analysis: Measuring associations between constructions and lexical elements*. Freiburg: Freiburg Institute for Advanced Studies.
Hoey, Michael. 2005. *Lexical priming: A new theory of words and language*. London: Routledge.
Huddleston, Rodney & Geoffrey K. Pullum. 2002. *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.
Jaeger, Florian. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology* 61, 23–62.
Kaatari, Henrik. 2012. Sampling the BNC – creating a randomly sampled subcorpus for comparing multiple genres. Poster presented at ICAME 33, Leuven, Belgium, 30 May – 3 June 2012.
Kaltenböck, Gunther. 2006. … *That* is the question: Complementizer omission in extraposed clauses. *English Language and Linguistics* 10(2), 371–96.
Kaltenböck, Gunther. 2011. Explaining diverging evidence: The case of clause-initial *I think*. In Doris Schönefeld (ed.), *Converging evidence: Methodological and theoretical issues for linguistic research*, 81–112. Amsterdam: John Benjamins.
Kaltenböck, Gunther. 2013. Development of comment clauses. In Bas Aarts, Joanne Close, Geoffrey Leech & Sean Wallis (eds.), *The English verb phrase: Investigating recent change with corpora*, 286–317. Cambridge: Cambridge University Press.
Mindt, Ilka. 2011. *Adjective complementation: An empirical analysis of adjectives followed by that-clauses*. Amsterdam: John Benjamins.
Mittlböck, Martina & Michael Schemper. 1999. Computing measures of explained variation for logistic regression models. *Computer Methods and Programs in Biomedicine* 58, 17–24.
R Development Core Team 2008. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. www.R-project.org/ (22 October 2014).
Rohdenburg, Günter. 1995. On the replacement of finite complement clauses by infinitives in English. *English Studies* 76, 367–88.

https://doi.org/10.1017/S1360674316000344 Published online by Cambridge University Press

Rohdenburg, Günter. 1996. Cognitive complexity and increased grammatical explicitness in English. *Cognitive Linguistics* 7, 149–82.
Stefanowitsch, Anatol & Stefan Th. Gries. 2003. Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8(2), 209–43.
Stefanowitsch, Anatol & Stefan Th. Gries. 2005. Co-varying collexemes. *Corpus Linguistics and Linguistic Theory* 1(1), 1–43.
Thompson, Sandra A. & Anthony Mulac. 1991a. A quantitative perspective on the grammaticization of epistemic parentheticals in English. In Elizabeth Closs Traugott & Bernd Heine (eds.), *Approaches to grammaticalization*, vol. 2, 313–39. Amsterdam: John Benjamins.
Thompson, Sandra A. & Anthony Mulac. 1991b. The discourse conditions for the use of the complementizer *that* in conversational English. *Journal of Pragmatics* 15(3), 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), 1–43.
Van Bogaert, Julie. 2010. A constructional taxonomy of *I think* and related expressions: Accounting for the variability of complement-taking mental predicates. *English Language and Linguistics* 14(3), 399–427.
Van Bogaert, Julie. 2011. *I think* and other complement-taking predicates: A case of and for constructional grammaticalization. *Linguistics* 49(2), 295–332.
Van linden, An. 2012. *Modal adjectives: English deontic and evaluative constructions in synchrony and diachrony*. Berlin: Mouton de Gruyter.

https://doi.org/10.1017/S1360674316000344 Published online by Cambridge University Press