New ways of analyzing complementizer drop in Montréal French: Exploration of cognitive factors

Yiming Liang, Pascal Amsili, Heather Burnett

TITLE: New ways of analyzing complementizer drop in Montréal French: Exploration of cognitive factors

AUTHORS:

Yiming Liang - Université de Paris, CNRS, Laboratoire de linguistique formelle

Pascal Amsili - Sorbonne Nouvelle, Laboratoire Lattice (CNRS/PSL-ENS/SN)

Heather Burnett - Université de Paris, CNRS, Laboratoire de linguistique formelle

CORRESPONDING AUTHOR: Yiming Liang - yiming.liang@etu.u-paris.fr

SHORT TITLE: *Que* drop in Montreal French

COMPETING INTERESTS: The authors declare none.

ABSTRACT

In this paper, we return to the well-studied yet still puzzling phenomenon of complementizer omission in a large spoken corpus of Quebec French, with the help of modern computational methods for annotation and mixed effects logistic regression models. Supporting previous work, our study reveals that complementizer 'que' omission is conditioned by social factors and grammatical factors; however, we also find that 'que' omission is conditioned by cognitive factors such as information density. Our paper thus illustrates an important way in which older variationist corpora can continue to be valuable resources for studying fine-grained patterns of variation, particularly in their cognitive aspects.

KEYWORDS: complementizer omission, information density, quantitative syntax, multifactorial analysis, Quebec French

ACKNOWLEDGMENTS: We would like to thank Yair Haendler for his help in the statistical analysis. We are also grateful to anonymous reviewers for their feedback on earlier versions of this paper. This research has received funding from the ERC under the European Union's Horizon 2020 research and innovation programme (grant agreement N°850539).

INTRODUCTION

This paper presents a new study of variable complementizer omission in Montréal French (also known as *que* drop). In most varieties of Canadian French, *que* may be optionally omitted when it introduces a complement, circumstantial or relative clause, as shown in (1)-(3).

(1) Bien je pense ~~que~~ c'est: c'est important ... (complement clause, speaker 128, *Corpus Montréal 84,* Thibault & Vincent [1990])

'Okay I think ~~that~~ it's: it's important...'

(2) ... parce ~~qu'~~ ici c'est bizarre. (circumstantial clause, speaker 2, *Corpus Montréal 84*)

'... because ~~that~~ here it is weird.'

(3) C'est là ~~que~~ ma mère à moi vivait. (relative clause, Roberge & Rosen [1999])

'It was there ~~that~~ my mom had lived.'

Que drop was first studied in the early days of variationist sociolinguistics by Sankoff, Sarrasin, and Cedergren (1971) and the social and grammatical conditions under which *que* is likely to be pronounced/omitted have since been widely investigated and given rise to debate in the variationist literature. Since the early 2000s, there have been advances both in our understanding of the social and cognitive factors that condition language variation and change (see for example Ferreira & Dell [2000] for *that* omission, Bresnan, Cueni, Nikitina, & Baayen [2007] for the dative alternation, and more recently Kleinschmidt, Weatherholtz, & Jaeger [2018] for phonological variants), and in available computational and statistical tools. Likewise, in the past 10 years, a rich line of research on variable complementizer drop in other languages has emerged, and we now have reason to believe that cognitive factors, particularly those related to the distribution of information across utterances (Jaeger, 2010; Levy & Jaeger, 2007), play a role. Our new investigation allows us to tease apart the syntactic and phonological factors conditioning complementizer *que* drop in a way that was not possible for previous studies, and brings to light the additional role that information-based reasoning plays in the variable production of complementizer in the French spoken in Montréal.

In the next two sections of the paper, we review the previous variationist literature on complementizer drop in French, and compare it to more recent work on complementizer drop in English and other languages. We argue that the literature on English complementizer omission has identified a number of new factors that are predicted to be relevant for the Canadian French variable. Then, we present the methodology of our study: the annotation, extraction, coding, and statistical analysis of *que* drop in the *Montréal 84* corpus. We then present our results and discuss how they relate both to the previous work on this variable in French and cross-linguistically. The final section concludes with future perspectives on variation in Montréal corpora.

COMPLEMENTIZER DROP IN FRENCH

The study of *que* drop in Canadian French has a rich history in sociolinguistics, but there has been very little agreement on what motivates the deletion of *que* in complement clauses (henceforth CC). Based on a first study of the Sankoff-Cedergren corpus of spoken Montréal French (Sankoff & Cedergren, 1972), Sankoff et al. (1971) and Sankoff (1980) suggested that the phonological contexts preceding and following *que* condition the omission. In particular, Sankoff and colleagues showed that sibilants favor *que* omission, compared to other sounds. Based on this result, they proposed that *que* omission may be conditioned by the sonority hierarchy (Clements, 1990), and hypothesized that omission could be motivated by consonant cluster simplification.

Sankoff et al.'s proposal is supported in Warren (1994)'s study of *Montréal 84*. Working on *que*-omission in complement, circumstantial and relative clauses, she reported the stability of *que* omission from the 1971 to the 1984 Montreal corpus. A simple syntactic structure preceding or following the complementizer favors *que* drop. Like Martineau's (1985) study of the Ottawa-Hull corpus (Poplack, 1989), Warren observed a correlation between the omission and some specific verbs and contexts, like *je pense* 'I think'and *disons* 'let's say', which should result from the grammaticalization of these contexts into 'epistemic phrases' (Thompson & Mulac, 1991a, 1991b).

The omission of *que* has also been argued conditioned by the type of the CC subject rather

than the phonological environment following *que*: Connors (1975) showed, also using the Sankoff-Cedergren corpus, that speakers tend to drop *que* when the subject of the CC is a pronoun rather than a NP. She contested Sankoff et al. (1971)'s analysis, arguing that frequent pronouns in French often begin with a sibilant (like [ʒ] in *je* 'I' and [s] in *ce* and *ça* 'it'), which makes sibilants more likely to favor omission.

Drawing on the *Français parlé à Ottawa-Hull (OH)* (Poplack, 1989) and *Récits du français Québécois d'Autrefois (RFQ)* (Poplack & St-Amand, 2007), Dion (2003), replicated the effect of the sonority hierarchy described by Sankoff and colleagues, reporting that omission occurs most before obstruents, less so before sonorants and least before vowels. Moreover, she argued that the syntactic effect observed by Connors (1975) should actually be analyzed as part of the phonological effect, since the pronouns starting with a vowel clearly disfavor omission. She also noticed a lexical effect, whereby certain verbs, like *rappeler* 'remind', *sembler* 'seem' and *penser* 'think', favor *que*-deletion.

These studies have some limitations. First, Connors (1975), Sankoff (1980), and Sankoff et al. (1971) could only analyze small subsets of their corpora (16 speakers). Similarly, Warren (1994) worked on 24 of 72 speakers from the *Montréal 84* corpus and Dion (2003) and Martineau (1985) studied only 30 and 14 speakers, which only represent a quarter and one-tenth of the total data found in the OH corpus respectively. Second, mixed-effects models which take into account random effects, such as inter-speaker variation (see Johnson, 2009), were not available. Finally, these previous studies have not investigated the cognitive aspects of this phenomenon.

THAT OMISSION IN ENGLISH

Although *que* omission is observed in spoken French, according to French prescriptive grammar, the use of the complementizer is obligatory in subordinate structures. In English, the omission of *that* is fully acceptable and widely observed. Linguistic and cognitive factors on choices between omitting and preserving *that* in a CC include: matrix subject (Thompson & Mulac, 1991a, 1991b; Torres

Cacoullos & Walker, 2009), subject of the embedded clause (Elsness, 1984; Ferreira & Dell, 2000), the distance of the subordinate clause from matrix verb (Elsness, 1984; Hawkins, 2001), and the presence of production difficulties at the beginning of CCs (Ferreira & Firato, 2002; Jaeger, 2005), among others. Building on this work, Jaeger (2010) shows that *information density* also plays a role in determining whether or not the complementizer *that* will be omitted.

Previous accounts of complementizer that *dropping*

We first present some of the most influential accounts for *that* dropping: availability-based accounts, dependency processing accounts, ambiguity avoidance accounts and grammaticalization accounts.

Availability-based accounts hold that the relative accessibility of referents affects speakers' syntactic choices in production. Here *accessibility* refers to the ease with which a word can be retrieved from the mental lexicon. According to *Principle of Immediate Mention* (Ferreira & Dell, 2000), given the time pressure of spontaneous speech, speakers tend to structure their message such that accessible words are pronounced first since they are available earlier for production. Therefore, speakers are assumed to pronounce the optional complementizer more often when the words at the CC onset (e.g., the CC subject) are more difficult to be retrieved from memory. Evidence for this hypothesis comes from both experiments and corpus studies: speakers prefer to insert *that* when the CC subject is a third-person pronoun or a lexical NP rather than the local pronoun *I* (Elsness, 1984; Ferreira & Dell, 2000; Jaeger, 2010); speakers use *that* more frequently when having production difficulties, such as repetition and disfluency, at the beginning of CC (Ferreira & Firato, 2002; Jaeger, 2005).

An alternative account predicts the ease of dependency processing to be the primary driving force behind speakers' preferences. Hawkins (2001, 2004) proposes the *Principle of Minimize Domains*, suggesting that speakers prefer syntactic options that lead to shorter dependencies. *Dependency processing accounts* have received support from studies reporting a correlation between an increased distance from the matrix verb to the CC and a higher rate of *that*-use

(Hawkins, 2001; Rohdenburg, 1998). In (4), where intervening material (*much too late* in (4)) appears, the use of complementizer shortens the length from the matrix verb *realize* to its VP domain, thus facilitating dependency processing.

(4) We realized much too late (that) Jill was not coming back. (Rohdenburg, 1998)

Ambiguity avoidance accounts, however, have attributed the use of the optional complementizer to temporary ambiguity avoidance (Hawkins, 2004). For example, speakers are assumed to insert *that* more frequently in (5a), as a way to reduce the chance that *you* would be temporarily interpreted as the direct object of *knew*. Nevertheless, this prediction has not been verified by any other studies (Ferreira & Dell, 2000; Jaeger, 2010; Roland et al., 2006; see also Jaeger [2011] for reducible subject relatives). One plausible explanation is that speakers only avoid ambiguity leading to severe garden path effects (as shown in (5b)).

(5) a. I knew (that) you missed the train.

b. The horse raced past the barn fell. (Bever, 1970:40)

Finally, the three processing accounts mentioned above have been contrasted with *grammaticalization accounts* (Thompson & Mulac, 1991a, 1991b), which hold that when the complementizer is absent, the matrix clause loses its primordial syntactic function in the sentence, and behaves like a parentherical that can float to different positions, as shown in (6). Epistemic main subjects (1st and 2nd person) and verbs (like *think* and *guess*), which are often used to express speakers' epistemic claims or degree of speaker's commitment, are more likely to be grammaticalized into epistemic phrases or discourse formulas, and hence should correlate with a lower rate of *that*-use (Thompson & Mulac, 1991b). However, this hypothesis seems to be contradictory to recent diachronic findings, where many frequent mental verbs, such as *think, suppose, know, believe,* and *understand,* undergo an overall increase in use of *that* from the 16th to 21st centuries (Shank & Plevoets, 2018).

(6) a. *I think* exercise is really beneficial, to anybody.

b. It's just your point of view you know what you like to do in your spare time *I think*.

(Thompson & Mulac, 1991b:313)

Uniform Information Density Hypothesis

Previous work has shown that more predictable words tend to be pronounced more quickly and with less phonetic and phonological details (e.g., Bell, Brenier, Gregory, Girand, & Jurafsky, 2009). Speakers' syntactic preferences are also driven by predictability of a syntactic structure. According to the *Uniform Information Density (UID)* hypothesis (Jaeger, 2006; Jaeger, 2010; Levy & Jaeger, 2007), human communication is viewed as information transmission over a capacity-limited noisy channel. The optimal strategy to transfer a message with high efficiency and with low risks of comprehension error is for speakers to try to distribute information uniformly across a message.

In traditional linguistic definitions, the information content of a sentence or a discourse is based on the compositional meaning of its words and constituents. However, in psycholinguistics and computational linguistics, *information* is used in its information-theoretic sense (i.e., *Shannon information*, Shannon, 1948), and *information density* means the amount of information conveyed per linguistic unit (e.g., phoneme, word, constituent, etc.). Associating information with *surprisal*, the more surprising a linguistic unit is in its context, the more information it conveys. For example, the more surprising the occurrence of a CC is in a given context, the more informative the CC onset is. Therefore, the UID hypothesis predicts that, where grammar permits, speakers will seek to structure their utterances so as to avoid peaks and troughs in information density. Peaks risk exceeding the channel's capacity, thus leading to comprehension difficulties; whereas, troughs would bring about redundancy and reduce transmission efficiency.

In the case of complementizer drop, Jaeger (2010) estimated the predictability of CC by the matrix verb's subcategorization frequency. If a CC is not predictable given the matrix verb, its appearance would be quite intense in information (because information and predictability are negatively correlated), and that information would have been added to the first words of the CC if

the complementizer had been omitted. In order to avoid such a peak, speakers are predicted to pronounce the complementizer to make the distribution of information more uniform. Inversely, if the appearance of a CC is highly expected, the complementizer will be redundant, and therefore its omission would be preferable. For example, since a CC is more predictable with *think* than with *confirm*, *that*-drop is more likely in (8) than in (7) [1]:

(7) My boss confirmed (that) we were absolutely crazy.

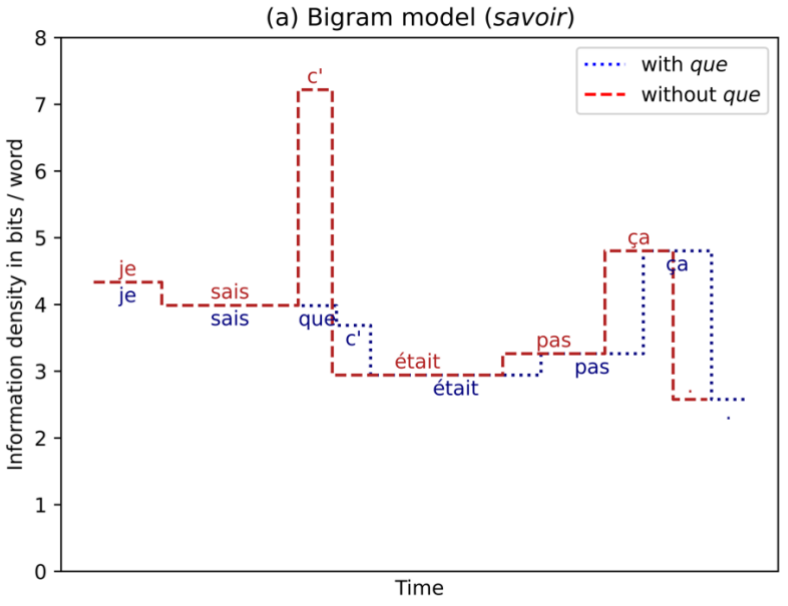
(8) My boss thinks (that) I am absolutely crazy. (Jaeger, 2010)

Figure 1 illustrates the idea of the UID hypothesis applied to complementizer omission in French: for each word of sentences (9) and (10), we have estimated a conditional probability using a bigram language model trained on the whole corpus *Montréal 84* (see Jurafsky & Martin, 2020: chapter 3 for a brief introduction to n-gram models), so that we can plot each word's Shannon information.[2] In Figure 1, we see that the CC onset (*c'* 'it') would be highly surprising if it immediately followed the verb *savoir* 'know', as shown by the dashed line in Figure 1(a), so that speakers will tend to use the complementizer, which is not only less surprising, but also makes the CC onset less surprising, so that the information density is more uniform with the complementizer. On the contrary, if the information density at CC onset is low (shown by the dotted line in Figure 1(b)), speakers are predicted to omit *que*, which would avoid a trough in information at CC onset and make the production more efficient (see example (10) and dashed line in Figure 1(b)).

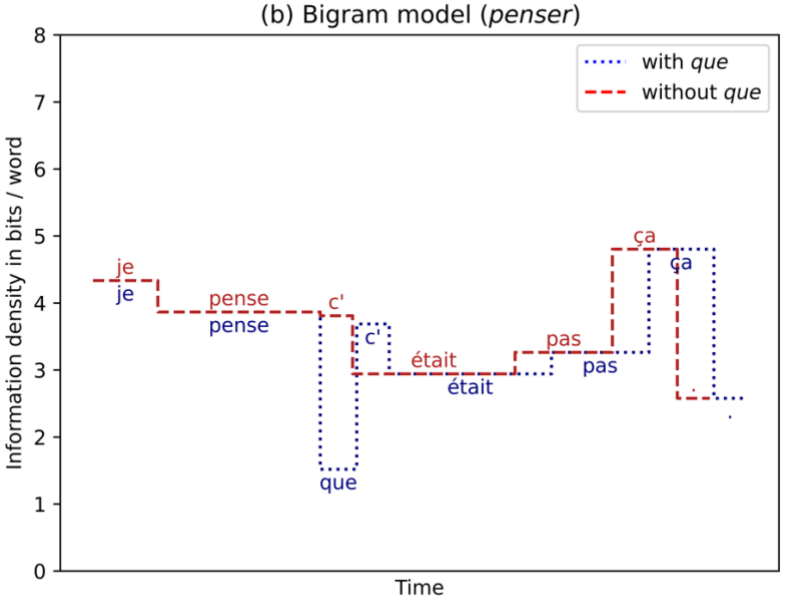
(9) Je sais (que) c'était pas ça.

(10) Je pense (que) c'était pas ça.

'I know/think (that) it wasn't that.'



(a)



(b)

FIGURE 1**.** Illustration of the information density in bits per word in time for two French CCs with (dotted lines) and without (dashed lines) the complementizer *que*. Figures (a) and (b) respectively show the information density of a non-predictable CC embedded by the matrix verb *savoir* "know" and a predictable CC embedded by *penser* "think".

UID plays a significant role in conditioning complementizer drop in the Switchboard corpus (Godfrey, Holliman, & McDaniel, 1992), on top of all the other factors (syntactic effects, disfluency

effects, etc.) that have been argued to play a role in previous work (Jaeger 2010) and successfully predicts other morphological and syntactic reduction phenomena in English such as auxiliary contraction, for example, *I have* versus *I've* (Frank & Jaeger, 2008) and *that*-relativizer omission (Levy & Jaeger, 2007). We therefore investigate the effect of information density on optional French complementizer *que*-omission by calculating the predictability of the CC given the matrix verb (i.e., subcategorization frequency of the matrix verb or CC-bias), to test the crosslinguistic validity of the UID hypothesis.

DATA AND METHOD

Corpus and semi-automatic annotation

The present study is based on *Montréal 84* (Thibault & Vincent, 1990), a French spoken corpus consisting of approximately 1.6 millions of words across sociolinguistic interviews with 72 Montréal natives of different genders, ages, education levels, and neighborhoods. All the interviews were transcribed and the main social characteristics of speakers such as genders and occupation are documented. A large corpus like this allows for the quantitative analysis of an infrequent syntactic phenomenon like *que*-omission in spontaneous speech.

We adopted a semi-automatic approach to annotate the corpus. The corpus was first POS-tagged with MElt (Denis & Sagot, 2012), and then, using a Python script, we extracted all the utterances containing a verb that can possibly embed a CC. We limited our attention in this study to verbs that occur more than 100 times in the corpus. As a matter of fact, due to less data, the estimation of the CC-bias (which is an approximation to information density of the CC onset in our study) of infrequent matrix verbs would be less reliable and the inclusion of these data might make the uneven distribution problem even worse. The extraction yielded a dataset with 24,635 sentences across 17 verbs (see Table 1). The script also identified the contexts preceding and following the verb. A second script subsequently coded complement clauses and *que*-omission cases along with all the factors (described in the next section) with the help of regular expressions. In case of

difficulty, we manually checked and annotated the token. At the end of this procedure, we coded 6113 observations as CCs, and more than three quarters of this data was coded by the script. However, 295 occurrences had to be removed because of missing values for some factors.

TABLE 1. *Verbs chosen for the study, ordered by CC-bias. Frequency = Frequency of the verb lemma in the corpus, CC = number of occurrences of CCs, CC-bias*[3] *= verb's subcategorization bias for a CC, O = number of* que *omissions*

Verb	**Frequency**	**CC**	**CC-bias**	**O**	**O/CC**
sembler *seem*	303	181	0.66	51	0.28
penser *think*	2456	1355	0.57	343	0.25
imaginer *imagine*	107	55	0.53	11	0.2
falloir *have to*	2406	1085	0.45	327	0.30
croire *believe*	195	76	0.40	7	0.09
remarquer *remark*	218	81	0.37	31	0.38
trouver *find*	2070	742	0.36	152	0.20
paraître *appear*	123	34	0.28	2	0.06
dire *say*	8322	1682	0.22	480	0.29
sentir *feel*	306	51	0.17	8	0.16
savoir *know*	3634	475	0.15	55	0.12
se souvenir *remember*	184	27	0.15	13	0.48
rappeler *remind*	205	24	0.12	12	0.5
vouloir *want*	2809	193	0.07	15	0.08
comprendre *understand*	690	40	0.06	3	0.08
demander *ask*	476	10	0.02	0	0
préférer *prefer*	131	2	0.02	0	0

The omission rate for the remaining 5818 complement clauses produced by 72 speakers is 24.7%[4] (1510 cases), which is similar to Sankoff (1980)'s findings (23%) in Montreal French, and

does not differ much from Martineau (1985) (32%) and Dion (2003) (37% for young speakers and 32% for older speakers) among complement clauses in Ottawa-Hull French. However, our omission rate differs from the 14% observed by Warren (1994) in the Montréal 84 corpus. This difference may result from different methodology: Warren looked at the first 400 hundred lines of each interview, whereas we used the entire interview but only for selected (high-frequency) verbs. Complementizer omission may vary over the course of the interview, because the first part of speech is generally one of the most formal parts, as observed by Martineau (1985). We are not in the position to say whether this variable is stable from 1971 to 1984 because previous work on the 1971 Montreal corpus only concentrates on the beginning of recordings.

Factors coded

Social factors

As shown by previous work, the omission of *que* is socially stratified (e.g., Dion, 2003; Martineau, 1985, 1988; Sankoff, 1980; Sankoff & Cedergren, 1971; Warren, 1994). The coding of social factors was mainly based on the classification made by the authors of the corpus:

1. SPEAKER AGE. Given the observation of Labov (1966) on linguistic variation in English, young speakers are more likely to use non-standard variants. This is consistent with Warren (1994)'s observations on *que* omission. However, such an effect is less clear in Martineau (1985) and in Dion (2003). We coded SPEAKER AGE in two ways: 1) a continuous variable ranging from 15 to 75 in the main statistical model, and 2) alternatively, a 3-level categorical variable: "under 25," "26-60," and "over 60". These groups correspond roughly to speakers' relation to the workplace (cf. Wagner & Sankoff, 2011): those below 25 years old have not, or have just entered into the working careers, while those beyond 60 years old should have retired and have no longer closed relation to the workplace.[5]

2. *SPEAKER GENDER*. Regarding *que*-omission, Warren (1994) shows that men are more likely

to omit *que* than women, but this holds only for young speakers. Other studies (Dion, 2003; Martineau, 1985; Martineau, 1988) have failed to detect a gender effect. SPEAKER GENDER was included as a binary variable in our model.

3. *SPEAKER EDUCATION*. Based on groups made in the corpus documentation, SPEAKER EDUCATION was coded as a three-level ordinal variable: low (some high school education), medium (high school graduates with no university degree), high (university graduates).

4. *SPEAKER OCCUPATION*. Optional *que* omission is also affected by speaker's occupation. Dion (2003) and Sankoff et al. (1971) found that workers without diplomas tend to make more omissions than other speakers. Following the classification and the ordering made by the authors of *Montréal 84*, SPEAKER OCCUPATION was coded as a six-level ordinal variable: professionals (liberal professionals and business leaders), graduates (other university graduates), technicians (technicians and foremen), white-collar, blue-collar, unemployed.

Since Canadian French is in close contact with English, some researchers raise question about the influence of bilingualism on *que* omission. However, Blondeau & Nagy's (2008) study of Anglo-Montrealers in both French and English reported that the rate of complementizer omission in French is 23%, which is the same as Sankoff (1980)'s observation. Since the information about speakers' bilingual status is not available in our data, we did not investigate this factor.

Linguistic factors

Four linguistic factors were included in the analysis, ranging from phonological to syntactic:

1. RIGHT PHONOLOGICAL CONTEXT. The omission of *que* is strongly driven by the right phonological environment of the complementizer (Dion, 2003; Martineau, 1985; Sankoff, 1980; Sankoff et al., 1971; Warren, 1994). Speakers show a higher preference for *que*-omission if the right phonological context is less sonorant, and this effect has been attributed to consonant cluster simplification in Quebec French. We used the Python module *epitran* (Mortensen, Dalmia, &

Littell, 2018) to transcribe automatically the first word following the complementizer into International Phonetic Alphabet (IPA). We checked manually the automatic transcription, then coded the first phoneme adjacent to the complementizer, i.e., RIGHT PHONOLOGICAL CONTEXT, as a three-level ordinal variable based on the sonority hierarchy: *obstruent* (*n* = 3608), *sonorant* (*n* = 616), and *vowel* (*n* = 1594). We do not distinguish sibilants from other types of obstruents, since in Canadian French, fricatives (including sibilants) and plosives are on the same sonority level (Côté, 2004).

We are aware of the potential non-orthogonality of the factors, especially between the following phonological context and the CC subject, which has triggered a long debate (cf. Connors, 1975; Dion, 2003). Note that the RIGHT PHONOLOGICAL CONTEXT is not necessarily the first phoneme of the CC subject, because of intervening material like prepositional phrases at the beginning of the CC, as illustrated by the following examples. For example, the right phonological context is coded as 'vowel' in (11a) and 'sonorant' in (11b). These represent 14.7% of cases (more than 800 observations). The use of the adjacent phonological context rather than the first segment of CC subject helps to dissociate these two factors to some extent. In addition, several statistical tests have been employed to ensure that there is no severe non-orthogonality between the phonological context and the CC subject.

(11)a. Je pense *pas* qu'*en* soixante-et-onze je travaillais là. (speaker 2)

'I think in the years of seventies I worked there.'

b. Puis on a toujours *pensé* que *les* cinq et demie on pouvait pas s'en acheter. (speaker 4)

'Then we were always thinking that we could not afford to buy five and a half.'

2. LEFT PHONOLOGICAL CONTEXT. Sankoff (1980) and Sankoff et al. (1971) found that the phonological environment preceding the complementizer also affects the omission of *que*; whereas Dion (2003), Martineau (1985), and Warren (1994) found no evidence for this effect. The last sound

preceding the CC, i.e., the LEFT PHONOLOGICAL CONTEXT, is therefore included in the model as a three-level ordinal variable based on the sonority hierarchy: *obstruent* (*n* = 1823), *sonorant* (*n* = 571), and *vowel* (*n* = 3424). Likewise, we use the left adjacent phonological context instead of the last segment of the embedding verb, so as to separate these two factors in variable coding.[6] As a matter of fact, the left phonological context is coded as 'vowel' in both (11a) and (11b). We found that in 13.8% of cases, the matrix subject is not adjacent to the CC.

3. MATRIX SUBJECT. According to grammaticalization accounts, 1st and 2nd person matrix subjects can express epistemicity (Thompson & Mulac, 1991b). It is observed that 1st and 2nd person matrix subjects correlate with lower rates of complementizer use than other types (Thompson & Mulac, 1991b), while Torres Cacoullos & Walker (2009) argued that when frequent main clause subject-verb collocations are excluded, pronominal matrix subjects favor zero-complementation more than full NPs. MATRIX SUBJECT was therefore coded as three ordered levels: *je_tu* (1st and 2nd person, *n* = 3362), *other type of pronoun* (*n* = 2407), and *lexical NP* (*n* = 49).

4. CC SUBJECT. Availability accounts (Ferreira & Dell, 2000) predict that more accessible CC subjects are associated with a higher rate of complementizer drop. This prediction is supported by previous work on English, which has reported that pronouns, especially those with local denotation like 1st person pronoun, are correlated with a higher rate of *that*-omission (Elsness, 1984; Ferreira & Dell, 2000; Jaeger, 2010). This pattern is, nevertheless, less clear for French, since Connors (1975) shows that pronouns differ from lexical NPs in favoring omission of *que*; whereas, Dion (2003) attributes it to phonological effects. CC SUBJECT was coded as a three-level ordinal variable based on the accessibility of its referential expression: *je_tu* (1st and 2nd person, *n* = 1897), other pronouns (*n* = 3408), and NP (*n* = 513). For example, CC subject is *je_tu* and 'other pronoun' in (11a) and in (11b), respectively. Moreover, in case of left dislocation (e.g., the lexical NP *mon père* 'my father' repeated by the 3rd person pronoun *il* in (12)), the CC subject was coded as 'lexical NP' (for

example in (12)). This notation is consistent with the proposition of Auger (1998), who argued that in Quebec French, the anaphoric pronoun can be analyzed as a clitic that agrees with the NP like inflections.[7]

(12) […] parce que je pense que *mon père* il buvait beaucoup […] (speaker 4)

'...because I think that my father he used to drink a lot...'

Cognitive factors

We investigated two factors that are associated with general cognition:

1. FREQUENCY OF MAIN VERB. A continuous variable ranging from 107 to 8322. It was calculated based on the frequencies of matrix verb observed in the *Montréal 84* corpus. Increasing production pressure could be attributed to less frequent words (Jescheniak & Levelt, 1994). According to availability accounts, if a CC is adjacent to the matrix verb, production pressure may spill over from the matrix verb to the CC onset, thus favoring the use of a complementizer. Several corpus studies have found a correlation between a less frequent matrix verb and a higher rate of *that*-use (Elsness, 1984; Jaeger, 2010; Roland, Elman, & Ferreira, 2006).

2. CC BIAS. The UID hypothesis (Jaeger, 2010) assumes an increasing preference for *que*-drop as the information density at the CC onset (i.e., the first word in the CC without the complementizer) lowers. We estimate the information density at the CC onset by the CC bias of the matrix verb (i.e., the number of CCs divided by the sum of CCs and non-CCs after removal of ambiguous cases). Therefore, the prediction is that speakers omit *que* more often if the CC is predictable given the matrix verb. For example, since the CC is more predictable with *penser* 'think' than with *savoir* 'know', we predict that speakers omit the complementizer more often with *penser* 'think'. The CC BIAS, ranging from 0.02 to 0.66 (see Table 1), was therefore included in the model.

Statistical modeling procedure

We employed a generalized linear mixed model (GLMM) to perform a multivariate regression

analysis on our data, which is highly unbalanced and clustered (mean number of observations per speaker = 80.8, median = 68.5, mode = 59, range = 8 - 280, SD = 54.8). We used the *glmer()* function of the *lme4* package (Bates, Mächler, Bolker, & Walker, 2015) in the statistical software R (R Core Team, 2020). Apart from all the fixed effects, the model also includes a random effect of SPEAKER to control inter-speaker variation.[8] The specification of the final model is shown as follows.[9] Omission or preservation of the complementizer *que* is respectively denoted by 1 and 0. Results will be visualized by effect graphs[10] in the following section.

The Model: *que*-omission modeled as depending on:

- Fixed effects: speaker age + speaker gender + speaker education + speaker occupation + matrix subject + CC subject + right phonological context + left phonological context + CC bias + frequency of main verb
- Random effect: speaker

After fitting the statistical model to the data, we evaluated the fitted model. The model correctly classifies 79% of the data overall. The estimated probabilities of the model shown in Figure 2 also show an acceptable fit.

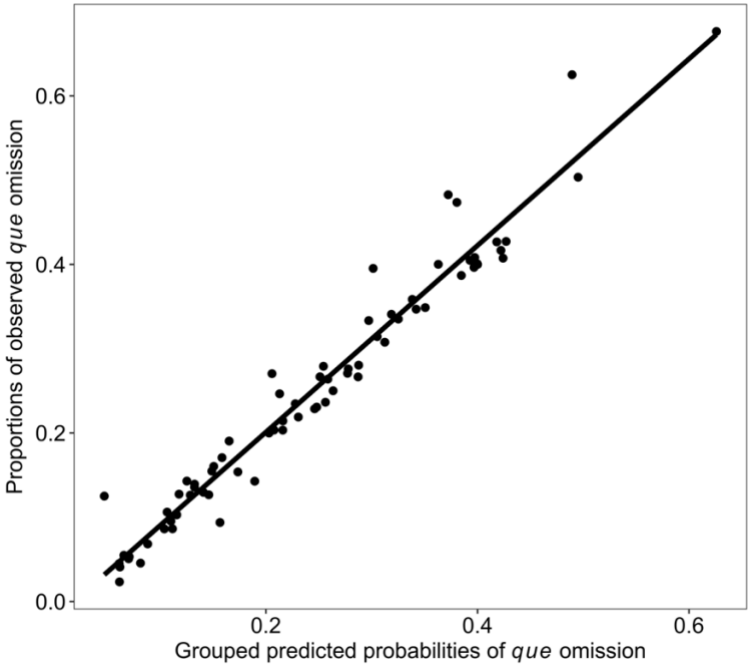


FIGURE 2. Mean predicted probabilities *vs.* observed proportions of omitted *que*. The data are grouped by speakers and the diagonal line represents a perfect match between predicted and actual proportions.

We also evaluated the eventual collinearity problem by calculating the variance inflation factor (VIF) of each variable. Collinearity is a severe modeling problem which appears when two independent variables are highly correlated with each other in the model (Allen, 1997). Since some variables in the model are polynomial, we applied the GVIF (General variance inflation factors) measure (Fox & Monette, 1992). The GVIF is proportional to the inflation due to collinearity in the confidence interval for the coefficient. We checked that GVIF^((1⁄2Df)) < 2 for each variable (Df is a variable's degree of freedom), which is usually interpreted as a low degree of collinearity (more or less corresponding to VIF < 4 for one-coefficient variables). The value for each variable is presented in Table 2, and we concluded that our model has no major concern of collinearity.

RESULTS AND DISCUSSION

Table 2 summarizes the results of all the fixed effects that were tested in the present study. In our regression model, for categorical predictors, we have compared each of the lower level with the higher one in the order clarified in Section DATA AND METHOD. For example, for polynomial variable like CC SUBJECT, we have contrasted 'other pronoun' against '*je_tu'*, and 'NP' against 'other pronoun'. The results show that *que* omission is conditioned by cognitive, linguistic and social factors.

TABLE 2. *Result summary: coefficient estimates* β, *standard errors SE*, z *value,* p *value and significance level indicated by stars * for all the variables in the model. A positive coefficient means that the first level correlates with a higher rate of* que*-omission than the second (number of CCs = 5818, number of* que*-omission cases = 1441)*

Predictor	**Coef.** β	**SE**	z	p		**GVIF^((1⁄2 Df))**
(Intercept)	-2.14	0.17	-12.49	< 2e-16	***	
SPEAKER AGE[11]	0.05	0.08	0.55	0.58		1.13

SPEAKER GENDER *F* vs. *M*	-0.05	0.20	-0.26	0.79		1.21
SPEAKER EDUCATION						1.36
= *medium* vs. *low*	0.20	0.21	0.89	0.37		
= *high* vs. *medium*	-1.01	0.32	-3.17	0.002	**	
SPEAKER OCCUPATION						1.18
= *graduates* vs. *professionals*	-0.09	0.32	-0.28	0.78		
= *technicians* vs. *graduates*	0.61	0.34	1.82	0.07	.	
= *white-collar* vs. *technicians*	-0.82	0.29	-2.80	0.005	**	
= *blue-collar* vs. *white-collar*	0.85	0.33	2.57	0.01	*	
= *unemployed* vs. *blue-collar*	-0.24	0.32	-0.74	0.46		
MATRIX SUBJECT						1.09
= *other pronoun* vs. *je_tu*	-0.09	0.09	-1.07	0.29		
= *NP* vs. *other pronoun*	-0.22	0.41	-0.54	0.59		
CC SUBJECT						1.12
= *other pronoun* vs. *je_tu*	-0.20	0.09	-2.52	0.01	*	
= *NP* vs. *other pronoun*	-0.64	0.19	-3.35	< 0.001	***	
RIGHT PHONOLOGICAL CONTEXT						1.10
= *sonorant* vs. *obstruent*	-1.11	0.17	-6.72	< 0.001	***	
= *vowel* vs. *sonorant*	-0.93	0.19	-4.76	< 0.001	***	
LEFT PHONOLOGICAL CONTEXT						1.16
= *sonorant* vs. *obstruent*	-0.28	0.15	-1.860	0.06	.	
= *vowel* vs. *sonorant*	0.12	0.12	0.94	0.35		
FREQUENCY OF MAIN VERB	0.40	0.05	8.35	< 0.001	***	1.37
CC BIAS	0.32	0.05	6.11	< 0.001	***	1.42

Table 3 provides the *que*-omission rate and number of tokens of each predictor level.

TABLE 3. *Que-omission rate, number of omission cases and number of CCs for each predictor level*

Predictor	**Level**	***que*-drop rate (%)**	***que*-drop tokens**	**CCs**
SPEAKER AGE	(continuous)	24.8	1441	5818

	<= 25	32.9	246	748
	26-59	24.8	1048	4234
	>= 60	17.6	147	836
SPEAKER GENDER	male	25.8	735	2846
	female	23.8	706	2972
SPEAKER EDUCATION	low	30.3	566	1871
	medium	30.8	635	2065
	high	12.8	240	1882
SPEAKER OCCUPATION	professionals	11.5	81	703
	graduates	12.5	106	845
	technicians	33.8	423	1252
	white-collar	19.9	303	1526
	blue-collar	34.9	252	723
	unemployed	35.9	276	769
MATRIX SUBJECT	*je_tu*	25	842	3362
	other pronoun	24.5	589	2407
	NP	20.4	10	49
CC SUBJECT	*je_tu*	36	683	1897
	other pronoun	20.8	709	3408
	NP	9.6	49	513
RIGHT PHONOLOGICAL CONTEXT	obstruent	35.2	1269	3608
	sonorant	10.6	65	616
	vowel	6.7	107	1594
LEFT PHONOLOGICAL CONTEXT	obstruent	25.6	467	1823
	sonorant	24.9	142	571
	vowel	24.3	832	3424
FREQUENCY OF MAIN VERB	(continuous)	24.8	1441	5818
CC BIAS	(continuous)	24.8	1441	5818
Total	(continuous)	**24.8**	**1441**	**5818**

Cognitive factors

As predicted by UID hypothesis, there was a clearly significant effect of information density on *que* omission: the more likely a verb is to appear with a CC, the more it favors the omission ($p < 0.001$) (cf. Figure 3). For example, it is more likely to observe *que* drop with *sembler* ‘seem’ in (13a) than

with *comprendre* 'understand' in (13b), given that *sembler* is more often followed by CCs than *comprendre*. Since CC-bias is an indicator of the information density at the CC onset, the result implies that speakers show a higher preference for *que*-drop if the CC onset is less informative, so as to avoid redundancy and thus increase communication efficiency.

(13) a. […] (il) me semble ~~que~~ la musique est belle […] (speaker 6)

'...(it) seems to me ~~that~~ the music is beautiful...'

b. Je comprends que c'est pas gros. (speaker 27)

'I understand that it is not much.'

Furthermore, χ^2 -tests were performed to compare the model described in Section DATA AND METHOD against the same one without one predictor. It turns out that the CC-bias has emerged as the third most important predictor in terms of its contribution to the improvement in model quality ($\chi^2_{\Delta(\Lambda)}(1) = 38.671$, $p = 5.017e-10 < 0.001$, after the RIGHT PHONOLOGICAL CONTEXT and the FREQUENCY OF MAIN VERB), which contrasts with Jaeger (2010)'s work where it is the strongest predictor. The effect of predictability on syntactic variation across languages is expected given that the UID makes predictions about general cognitive and communication principles which should not differ much among speakers. However, the importance of information density with regard to syntactic reduction may vary across languages, since speakers' preferences may be the result of competition among different linguistic rules (in particular phonological rules) and communication strategies in different ways. For example, phonological constraints do not play an important role in governing complementizer drop in English, while information density has a dominant role; whereas in French, the sonority hierarchy is more influential, so information density has a weaker role.

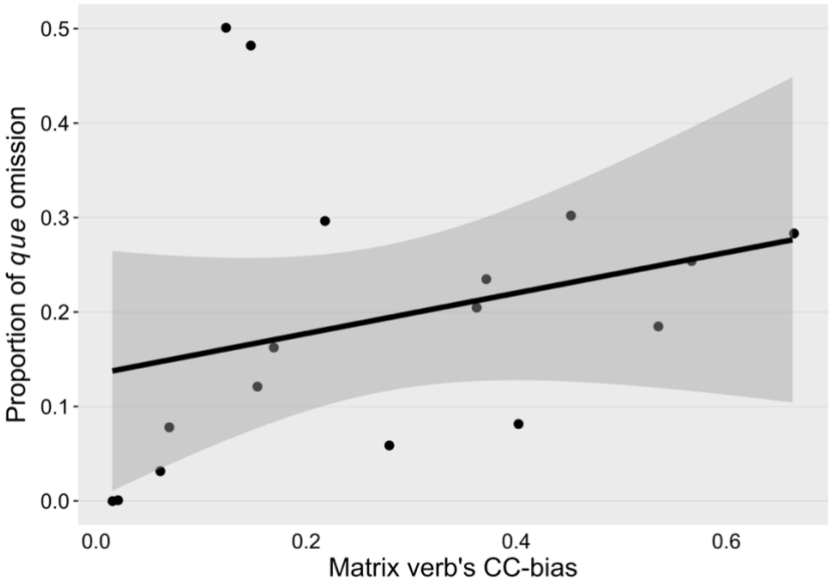


FIGURE 3. Effect of the matrix verb's CC bias on the omission of the complementizer *que*, along with 95% confidence interval (shaded area). The dots represent matrix verbs and the line indicates the linear model.[12]

In line with previous findings on *that*-use (Jaeger, 2010; Roland et al., 2006), our results also show that matrix verb frequency is a highly significant predictor: the more frequent a verb is in the corpus, the more likely *que* is to be omitted when it introduces a CC ($p < 0.001$). Jaeger (2010) has attributed this effect to the availability accounts, since less frequent matrix verbs are less accessible and may lead to production pressure that can spill over into the upcoming adjacent complement clause, thus encouraging the use of a complementizer. However, since a word's probability can be simply estimated by its frequency when ignoring context (Jaeger, 2011), it is possible that the frequency effect could be related to the information density.[13] Hence, further work is needed to flesh out how exactly verb frequency and CC onset informativity are related and whether this effect is related to availability accounts or to the UID hypothesis.

Linguistic factors

As Table 2 shows, the segment following the site of *que* has a significant effect on whether or not it

will be omitted ($p < 0.001$). Confirming Dion (2003), Sankoff (1980) and Warren (1994)'s findings, the highest rate of *que* omission is when it would be followed by an obstruent (35.2%[14]), then a sonorant (10.6%), and finally a vowel (6.7%) (see Figure 4 and examples in (14)). Chi-squared tests for partial effects show that the effect of RIGHT PHONOLOGICAL CONTEXT is larger than any other variable in the model ($\chi^2_{\Delta(\Lambda)}(2) = 436.3$, $p = 2.2e-16 < 0.001$). However, the effect of the LEFT PHONOLOGICAL CONTEXT has not been observed in our study.

(14) a. Bon bien je pense ~~que~~ *j*'ai tout. (obstruent, speaker 27)

'Well I think ~~that~~ I have all.'

b. Disons ~~que~~ *moi* j'ai été refusée comme di distributeur [...] (sonorant, speaker 7)

'Let's say ~~that~~ me I was refused as a distributor […]'

c. Ici on dit ~~qu'~~ *on* va étirer ça. (vowel, speaker 1)

'Here we say ~~that~~ we will avoid this.'

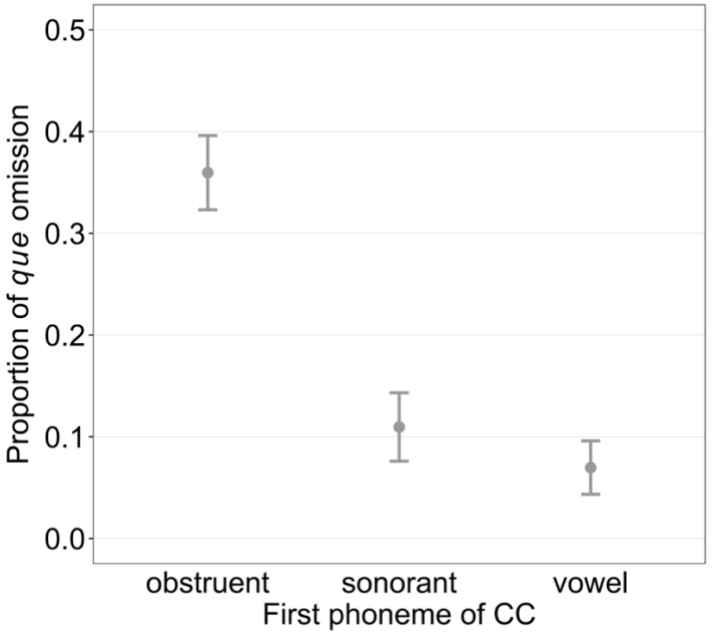


FIGURE 4. Right phonological context *vs. que* omission (with 95% confidence interval).

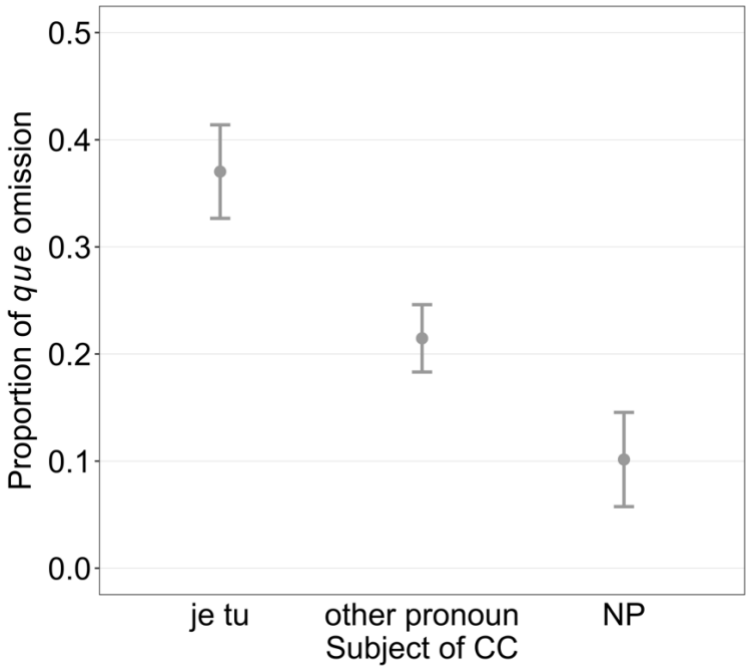


FIGURE 5. CC subject *vs. que* omission (with 95% confidence interval).

This pattern can be understood if we view *que* omission as the optimal strategy in Quebec French for repairing disfavored consonant clusters at the beginning of CC. Since, in modern French, the complementizer *que* (pronounced [k][15]) is the head of the CP domain (complementizer phrase) and should first merge with its complement clause (Kayne, 1976), it must be syllabified with the phonological material to its right. When the material following the obstruent [k] starts also with an obstruent, like [t] in the sequence *je pense que tu dors* 'I think that you sleep', it creates a cluster [kt] which violates the *Sonority Sequencing Principle* (SSP, particularly that onsets must increase in sonority, see Clements, 1990; Dell, 1995). European dialects of French often insert a schwa to repair such clusters; however, Quebec French prefers consonant cluster simplification: for example, evidence shows that speakers tend to simplify complex clusters in coda (Côté, 2012). Therefore, the best option here is to simply delete *que*, particularly if its information is low. In the case the SSP is not violated, we observed that speakers still omit *que*, and that happens more often when *que* is followed by a sonorant (e.g., [m] in (14b)) than a vowel (e.g., [ɔ̃] in (14c)), which suggests that Quebec French prefers larger intervals on the sonority scale. Hence, the less sonorous the right

phoneme, the more likely the *que* is to be omitted.

As for CC subject, our study shows that the most accessible CC subjects differ significantly from other pronouns ($p = 0.01$), and the contrast between other pronouns *vs.* lexical CC subjects also reaches significance ($p < 0.001$): *je_tu* (36%) > other pronoun (20.8%) > NPs (9.6%) (cf. Figure 5 and examples in (15)). These results are partially comparable with Connors (1975) and Dion (2003) in Canadian French and Jaeger (2010) in English, who found that pronouns and NPs behave differently with regard to complementizer drop. Moreover, we found an additional contrast between 1st and 2nd pronouns *vs.* other pronouns, which provides further support for availability-based accounts. In line with findings on English, *que*-omission is also driven by the accessibility of the CC subjects.

(15) a. Faut ~~que~~ *tu* regardes le positif dans ça. (*je_tu*, speaker 1)

'You have to look at the positive side of it.'

b. Quand je sais ~~que~~ *quelqu-un* parle très bien le français […] (other pronoun, speaker 77)

'When I know ~~that~~ someone speaker French very well […]'

c. Je sais ~~que~~ *mon père* a fait la Polytechnique. (NP, speaker 123)

'I know ~~that~~ my father went to the Polytechnique.'

Although the GVIF measure has shown no collinearity concern in the model, given the debate on whether CC SUBJECT and RIGHT PHONOLOGICAL CONTEXT have independent effects on *que* omission, we performed stepwise regression to further study this issue. More concretely, we use ANOVA to compare the model with all fixed factors (*m1*) and the identical model without CC subject (*m3*) or right phonological context (*m2*). Results shown in Tables 4 and 5 show that the model *m1*, the one including both CC subject and right phonological context, significantly fits the data better, with a lower AIC than the identical model without one of these variables like *m2* or *m3*, meaning a preference to include both variables. Besides, the cross-tabulation of right phonological

context and CC subject (cf. Table 6) also shows that these two factors are quite different, given the non-negligible number of tokens in each cell.

TABLE 4. *Results of comparison between the full model (m1) and the model without right phonological context (m2)*

Models	**Number of free parameters**	**AIC**	**BIC**	**Chisq**	**Df**	**Pr(>Chisq)**
m2	19	5746.5	5873.2			
m1	21	5314.1	5454.1	436.45	2	< 2.2e-16 ***

TABLE 5. *Results of comparison between the full model (m1) and the model without CC subject (m3)*

Models	**Number of free parameters**	**AIC**	**BIC**	**Chisq**	**Df**	**Pr(>Chisq)**
m3	19	5332.4	5459.1			
m1	21	5314.1	5454.1	22.259	2	1.468e-05 ***

TABLE 6. *Cross-tabulation of right phonological context and CC subject*

Proportion of *que* omission		**Right phonological context**			
		obstruent	sonorant	vowel	Total
CC subject	*je_tu*	658/1706 (38.6%)	15/114 (13.2%)	10/77 (13.0%)	683/1897 (36.0%)
	other_pronoun	596/1799 (33.1%)	18/137 (13.1%)	95/1472 (6.5%)	709/3408 (20.8%)
	NP	15/103 (14.6%)	32/365 (8.8%)	2/45 (4.4%)	49/513 (9.6%)
	Total	1269/3608 (35.2%)	65/616 (10.6%)	107/1594 (6.7%)	1441/5818 (24.8%)

Figure 6 further shows that pronouns behave quite differently from NPs within the same phonological group, especially when the CC begins with an obstruent; we have tried to encode the interaction between these two variables and refitted the model, but the interaction effect is not significant. Hence, unlike previous work which reduces phonotactic constraints and syntactic effect one to another (e.g., Connors, 1975; Dion, 2003), we conclude that both effects are independently important for *que* omission.

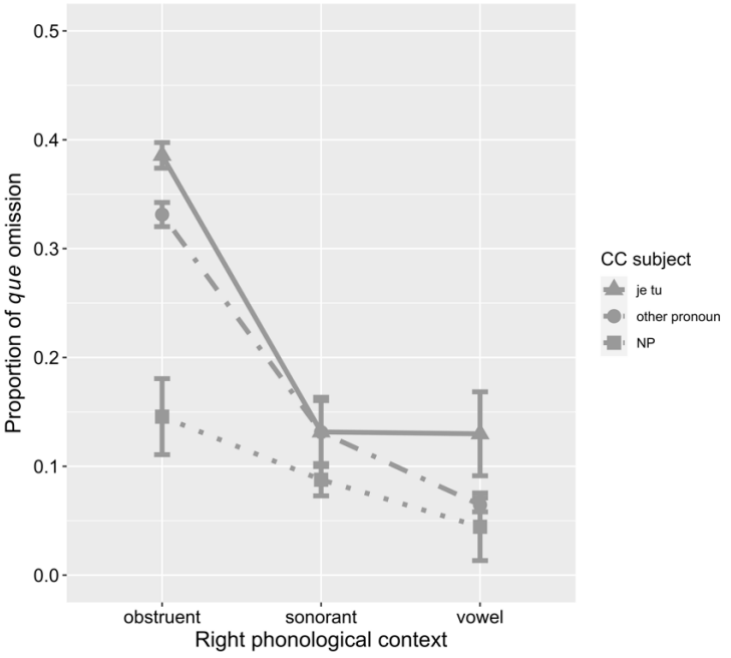


FIGURE 6. *que*-drop rate across CC subject and right phonological context groupings.[16]

Social factors

Both profession and education have significant effects on *que* omission (cf. Figures 7 and 8). In particular, we find that "speakers whose economic activity […] requires or is necessarily associated with competence in the *legitimized* language (or standard, elite, educated, etc. language)" (D. Sankoff & Laberge, 1978:239), that is, liberal professionals, white-collar workers, and other university graduates omit *que* less often than do the other members of the community, whose economic success does not depend so much on language (technicians and foremen, blue-collar workers, and the unemployed). These results partially contradict Warren (1994) on the same data, who groups each of the two adjacent groups into one group, and shows that the new "blue-collar and unemployed" and "technicians and white-collar" groups prefer omission whereas the new "professionals and graduates" group tends to retain *que*. We doubt whether it is pertinent to combine technicians and white-collar, since their working environments are quite different. The statistical results in Table 2 also reveal a higher omission rate associated with technicians and forepeople than with white-collar workers ($p < 0.01$). As for education, speakers having received a university

degree (i.e., high) are more likely to use the complementizer than those who have a medium or low education level.

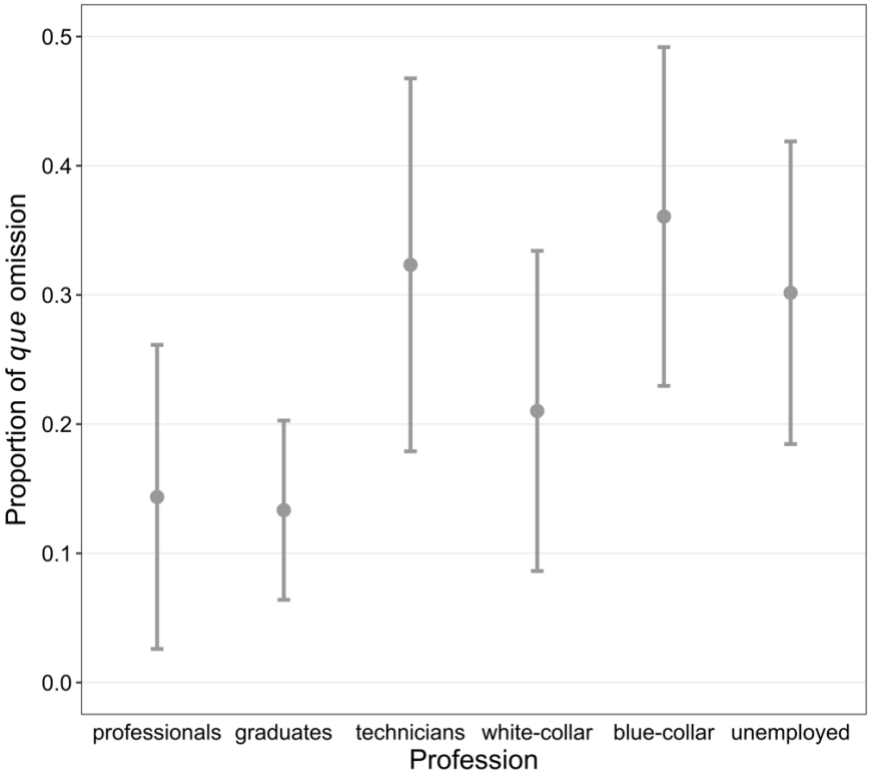


FIGURE 7. Profession *vs. que* omission.

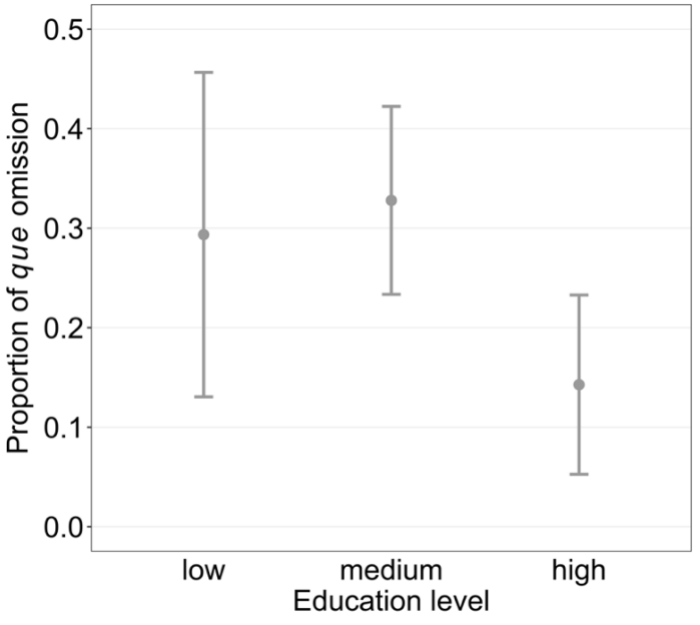


FIGURE 8. Education level *vs. que* omission.

These results can be explained by the *linguistic market* or *linguistic marketplace*[17] (in French *le marché linguistique*, Bourdieu & Boltanski, 1975), which refers to the symbolic market where linguistic exchange takes place. In order to achieve the highest profit from linguistic exchange, speakers should "produce the adequate speech in a given situation" (Bourdieu, 1977:647), that is, choose the linguistic variety with the best value in the linguistic market, which varies, for example, by the social relation between speaker and listener. Since the linguistic market related to higher education context (universities) and occupations such as those performed by white collar workers, liberal professionals and business leaders favor the use of the standard language, speakers occupying these social roles are under more pressure to pronounce the *que* (the standard variant) for the linguistic success in school or working life.

CONCLUSION

In this paper, we have provided an empirical study of *que* drop in the Montréal 84 corpus of spoken Montréal French. Our study is a contribution to the existing literature on this well-known variable, one which takes advantage of recent advances in computational linguistics, statistics, and cognitive science, and reveals that the *que* omission variable is conditioned by linguistic, cognitive and social factors. We show that many of the earlier results obtained by previous researchers on smaller subsets of corpora, including the Montréal 84 corpus, still hold at the level of the whole variationist corpus. Most importantly, the least sonorant right phonological context is shown to be the most important factor favoring *que* omission, but the accessibility of the syntactic subject of CC also has an independent role. In particular, we find that not all pronominal CC subjects behave in the same way: more accessible pronouns like *je_tu* favor the omission more than less accessible ones. In addition, two of the three most important factors conditioning the variable had not been tested before: frequency and information density. These two factors have been extensively studied in psycholinguistics, and our study shows that they are also useful for understanding how linguistic, social and general cognitive factors interact in variation. However, the importance of either of these

factors is weaker than the phonological constraint in French, which suggests that the importance of different types of constraints on speakers' preferences may vary in different languages.

More generally, we believe that our study illustrates an important way in which variationist corpora can continue to be valuable resources for studying fine-grained patterns of variation, particularly in their cognitive aspects. Currently, such corpora have been shown to be useful for diachronic comparative work (see Blondeau, Mougeon, & Tremblay [2019] for a recent example); however, given the greater *rapprochement* between variationist sociolinguistics and psycholinguistics in the past 15 years (see Tamminga, MacKenzie, & Embick, 2016), these corpora continue to be valuable resources for studying language in its social and cognitive aspects.

NOTES

1. The correlation between predictability of a structure and zero complementizer is also observed by Torres Cacoullos & Walker (2009), who report that *think*, co-occurring more often with a complement structure than *say* or *know*, is associated with a higher rate of zero complementizer.
2. The information of a linguistic unit, I($unit_i$), is defined by using its conditional probability given the context:

$$I(unit_i) = log \frac{1}{P(unit_i|context)}$$

3. Note that the denominator of CC-bias is not necessarily the frequency of matrix verbs. When manually identifying CCs, we had to leave out 1447 ambiguous cases (5.9%), for example "je pense que/∅ oui", or "Ils disaient on bien on va en profiter". Therefore, the CC-bias is calculated only among unambiguous cases.
4. Since it will be revealed later in this study that the frequency of the matrix verb is positively correlated with the number of que omissions, our selection of the 17 most frequent verbs may lead to an overestimate of the real omission rate.
5. We did not exactly follow the three age groups made by Warren (1994): 29-33, 35-64, and 66-73 years old. In fact, since we worked on the entire corpus, a noticeable number of speakers are too young to enter the working place. Similarly, in the 1980s, retirements often took place after the age of 60.
6. Dion (2003) also raised questions about the non-orthogonality between the preceding phonological context and the embedding verbs, as the last segment of verbs could be bound to a certain phonological category.
7. We also considered whether certain semantic classes of verbs favor the presence of *que* (such as verbs selecting the subjunctive or true factive verbs). However, as observed by Poplack, Lealess, & Dion (2013) and Kastronic (2016), the subjunctive mood has a very limited distribution in Canadian French, appearing with only two verbs in our list of frequent verbs in Montréal 84: *falloir* 'have to' and *vouloir* 'want' and there is only one true factive verb in our list: *savoir* 'know'. Therefore, we could not test them in this study.
8. Since we only have 17 different verbs, it is possible that the variation among verbs may be great if they were included as a random effect, and the model would raise its standard to a great extent, especially for factors closely linked to verb identity, such as the frequency and CC-bias of verbs. Since this is the first investigation of the *Uniform Information Density* hypothesis in French, we preferred not to include them as random effect in this study (following Jaeger, 2010).
9. The entire R script for statistical modeling is accessible through: https://osf.io/47e5r/?view_only=bd108eb1b4874c7d9215f0f0feffc96f.
10. Figures were generated by *plyr* (Wickham, 2011) and *ggplot2* (Wickham, 2016).
11. Grouping ages into three groups also reveals no significant effect on *que*-drop.
12. The two outliers in the upper-left corner of Figure 3 correspond to *rappeler* 'remind' and *se souvenir* 'remember' respectively. They have similar meanings and seem to belong to the same semantic class. However, given their restricted number of CCs in our corpus data (24 and 27 respectively), further study investigating these verbs in other corpora would be preferable.
13. A frequent verb is low in information, thus making the information level before CC onset already low. Therefore, the UID predicts the omission of the complementizer to avoid an information trough.
14. The percentage between brackets here and in the following paragraphs means the proportion of *que* omission.
15. The pronunciation of *que* is documented as [kə] in dictionaries. But in spoken French, the schwa of *que* is often dropped and thus realized as [k].

16. This figure was computed by *dplyr* (Wickham, François, Henry & Müller, 2020) and *ggplot2*.
17. The linguistic marketplace measure was developed for the 1971 Montreal corpus and coded speakers according to their professions in 1971 (Sankoff & Laberge, 1978). Between1971 and 1984, a number of speakers change their occupations (see, for example, Sankoff & Blondeau, 2007), so the 1971 linguistic marketplace characterization is out of date for the 1984 data. For this reason, we cannot code our social variables directly based on linguistic marketplace. Instead, we chose to measure socio-economic status using the speakers' education and occupation in 1984.

REFERENCES

Allen, Michael P. (1997). *Understanding Regression Analysis*. New York: Springer US.

Auger, Julie. (1998). Le redoublement des sujets en français informel québécois: une approche variationniste. *Canadian Journal of Linguistics/Revue canadienne de linguistique*, *43*(1), 37–63.

Bates, Douglas, Mächler, Martin, Bolker, Ben, & Walker, Steve. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1–48.

Bell, Alan, Brenier, Jason M., Gregory, Michelle, Girand, Cynthia, & Jurafsky, Dan. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, *60*(1), 92–111.

Bever, Thomas G. (1970). The cognitive basis for linguistic structures. In J. R. Hayes (ed.), *Cognition and the Development of Language* (pp. 279-352). New York: Wiley.

Blondeau, Hélène, Mougeon, Raymond, & Tremblay, Mireille. (2019). Analyse comparative de ça fait que, alors, donc et so à Montréal et à Welland : mutations sociales, convergences, divergences en français laurentien. *Journal of French Language Studies*, *29*(1), 35–65.

Blondeau, Hélène, & Nagy, Naomi. (2008). Subordinate clause marking in Montreal Anglophone French and English. In M. Meyerhoff & N. Nagy (eds.), *Social Lives in Languages. Sociolinguistics and Multilingual Speech Communities. Celebrating the work of Gillian Sankoff* (pp. 273–313). Amsterdam and Philadelphia, PA: John Benjamins.

Bourdieu, Pierre. (1977). The economics of linguistic exchanges. *Social Science Information*, *16*(6), 645–668.

Bourdieu, Pierre, & Boltanski, Luc. (1975). Le fétichisme de la langue. *Actes de la recherche en sciences sociales*, *1*(4), 2–32.

Bresnan, Joan, Cueni, Anna, Nikitina, Tatiana, & Baayen, R. Harald. (2007). Predicting the dative alternation. In G. Bouma, I. Kraemer, & J. Zwarts (eds.), *Cognitive foundations of interpretation* (pp. 69-94). Amsterdam: KNAW.

Clements, George N. (1990). The role of the sonority cycle in core syllabification. In J. Kingston & M. E. Beckman (eds.), *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech* (pp. 283–333). Cambridge: Cambridge University Press.

Connors, Kathleen. (1975). L'effacement de que—règle syntaxique. *Recherches linguistiques à Montréal, 4*, 17–33.

Côté, Marie-Hélène. (2004). Consonant cluster simplification in Québec French. *Probus*, *16*(2), 151–201.

Côté, Marie-Hélène. (2012). Laurentian French (Quebec) extra vowels, missing schwas. In R. Gess, C. Lyche, & T. Meisenburg. (eds.), *Phonological variation in French: Illustrations from three continents* (pp. 235-274). Amsterdam: John Benjamins Publishing.

Dell, Francois. (1995). Consonant clusters and phonological syllables in French. *Lingua*, *95*(1-3), 5–26.

Denis, Pascal, & Sagot, Benoît. (2012). Coupling an annotated corpus and a lexicon for state-of-the-art pos tagging. *Language Resources and Evaluation*, *46*(4), 721–736.

Dion, Nathalie. (2003). L'effacement du *que* en français canadien : Une étude en temps réel. *MA mémoire, University of Ottawa*.

Elsness, Johan. (1984). That or zero? A look at the choice of object clause connective in a corpus of American English. *English studies, 65*(6), 519-533.

Ferreira, Victor S., & Dell, Gary S. (2000). Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive psychology*, *40*(4), 296–340.

Ferreira, Victor S., & Firato, Carla E. (2002). Proactive interference effects on sentence production. *Psychonomic bulletin & review*, *9*(4), 795–800.

Fox, John, & Monette, Georges. (1992). Generalized collinearity diagnostics. *Journal of the American Statistical Association*, *87*(417), 178–183.

Frank, Austin F., & Jaeger, T. Florian. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 30, pp. 939-944). Washington, DC.

Godfrey, John J., Holliman, Edward C., & McDaniel, Jane. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of the 1992 IEEE international conference on Acoustics, speech, and signal processing* (Vol. 1, pp. 517–520). San Francisco, CA.

Hawkins, John A. (2001). Why are categories adjacent? *Journal of linguistics*, *37*(1), 1–34.

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

Jaeger, Florian T. (2005). Optional that indicates production difficulty: Evidence from disfluencies. In *DiSS-2005* (pp. 103-108). Aix-en-Provence.

Jaeger, Florian T. (2006). *Redundancy and syntactic reduction in spontaneous speech*. Stanford University. Stanford, CA.

Jaeger, Florian T. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive psychology*, *61*(1), 23–62.

Jaeger, Florian T. (2011). Corpus-based research on language production: Information density and reducible subject relatives. In E. M. Bender, J. E. Arnold (eds.), *Language from a cognitive perspective: Grammar, usage, and processing: Studies in honor of Tom Wasow* (pp. 161–197). Stanford, CA: CSLI Publications.

Jescheniak, Jorg D., & Levelt, Willem J. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4), 824.

Johnson, Daniel E. (2009). Getting off the Goldvarb standard: Introducing Rbrul for mixed-effects variable rule analysis. *Language and linguistics compass*, *3*(1), 359–383.

Jurafsky, Daniel, & Martin, James H. (3rd ed. draft, 2020). *Speech and language processing*. *URL:*

https://web.stanford.edu/~jurafsky/slp3/

Kastronic, Laura. (2016). *A comparative variationist approach to morphosyntactic variation in hexagonal and Quebec French*. PhD Thesis. Université d'Ottawa.

Kayne, Richard S. (1976). French relative que. In F. Hensey and M. Luján (Eds.), *Current Studies in Romance Linguistics* (pp. 255-299). Washington: Georgetown University Press.

Kleinschmidt, Dave F, Weatherholtz, Kodi, & Jaeger, T. Florian. (2018). Sociolinguistic Perception as Inference Under Uncertainty. *Topics in Cognitive Science*, 10(4), 818–834.

Labov, William. (1966). *The social stratification of English in New York city.* Washington, DC: Center for Applied Linguistics.

Levy, Roger, & Jaeger, T. Florian. (2007). Speakers optimize information density through syntactic reduction. In B. Schlökopf, J. Platt, & T. Hoffman (eds.), *Advances in Neural Information Processing Systems 19* (pp. 849-856), Cambridge, MA: MIT Press.

Martineau, France. (1985). *Elision variable de (que) dans le parler d'Ottawa-Hull*. University of Ottawa.

Martineau, France. (1988). Variable deletion of que in the spoken French of Ottawa-Hull. In J.-P. Montreuil, & D. Birdsong (eds.), *Advances in Romance linguistics* (pp. 275–287). Dordrecht: Foris.

Mortensen, David R., Dalmia, Siddharth, & Littell, Patrick (2018). Epitran: Precision G2P for many languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation.* Miyazaki.

Poplack, Shana. (1989). The care and handling of a megacorpus: The Ottawa-Hull French project. In R. Fasold & D. Schiffrin (eds.), *Language change and variation*. Amsterdam: John Benjamins Publishing.

Poplack, Shana, Lealess, Allison, & Dion, Nathalie. (2013). The evolving grammar of the French subjunctive. *Probus*, *25*(1), 139–195.

Poplack, Shana, & St-Amand, Anne. (2007). A real-time window on 19th-century vernacular French: The Récits du français québécois d'autrefois. *Language in Society*, 36(5), 707-734.

R Core Team (2020). R: A language and environment for statistical computing. R foundation for statistical computing. Vienna, Austria. *URL http://www. r-project. org*.

Roberge, Yves, & Rosen, Nicole. (1999). Preposition stranding and que-deletion in varieties of north American French. *Linguistica atlantica*, *21*, 153–168.

Rohdenburg, Günter. (1998). Clausal Complementation and Cognitive Complexity in English. In F.-W. Neumann, S. Schülting (eds.), *Anglistentag 1998 Erfurt* (pp. 101–12). Trier: Wissenschaftlicher Verlag Trier.

Roland, Douglas, Elman, Jeffrey L, & Ferreira, Victor S. (2006). Why is that? Structural prediction and ambiguity resolution in a very large corpus of English sentences. *Cognition*, *98*(3), 245–272.

Sankoff, David, & Laberge, Suzanne. (1978). The linguistic market and the statistical explanation of variability. In D. Sankoff (ed.), *Linguistic variation: Models and methods* (pp. 239–250). Cambridge, MA: Academic Press.

Sankoff, Gillian. (1980). *The social life of language*. University of Pennsylvania Press.

Sankoff, Gillian, & Blondeau, Hélène. (2007). Language change across the lifespan: /r/ in Montreal French. *Language*, *83*(3), 560–588.

Sankoff, Gillian, & Cedergren, Henrietta. (1971). Some results of a sociolinguistic study of Montreal French. *Linguistic diversity in Canadian society*, 61–87.

Sankoff, Gillian, & Cedergren, Henrietta. (1972). Sociolinguistic research on French in Montreal. *Language in Society*, *1*(1), 173–174.

Sankoff, Gillian, Sarrasin, Robert, & Cedergren, Henrietta. (1971). *Quelques considérations sur la distribution de la variable que dans le français de Montréal.* Communication presented at the *Congrès de l'Association-française pour l'Avancement des Sciences*. Sherbrooke.

Shank, Christopher & Plevoets, Koen. (2018). Investigating the impact of structural factors upon

that/zero complementizer alternation patterns in verbs of cognition: a diachronic corpus-based multifactorial analysis. *Research in Corpus Linguistics 6*, 83-112.

Shannon, Claude E. (1948). A mathematical theory of communication. *Bell system technical journal*, *27*(3), 379–423.

Tamminga, Meredith, MacKenzie, Laurel, & Embick, David. (2016). The dynamics of variation in individuals. *Linguistic Variation*, *16*(2), 300–336.

Thibault, Pierrette, & Vincent, Diane. (1990). *Un corpus de français parlé : Montréal 84 : Historique, méthodes et perspectives de recherche.* Université Laval. Québec.

Thompson, Sandra A., & Mulac, Anthony. (1991a). The discourse conditions for the use of the complementizer that in conversational English. *Journal of pragmatics*, *15*(3), 237–251.

Thompson, Sandra A., & Mulac, Anthony. (1991b). A quantitative perspective on the grammaticization of epistemic parentheticals in English. *Approaches to grammaticalization, Volume II. Types of grammatical markers* (pp. 313–329). Amsterdam: John Benjamins Publishing.

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

Warren, Jane. (1994). Plus ça change, plus c'est pareil: The case of 'que' in Montreal French. *Culture*, *14*(2), 39–49.

Wagner, Suzanne Evans, & Sankoff, Gillian. (2011). Age grading in the Montréal French inflected future. *Language Variation and Change*, *23*(3), 275–313.

Wickham, Hadley. (2011). The Split-Apply-Combine Strategy for Data Analysis. *Journal of Statistical Software*, *40*(1), 1–29.

Wickham, Hadley. (2016). *ggplot2: Elegant Graphics for Data Analysis.* New York: Springer-Verlag. URL: https://ggplot2.tidyverse.org.

Wickham, Hadley, François, Romain, Henry, Lionel, & Müller, Kirill. (2020). *dplyr: A grammar of*

data manipulation. R package version 1.0.0. URL: https://dplyr.tidyverse.org.