Faut je t'apporte quelque chose: Complementizer drop in Quebec French

Yiming Liang Pascal Amsili Heather Burnett

Laboratoire de Linguistique Formelle
Université Paris Diderot & CNRS
yiming.liang@etu.univ-paris-diderot.fr,
pascal.amsili@gmx.fr,heather.susan.burnett@gmail.com

1 Introduction

Natural languages often provide speakers with various options to express the same meaning, and identifying which social, linguistic and cognitive factors determine speakers' preferences is an active area of theoretical and empirical linguistics. In this paper, we test the *Uniform Information Density* hypothesis (Jaeger and Levy, 2007; Jaeger, 2010) through a study on complementizer drop in French complement clauses, aiming to bring more crosslinguistic evidence for this account.

Complementizer drop (1) is a well-studied feature of English ; however, it also appears in many dialects of French, such as those spoken in Quebec, Canada (2).

(1) My boss thinks (that) you're absolutely right. (Switchboard Corpus)

(2) Faut (que) je t'apporte quelque chose. (Français parlé à Ottawa-Hull Corpus)

In the present study, we investigate which factors condition speakers' choice between complementizer retention and omission in spoken Quebec French, by using the *Montreal 84* corpus of spoken Montréal French (Thibault and Vincent, 1990).

2 Complementizer drop in English

Previous work on complementizer drop has identified a number of linguistic and cognitive factors that influence the probability that a speaker will omit or pronounce *that* in a complement clause. These include the matrix subject (*I*, *you*, other pronouns *vs.* lexical NPs (Thompson et al., 1991; Cacoullos and Walker, 2009)), the subject of the embedded clause (*I*, other pronouns *vs.* lexical NPs (Elsness, 1984; Ferreira and Dell, 2000)), the distance of the complement clause from matrix verb (Elsness, 1984; Hawkins, 2001), the presence or absence of a disfluency (Ferreira and Firato, 2002; Jaeger, 2005), among others. Building on this work, Jaeger (2010) shows that, in addition to the previously identified factors, *information density* also plays a role in determining whether or not *that* will be omitted. Jaeger investigated this factor in the context of the *Uniform Information Density (UID)* hypothesis (Jaeger and Levy, 2007; Jaeger, 2010), which views communication as information transmission over a capacity-limited noisy channel. The UID can be stated as follows :

> Within the bounds defined by grammar, speakers prefer utterances that distribute information uniformly across the signal (information density). Where speakers have a choice between several variants to encode their message, they prefer the variant with more uniform information density. (Jaeger, 2010, 25)

The information of a word w_i, $I(w_i)$, is defined using the logarithm-transformed inverse of its conditional probability in the context. When applied to the phenomenon of complementizer drop in complement clauses (CCs), the UID predicts that speakers will be more likely to omit *that* in CCs with verbs that are more likely to appear with CCs overall. For example, since, as Jaeger shows, the verb *think* appears more often with a CC than the verb *confirm*, the information of *that* is lower when it follows *think* than when it follows *confirm*. Therefore, the UID predicts that *that* will be dropped at a higher rate after *think* than after *confirm*.

(3) My boss thinks (that) I am absolutely crazy.

(4) My boss confirmed that we were absolutely crazy. (Jaeger, 2010)

Using mixed effects statistical analysis (*multilevel logit model*), Jaeger shows that UID plays a significant role in conditioning complementizer

drop in the Switchboard corpus (Godfrey et al., 1992), **on top of** all the other factors that have been argued to play a role (syntactic effects, disfluency effects, etc.) in previous works.

3 Complementizer drop in French

The UID aims to be a hypothesis about human's communication and should generalize to more syntactic reduction phenomena crosslinguistically ; however, with respect to complementizer drop, its predictions have only been tested on English. Although complementizer drop is limited in Standard European French, it is a robust syntactic phenomenon in dialects of French spoken in Canada, as shown by the examples in (5-7).

(5) Ils disaient (que) c'était pas bon.
(CC, Français parlé à Ottawa-Hull)

(6) ... parce (qu') ici c'est bizarre.
(circumstantial clause, Montreal 84)

(7) C'est là (que) ma mère à moi vivait.
(relative, Roberge and Rosen (1999))

We therefore test the crosslinguistic validity of the UID through a corpus study of complementizer drop in Montréal French.

The study of complementizer drop in Canadian French has a rich history in sociolinguistics. Sankoff and Cedergren (1971) and Sankoff (1980) argue that phonological factors condition *que* omission in the Sankoff-Cedergren corpus of spoken Montréal French (Sankoff and Cedergren, 1972). In particular, Sankoff and colleagues show that sibilants favour *que* omission, compared with other sounds. As a result, they propose that the omission may be conditioned by the sonority hierarchy (Clements, 1990). The study of Warren (1994) further supported Sankoff et al.'s proposal, reporting that omission is preferred when the sound of the CC onset is less sonorant. In addition, Warren observes that omission often cooccurs with some specific verbs and contexts, like *je pense (I think)*. He attributes this lexical effect to the fact that these verbs are often used in epistemic phrases. Finally, Connors (1975) argues that a syntactic factor plays a role : she shows, also using the Sankoff-Cedergren corpus, that speakers tend to drop *que* when the subject of CC is a pronoun rather than a NP.

The earlier studies focused on individual conditioning factors separately ; however, Dion (2003) presented a multifactorial study of two spoken corpora, *Français parlé à Ottawa-Hull (OH)* (Poplack, 1989) and *Récits du français Québécois d'Autrefois (RFQ)* (Poplack and St-Amand, 2007). After studying a list of factors ranging from social to linguistic, she found that *que* omission was significantly conditioned by social factors : unskilled workers delete *que* more often than other members of the community. She replicates the effect of the sonority hierarchy described by Sankoff and colleagues, finding that omission occurs most before obstruents, then sonorants, then least before vowels. Moreover, Dion argues that the syntactic effect observed by Connors (1975) should actually be analyzed as part of the phonological effect, since, in her study, the pronouns starting with a vowel clearly disfavour omission. She also noticed that certain verbs, like *rappeler* 'remind', *sembler* 'seem' and *penser* 'think', favour *que*-deletion, but she has no clear interpretation for this lexical effect.

Although previous work on *que* omission in Canadian French has provided valuable insights into the source of this phenomenon, this work is limited for a number of reasons. First, Sankoff, Connors and Warren's research are based on small subsets of their corpora (16 speakers) without any modern statistic analysis, which makes their results less reliable. Likewise, Dion studied only 14 speakers, which represents only half of the total data found in the Ottawa-Hull corpus. Furthermore, her multivariate analysis was limited to social factors and was carried out in *Goldvarb*, a program which assumes a balanced distribution of data. However, in corpus studies, it is often the case that a few speakers contribute most of the data while many speakers contribute little data. Therefore, more sophisticated statistical modelling techniques, such as mixed methods, should be used (see (Johnson, 2009)). Finally, unlike research on English complementizer drop, none of the previous work on French investigated more general cognitive factors such as information density. We therefore present a new study of social, linguistic and cognitive factors conditioning *que* omission based on the entire Montréal 84 corpus.

4 Data and methods

4.1 Data source and method

Montreal 84 (Thibault and Vincent, 1990) is a spoken corpus consisting of approximately 1.6

millions of words across 72 sociolinguistic interviews with Montréal natives of different genders, ages, education levels and neighbourhoods. The interviews were transcribed ; however, the corpus does not contain any morphological or syntactic annotation. Therefore, in order to extract the CCs and code them for the presence/absence of *que*, we had to adopt a semi-automatic approach and add these annotations to the corpus. The corpus was first POS-tagged with MElt (Denis and Sagot, 2012), and then, using a Python script, we extracted all the utterances containing a verb that can possibly embed a complement clause. We limited our attention in this study to verbs that occur more than 100 times in the corpus. This yielded a dataset with more than 24,000 observations across 17 verbs. The script also identified the contexts preceding and following the verb. A second script subsequently coded complement clauses along with all the factors. The automatic annotation was then completed by a manual check. At the end of this procedure, we coded about 6,000 observations as complement clauses, and more than three quarters of this data was coded by the script.

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

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

4.2 Factors

We first conducted a pilot study where we coded on the highly frequent verb *penser* all the factors that had been previously argued to play a role in complementizer drop in English or French. Based on this study, we identified 10 factors which seemed likely to be active in our dataset.

4.2.1 Social factors

We investigated four social factors : 1) SPEAKER AGE : a continuous factor ranging from 15 to 75 ; 2) SPEAKER GENDER : male *vs.* female ; 3) SPEAKER EDUCATION : a three level factor : 1 (some high school education), 2 (high school graduates with no university degree), 3 (university graduates) ; and 4) SPEAKER OCCUPATION : a six level factor : 1 (liberal professionals and business leaders), 2 (other university graduates), 3 (technicians and foremen), 4 (white collar workers), 5 (blue collar workers), 6 (unemployed people).

4.2.2 Linguistic factors

We investigated 6 linguistic factors, ranging from phonological to syntactic :

1. MATRIX SUBJECT : according to grammaticalization accounts (Thompson et al., 1991), 1st and 2nd person matrix subjects should correlate with the lowest rates of complementizer mentioning because of their epistemic use, followed by other pronouns and NPs. MATRIX SUBJECT was therefore coded as three ordered levels : *je_tu* (1st and 2nd person), other types of pronouns and NPs.
2. CC SUBJECT : based on availability accounts (Ferreira and Dell, 2000), more accessible CC subjects should correlate with a lower rate of complementizer mentioning. Hence, two groups of CC subject were included in the statistical model : accessible subjects regrouping *je, tu, il expletif, ce* and *ça*, and less accessible subjects involving all the remaining types of pronouns and NPs.
3. LEFT PHONOLOGICAL CONTEXT : we coded the type of the last sound before the CC in 3 ordered levels : obstruent, sonorant and vowel.
4. RIGHT PHONOLOGICAL CONTEXT : the same 3-way distinction of the first sound of the CC as LEFT PHONOLOGICAL CONTEXT. Notice that the complementizer is not necessarily followed by the CC subject, since PPs or other intervening words could appear at the beginning of CC, like in *je pense pas qu'en soixante-et-onze je travaillais là. (Corpus montreal 84)*.

We also considered whether certain semantic classes of verbs favour or disfavour the presence of

que. In particular, perhaps verbs selecting the subjunctive would be more likely to preserve *que* than those selecting the indicative ? Or perhaps true factive verbs would be more likely to disfavour *que* than non-factives ? Unfortunately, as observed by Poplack et al. (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* and *vouloir*. Therefore, it was not possible to include this factor in the statistical analysis. Likewise, as shown in TABLE 1, there is only one true factive verb in our list : *savoir*. Therefore, the influence of factivity could not be tested statistically either.

4.3 Cognitive factors

We investigated two factors that are associated with general cognition.

1. FREQUENCY OF MAIN VERB : a continuous factor ranging from 107 to 8322. It was calculated based on the frequencies found within the Montréal 84 corpus.
2. PROBABILITY OF CC : a continuous factor. Following Jaeger (2010), we calculated the probability of having a CC for each verb (by dividing the number of occurrences of CCs by the number of occurrences of verb lemma in the final database). Notice that some previous study on syntactic reduction has employed more sophistic models to evaluate the information density (see for example (Resnik, 1996; Jaeger and Levy, 2007)). Given the time constraints, we focused on Jaeger (2010)'s method in this paper, which is to estimate the information density based on verb's subcategorization frequency.

4.4 Statistical analysis

We employed the *multilevel logit model* to examine the effect of each factor. One advantage of using this model is that it can reliably analyse even highly unbalanced and clustered data like ours. We used the *glmer()* function of the *lme4* package in the statistical software R (Team, 2011). The model is used to test the partial effect of information density (noted as probability of CC in our data), while controlling other 9 variables. Apart from all the fixed effects, the model also includes a random effect of speaker. In order to make our results more reliable, we evaluated the fitted model as in TABLE 2, which shows an acceptable fit of the model.

		Predicted		% correct
		0	1	
Observed	0	4043	334	92%
	1	875	566	39%
			Exactitude	79%

TABLE 2 – Model accuracy. 1 = *que* omission ; cut value = 0.50

5 Results and discussion

TABLE 3 lists the results of all the fixed effects that were tested in the present study. It shows that *que* omission is conditioned by cognitive, linguistic and social factors.

Predictor	Coef. β	**p value**
(Intercept)	-1.69	< 0.0001 ***
age	0.07	$= 0.41$
gender *M vs. F*	0.007	$= 0.97$
scolarity2-1	0.16	$= 0.47$
scolarity3-2	-0.87	< 0.01 **
profession2-1	0.03	$= 0.94$
profession3-2	0.74	< 0.05 *
profession4-3	-0.79	< 0.01 **
profession5-4	0.83	< 0.05 *
profession6-5	-0.22	$= 0.50$
matrix subject		
=other pro vs. je_tu	-0.05	$= 0.57$
=NPs vs. other pro	-0.24	$= 0.56$
CC subject		
=less vs. more accessible	-0.91	< 0.001 ***
right phono context		
=son. vs. obs.	-0.97	< 0.0001 ***
=vow. vs. son.	-0.62	< 0.0001 ***
left phono context		
=son. vs. obs.	-0.25	$= 0.09$.
=vow. vs. son.	0.13	$= 0.31$
verb frequency	0.39	< 0.0001 ***
probability of CC	0.32	$<$ **0.0001 *** **

TABLE 3 – Results summary

5.1 Information density and frequency

As predicted by UID, there was a clearly significant effect of information density : the more likely a verb is to appear with a CC, the more it favours *que* omission ($p < 0.0001$), as shown by FIGURE 1. We therefore conclude that the association between predictability and omission proposed by the UID also appears in French, which is expected given that it is thought to result from general cognitive principles that do not vary across languages.

Our results also show that matrix verb frequency is a highly significant predictor : the more frequent a verb is in the Montréal 84 corpus, the more likely *que* is to be omitted when it introduces

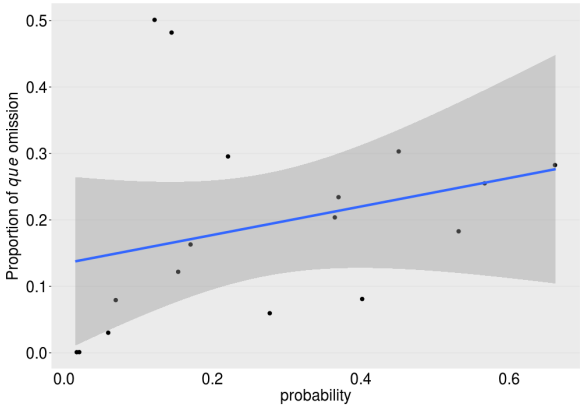

FIGURE 1 – Probability of CC *vs. que* omission

a CC. Since frequency often correlates with predictability, it is possible that the frequency effect is actually related to the information density effect. However, further work is needed to flesh out how exactly verb frequency and CC onset informativity are related.

5.2 Linguistic factors

As TABLE 3 shows, the segment following the site of *que* has a significant effect on whether or not it will be omitted. Confirming Sankoff, Warren and Dion's findings, the highest rate of *que* omission is when it would be followed by an obstruent, then a sonorant, and then a vowel.

This pattern can be understood if we view *que* omission as the optimal strategy in Quebec French for repairing disfavoured consonant clusters at the beginning of CC. *Que* (pronounced [k]) is a clitic that must be syllabified with the phonological material to its right. When the material following [k] starts with an obstruent, like [t] in the sequence *je pense que tu dors*, it creates a cluster [kt] which violates the Sonority Sequencing Principle (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 to simplify abnormal clusters (see (Côté, 2012)). Therefore, the best option here is to simply delete *que*, particularly if its information is low.

As for CC subject, our study shows that more accessible subjects, like 1st and 2nd pronouns and expletives, differ significantly from less accessible subjects such as other pronouns and NPs ($p < 0.001$). Following availability accounts, we suggest that when the CC subject is harder to be retrieved from memory, speakers prefer to utter the complementizer to keep their production fluent.

Furthermore, in order to test whether CC SUBJECT and RIGHT PHONOLOGICAL CONTEXT have independent effects on *que* omission, we have repeated the multilevel logit analysis on subsets of data after removing either phonological or syntactic effects. Results show that the same phonological or syntactic patterns of the right context found in TABLE 3 were maintained in each data subset. Hence, unlike previous work which reduces either phonotactic constraints to syntactic effects or the latter to the former, we conclude that both effects are independently significant on *que* omission.

5.3 Social factors

Both profession and education have significant effects on *que* omission. 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)" (Sankoff and Laberge, 1978, 238-9) *i.e.* liberal professionals, white collar workers and other university graduates (groups 1, 2 and 4) omit *que* less often than do the other members of the community, whose economic success does not depend so much on language (technicians (group 3), blue collar workers (5) and the unemployed (6)).

6 Conclusion

In this paper, we investigated which factors influence *que*-omission in complement clauses in a spoken corpus of Québec French. Mixed effects statistical modelling revealed that omission is conditioned by social factors (speaker profession), grammatical factors (phonotactic and syntactic constraints), as well as *information density*. Our study therefore provides support for the role of informativity in syntactic variation cross-linguistically, and, more generally, for the interaction of grammatical, social and cognitive factors in variable syntactic phenomena.

Acknowledgements

This work is partially supported by a public grant overseen by the French National Research Agency (ANR) as part of the program "Investissements d'Avenir" (reference : ANR-10-LABX-0083). It contributes to the IdEx Université de Paris - ANR-18-IDEX-0001. We would like to thank Yair Haendler for his help on statistical analysis.

References

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

George N Clements. 1990. The role of the sonority cycle in core syllabification. *Papers in laboratory phonology*, 1 :283–333.

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

Marie-Hélène Côté. 2012. Laurentian french (quebec) extra vowels, missing schwas. *Phonological variation in French : Illustrations from three continents*, 11 :235.

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

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

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

Johan Elsness. 1984. That or zero ? a look at the choice of object clause connective in a corpus of american english.

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

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

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard : Telephone speech corpus for research and development. In *[Proceedings] ICASSP-92 : 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520. IEEE.

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

T Florian Jaeger and Roger P Levy. 2007. Speakers optimize information density through syntactic reduction. In *Advances in neural information processing systems*, pages 849–856.

T Florian Jaeger. 2005. Optional that indicates production difficulty : Evidence from disfluencies. In *Disfluency in spontaneous speech*.

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

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

Laura Kastronic. 2016. *A Comparative Variationist Approach to Morphosyntactic Variation in Hexagonal and Quebec French*. Ph.D. thesis, Université d'Ottawa/University of Ottawa.

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

Shana Poplack, Allison Lealess, and Nathalie Dion. 2013. The evolving grammar of the french subjunctive. *International Journal of Latin and Romance Linguistics*, 25(1) :139–195.

Shana Poplack. 1989. The care and handling of a megacorpus : The ottawa-hull french project. *Language change and variation*, 4.

Philip Resnik. 1996. Selectional constraints : An information-theoretic model and its computational realization. *Cognition*, 61(1-2) :127–159.

Yves Roberge and Nicole Rosen. 1999. Preposition stranding and que-deletion in varieties of north american french. *Linguistica atlantica*, 21 :153–168.

Gillian Sankoff and Henrietta Cedergren. 1971. Some results of a sociolinguistic study of montreal french. *Linguistic diversity in Canadian society*, pages 61–87.

Gillian Sankoff and Henrietta Cedergren. 1972. Sociolinguistic research on french in montreal. *Language in Society*, 1(1) :173–174.

David Sankoff and Suzanne Laberge. 1978. The linguistic market and the statistical explanation of variability. *Linguistic variation : Models and methods*, pages 239–250.

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

R Core Team. 2011. R : A language and enviroment for statistical computing. r foundation for statistical computing. vienna, austria. *URLhttp ://www. r-project. org*.

Pierrette Thibault and Diane Vincent. 1990. Un corpus de français parlé : Montréal 84. *Université Laval, Québec*.

Sandra A Thompson, Anthony Mulac, et al. 1991. A quantitative perspective on the grammaticization of epistemic parentheticals in english. *Approaches to grammaticalization*, 2 :313–329.

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