French subject doubling: A third path

Yiming Liang, Caterina Donati, Heather Burnett

▶ **To cite this version:**

Yiming Liang, Caterina Donati, Heather Burnett. French subject doubling: A third path. Isogloss. Open Journal of Romance Linguistics, 2024, 10 (7), pp.1-28. 10.5565/rev/isogloss.420 . hal-04841794

HAL Id: hal-04841794
https://hal.science/hal-04841794v1

Submitted on 17 Dec 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Isogloss
Open Journal of Romance Linguistics

Received: 29-02-2024
Accepted: 17-09-2024
Published: 12-12-2024

French Subject Doubling: A Third Path

Yiming Liang
Universiteit Gent
yiming.liang@ugent.be

Caterina Donati
Université Paris Cité
caterina.donati@u-paris.fr

Heather Burnett
Centre national de la recherche scientifique, Université Paris Cité
heather.susan.burnett@gmail.com

How to cite: Liang, Yiming, Donati, Caterina & Heather Burnett. 2024. French Subject Doubling: A Third Path. RLLT23, eds Lisa Brunetti, Ioana Chitoran & Alexandru Mardale. Special Issue of *Isogloss. Open Journal of Romance Linguistics* 10(7)/12, 1-29.
DOI: https://doi.org/10.5565/rev/isogloss.420

Abstract

This paper revisits the status of subject clitics in Spoken French by studying subject doubling as a sociolinguistic variable. In the literature, two influential analyses have been proposed to account for French subject doubling. Based on new evidence from a corpus study on the large *Multicultural Paris French* (MPF) corpus, we argue for an analysis reconciling these two competing views of the construction in Spoken (colloquial) French. On one hand, we provide further support to the morphological approach (Auger, 1994; Culbertson, 2010) in which subject clitics are morphological agreement markers on the verb. On the other hand, we argue based on new evidence that lexical subjects are topicalized, as in the dislocation analysis. Furthermore, we argue that Spoken French is in a diglossia situation where speakers alternate structures provided by both Standard French and Colloquial French grammars. This paper provides further evidence of how quantitative studies of language use can shed light on long-standing theoretical debates.

Keywords: subject doubling, dislocation, null subject, Colloquial French, quantitative syntax.

1. Introduction

Subject doubling, where a subject noun phrase and a coreferential subject clitic co-occur, is a common phenomenon in spoken French.[1] An example is shown in (2), and contrasts with the non-doubling variant shown in (1). This phenomenon is also found in other languages such as Picard (Auger, 2003a) and most Northern Italian dialects (Brandi & Cordin, 1989).

(1) Marie mange. (canonical order)
'Mary eats.'

(2) Marie$_i$ elle$_i$ mange. (preverbal subject doubling)
'Mary$_i$ she$_i$ eats.'

In addition to using a co-referential subject clitic to refer to a preverbal subject noun phrase (referred to as *preverbal subject doubling*), another option is to place the nominal subject towards the end of the sentence, in which case the use of a subject clitic is obligatory. This type of subject doubling, referred to as *postverbal subject doubling*, is illustrated by (3). Since most previous studies have focused on preverbal subject doubling (e.g., De Cat, 2005; Nadasdi, 1995; Zahler, 2014, among others), this paper will also concentrate on preverbal subject doubling to facilitate comparisons with other research.

(3) Elle$_i$ mange, Marie$_i$. (postverbal subject doubling)[2]
'She$_i$ eats, Mary$_i$'

Most previous studies of subject doubling investigate whether a DP subject is doubled by a subject clitic when the DP subject is present. However, an alternative perspective can also be adopted, focusing on whether speakers opt to double a subject clitic with a co-referential DP subject when the clitic is present. This alternative analysis involves a choice between the unmarked variant (4) and the doubling variant (5). In the context of this study, we have chosen not to pursue this alternative perspective. Instead, we concentrate solely on cases involving a DP subject, specifically the variation between sentences (1) and (2), following most previous studies on subject doubling. This decision is also motivated by the following reason: to know whether subject clitics function as agreement markers in Colloquial French,

[1] It has been argued that a distinction should be made between subject doubling and left dislocation, where the subject clitic functions as an agreement marker in the former and as a syntactic argument in the latter (Nadasdi, 1995; De Cat, 2005). However, given the long-standing debate on the status of subject clitics in Colloquial French (see Section 2), there is little consensus on which sentences resembling (2) should be categorized as subject doubling, and *subject doubling* may be used as a broad term to include dislocation, as stated by Coveney (2005: 96): "Linguists have used a bewildering range of names for this and related structures, among them *reprise* and 'left dislocation,' but the one I will generally employ here is 'subject doubling'". Therefore, we will refer to all instances of co-occurrence of a subject NP and a coreferential subject clitic as *subject doubling*, and use "dislocation" and "morphological analysis" to refer to different analyses of the phenomenon.

[2] This kind of doubling is usually called as "clitic right dislocation" in the literature.

it is crucial to know how often they appear where standard French grammar prohibits their presence, i.e. when a DP subject is already present in the sentence.

(4) Elle est partie.
'She has left.'

(5) Marie$_i$ elle$_i$ est partie.
'Mary$_i$ she$_i$ has left.'

Numerous studies have focused on the variation between the non-doubling (1) and the doubling (2) constructions, and have identified some social and linguistic factors that condition speakers' preference between these two variants. These factors include the nature of the DP subject (Auger, 1998; Auger & Villeneuve, 2010; Nadasdi, 1995), clause type (Auger & Villeneuve, 2010), presence of intervening elements (Zahler, 2014), information status (Pabst et al., 2020), and verb type (Auger & Villeneuve, 2010; Zahler, 2014), among others. While subject doubling has received substantial attention in sociolinguistics, it has also intrigued researchers from formal syntactic theory. The structure of subject doubling, along with the syntactic status of subject clitics in Colloquial French, has been heavily debated in the literature. Two opposing proposals have gained the most attention: 1) the *dislocation* analysis (De Cat, 2005; Kayne, 1975; Rizzi, 1986), which considers subject doubling as an instance of left dislocation and subject clitics as argument-bearing pronouns receiving a θ role, and 2) the *morphological* analysis (Auger, 1995, 2003a; Culbertson, 2010), whereby the DP subject is the real subject of the sentence and the argument of the verb, while subject clitics are agreement markers base-generated and merged at T. Since both accounts have received empirical support, it remains uncertain which one provides a better analysis of the phenomenon. Besides, although some studies are based on corpus studies, only a subset of the relevant aspects of this complex and puzzling linguistic phenomenon have been investigated, and factors such as word frequency have not been considered or controlled in the statistical analysis. To have a better understanding of the syntactic structure of this common phenomenon in Spoken French, we conduct a new corpus study of multiple factors including frequency in a large-scale spoken corpus, and propose a new analysis based on new results we obtain from corpus study. The goal of this paper is thus two-fold: 1) to gain a more comprehensive understanding of this variationist phenomenon by incorporating frequency factors and more fine-grained grammatical factors in the corpus study; 2) to examine which formal analysis of subject doubling is more consistent with the patterns observed.

This paper is organized as follows: in Section 2, we present arguments that have been advanced by previous studies in favor of each of the two influential approaches. Section 3 describes how our corpus study was conducted, including data extraction, annotation and statistical analysis. The results of the corpus study are presented in Section 4. Section 5 focuses on a new structure proposed in light of results from our corpus study. We argue that this new analysis, which is a reconciliation of the two opposing previous accounts, better accounts for the results from our study, but also for evidence that has been put forward in the literature. Section 6 concludes the paper with a summary.

2. Literature: debates on two structures

Two analyses have been proposed to account for the structure of subject doubling:

(6) a. Dislocation: $[_{TOP}$ Marie $[_{TP}$ elle $[_{T}$ mange]]]
b. Morphological: $[_{TP}$ Marie $[_{T}$ elle-mange]]

Researchers like Kayne (1975), Rizzi (1986) and De Cat (2005), among others, analyze the construction as a case of *dislocation*, where the DP subject is dislocated into a topic position in the left periphery, while the subject clitic is a syntactically argument-bearing pronoun merged in Spec,TP and phonologically cliticized onto the inflected verb (cf. (6)). While this analysis enjoys a wide following in particular to analyze Standard French, De Cat (2005) argues that it is also valid for spoken French, based on evidence such as: 1) In elicitation studies, the subject clitic does not systematically co-occur with a nominal subject, hence behaving like an argument instead of an agreement marker which is generally obligatory; 2) Subject clitics are available for syntactic operations like movement. De Cat demonstrates that the inversion of subject clitics in interrogatives is productive in spontaneous speech production of speakers from Belgium, Canada and France, suggesting that subject clitics are not prefixes but syntactically independent entities; 3) Other clitics, like the negation particle *ne*, or object clitics, can intervene between the subject clitic and the verb, and those data are attested both in written and spoken French, as shown by (7). According to Zwicky & Pullum (1983), affixes can only be attached to a bare word or a word containing only affixes, while clitics can be attached to words already containing clitics. Therefore, if subject clitics were analyzed as agreement markers, other clitics that can intervene between them and the verb should also be analyzed as agreement markers. However, De Cat (2005) and Rowlett (1998) both show that *ne* is sensitive to syntactic locality constraints. For example, *ne* can appear in a different clause from a negative expression (*personne* 'nobody' in (8a)). However, this long-distance relation between *ne* and the negative expression is only possible when the embedded clause is non-finite (as shown in (8b)), and cannot hold across the boundary of a complex DP (as in (8c)), which is a strong island. Given that the distribution of *ne* is controlled by syntactic constraints, De Cat (2005) concludes that it cannot be analyzed as an affix; 4) It is difficult for the DP to receive a focus reading when the subject clitic is present. De Cat (2005) use an acceptability rating task to demonstrate that when a focus reading of the subject is forced, for example in (9), the presence of the subject clitic is rarely acceptable, showing that the lexical DP is in a Topic position when a subject clitic is present; 5) Subject doubling obeys the topicality hierarchy: it appears to be incompatible with indefinite and quantified noun phrases with an existential reading. This is expected under a dislocation analysis which views the DP as a topic (Rizzi, 1986). In cases where an indefinite DP subject co-occurs with a subject clitic, the indefinite DP must have a generic interpretation under which indefinite can be topics (Côté, 2001).

(7) a. Il n'est pas là.
'He is not there.'
b. Il me l'a donné.
'He gave it to me.'

(8) De Cat (2005: 9-10)
a. Paul n'accepte de renvoyer personne.
'Paul doesn't agree to dismiss anybody.'
b. *Paul n'accepte qu'on renvoie personne.
Intended: "Paul doesn't agree do dismiss anybody"
c. *Il ne reste [de potager [avec aucun arbre fruitier]].
Intended: "There is no allotment with fruit trees left."

(9) De Cat (2005: 16)
Context: La voiture bleue est foutue.
'The blue car's knackered.'
Follow-up:
1. ?Non, la voiture ROUGE elle est foutue.
'No, the RED car it's knackered.'
2. Non, la voiture ROUGE est foutue.
'No, the RED car's knackered.'

Other researchers, in particular Auger (1995, 2003a, 2003b), Roberge (1990), and more recently Culbertson (2010), among others, argue for a *morphological* analysis, according to which the DP subject occupies the canonical subject position and the subject clitic is an agreement marker base-generated in T (cf. **Error! Reference source not found.**). This analysis has been argued to be applied exclusively to spoken French. Evidence for this analysis includes the following arguments. 1) Colloquial French subject clitics are subject to phonological reduction phenomena, involving both vowel and consonant elisions. These idiosyncratic morphophonological properties are compatible with an affix status (Auger, 1993, 1994; Culbertson, 2010); 2) In many corpus studies of spoken French, subject doubling is nearly categorical: for example, 80.6% among Lyon child-directed speech (Culbertson, 2010); over 80% in Marseille French speech (Sankoff, 1982), 96% in adolescent speech from Villejuif (Campion, 1984) and 70% in Montreal French speech (Auger, 1991), all cited in Auger (1994: 116)). On the contrary, *ne* in negative contexts is rarely attested (e.g., 7.5% in the Lyon corpus, Culbertson, 2010), nor is the subject-verb inversion in interrogatives (e.g., 0.1% in yes-no questions and 1.4% in wh-questions in the Lyon corpus, Culbertson, 2010); 3) Regardless of whether it is followed by a subject clitic, no phonological or prosodic features single out the subject DP as being dislocated; 4) Culbertson (2010) demonstrates through acceptability rating experiments that the subject clitic is acceptable when the sentence is in broad-focus contexts, suggesting that the DP subject is not necessarily interpreted as a topic. In order to account for the restriction to definites, Culbertson (2010) proposes that French subject doubling is subject to Suñer (1988)'s 'matching hypothesis', whereby agreement markers and their specifier must match featurally. Since subject clitics have the feature [+definite, +accessible], their DP specifier must also bear these features.

Although some of the arguments we have just reviewed have come from linguists' intuitions or experiments, corpus studies have played an enormous role in the development of syntactic analyses for subject doubling (e.g., Culbertson, 2010; De Cat, 2005, among others). However, theoretical research often draws upon corpus examples or doubling rates to support their analyses, while not fully tapping into the rich insights provided by comprehensive variationist studies that simultaneously

assess multiple factors impacting the phenomenon through statistical modeling. For example, Culbertson (2010) shows that the doubling rates are high both for declaratives and interrogatives in the Lyon corpus (around 80%), but the grammatical categories she studied are quite broad, not distinguishing, for example, between different kinds of declarative clauses. Yet, the doubling rate has been shown to vary across different types of declaratives by quantitative studies, a fact that could lead to a fine-grained understanding of subject doubling (Auger & Villeneuve, 2010; Zahler, 2014). Furthermore, although many variationist sociolinguistic studies (Auger & Villeneuve, 2010; Coveney, 1996; Zahler, 2014, among others) investigate both social factors and grammatical aspects, they do not take into account processing factors, such as subject informativity and verb frequency. In order to get a fuller picture, we therefore decided to track the contours of this phenomenon in one of the most recent corpora of Spoken French: the *Multicultural Paris French* corpus (Gadet & Guerin 2016).

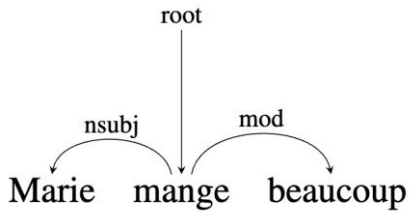
3. Corpus study

3.1. Data extraction

Consisting of 66 interviews with 790,000 transcribed words on the day of investigation, the *Multicultural Paris French* corpus (MPF) (Gadet, 2017; Gadet & Guerin, 2016) documents the oral language of young individuals aged 12 to 37. Speakers reside either in the Northern part of Paris or in its suburbs, and all have a multicultural family background, which means that at least one of their parents was born outside of France, or they have regular contact with other cultures. The register of the corpus is informal, as the interviews are in-person conversations between friends or acquaintances, covering various topics such as family, daily life, language change, among others.

As the corpus does not contain any linguistic annotations, we used *Stanza* (Qi et al., 2020) to pretokenize, POS-tag and lemmatize the corpus, and employed the HOPS parser (Grobol & Crabbé, 2021) to obtain syntactic dependencies between words. Syntactic dependency is a binary and antisymmetrical relation between two words in an utterance indicating the syntactic relationship between them. An example of a syntactic dependency tree of an utterance is shown in Figure 1.

Figure 1. Dependency parsing tree of Marie mange beaucoup ‘Mary eats a lot’.



Source: generated by the Authors.

Once the corpus preprocessing was completed, we extracted all utterances containing a preverbal nominal subject (e.g., DP subject like *mon père* ‘my father’ and

un garçon ‘a boy’, quantified subject like *certains* ‘certain people’, *tout le monde* ‘everybody’, proper names like *Marie* ‘Mary’, etc.) from the entire corpus and annotated whether the nominal subject is doubled by a subject clitic (e.g., *il(s), elle(s), ce, ça*), or not. Data extraction and the annotation of independent variables were conducted through a semi-automatic process. More specifically, a Python script was used to extract and annotate relevant tokens based on morphological and syntactic annotations obtained by automatic preprocessing of the corpus, and manual verification was performed for cases that were difficult for automatic processing. Only *preverbal third-person subjects* were considered. First- and second-person subjects were excluded because when first- and second-person strong pronouns are employed, doubling by a subject clitic is obligatory, thereby not subject to variation. Although third-person strong pronouns like *lui* and *eux* can occur without subject clitics, we did not include them in the current study. This exclusion is motivated by two reasons: 1) only a limited subset of third-person pronouns can be considered, as *elle* and *elles* are ambiguous between strong pronouns and subject clitics; 2) strong personal pronouns are notably inclined towards favoring subject doubling. For example, third-person strong pronouns are doubled 70% of the time, which is much higher than the overall subject doubling rate of 22% in the CFPP (Paris) corpus (Zahler, 2014). Likewise, in Saguenay French (Quebec), while the overall doubling rate is 45%, the doubling rate associated with third-person strong pronoun subjects is 78% (Auger & Villeneuve, 2010). In Ontario French, 74% of third-person pronoun subjects are doubled, also a much higher rate than 27% which is the overall doubling rate (Nadasdi, 1995). Since the doubling rate is a crucial indicator for determining the syntactic status of subject clitics and strong personal pronouns may possess inherent characteristics that contribute to a higher rate of doubling (for example due to their potential association with emphasis, making them structurally aligned with left dislocation structures that emphasize content), we have opted to only include DP subjects. In addition, we excluded DP subjects that contain coordination, because it is impossible to code one of the fixed factors - DP subject head frequency - in such cases. For the same reason, we only included DP subjects whose head noun consists of one word (for example *Anne Dupont* is excluded as the head noun contains two words). Furthermore, we excluded utterances in which the verb is not completely pronounced before interrupting the utterance. Additionally, tokens with missing values for any fixed factors, which is necessary for statistical modeling, were also excluded. Most of these exclusions are due to missing social information (n=711). The whole process yielded a dataset of 3,543 occurrences,[3] with a doubling rate of 74%.

3.2. Factor coding

To obtain the most complete corpus study of subject doubling to date, we coded all the extracted tokens for the following factors.

[3] An anonymous reviewer suggested excluding clear cases of left dislocation, such as *Marie, je me souviens quand elle s’est cassé la jambe* ‘Marie, I remember when she broke her leg’. Given the lack of consensus on what qualifies as subject doubling (see also Footnote 1), this study is focused on all instances of co-occurrence between DP subjects and subject clitics, so we did not exclude these cases from our statistical analysis. To address the reviewer's concern, we conducted a supplementary analysis excluding 32 clear cases of left dislocation, and all results remained unchanged.

3.2.1. Social factors

We included speaker AGE, GENDER, EDUCATION LEVEL and PROFESSION as fixed effects. More specifically, speaker age was coded as a numeric factor and speaker gender as a binary variable, with men contrasting with women. Speaker education was included as a 3-level ordinal variable: <BAC,[4] BAC, >BAC (i.e., university degree). Speakers' socio-professional groups were coded as a 7-level ordinal variable based on the French government's classification (Insee, 2003) in this order: *chômeurs* ('unemployed'), *élèves/étudiants* ('students'), *ouvriers* ('workers'), *employés* ('employees'), *professions intermédiaires* ('intermediate professions'), *cadres et professions intellectuelles supérieures* ('managers and higher intellectual occupations'), *artisans, commerçants et chef d'entreprise* ('craftsmen, merchants and entrepreneurs').

Age has been shown to condition subject doubling in Picardy spoken French, where the doubling rate tends to decrease with older speakers (Coveney, 1996). Auger & Villeneuve (2010) observe a weaker age effect in French spoken in the Saguenay (Quebec, Canada), where only the youngest group (below 25 years old) double subject more often than the oldest group of speakers (above 64 y.o.). However, Zahler (2014) reports a completely inverse trend in Parisian French, where the oldest age group (> 56 y.o.) double the subject most often, the middle-aged group disfavour subject doubling, and the youngest group (<30 y.o.) double the subject the least. Speakers' gender has also been found as a significant predictor of subject doubling. For instance, Zahler (2014) reports that female speakers favour subject doubling whereas male speakers disfavour it. Although Auger & Villeneuve (2010) report an absence of a significant overall gender effect, they observe a difference in the context of doubling with *il(s)/elle(s)* and with *ça/ce*: while no particular tendency concerning age or gender is revealed for doubling with *ça/ce*, young female speakers notably produce more subject doubling with *il(s)/elle(s)* compared to other groups. As for social class, Nadasdi (1995) observes that speakers from working-class double the subject the most, and middle-class speakers double the subject the least.

3.2.2. Linguistic factors

We included four linguistic factors as described below.

SENTENTIAL POLARITY: Coveney (1996) finds an association between a speaker's rate of *ne* retention and their rate of subject doubling. Zahler (2014) and Roberts (2014) also reports an effect of *ne* retention, demonstrating that negative utterances containing *ne* disfavour subject doubling, while both negative utterances without *ne* and affirmative contexts favour doubling in Parisian French and Martinique French. Following Zahler (2014), we thus distinguished three categories: affirmative, negation with *ne*, and negation without *ne*.

DP SUBJECT TYPE: Zahler (2014) observes that the subject is doubled most often with strong pronouns (70%); proper nouns have an intermediate doubling rate (32%); while common nouns and other pronouns have the lowest doubling rate (20%). In the current study, we included DP subject type as a three-way categorical factor:

[4] The BAC, short for *baccalauréat*, is a French national academic qualification that students can obtain upon completing their secondary education, typically at the end of high school.

definite (e.g., *Marie* 'Mary', *mon père* 'my father'), indefinite (e.g., *un garçon* 'a boy'), and quantified subjects (e.g., *tout le monde* 'everybody').

CLAUSE TYPE: Several studies report that main clauses favor subject doubling in comparison with subordinate clauses (Auger & Villeneuve, 2010; Zahler, 2014). Within subordinate clauses, relative clauses disfavor subject doubling compared with other types (Auger & Villeneuve, 2010). Therefore, we included clause type as a 3-level categorical variable: root clauses including in-situ interrogatives, other subordinate, and relative clauses.

VERB FREQUENCY: Although this factor has never been investigated in previous studies on subject doubling, verb frequency has been argued to condition the evolution of morphosyntactic patterns. As Bybee (2003) notes, highly frequent (and irregular) verbs tend to preserve conservative morphosyntactic characteristics. This trend has been attested across various sociolinguistic variables, such as the selection between subjunctive and indicative forms (Poplack, 2001; Poplack et al., 2013), the choice between passé simple and passé composé (Engel, 1990), and the preference for synthetic future over periphrastic future (Blondeau & Labeau, 2016; Tristram, 2020). If subject clitics are agreement markers, they are part of the verbal inflection and therefore their distribution should be conditioned by verb frequency. In the current study, verb frequency was measured in the MPF corpus[5] and log-transformed.

3.2.3. Processing factors

We included two factors that may condition the variation of subject doubling motivated by processing accounts.

DISTANCE between the DP subject head and the verb: previous studies have found that the presence of all types of intervening elements between DP subject and the verb favor the usage of subject doubling (Auger & Villeneuve, 2010; Roberts, 2014; Zahler, 2014). More specifically, Zahler (2014) finds that emphatic and parenthetical elements, *oui, non,* and parenthetical clauses are correlated with the highest doubling rate. Furthermore, hesitation, adverbials and a mixture type of intervention slightly favor subject doubling. The doubling rates associated with different types of intervening elements are slightly different in Auger & Villeneuve (2010), where emphatic pronouns and parenthetic expressions favor subject doubling the most, whereas parenthetical clauses favor it the least. Zahler (2014) attributes this effect of intervening elements to a processing factor associated to linear distance and possibly memory, but does not explain why subject doubling varies according to different types of intervening elements.

In this study, we do not distinguish different types of interveners, as it remains unclear why the type of intervening elements would influence subject doubling and how different types should be grouped for statistical analysis. For example, Zahler (2014) grouped hesitation and adverbials together, but no explanation or motivation was provided for this grouping. Instead, we pursue Zahler's idea that these phenomena of intervention are due to a processing factor akin to distance, and measure the length of interveners, assuming based on psycholinguistic evidence that speakers would prefer the linguistic variant which shortens dependencies (Gibson, 2000; Hawkins,

[5] Word frequency was measured using our corpus rather than a larger, more representative one because our corpus reflects the French spoken by young speakers from multicultural backgrounds in the suburbs. As a result, the lexicon, characterized by loanwords and *verlan*, is likely to differ from that of other corpora.

2001, 2004). When applied to our case, the prediction is that the farther the verb is from the nominal subject, the more likely the speaker is to use the co-referential subject clitic to shorten the subject-verb dependency. More specifically, we included the number of words intervening between the DP subject head and the verb (subject clitic excluded). If the subject head noun and the verb are adjacent, the distance was counted as 1, as in (10a). If *n* intervening words are present, the distance was counted as n+1. For example, the distance was coded as 2 for (10). The distance was thus coded as a numeric variable, varying from 1 to 34.[6]

(10) Juliet2, MPF
   a. Même moi au bled il mon père il a fait la maison quand il est reparti là.
      'Even me, at home, my father he built the house when he went back here.'
   b. Ben déjà euh le projet il s'intitule nous sommes un musée.
      'Well already uh the project it is titled *We are a Museum*.'

DP SUBJECT HEAD FREQUENCY: Multiple studies have demonstrated a trade-off between syntactic redundancy and information density in diverse variationist phenomena (Frank & Jaeger, 2008; Jaeger, 2011; Schäfer et al., 2021, among others), aligning with the *Uniform Information Density* (UID) hypothesis (Jaeger, 2010; Levy & Jaeger, 2007). The UID hypothesis states that speakers tend to choose the linguistic variant which allows for a more uniform distribution of information across the sentence. Applied to syntactically redundant phenomena, the hypothesis predicts that words or structures with low information density (i.e., redundant) tend to be omitted or simplified to enhance communication efficiency. Given that the insertion of the subject clitic does not bring new information to the sentence, subject doubling can be viewed as a case of syntactic redundancy, possibly conditioned by the information density of the DP subject. Information of a word, as defined by Shannon (1948), is its negative log-probability. This definition captures the idea that the more predictable a word is, the less information it conveys.

How to estimate the information density of a word remains an open question (see Meister et al., 2021, for a discussion) which exceeds the scope of this paper. Since we seek to control for processing factors that may influence subject doubling by including subject informativity, we used a simplistic measure that considers the frequency of the subject head as a proxy for the information density of the DP subject. The intuition behind this is that the more frequent a word is, the more predictable it is, and thus the less information it conveys. The subject head frequency was thus measured within the MPF corpus and log-transformed before being integrated into the statistical model.

[6] An anonymous reviewer suggested investigating structural distance as a factor in addition to the linear distance, as it has been shown to affect some variation phenomena like gender agreement in Spanish (Alemán Bañón et al., 2012). Although it could also play a role in subject doubling, calculating structural distance in corpus studies is more challenging due to the diverse nature of intervening elements which may consist of different types of linguistic segments. Furthermore, the current study is mainly focused on linguistic factors. Therefore, we suggest that the intriguing question of how distance affects subject doubling be left to future research specifically dedicated to processing factors.

3.3. Statistical analysis

The statistical modelling was conducted using the generalized logistic mixed model with R (R Core Team, 2022) under the *lme4* package (Bates et al., 2015).[7] The alternation between doubled DP subject (coded as 1) and non-doubled DP subject (coded as 0) was modelled as depending on the 10 fixed effects that are presented above and 3 random intercepts (speaker n=88, verb lemma n=374, DP subject head lemma n=988), as specified below:

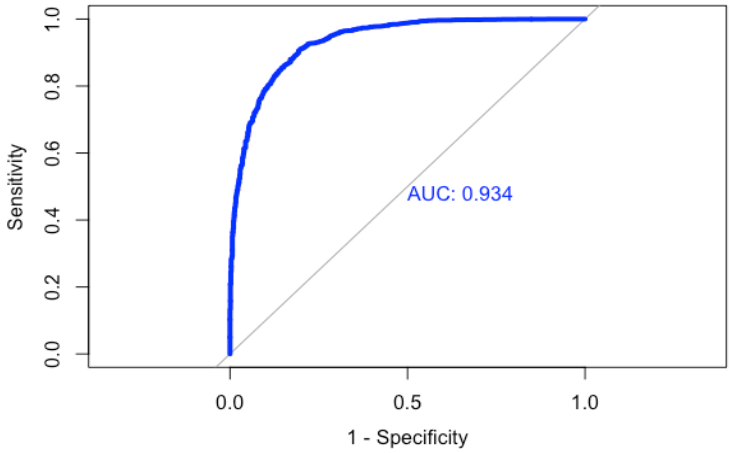
DP subject doubling is dependent on:
age + gender + education + profession
+ sentential polarity + DP subject type
+ clause type + verb frequency
+ distance + DP subject head frequency
+ (1|speaker) + (1| verb lemma) + (1| DP head lemma)

A backward difference coding method was applied to all categorical variables in the order defined above. All numeric predictors have been centralized and standardized. The GVIF measure (General Variance Inflation Factors, Fox & Monette, 1992) shows no major concern of collinearity in the model, as each variable has a $GVIF^{1/(2Df)}$ inferior to 2 (cf. Table 1). The model achieves an accuracy of 88.3%, higher than the baseline accuracy 74.3%. The ROC curve (receiver operating characteristic curve) shown in Figure 2 with an AUC (area under the ROC curve)[8] of 0.934 also shows an outstanding discriminatory ability of the model to distinguish subject doubling from non-doubling cases (Mandrekar, 2010).

[7] Data and Rscript for this study are available at this link: https://osf.io/dzyrq/?view_only=2da72e261b5548ef9b58623bd7629bfb

[8] An ROC curve shows the performance of a classification model at all classification thresholds. AUC provides an aggregate measure of performance across all possible classification thresholds, equivalent to the probability that a classifier will rank a random positive example higher than a randomly chosen negative instance. See Fawcett (2006) for example for a presentation of ROC and AUC.

Figure 2. ROC curve of the model of subject doubling.



Source: generated by pROC package in Rstudio (R Core Team, 2022).

4. Results: evidence for both analyses

Mixed effects logistic regression analysis of the data reveals significant effects that provide support both to the dislocation analysis and to the morphological approach (cf. Table 1). Furthermore, subject doubling is also conditioned by social and processing factors. We start by presenting linguistic effects.

Table 1. Mixed-effect analysis of the data predicting the doubling of lexical subjects in the MPF corpus (n=3,543). Italics highlight factors that contribute to doubling the lexical subject. A positive coefficient means that subject doubling is more frequent in the first category over the second.

Predictor	Coef.	SE	*z*	*p*	Sig.	GVIF$^{1/(2Df)}$
(Intercept)	-7.01804	11.58985	-0.606	0.544825		
Age (numeric)	0.24122	0.15352	1.571	0.116120		1.55
Gender: *m. vs. f.*	0.05401	0.25667	0.210	0.833330		1.12
Education:						1.22
= BAC *vs.* < BAC	-0.43583	0.42923	-1.015	0.309925		
= > BAC *vs.* BAC	-0.85466	0.41064	-2.081	0.037409	*	
Profession:						1.10
= student *vs.* unemployed	-0.64269	0.58335	-1.102	0.270578		
= worker *vs.* student	-0.34296	0.64765	-0.530	0.596429		
= employed *vs.* worker	-0.54764	0.76189	-0.719	0.472271		
= intermediate *vs.* employed	-0.25476	0.58989	-0.432	0.665838		
= manager *vs.* intermediate	0.74784	0.48581	1.539	0.123716		
= entrepreneur *vs.* manager	-0.47324	0.88960	-0.532	0.594746		
Sentential polarity:						1.00
= *ne vs.* aff.	-18.67336	34.74942	-0.537	0.591011		
= without *ne vs. ne*	19.07211	34.74918	0.549	0.583108		
Verb frequency	0.31677	0.08813	3.594	0.000325	***	1.00
DP head frequency	-0.22194	0.09562	-2.321	0.020290	*	1.01
Distance	0.34204	0.06369	5.370	7.87e-08	***	1.01
Subject:						1.01
= indefinite *vs.* universal	2.20039	0.65582	3.355	0.000793	***	

= definite *vs.* indefinite	1.80070	0.54676	3.293	0.000990	***	
Clause						1.00
= other subord. *vs.* relative	1.57580	0.35806	4.401	1.08e-05	***	
= main *vs.* other subord.	0.81584	0.13316	6.127	8.96e-10	***	

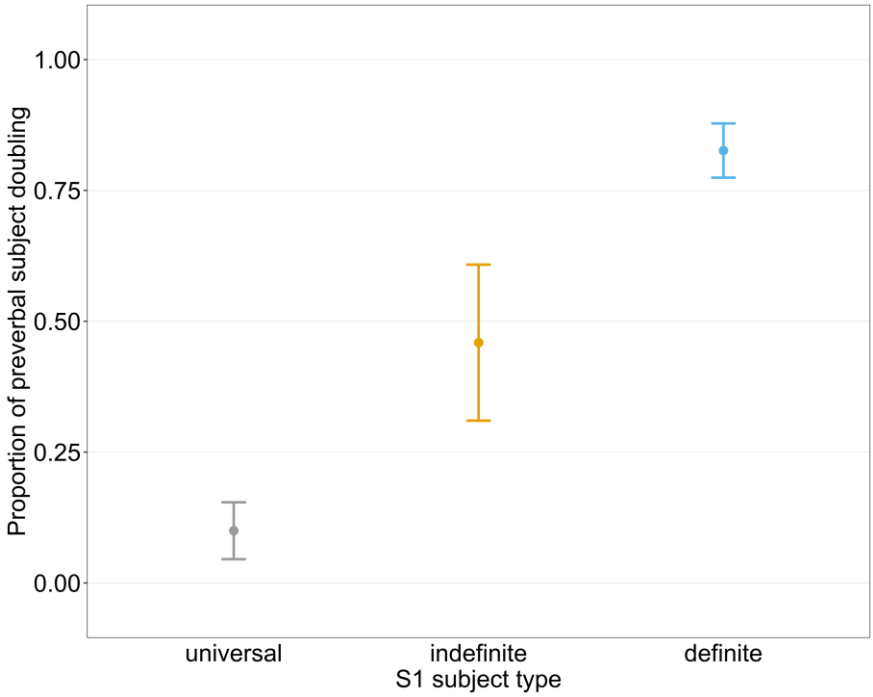
Source: generated by mixed-effect logistic regression modeling with Rstudio.

4.1. Supporting the dislocation analysis

SUBJECT TYPE We find that doubling is governed by subject type in terms of *topicality hierarchy* (cf. Figure 3): definite DPs (rate of doubling: 81%) > indefinite DPs (43%) > universal QPs (7%), differences between two adjacent categories being significant ($p < 0.001$). As an illustration, the subject is more often doubled in cases like (11), followed by (11) and the least often in examples as (11). This result is in line with the idea that the subject occupies a topic position in the left periphery. However, recall that a similar pattern was also found by Culbertson (2010) and Nadasdi (1995), who do not consider this constraint as evidence of dislocation, but rather due to feature matching. While the effect of subject type can be accounted for by both accounts, we argue that feature matching is not sufficient to explain another significant effect: clause type.

(11) a. Le garçon il est là. (definite subject)
‘The boy he is there.’
b. Quand une meuf elle a plein de frères... (indefinite subject, Zakia3, MPF)
‘When a woman she has plenty of brothers...’
c. Tout le monde ici ils sont de Maghnia. (quantified subject, Nacer2, MPF)
‘Everyone here they are from Maghnia.’

Figure 3. Effect of subject type on preverbal subject doubling.

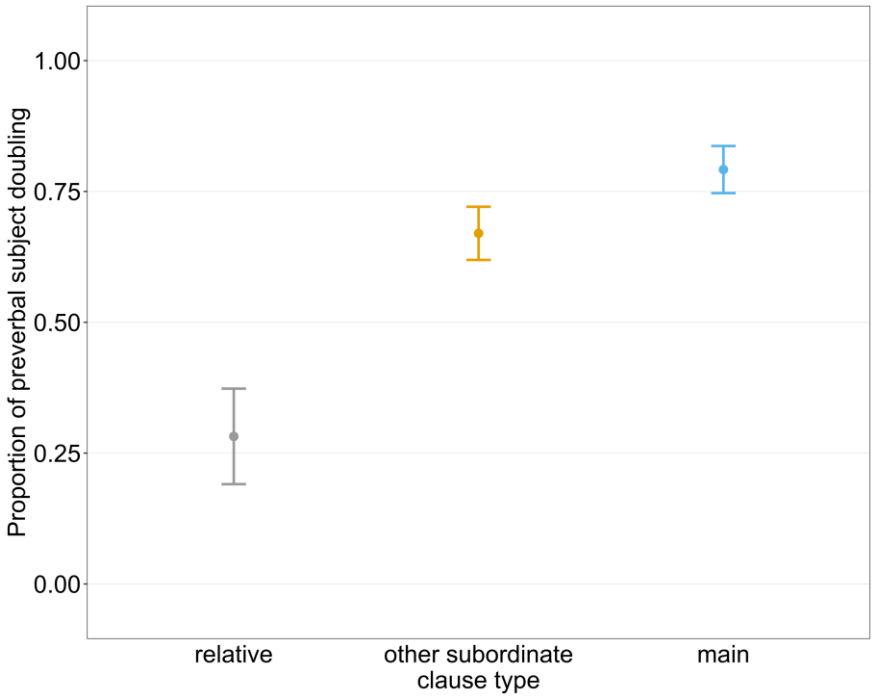


Source: generated by Rstudio with dplyr and ggplot2 packages.

CLAUSE TYPE We find that root clauses, including root interrogatives, as in (12), are associated with the highest rate of subject doubling; whereas subordinate clauses disfavour it. Among them, relative clauses (12) disfavour doubling the most, while other subordinate clauses (12) are in-between (cf. Figure 4): root (78%) > other subordinates (62%) > relatives (30%), differences between two adjacent categories being significant ($p < 0.001$). This pattern replicates anologous findings in Saguenay French (Auger & Villeneuve, 2010) and Parisian French (Zahler, 2014). Nonetheless, to the best of our knowledge, no explanation has yet been provided for this effect.

(12) a. Le garçon il a dit quoi? (root clause, Joanne11, MPF)
'The boy he said what?'
b. Quand une meuf elle a plein de frères...
(other subordinate clause, Zakia3, MPF)
'When a woman she has plenty of brothers...'
c. Un mec que Sylvie$_i$ elle$_i$ a rencontré dans le métro elle lui a dit ...
(relative, Aristide4, MPF)
'A man that Sylvie she met in the metro she said to him...'

Figure 4. Effect of clause type on preverbal subject doubling.



Source: generated by Rstudio with dplyr and ggplot2 packages.

We argue that the impact of clause structure on subject doubling suggests that the DP subject occupies a topic position in the left periphery in the doubling construction. It is indeed well-known that a number of left-peripheral phenomena, such as topicalization and focalization, are main clause phenomena (MCP) that are only available in root clauses and a subset of embedded clauses (more specifically asserted clauses) (Aelbrecht et al., 2012a, 2012b; Hooper & Thompson, 1973). The absence of MCP in embedded clauses has been attributed to a truncated left periphery of certain

types of embedded clauses, such as presupposed embedded clauses (Haegeman, 2006). Consequently, some topic positions are not available in these types of subordinate clauses.

More specifically, Frascarelli & Hinterhölzl (2007) (see also Bianchi & Frascarelli, 2010) demonstrate a systematic correlation between the syntactic distribution of topics, their intonational properties and their function in the discourse, and distinguished three types of topics in Italian and German: 1) *Aboutness Topics* (also called *Aboutness-shift Topics* or *A-Topics* for short, Bianchi & Frascarelli, 2010), used to newly propose or reintroduce a topic in the discourse; 2) *Contrastive Topics* (or *C-Topics* for short), used to oppose two topics; 3) *Familiarity Topics* (also known as *Familiar Topics*, *Given Topic* or *G-Topics*), used to resume background information or for topic continuity (Givón, 1983). Accordingly, a complex cartography for topic structures as shown in (13) in the left periphery is proposed to account for the different distributions of these different topics (Bianchi & Frascarelli, 2010; Frascarelli & Hinterhölzl, 2007; Haegeman, 2006): both Aboutness and Contrastive Topics are realized in a topic position higher than FocP and are restricted to root clauses and some subordinate clauses (Haegeman, 2006). On the other hand, the *Familiarity Topic* occupies the lowest topic projection which is available in all clauses. Furthermore, some topics, like Hanging Topics (Cinque, 1977), are restricted to root clauses. As for French, although the mapping between prosodic properties and pragmatic function of topics is less clear-cut (Brunetti et al., 2012; Brunetti & Avanzi, 2017), Brunetti and colleagues also distinguish Aboutness Topics and Familiarity Topics, but consider Contrastive Topics as an independent category that can combine with those two types of topics. Going back to our data, we postulate that a topic structure similar to (13) also exists in French, and that some topic positions are not available in certain types of subordinate clauses.

(13) Bianchi & Frascarelli (2010: 59), but also in Frascarelli & Hinterhölzl (2007), Haegeman (2006)[9]
[$_{ShiftP}$ A-Topic [$_{ContrP}$ C-Topic [$_{FocP}$ [$_{FamP*}$ G-Topic [$_{FinP}$ [$_{IP}$

If DP subjects were not topics in doubling constructions, there should be no correlation between clause type and subject doubling. However, an analysis in which the subject DP is analyzed as some kind of topic does predict such a relation: root clauses do not have any restrictions on topic positions, which leads to no limitations on subject doubling. In contrast, subordinate clauses do show restrictions on topics, as non-asserted subordinate clauses, like central adverbial clauses and complement clauses introduced by emotive verbs, only allow low topics like *Familiarity* topic. Therefore, embedded clauses are thus predicted to display a slightly lower rate of subject doubling compared with root clauses.

The fact that relative clauses further disfavour subject doubling is also compatible with the hypothesis that the DP subject occupies a topic position. As illustrated by (14), since a topic is an A' constituent, it is expected to intervene in the A'-dependency relating the antecedent *la giraffe* and the gap within the relative clause. To avoid this intervention effect, positioning the DP subject in the A position Spec, TP is preferable, as intervention is known to be sensitive to the A/A' distinction.

[9] The asterisk on the functional category FamP indicates recursion.

Consequently, the option of subject doubling is expected to be disfavored in relative clauses compared to other subordinate clauses where subject doubling does not trigger an intervention effect, explaining why the rate of doubling is remarkably low in relative clauses.

(14) La giraffe$_i$ [$_{RC}$ que [$_{TOP}$ l'éléphant$_j$ [$_{TP}$ il$_j$ a arrosée ___$_i$]]]
'The giraffe that the elephant$_j$ it$_j$ watered...'

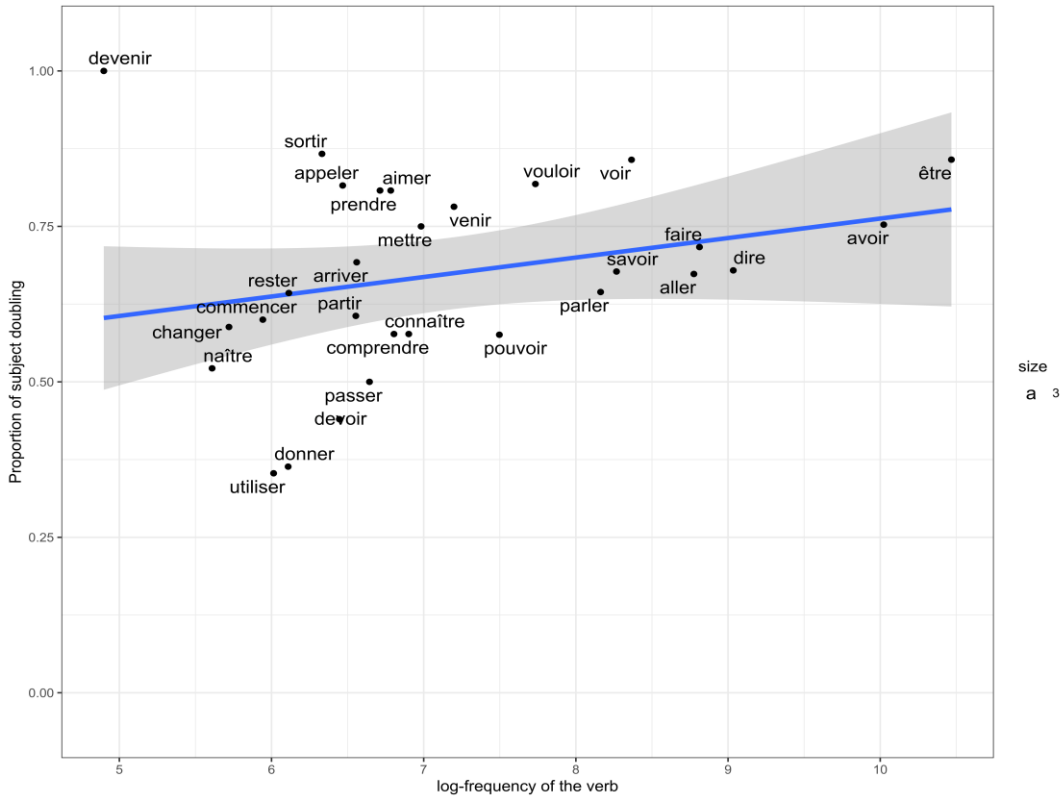
4.2. Supporting the morphological approach

Meanwhile, our results also provide support for the morphological approach. Firstly, we find a very high rate of subject doubling (74%). In fact, in a secondary analysis, we included postverbal subject doubling cases (e.g., *Il est là mon père* 'He is there our father'), and the rate of doubling increased to 77%. Furthermore, within the remaining 23% of occurrences without subject clitics, many involve *ne* retention or are about education and religion, topics that are typically associated with the formal Standard French register. Overall, this high doubling rate suggests that the presence of subject clitics is nearly categorical in the informal register of Spoken French, an argument in favor of a morphological analysis of subject clitics as inflection markers. Since non-doubling structures tend to be associated with the formal register, speakers may use the option without a subject clitic, as this aligns with Standard French conventions.

Moreover, we found an extremely low rate of doubling with negative utterances with *ne* (0%, or 2% if tokens with missing social information are retained). Although the polarity effect was not revealed to be significant by the mixed effect model, this doubling rate is notably lower than rates for affirmatives and negative contexts without *ne*, which are 75.8% and 75.1% respectively. This low rate of doubling with *ne* aligns with Culbertson (2010)'s findings. Following Culbertson, we interpreted this low rate of doubling with *ne* as supplementary indirect support to the morphological analysis. Given that *ne* is not an agreement marker, it would disrupt the morphological combination of the subject clitic and the verb.

We also find that VERB FREQUENCY is positively correlated with subject doubling ($p < 0.001$): more frequent verbs are associated with a higher rate of subject doubling, as shown by Figure 5. We argue that this effect provides more indirect evidence in favor of the morphological approach given that highly frequent verbs tend to preserve old inflectional paradigms (Bybee, 2003). Old French used to display a rich verbal subject agreement system, where verbal inflections unambiguously reflected subject's person and number. As phonological erosion began to take place during the medieval period, verbal endings gradually became synthetic (Bettens, 2023; Simonenko et al., 2019). Although subject doubling is a relatively recent phenomenon, we interpret Bybee's proposal more broadly: we propose that highly frequent verbs are more likely to preserve the older features of French's rich verbal agreement, where the subject clitic has evolved into an agreement marker on the verb, enriching verbal inflection and marking the subject's person and gender.

Figure 5. Effect of verb frequency on preverbal subject doubling.



Source: generated by Rstudio with dplyr and ggplot2 packages.

4.3. Social and processing effects

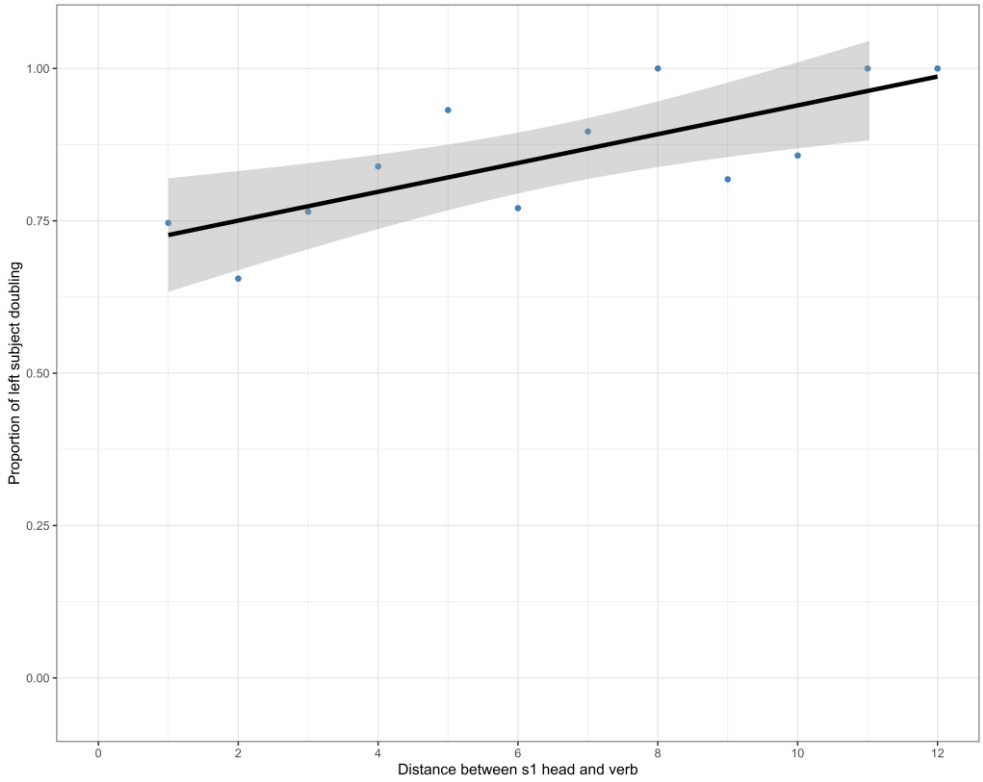
Among all the tested social factors, only EDUCATION affects subject doubling: we find that speakers with the highest education level, i.e., having entered the university, doubled DP subjects less often than speakers who only obtained BAC or have a lower education level ($p < 0.05$). Although education level has rarely been investigated in previous studies, this result is in line with the overall observation that subject doubling is related to an informal register. It is thus not surprising that speakers from a higher prestige class, for instance with a higher education level or from a high social class (Nadasdi, 1995) tend to not produce subject doubling. Furthermore, as presented above, negative utterances with *ne* clearly disfavor subject doubling. This effect has been interpreted as a linguistic argument in favor of a morphological analysis of subject doubling; however, it can also be seen as a social effect, thereby further demonstrating that subject doubling is associated with an informal register.

Unlike prior research which revealed an effect of age, our study does not find any influence of age on subject doubling. This could be due to the nature of the studied corpus, where most utterances were produced by speakers under 30 years old, thereby constraining the age range and potentially weakening age-related effects. Likewise, as young speakers are predominantly students, their occupations are not varied and therefore it is not surprising that profession is not a significant predictor. It is worth noting that the doubling rate in the MPF corpus is 74%, which is higher than 22% reported by Zahler (2014) in the CFPP corpus, which includes older speakers. While this difference in doubling rates could potentially imply an age effect, it could also be due to foundational differences between the two corpora, particularly concerning register and bilingualism. Consequently, these findings lead us to conclude that while

subject doubling is prevalent in an informal register, the precise social implications of this phenomenon remain unclear due to the inherent limitations of the studied corpus.

On the other hand, it turns out that subject doubling is conditioned by processing factors: in line with our prediction, DISTANCE plays an important role in subject doubling: as illustrated by Figure 6, the more distant the verb is from the head of the DP subject, the more likely are speakers to produce the subject clitic, so as to shorten the dependency between subject and verb. While this effect is consistent with prior research which found that the presence of intervening elements favor subject doubling, our study provides a more precise picture of the effect of distance. Furthermore, subject doubling is also affected by DP SUBJECT HEAD FREQUENCY, as a more frequent DP subject tends to be doubled less often ($p < 0.05$). This aligns with the prediction of the *Uniform Information Density* hypothesis, which predicts that a more predictable subject (in general more frequent) disfavors the insertion of the redundant subject clitic to enhance communication efficiency.

Figure 6. The effect of subject-verb distance on subject doubling.



Source: generated by Rstudio with dplyr and ggplot2 packages.

5. New proposal reconciling two analyses

To account for both the evidence that Colloquial French subject clitics are agreement markers, and the evidence that subject DPs are topics in doubling construction, we propose an analysis that reconciles the dislocation and the morphological analyses. We propose that subject doubling in Colloquial French (i.e., informal register) involves an agreement marker generated in T which is doubled by a DP located in topic position, as shown by the structure in (15) below.

(15) Our analysis: [$_{TopP}$ Marie [$_{TP}$ *pro* $_{T}$ elle-mange]]

This analysis correctly predicts that the subject DP will obey the topicality hierarchy. Moreover, it explains that the clause type affects subject doubling due to the truncated left periphery of certain subordinate clauses (Haegeman, 2006). Since some (high) topic positions are only available in root and a subset of embedded clauses, subordinate clauses have more restrictions on subject doubling than root clauses, leading to a slightly but statistically significant lower rate of subject doubling compared with root clauses. The analysis of the subject DP as topic is also in line with the observed fact that doubling is strongly disfavored with relative clauses, since the topicalized subject DP would act as an intervener for the relative A-bar dependency.

A corollary of our analysis is that the informal register of spoken French is a null-subject language, as already proposed by Roberge (1990), Culbertson (2010) and others, and claimed for other Romance languages like Picard (Auger, 2003a, 2003b) or Northern Italian dialects (Poletto, 2000) which display doubling. The analysis that lexical DPs occupy the Topic position in Colloquial French parallels findings in other null-subject languages, such as Spanish (Olarrea, 1998), where subjects are similarly analyzed as occupying topic positions rather than subject positions.

Our analysis of the subject clitic as an agreement marker and Colloquial French as a null-subject language is in line with Taraldsen's Generalisation (Rizzi, 1986; Taraldsen, 1978), which posits a relation between the richness of verbal subject agreement and the availability of null-subjects for synchronic variation and diachronic development: as the verbal inflection integrating subject clitic prefix is sufficient to identify the subject's person, Colloquial French has the possibility of not expressing the subject, thereby evolving into a null-subject language. Moreover, our analysis of the subject clitic as an agreement marker coupled with a topic-like preverbal subject indeed echoes several influential hypotheses, such as Alexiadou & Anagnostopoulou (1998) and Manzini & Savoia (2005), among others, which argue that the verbal morphology in null subject languages is itself pronominal and satisfies the Extended Projection Principle (EPP).

Furthermore, we propose that Spoken French exhibits a diglossia, incorporating two distinct grammars, one corresponding to Standard French and the second to Colloquial French, and French speakers might be switching between two grammars when talking. This proposal is in line with other work which argues that French is in a diglossia situation (Massot, 2010; Massot & Rowlett, 2013; Zribi-Hertz, 2011, 2013), and that variation reflects grammar competition (Kroch, 1989, 1994). More specifically, when the DP subject is not doubled by a clitic, we suggest that the structure involved is [$_{TP}$ Marie [$_{T}$ mange]], as prescribed by the grammar of Standard French. This hypothesis is supported by the observation that many occurrences of non-doubling constructions involve *ne*-retention and conversational topics like education, religion and history. We propose that this structure is always available, and will be chosen by speakers especially when the doubling structure is disfavored, for example when the DP subject is quantified or when the clause type is relative. This availability of two grammars in Spoken French has the advantage of reconciling evidence found in prior research against both the morphological and the dislocation accounts. For example, as mentioned in Section 2, subject clitics sometimes behave like syntactic entities in Spoken French, as 1) they are not present all the time, 2) *ne* can intervene,

and 3) subject clitics can appear in a postverbal position in interrogatives with subject-verb inversion (De Cat, 2005). We argue, following Culbertson (2010), that in those cases, subject clitics are indeed syntactic arguments as the Standard French grammar is active. Additionally, given the overwhelming evidence supporting both the dislocation and the morphological analyses found in the present study as well as in the existing literature debate (see Section 2), we suggest that only an analysis that reconciles both approaches as (15) for subject doubling can account for this contradictory evidence. For example, as mentioned in Section 2, De Cat (2005) highlights the unavailability of a narrow focus reading for a DP subject within a doubling structure, showing that the DP is in Topic position when the co-referential subject clitic is present. This challenges the morphological analysis, which posits that the DP subject is not topicalized. However, it is compatible with our analysis, which assumes the DP subject in Topic in cases of subject doubling.

On the other hand, Culbertson (2010) shows that phonological and prosodic features, like pauses, duration, and resyllabification, following the DP are comparable in subject doubling and non-doubling structures, but more pronounced after a dislocated DP object, such as *David$_i$ il$_j$ l$_i$'a déjà invité* 'David$_i$, he$_j$ already invited him$_i$'. This can be interpreted as a strong argument against the dislocated DP subject analysis. Nonetheless, although Culbertson (2010) carefully created minimal pairs of the test sentence in her experiment, she did not seem to control for the type of topics when comparing doubled subject DPs with doubled object DPs. Numerous studies in Italian, German, and French have indicated that different topics exhibit varying degrees of phonological and prosodic marking (Bianchi & Frascarelli, 2010; Brunetti et al., 2012; Brunetti & Avanzi, 2017; Frascarelli & Hinterhölzl, 2007, among others). For example, Brunetti & Avanzi (2017) analyze nearly 250 clitic left-dislocated DPs, including both subject and object DPs in Spoken French. Their findings indicate that within new topics, contrastive topics exhibit more prominent prosodic boundaries than non-contrastive ones. Since an object DP is not a prototypical topic, its dislocation to the topic position is often required by specific contexts, particularly those involving contrastive topics (see Brunetti & Avanzi 2017). Riou & Hemforth (2015) also report that doubled objects exhibit a significantly higher frequency of contrastive topics than doubled subjects. The example stimuli that are available in Culbertson (2010)'s paper indeed show that in both subject doubling and non-subject-doubling conditions, the lexical DP subject is a continuing (corresponding to *Familiar Topics* in Bianchi & Frascarelli (2010)'s terms) non-contrastive topic, while in object doubling condition, the lexical DP object is a new contrastive topic. Therefore, Culbertson's observation that doubled object DPs are more phonologically prominent than doubled subject DPs could potentially be due to the fact that new contrastive topics often exhibit more marked prosodic boundaries compared to non-contrastive topics.

Furthermore, Culbertson seems to only consider DP subjects serving as Familiar Topics, as suggested by the available example stimuli. Nevertheless, Brunetti et al. (2012) investigate the interaction between the pragmatic roles of DP subjects and prosodic boundaries in subject doubling and non-doubling constructions. They find that "topique actif" (active topic, corresponding to *Familiar* (or continuing) *Topics*) and new topics have comparable prosodic boundaries in both subject doubling and non-doubling constructions. However, resumptive (i.e., reintroduced) topics have a significantly more prominent prosodic boundary in subject doubling constructions compared to subject non-doubling constructions. Therefore, it is not surprising that

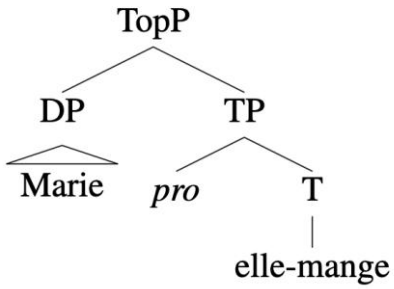
Culbertson finds that doubled and non-doubled DP subjects exhibit similar prosodic boundaries in her experiment, as it appears that the experiment mainly involves Familiar Topics. However, this does not mean that non-doubled and doubled subject DPs do not differ at all in all cases. Hence, we argue that Culbertson's prosodic findings are not necessarily a counter-argument to the dislocation analysis of DP subjects. Further experiments, conducted with a more meticulous control for subject topic types, are needed to explore this aspect more deeply.

6. Conclusion

In this paper, we presented a quantitative investigation into DP subject doubling using the MPF corpus. The linguistic effects we observed provide support for both dislocation and morphological accounts, two prominent analyses of the phenomenon that have been subjects of intense debate in the literature. We found that subject doubling is conditioned by subject type in terms of topicality hierarchy and by clause type, which, we argue, points to an analysis wherein DP subjects occupy a topic position. On the other hand, we found a very high rate of subject doubling and an extremely low rate of doubling with *ne*, pointing to the affixal status of subject clitics. Furthermore, verb frequency is positively correlated with subject doubling, which could suggest that subject clitics are an integral part of verbal inflection, since high frequency can impact verbal morphosyntax (Bybee, 2006).[10]

Based on these results, we propose a new account of Colloquial French subject doubling that is a hybrid of the dislocation and morphological analyses. In this new analysis, the doubled DP subject occupies a Topic position and the subject clitic is an agreement marker merged in T, schematized as follows (cf. Figure 7):

Figure 7. Our analysis of subject doubling in Colloquial French.



Source: generated by the Authors.

This new account aligns with observations that support both the dislocation and the morphological analyses, as discussed in the present study and the literature. Our analysis further suggests that Colloquial French is a null-subject language, a proposition that has been put forth by scholars such as Roberge (1990), Culbertson (2010) and others, and proposed for other Romance languages like Picard (Auger, 2003a, 2003b) or Northern Italian dialects (Poletto, 2000). In conjunction with the

[10] As pointed out by one anonymous reviewer and the editor, this is not a straightforward interpretation of Bybee (2006), as subject doubling is a new phenomenon. Please refer to Section 4.2 for a discussion.

morphological analysis of subject clitics, this new perspective resonates with Taraldsen's Generalisation (Rizzi, 1986; Taraldsen, 1978), which posits a correlation between the richness of verbal subject agreement and the availability of null-subjects, both in terms of synchronic variation and diachronic development.

Furthermore, we argue that Spoken French involves an alternation of two grammars. One corresponds to Standard French, where subject DPs are real subjects occupying the Spec,TP position and thus in competition with subject clitics which also function as syntactic arguments. This configuration typically leads to a non-doubling structure. In contrast, the second grammar corresponds to Colloquial French, in which subject DPs are topics in the left periphery, while subject clitics function as agreement markers on the verb. This grammar allows the option of subject doubling observed in Spoken French. This proposition aligns with the observation that subject doubling frequently co-occurs with other sociolinguistic variables, such as in-situ interrogatives and *ne* omission.

More generally, our study underscores the valuable insights that quantitative studies of linguistic variation offer to the formal analysis of syntactic phenomena. Our research makes a step forward in understanding the intricate relationship between theoretical syntax and quantitative variation. By leveraging intra-speaker variation data, we have refined and expanded syntactic models, contributing to the growing body of research at the intersection of theoretical syntax and quantitative variation analysis explored already by Adger & Smith (2010), Adger (2014), Thoms et al., (2019), and others.

Acknowledgments

The authors would like to thank two anonymous reviewers and the editor for their valuable feedback and comments. The work received funding from the ERC SMIC project (under the European Union's Horizon 2020 research and innovation programme, grant agreement N°850539) and Labex EFL (ANR-10-LABX-0083).

References

Adger, David. 2014. Variability and grammatical architecture. In C. Picallo (ed.), *Linguistic variation in the minimalist framework*, 179–196. Oxford: Oxford University Press.

Adger, David, & Jennifer Smith. 2010. Variation in agreement: A lexical feature-based approach. *Lingua* 120: 1109–1134.

Aelbrecht, Lobke, Haegeman, Liliane, & Rachel Nye. 2012a. Main Clause phenomena and the privilege of the root. In Aelbrecht, Lobke, Haegeman, Liliane, & Rachel Nye (eds), *Main Clause phenomena: New horizons*, 1–19. Amsterdam: John Benjamins.

Aelbrecht, Lobke, Haegeman, Liliane, & Rachel Nye. 2012b. *Main Clause Phenomena: New Horizons.* Amsterdam: John Benjamins.

Alemán Bañón, José, Fiorentino, Robert, & Alison Gabriele. 2012. The processing of number and gender agreement in Spanish: An event-related potential investigation of the effects of structural distance. *Brain Research* 1456: 49–63. https://doi.org/10.1016/j.brainres.2012.03.057

Alexiadou, Artemis, & Elena Anagnostopoulou. 1998. Parametrizing AGR: Word order, V-movement and EPP-checking. *Natural Language & Linguistic Theory* 16(3): 491–539.

Auger, Julie. 1991. Variation and syntactic theory: Agreement-marking vs. Dislocation in Québec Colloquial French. *NWAVE XX Meeting, Washington, DC*.

Auger, Julie. 1993. More evidence for verbal agreement-marking in colloquial French. In W. J. Ashby, M. Mithun, & G. Perissinotto (eds), *Linguistic Perspectives on Romance Languages: Selected Papers from the XXI Linguistic Symposium on Romance Languages, Santa Barbara, February 21, 1991*, 177–198. Amsterdam: John Benjamins. https://doi.org/10.1075/cilt.103.20aug

Auger, Julie. 1994. *Pronominal Clitics in Québec Colloquial French: A Morphological Analysis*. Ph.D. thesis, University of Pennsylvania.

Auger, Julie. 1995. Les clitiques pronominaux en français parlé informel: une approche morphologique. *Revue québécoise de linguistique* 24(1): 21–60.

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. https://doi.org/10.1017/S0008413100020429

Auger, Julie. 2003a. Le redoublement des sujets en picard. *Journal of French Language Studies* 13(3): 381–404. https://doi.org/10.1017/S0959269503001200

Auger, Julie. 2003b. Les pronoms clitiques sujets en picard: Une analyse au confluent de la phonologie, de la morphologie et de la syntaxe. *Journal of French Language Studies* 13(1): 1–22. https://doi.org/10.1017/S0959269503001066

Auger, Julie, & Anne-José Villeneuve. 2010. La double expression des sujets en français saguenéen: Étude variationniste. In W. Remysen, & D. Vincent (eds), *Hétérogénéité Et Homogénéité Dans Les Pratiques Langagieres: Mélanges Offerts à Denise Deshaies*, 67–86. Quebec: Presses de l'Université Laval.

Bates, Douglas, Mächler, Martin, Bolker, Ben, & Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using Lme4. *Journal of Statistical Software* 67(1): 1–48. https://doi.org/10.18637/jss.v067.i01

Bettens, Olivier. 2023. *Chantez-vous français?* Retrieved on 2023-09-03 from https://virga.org/cvf/.

Bianchi, Valentina, & Mara Frascarelli. 2010. Is Topic a Root Phenomenon? *IBERIA: An International Journal of Theoretical Linguistics*. 2(1): 43–88.

Blondeau, Hélène, & Emmanuelle Labeau. 2016. La référence temporelle au futur dans les bulletins météo en France et au Québec: Regard variationniste sur l'oral préparé. *Canadian Journal of Linguistics/Revue Canadienne de Linguistique* 61(3): 240–258. https://doi.org/10.1017/cnj.2016.26

Brandi, Luciana, & Patrizia Cordin. 1989. Two Italian Dialects and the Null Subject Parameter. In O. A. Jaeggli, & K. J. Safir (eds), *The Null Subject Parameter*, 111–142. Dordrecht: Springer. https://doi.org/10.1007/978-94-009-2540-3_4

Brunetti, Lisa, & Mathieu Avanzi. 2017. *Discourse properties of French clitic left dislocated NPs and their effect on prosody*. Ms. Université Paris Diderot/CNRS.

Brunetti, Lisa, Avanzi, Mathieu, & Cédric Gendrot. 2012. Entre syntaxe, prosodie et discours: les topiques sujets en français parlé. In F. Neveu, V. Muni Toke, P. Blumenthal, T. Klingler, P. Ligas, S. Prévost, & S. Teston-Bonnard (eds), *SHS Web of Conferences, 3e Congrès Mondial de Linguistique Française, Lyon, France, July 4-7, 2012*, Vol 1: 2041–2054. https://doi.org/10.1051/shsconf/20120100209

Bybee, Joan. 2003. Mechanisms of Change in Grammaticization: The Role of Frequency. In B. D. Joseph, & R. D. Janda (eds), *The Handbook of Historical Linguistics*. 602–623. Malden, MA / London : John Wiley & Sons https://doi.org/10.1002/9780470756393.ch19

Bybee, Joan. 2006. From Usage to Grammar: The Mind's Response to Repetition. *Language* 82(4): 711–733. https://www.jstor.org/stable/4490266

Campion, Elizabeth. 1984. *Left dislocation in Montréal French*. Ph.D. thesis, University of Pennsylvania.

Cinque, Guglielmo. 1977. The Movement Nature of Left Dislocation. *Linguistic Inquiry* 8(2): 397-412.

Côté, Marie-Hélène. 2001. On the status of subject clitics in Child French. *Research on Child Language Acquisition: Proceedings of the 8th Conference of the International Association for the Study of Child Language*, 1314–1330. Somerville, MA: Cascadilla Press.

Coveney, Aidan. 1996. *Variability in Spoken French: A Sociolinguistic Study of Interrogation and Negation*. Exeter, UK: Elm Bank.

Coveney, Aidan. 2005. Subject Doubling in Spoken French: A Sociolinguistic Approach. *The French Review* 79(1): 96–111.

Culbertson, Jennifer. 2010. Convergent evidence for categorial change in French: From subject clitic to agreement marker. *Language* 86(1): 85–132. https://doi.org/10.1353/lan.0.0183

De Cat, Cécile. 2005. French subject clitics are not agreement markers. *Lingua* 115(9): 1195–1219.

Engel, Dulcie. 1990. *Tense and Text: A Study of French Past Tenses*. London: Routledge.

Fawcett, Tom. 2006. Introduction to ROC analysis. *Pattern Recognition Letters* 27: 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Fox, John, & Georges Monette. 1992. Generalized Collinearity Diagnostics. *Journal of the American Statistical Association* 87(417): 178–183. https://doi.org/10.2307/2290467

Frank, Austin, & Florian Jaeger. 2008. Speaking Rationally: Uniform Information Density as an Optimal Strategy for Language Production. *The 30th Annual Meeting of the Cognitive Science Society (CogSci08)*, 939–944.

Frascarelli, Mara, & Roland Hinterhölzl. 2007. Types of topics in German and Italian. In K. Schwabe, & S. Winkler (eds), *On Information Structure, Meaning and Form*, 87–116. Amsterdam: John Benjamins. https://doi.org/10.1075/la.100.07fra

Gadet, Françoise (ed). 2017. *Les parlers jeunes dans l'île-de-France multiculturelle*. Paris: Ophrys.

Gadet, Françoise, & Emmanuelle Guerin. 2016. Construire un corpus pour des façons de parler non standard: Multicultural Paris French. *Corpus*. https://doi.org/10.4000/corpus.3049

Gibson, Edward. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In A. Marantz, Y. Miyashita, & W. O'Neil (eds), *Image, language, brain: Papers from the first mind articulation project symposium*, 94–126. Cambridge, MA: The MIT Press.

Givón, Thomas. 1983. *Topic Continuity in Discourse: A quantitative cross-language study*. Amsterdam / Philadelphia: John Benjamins. https://doi.org/10.1075/tsl.3

Grobol, Loïc, & Benoît Crabbé. 2021. Analyse en dépendances du français avec des plongements contextualisés (French dependency parsing with contextualized embeddings). *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Vol. 1: conférence principale*, 106–114.

Haegeman, Liliane. 2006. Argument fronting in English, Romance CLLD, and the left periphery. In R. Zanuttini, H. Campos, E. Herburger, & P. Portner (eds),

Crosslinguistic Research in Syntax and Semantics: Negation, Tense, and Clausal Architecture, 27–52. Washington, DC: Georgetown University Press

Hawkins, John. 2001. Why are categories adjacent? *Journal of Linguistics* 37(1): 1–34. https://doi.org/10.1017/S002222670100860X

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

Hooper, Joan B., & Sandra A Thompson. 1973. On the Applicability of Root Transformations. *Linguistic Inquiry* 4(4): 465–497.

Insee. 2003. *Professions et Catégories Socioprofessionnelles*. https://www.insee.fr/fr/metadonnees/pcs2003/categorieSocioprofessionnelleAgregee/1?champRecherche=true

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

Jaeger, Florian. 2011. Corpus-based research on language production: Information density and reducible subject relatives. *Language from a Cognitive Perspective: Grammar, Usage and Processing. Studies in Honor of Tom Wasow*, 161–198. Stanford, CA: CSLI Publication Stanford.

Kayne, Richard S. 1975. *French syntax: The transformational cycle*. Cambridge, MA: The MIT Press.

Kroch, Anthony. 1989. Reflexes of grammar in patterns of language change. *Language Variation and Change* 1(3): 199–244. https://doi.org/10.1017/S0954394500000168

Kroch, Anthony. 1994. Morphosyntactic variation. In K. Beals (ed.), *Papers from the 30th Regional Meeting of the Chicago Linguistics Society: Parasession on Variation and Linguistic Theory.* 180–201. Chicago, IL: Chicago Linguistic Society.

Levy, Roger, & Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. In B. Schölkopf, J. Platt, & T. Hofmann (eds), *Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference*, 849–856. Cambridge, MA: The MIT Press.

Mandrekar, Jayawant N. 2010. Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology* 5(9): 1315–1316.

Manzini, M. Rita, & Leonardo Savoia. 2005. *I dialetti italiani e romanci. Morfosintassi generativa.* Alessandria: Edizioni dell'Orso.

Massot, Benjamin. 2010. Le patron diglossique de variation grammaticale en français. *Langue française* 168(4): 87–106. https://doi.org/10.3917/lf.168.0087

Massot, Benjamin, & Paul Rowlett. 2013. Le débat sur la diglossie en France: Aspects scientifiques et politiques. *Journal of French Language Studies* 23(1): 1–16. https://doi.org/10.1017/S0959269512000336

Meister, Clara, Pimentel, Tiago, Haller, Patrick, Jäger, Lena, Cotterell, Ryan, & Roger Levy. 2021. Revisiting the Uniform Information Density Hypothesis. *arXiv:2109.11635 [Cs]*. https://arxiv.org/abs/2109.11635

Nadasdi, Terry. 1995. Subject NP doubling, matching, and minority French. *Language Variation and Change* 7(1): 1–14. https://doi.org/10.1017/S0954394500000879

Olarrea, Antxon. 1998. On the Position of Subjects in Spanish. *ASjU* 32(1) : 47–108.

Pabst, Katharina, Konnelly, Lex, Wilson, Fiona, & Naomi Nagy. 2020. Variation in subject doubling in Homeland and Heritage Faetar. *Toronto Working Papers in Linguistics* 42: 1-15.

Poletto, Cecilia. 2000. *The higher functional field: Evidence from northern Italian dialects*. Oxford University Press.

Poplack, Shana. 2001. Variability, frequency, and productivity in the irrealis domain of French. In J. Bybee, & P. Hopper (eds), *Frequency and the Emergence of Linguistic Structure*, Vol. 45, 405–428. Amsterdam: John Benjamins.

Poplack, Shana, Lealess, Allison, & Nathalie Dion. 2013. The evolving grammar of the French subjunctive. *Probus* 25(1): 139–195. https://doi.org/10.1515/probus-2013-0005

Qi, Peng, Zhang, Yuhao, Zhang, Yuhui, Bolton, Jason, & Christopher Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. *arXiv:2003.07082 [Cs]*. https://arxiv.org/abs/2003.07082

R Core Team. 2022. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Riou, Étienne, & Barbara Hemforth. 2015. Dislocation clitique de l'objet à gauche en français écrit. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics* 16. https://doi.org/10.4000/discours.9037

Rizzi, Luigi. 1986. Null Objects in Italian and the Theory of pro. *Linguistic Inquiry* 17(3): 501–557. https://www.jstor.org/stable/4178501

Roberge, Yves. 1990. *Syntactic Recoverability of Null Arguments*. McGill-Queen's University Press. https://www.jstor.org/stable/j.ctt7znfn

Roberts, Nicholas. 2014. *A sociolinguistic study of grammatical variation in Martinique French*. Ph.D. thesis, Newcastle University.

Rowlett, Paul. 1998. *Sentential Negation in French*. Oxford University Press.

Sankoff, Gillian. 1982. Usage linguistique et grammaticalisation: Les clitiques sujets en français. *La Sociolinguistique Dans Les Pays de Langue Romane*, 81–85. Tübingen: Gunter Narr Verlag.

Schäfer, Lisa, Lemke, Robin, Drenhaus, Heiner, & Ingo Reich. 2021. The Role of UID for the Usage of Verb Phrase Ellipsis: Psycholinguistic Evidence From Length and Context Effects. *Frontiers in Psychology* 12 : article 661087. https://doi.org/10.3389/fpsyg.2021.661087

Shannon, Claude E. 1948. A Mathematical Theory of Communication. *The Bell System Technical Journal* 27(3): 379–423.

Simonenko, Alexandra, Crabbé, Benoit, & Sophie Prévost. 2019. Agreement syncretization and the loss of null subjects: Quantificational models for Medieval French. *Language Variation and Change* 31(3): 275–301. https://doi.org/10.1017/S0954394519000188

Taraldsen, Tarald. 1978. On the NIC, vacuous application and the that-trace filter. Ms. Indiana University Linguistics Club, Bloomington.

Thoms, Gary, Adger, David, Heycock, Caroline, & Jennifer Smith. 2019. Syntactic variation and auxiliary contraction: The surprising case of Scots. *Language* 95(3) : 421–455.

Tristram, Anna. 2020. Variation and change in future temporal reference with *avoir* and *être*. *Journal of French Language Studies* 31(1): 1–25. https://doi.org/10.1017/S0959269520000174

Zahler, Sara. 2014. Variable subject doubling in spoken Parisian French. *University of Pennsylvania Working Papers in Linguistics* 20(1): 360–371.

Zribi-Hertz, Anne. 2011. Pour un modèle diglossique de description du français: Quelques implications théoriques, didactiques et méthodologiques. *Journal of French Language Studies* 21(2): 231–256. https://doi.org/10.1017/S0959269510000323

Zribi-Hertz, Anne. 2013. De la notion de grammaire standard dans une optique diglossique du français. *Journal of French Language Studies* 23(1): 59–85. https://doi.org/10.1017/S0959269512000361

Zwicky, Arnold, & Geoffrey Pullum. 1983. Cliticization vs. Inflection: English *n't*. *Language* 59(3): 502–513.