Uniform information density explains subject doubling in French

Yiming Liang (yiming.liang@ugent.be)
Department of Linguistics, Universiteit Gent
Gent, Belgium

Pascal Amsili (pascal.amsili@ens.fr)
ILPGA, Université Sorbonne Nouvelle
Paris, France

Heather Burnett (heather.susan.burnett@gmail.com)
Centre national de la recherche scientifique (CNRS)
Paris, France

Vera Demberg (vera@coli.uni-saarland.de)
Saarland University
Saarbrücken, Germany

Abstract

In this paper we investigate whether subject doubling in French is affected by the Uniform Information Density (UID) principle, which states that speakers prefer language encoding that minimizes fluctuations in information density. We show that, other factors being controlled, speakers are more likely to double the NP subject when it has a high surprisal, thus providing further empirical evidence to the UID principle which predicts a surprisal-redundancy trade-off as a property of natural languages. We argue for the importance of employing GPT-2 to investigate complex linguistic phenomena such as subject doubling, as it enables the estimation of subject surprisal by considering a rather large conversational context, a task made possible by powerful language models that incorporate linguistic knowledge through pre-training on extensive datasets.

Keywords: Uniform Information Density; subject doubling; spoken French; syntactic redundancy; surprisal

Introduction

Subject doubling, where a nominal subject and a coreferential subject clitic co-occur, as shown by (1a), is a prominent characteristic of Spoken French. Several studies have demonstrated a high doubling rate with certain groups of people, e.g., more than 80% doubling in Marseille French speech (Sankoff, 1982); 96% in adolescent speech from Villejuif (Campion, 1984), reported in (Auger, 1994, p.116)[1]. Numerous factors have been proposed to explain the variation between the two constructions (doubling (1a) *vs.* non-doubling (1b)), such as the NP subject type (Nadasdi, 1995; Auger, 1998; Auger & Villeneuve, 2010), clause type (Auger & Villeneuve, 2010), presence of intervening elements (Zahler, 2014), information status (Pabst et al., 2020), among others.

(1) a. Marie$_i$ elle$_i$ veut une pomme.
b. Marie veut une pomme.
'Marie$_i$ (she$_i$) wants an apple'

This phenomenon is also characterized by the fact that the insertion of the subject clitic is completely optional and does not add any new information to the sentence (i.e., redundant). Thus, speakers have a choice between two semantically comparable constructions. This type of alternation, which can also be observed with the omission of functional words, like complementizer 'that' omission in English (2a) *vs.* (2b) and article deletion in German (3a) *vs.* (3b), provides a privileged domain for investigating the impact of a particular cognitive hypothesis on language encoding: the hypothesis stating that speakers tend to choose the linguistic variant that minimizes fluctuations in information distribution across the utterance (the *Uniform Information Density* (UID) hypothesis, Jaeger, 2010).

(2) a. I think my boss is crazy.
b. I think that my boss is crazy.
(Jaeger, 2010)

(3) a. Niederlage für die ganze Gesellschaft
b. Eine Niederlage für die ganze Gesellschaft
'(A) Defeat for the whole society'
(Lemke et al., 2017)

The present work contributes to this line of research and aims to investigate how this principle may contribute to explaining speakers' language encoding choices with the help of large language models. Using a recent oral corpus of Spoken French, we demonstrate that, when other factors are controlled for, the surprisal of the NP subject has a significant impact on subject doubling, thereby providing further evidence for the UID hypothesis from a French syntactic redundancy phenomenon[2].

This paper is structured as follows. We present the UID hypothesis and research supporting it in the next section. The "Data and Predictions" section presents our corpus data and the predictions made by the UID principle. We further demonstrate in the "Measuring surprisal of NP subject" section how we use a generative transformer model to compute surprisal estimates of NP subjects. The following section presents our statistical modeling procedure and all the control factors. The "Results and discussion" section discusses the results of our study and outlines potential future directions. And finally, the "Conclusion" section provides some concluding remarks on the paper.

[1] Although subject doubling is frequent in informal French, it has been stigmatized in formal speech since at least the 17th century (Auger, 1994) and continues to be considered inappropriate in formal registers to this day.

[2] Code, statistical analysis scripts and data are available at https://osf.io/429zu/?view_only=b41cb68492224a3d80474fb4bb6982bb.

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

The Uniform Information Density Hypothesis

The *Uniform Information Density* (UID) hypothesis posits that, when grammar permits, speakers prefer utterances that distribute information as evenly as possible across the message (Levy & Jaeger, 2007; Jaeger, 2010). This hypothesis is based on a dual intuition. Firstly, communication can be seen as the transmission of information through a noisy channel with a limited capacity. To optimize the chances of successful communication, it is preferable to make sure that the information conveyed by each linguistic unit does not exceed the channel's capacity. Secondly, effective communication assumes that the communication channel is not underused, therefore a minimum level of informativeness should be guaranteed. Hence, successful and efficient communication should have an information profile that avoids peaks and troughs in information density, while keeping the transmission rate close to channel capacity.

Information density refers to the amount of information conveyed per linguistic unit (e.g., phoneme, word, constituent, etc.). Conceptualized within the framework of information theory (Shannon, 1948), the information of a word, w_i, is defined as the negative logarithm of the conditional probability of the word given the previous context (Equation 1). Also known as ***surprisal***, this measurement captures the idea that words with higher surprisal, thus less predictable, convey a greater amount of information. Psycholinguistic studies have shown that high surprisal results in an increase in reading time and comprehension processing effort (Hale, 2001; Demberg & Keller, 2008; Wilcox et al., 2020; Futrell et al., 2020; Frank et al., 2015).

$$I(w_i) = -\log P(w_i|w_0 \dots w_{i-1}) \quad (1)$$

The UID hypothesis has received compelling support from various levels of linguistic variation. To name a few examples: at the phonetic level: sounds or words with higher surprisal are pronounced more slowly across languages (Demberg et al., 2012; Pimentel et al., 2021); at syntactic level: functional words tend to be omitted when the structure they introduce is more predictable (e.g., omission of "that" in complement clauses (Jaeger, 2010) and relative clauses (Jaeger, 2011) in English, omission of articles in German newspaper headlines (Lemke et al., 2017)); at the discourse level: logical connectors are more likely to be omitted when the causal relation between two sentences is less surprising (Torabi Asr & Demberg, 2015), among others.

These works convincingly demonstrate the importance of informational density on linguistic variation, but the operationalization of the hypothesis remains delicate. Two aspects are especially discussed in the literature: on one hand, it is important to consider how the UID principle is interpreted, in particular, whether speakers try to transmit information at a roughly constant rate (i.e., minimizing global variability) or avoid rapidly shifting between information dense and sparse components of an utterance (i.e., minimizing local variability) (see Meister et al., 2021, for an in-depth discussion). In the present study, we interpret the UID hypothesis as minimizing the deviation from the global mean, which is in line with studies as Jaeger (2010). On the other hand, it involves determining the methods through which the informativeness level of a linguistic unit (typically a token) will be estimated. Corpus-based studies often rely on bigram models, in particular, verb subcategorization frequencies (Jaeger, 2010, 2011; Lemke et al., 2017). For instance, in Jaeger (2010)'s work on the omission of *that*, as in "I think (that) my boss is crazy," complementizer omission is viewed as a function of the predictability of the complement clause given the matrix verb lemma. However, in real conversations, people can naturally infer the upcoming structure from a larger conversational context. Therefore, even for the same verb, the predictability of the complement clause may vary depending on the context: if, in the previous context, someone asks "What do you think of this movie" and the speaker's answer begins with "I think," we will highly expect a complement clause to appear. Nevertheless, if someone asks "What are you thinking about?", the probability of "I am thinking" followed by a complement clause in the answer becomes low, because the answer would probably be "I am thinking about ...". Other corpus-based studies use linguistically motivated cues. For instance, Torabi Asr & Demberg (2015) use the presence of negation in the first sentence as an approximate indicator of the predictability of a causal relation with the second sentence. Although these measurements take into consideration a larger conversational context, they rely on very specific cues that are not easily generalizable to other phenomena, such as subject doubling, which constitutes the focal point of the current investigation.

In fact, subject doubling requires predicting the subject of a new sentence from the preceding context. Given that the subject typically occupies the initial position of a sentence and introduces the topic of the whole sentence, it is necessary to read a long section of the preceding context in order to understand the story and adequately predict the subject. Moreover, human language processing often relies on world knowledge and general linguistic competence to anticipate upcoming words. To replicate human predictive behaviors, computational models must possess a sophisticated architecture and be trained on extensive datasets. This was not feasible until the advent of large language models in recent years. In fact, it has been reported that Transformer-based models tend to capture human reading behaviour, including reading time and neural activity, better than n-gram models and Recurrent Neural Networks (RNNs) (Merkx & Frank, 2021; Wilcox et al., 2020).

On the other hand, when undertaking prediction tasks, models must adhere to a strictly left-to-right, incremental reading process to simulate human processing of spoken language. Consequently, **Generative Pre-trained Transformer (GPT)** models emerge as the most suitable candidates, compared with other transformer architectures like Bidirectional Encoder Representations from Transformers (BERTs), for

providing surprisal estimates of the subject in our study, given their training on next-word prediction. Indeed, GPT models have already been employed to examine the uniformity of information density in dialogues (Giulianelli et al., 2021), demonstrating their ability to provide probability estimates in conversational settings. However, they are rarely used to investigate the role of surprisal in syntactic variation. Therefore, we use a GPT model to estimate the surprisal (i.e., information) of subjects, in order to explore the influence of information density on subject doubling in French.

Data and predictions

The UID principle is assumed to hold for phenomena of syntactic redundancy, of which subject doubling is a special case, as the presence of the co-referential subject clitic is redundant while the NP subject already conveys all information about the subject. Therefore, we hypothesize that French speakers would use the subject clitic as a way of smoothing information density. The reasoning is the following: since the subject clitic is highly predictable, its insertion will lead to a local decrease in the information density of the utterance. Therefore, if the NP subject has a high surprisal, speakers should have a greater tendency to add the subject clitic, to lower the information density at the subject site. On the contrary, when the NP subject is associated with a low surprisal, the insertion of the redundant subject clitic would be dispreferred, as it would result in a trough in information density.

To test these predictions, we used one of a recent corpus of spoken French, the *Multicultural Parisian French* corpus (MPF) (Gadet & Guerin, 2016; Gadet, 2017). This corpus was chosen because of its substantial size and its comprehensive documentation of spontaneous, informal speech, which enables us to build a large-scale dataset of subject doubling. Consisting of 66 interviews with 790,000 transcribed words to date, the corpus aims to document the oral language of young individuals aged 12 to 37, from multicultural family backgrounds, residing in the suburbs of Paris. These interviews and their transcriptions are freely accessible on the corpus website (Multicultural Parisian French [corpus], 2019)[3]. The interviews are in-person conversations between friends or acquaintances, covering various topics such as family, daily life, language evolution, among others. As the corpus does not contain any linguistic annotations, we used Stanza (Qi et al., 2020) to pretokenize and POS-tag the corpus, and parsed it with the HOPS parser (Grobol & Crabbé, 2021).

Once the corpus preprocessing was completed, we extracted all utterances containing a preverbal nominal subject (e.g., mon père 'my father', un garçon 'a boy', certains 'certain people', tout le monde 'everybody', Marie 'Mary', etc.) from the entire corpus and annotated whether the nominal subject was doubled by a clitic (e.g., *il(s)*, *elle(s)*, *ce*, *ça*, as shown by example (1a)) or not (1b). Only preverbal third-person subjects were considered. Strong pronouns like *lui* and *eux* were excluded. The extraction was performed using a Python script based on morpho-syntactic annotations of the corpus, supplemented by manual verification. Cases where the NP subject contains a disfluency, marked by a disfluency marker (e.g., *le cou-*), a filled pause (e.g., *euh*), or repetition (e.g., *le le cours* 'the the course'), were also removed from the study, as it is not clear whether the UID hypothesis holds in these cases, and other cognitive principles could play a more important role in them. This process resulted in a dataset of 4,057 occurrences, with a doubling rate of 74.8%.

Measuring surprisal of NP subject

In order to compute the surprisal of the noun phrase subject, we used GPTfr-124M (Simoulin & Crabbé, 2021), a pre-trained generative Transformer language model proposed for French, adapted from the OpenAI GPT-2 model (Radford, Narasimhan, et al., 2019; Radford, Wu, et al., 2019). A GPT model was chosen for the present study because 1) it is more robust in computing reliable probability estimates by considering an extensive context and incorporating general knowledge compared to other types of models such as n-grams and LSTMs, and 2) it adheres to an incremental processing of sentences, unlike BERTs. However, it has been observed that GPT models with more parameters are less effective in producing word surprisals that correlate with reading times, compared to smaller versions of the model (Oh et al., 2022; Oh & Schuler, 2023). Therefore, we chose the smallest model available for French, which contains 124 million parameters.

As GPTfr-124M was pre-trained mainly on written texts, it is less competent at capturing characteristics of spoken data such as hesitations, repetitions, reformulations, and familiar lexical registers. Therefore, we fine-tuned the model on another corpus of spoken Parisian French, the Corpus de Français Parlé Parisien des années 2000 (CFPP2000) (Branca-Rosoff et al., 2012). Collected from 2005-2006, it consists of 51 interviews conducted in various neighborhoods of Paris and its close suburbs, totaling 750,000 tokens. The selection of this particular corpus for the fine-tuning process was based on its comparable size to our target corpus, MPF, as well as the shared characteristic of both corpora involving Spoken Parisian French[4]. During fine-tuning, we used the same tokenizer as GPTfr-124M (Simoulin & Crabbé, 2021) and constructed training examples in the following way: each interview was divided into speech turns, and a training example was constructed by concatenating successive speech turns until reaching a maximum of 1,024 tokens. Once this limit was reached, a new example was constructed from the next speech turn, ensuring that no examples spanned across interview boundaries. The last example of each interview was padded. The model was fine-tuned on 90% of the CFPP2000 corpus for 30 epochs using hyperparameters by default. We

[3] https://www.ortolang.fr/market/corpora/mpf/v3

[4] The CFPP corpus is accessible at http://cfpp2000.univ-paris3.fr/. Although the CFPP corpus also documents spontaneous speech, the used language is more formal than the vernacular French (Branca-Rosoff et al., 2012), with a rather low rate of subject doubling (22%) (Zahler, 2014). For this reason, we did not choose it as our target corpus.

computed the perplexity of the original model and the fine-tuned model on the remaining 10% of the CFPP2000 corpus and 50% of the MPF corpus. As shown in Table 1, fine-tuning yields a significant reduction in the model's perplexity, showing that the fine-tuned model is better adapted to oral data compared to the original one.

	CFPP2000	**MPF**
GPTfr-124M original	42.25	42.61
GPTfr-124M fine-tuned	**28.93**	**40.97**
p value of t-test	$<$ 2.2e-16	0.043

Table 1: Perplexity of the GPTfr-124M models on evaluation corpora.

Then, we use the fine-tuned GPT model to estimate the surprisal (Equation 1) of the NP subject of all occurrences extracted from the whole MPF corpus. Here "NP subject" means the sequence starting from the determiner of the NP subject until the subject head noun. 93% of the utterances involve NP subjects composed of one or two words, which respectively correspond to a proper name or an NP constituent composed of a determiner and a head noun. Another option would be to only consider the surprisal of the head noun. However, this approach would introduce an important disparity between language model predictions and human behavior: the presence of a determiner in French facilitates the language model's prediction of the subsequent noun; humans, on the other hand, do not have sufficient time to incorporate it into the previous context when predicting the upcoming subject head noun, as the determiner is very short. To avoid this problem, we compute the surprisal of the NP constituent by summing the logarithmic probabilities of its constituent words and subwords, following the chain rule. This method aligns with the practices of several studies investigating the cognitive plausibility of neural language models, which use addition rather than averaging when computing the surprisal of words split into subwords (Wilcox et al., 2020; Kuribayashi et al., 2021; Oh & Schuler, 2022).

NP subjects tend to appear at the sentence outset and their predictability is determined by the previous context of the conversation. It is however an open question, how much previous context is needed exactly to approximate best the context length that contributes to estimating the predictability from a human perspective. We therefore experiment with different context lengths, ranging from 1 to 11 preceding speech turns[5], and found that they all yield comparable results (cf. the next section). Hence, we decide to report the surprisal estimates from the model with just a single speech turn as context in the analyses below. A speech turn contains an average of 10 tokens. Table 2 shows an example of estimating the NP surprisal considering one preceding speech turn. The GPT model will compute the joint probability of the NP subject (bolded) by considering one preceding speech turn and the sequence preceding the subject in the target turn. The surprisal of the NP subject, $-\log P(\text{NP}|\text{context})$, will be considered as a fixed effect in the statistical model. We also experiment with different context lengths and will discuss these results in the next section.

[5]We have set the maximal context to 11 preceding speech turns, as extending beyond this limit would result in some utterances with a context exceeding the 1,024-token constraint of the GPT model.

Statistical modeling

Mixed-effects logistic regression modeling was performed to investigate the impact of surprisal on subject doubling, using the R Studio software (Team, 2022), implemented through the `glmer()` function from the lme4 package (Bates et al., 2015). In addition to the surprisal of the NP subject, the statistical model also included the following fixed effects, so as to control other factors that have been found to condition subject doubling in the literature. These factors were annotated in the following way using a second Python script, supplemented by a manual check:

- Sentential polarity (Zahler, 2014): affirmative, negative with negative marker "ne", negative without "ne".
- Type of the NP subject (Nadasdi, 1995; Auger & Villeneuve, 2010): universally quantified subject (e.g., tout le monde 'everybody', rien 'nothing'), indefinite NP (e.g., un garçon 'a boy'), definite NP (e.g., le garçon 'the boy', Daniel).
- Clause type (Zahler, 2014; Auger & Villeneuve, 2010): main clause, other subordinate clause, relative clause
- Verb frequency (log-transformed) measured in the same corpus (Liang et al., 2023)
- Distance in words (log-transformed) between the subject head and the verb (after eliminating the subject clitic): Zahler (2014) found that the presence of intervening elements between subject and verb favour subject doubling. Here we use the distance in words instead to have a more accurate measurement.
- Length of the NP subject (in words): to control for the confounding effect of subject length, the model incorporates the length of the NP subject. This adjustment is important because longer NP subjects tend to yield higher surprisal, as the surprisal is computed by summing the log probabilities of words comprising the NP subject.

All numeric factors have been centered and standardized. For categorical factors, the Backward Difference coding was employed, comparing adjacent levels on the defined scale, with the higher level compared to the previous one. In addition to the fixed effects, the model takes into account two random intercepts: speaker and verb lemma. We did not include the subject head lemma as a third random intercept, because in our data, 61.7% of the (lemmatized) nouns appear only once as subjects, and 91.9% of the (lemmatized) nouns appear less than 5 times as subjects (out of 1,071 different subject lemmas in total). For these cases, the subject head lemma

Context		NP subject	P(NP|context)	$-\log P$(NP|context)
Previous turn	*Target*			
Pourquoi ils sont partis tes parents ?	Ben parce que ils pensaient que	**la vie**	0.00013	**8.93**

Table 2: Example: estimating NP subject surprisal by considering one previous speech turn.

random intercept would better explain the subject variance in relation to subject doubling, thereby reducing the explanatory power of the subject surprisal effect in general. Hence, the statistical model is defined as follows (doubling = 1, non-doubling = 0). The statistical model reaches an accuracy of 86.68% and shows no concern of multicollinearity, as each variable demonstrates a GVIF$^{(1/(2*Df))}$ inferior to 1.05[6].

```
Subject doubling ~ NP subject surprisal
    + sentential polarity
    + NP subject type + clause type
    + verb frequency + distance
    + NP subject length
    + (1 | speaker) + (1 | verb lemma)
```

Results and discussion

We test whether speakers use the redundant subject clitic to reduce fluctuation in information density at the subject level. As shown by Table 3, which summarizes the results of statistical modeling, while other factors being controlled, the surprisal of the NP subject has a significant impact on subject doubling. An ANOVA test comparing the full model with the model without NP subject surprisal also confirms that the model containing surprisal as a predictor fits significantly better the data than the model without (χ^2 = 14.718, $p < 0.001$), with a decrease in both AIC and BIC values. To better interpret the coefficient of NP subject surprisal, we compute the average marginal probability across all groups in a sample of 100 data points. The findings show that a shift in surprisal from minimum to maximum leads to an average probability change in subject doubling from 0.71 to 0.87, showing the important impact of surprisal on the outcome probability in the mixed model.

The effect of surprisal can also be illustrated by Figure 1, which demonstrates the relationship between NP subject surprisal and the proportion of subject doubling. It can be observed that, as predicted by the UID hypothesis, the more surprising the NP subject is, the more likely it is to be doubled by a redundant subject clitic.

To further test the robustness of the effect of NP subject surprisal, we conducted tests on NP subject surprisal using different lengths for the preceding context, ranging from 1 to 11 speech turns. Table 4 reports the coefficient of the NP subject surprisal in the logistic regression model described previously, across various context sizes. We can see that NP

[6]The GVIF measure (General Variance Inflation Factors) shows no major concern of collinearity in the model if each variable has a GVIF$^{1/(2Df)}$ inferior to 2 (Fox & Monette, 1992).

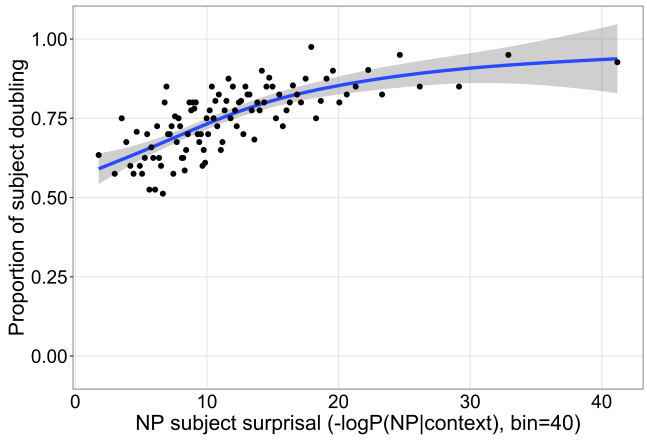

Figure 1: Subject doubling rate as a function of NP subject surprisal. Each point represents a group of 40 observations grouped by approximate surprisal estimates. The curve is generated by GAM (Generalized Additive Model). Subject surprisals are estimated on a single preceding turn.

subject surprisal always exhibits a significant effect on subject doubling in all these experiments. In addition, we also use the GPTfr-124M original model to estimate NP subject surprisal across varying context lengths, and once again, the effect of NP subject surprisal remains significant in all manipulations, with the effect direction aligned with UID expectations. These results underscore the robustness of NP subject surprisal, hardly affected by context size or the GPT model used for surprisal estimation.

Therefore, we conclude that our study provides further evidence for the UID hypothesis from French subject doubling: when the NP subject is not predictable, resulting in a possible exceedance of the channel capacity, speakers tend to use a redundant subject clitic to reduce the information density. On the other hand, when the subject noun is highly predictable, the addition of a clitic would result in a lower density, explaining why the version without the clitic is preferred. This result shows that the UID principle can be extended to French, a language that has received little attention in this regard.

One may wonder whether subject doubling constitutes a suitable case for the UID hypothesis, as subject doubling might be considered as a form of topicalization. If this is the case, information structure and other discourse factors would play a larger role in determining whether the NP subject is topicalized or not. Indeed, the topicalization analysis of subject doubling, where the lexical subject is the topic dislocated in the left periphery, and the subject clitic is the real

Fixed effects:	Coef.	Std. Error	z	p	Sig.
(Intercept)	-3.287674	0.411858	-7.983	1.43e-15	***
polarity (*ne* vs. aff.)	-5.967225	1.055523	-5.653	1.57e-08	***
polarity (without *ne* vs. *ne*)	6.215579	1.066526	5.828	5.61e-09	***
subject (indefinite vs. universal)	2.289983	0.402894	5.684	1.32e-08	***
subject (definite vs. indefinite)	2.436020	0.330875	7.362	1.81e-13	***
clause (other subordinate vs. relative)	1.522349	0.302683	5.030	4.92e-07	***
clause (main vs. other subordinate)	0.961943	0.112814	8.527	$<$ 2e-16	***
verb frequency	0.270007	0.079023	3.417	0.000634	***
distance between subject head and verb	0.237333	0.053328	4.450	8.57e-06	***
NP subject length	0.008452	0.055579	0.152	0.879132	
NP subject surprisal	0.212546	0.056373	3.770	0.000163	***

Table 3: Mixed-effects logistic regression model of subject doubling, with coefficients, standard error, z values, p values, and significance levels of fixed effects. Speaker and verb lemma were included as random intercepts.

Cont. length.	**Coef. surprisal**	p	**N**
1	0.21	< 0.001	4057
2	0.15	< 0.01	4056
3	0.14	< 0.01	4053
4	0.13	< 0.05	4049
5	0.13	< 0.05	4047
6	0.14	< 0.01	4043
7	0.13	< 0.05	4043
8	0.12	< 0.05	4042
9	0.11	< 0.05	4039
10	0.12	< 0.05	4036
11	0.12	< 0.05	4033

Table 4: Context length in terms of the number of preceding speech turns, coefficient of surprisal in the mixed-model, its p value and number of observations. NP subject surprisal estimates are obtained from the fine-tuned GPTfr-124M model.

subject of the sentence (Kayne, 1975; Rizzi, 1986; De Cat, 2005), has been influential in the formal syntactic literature. However, there is an alternative analysis of subject doubling, known as the morphological analysis, which has sparked deep debate in the literature alongside the topicalization analysis. This analysis proposes that the lexical subject functions as the subject of the sentence, while the subject clitic acts as an agreement marker on the verb (Roberge, 1990; Auger, 1995; Culbertson, 2010), and has been argued to be particularly relevant for vernacular French. In our corpus, the remarkably high rate of lexical subject doubling, reaching 75%, suggests that the topicalization analysis may not be able to account for most of the cases, as it is unlikely that lexical subjects are topicalized at such a high rate. Therefore, we argue that subject doubling is a case of redundancy, and it is not surprising that the UID hypothesis can influence speakers' choices when both options are available in the language. Furthermore, a growing body of research has argued for a correlation between information structure and information theory, showing that information theory offers a more general explanation that encompasses a wider array of phenomena, including those traditionally explained by information structure. For example, it is well-known that information structure has an impact on word order variation cross-linguistically, resulting in a theme-first pattern (*old things first*) in many languages like English and Czech. However, this approach cannot explain the rheme-first tendency observed in a small number of languages like Iroquoian and Caddoan (Mithun, 1995; Creider & Creider, 1983). To reconcile these facts, Komagata (2003) proposes an information-theoretical definition of information structure and introduces the hypothesis of *information balance*, which resonates with the UID principle and explains both patterns. Another example concerns fragments, a classical case of ellipsis. While the information-structural approach requires omitted expressions to be given (Reich, 2007; Weir, 2014), psycholinguistic experiments have shown that speakers are sensitive to the predictability of words in the usage of fragments, aligning with the information-theoretical approach (Lemke, 2021; Lemke et al., 2021). The two approaches correlate in this regard, as predictable words are often given. Furthermore, the information-theoretical approach can explain why a word is preferably omitted at a particular site and why the surprisal of surrounding words would influence the omission of the target words – aspects that information structure is less capable of explaining.

Conclusion

In this study, we examine the applicability of the *Uniform Information Density* (UID) principle to the subject doubling phenomenon in French. Our findings reveal a significant effect of NP subject surprisal on subject doubling, providing additional support for the UID hypothesis in French. We argue for the usage of GPT models to obtain surprisal estimates, as they offer the advantage of incremental sentence processing, similar to human processing, and incorporate linguistic knowledge and substantial previous context when estimating surprisals. Therefore, they offer a promising approach to studying various linguistic phenomena and exploring the role of surprise in language processing.

Acknowledgments

This work was funded by Heather Burnett's ERC SMIC project (under the European Union's Horizon 2020 research and innovation programme, grant agreement N°850539), Labex EFL ANR-10-LABX-0083 and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) — Project-ID 232722074 — SFB/CRC 1102.

References

Auger, J. (1994). *Pronominal Clitics in Québec Colloquial French: A Morphological Analysis*. Unpublished doctoral dissertation, University of Pennsylvania.

Auger, J. (1995). Les clitiques pronominaux en français parlé informel : une approche morphologique. *Revue québécoise de linguistique*, *24*(1), 21–60. doi: 10.7202/603102ar

Auger, J. (1998). Le redoublement des sujets en français informel québécois: Une approche variationniste. *Canadian Journal of Linguistics/Revue canadienne de linguistique*, *43*(1), 37–63. doi: 10.1017/S0008413100020429

Auger, J., & Villeneuve, A.-J. (2010). La double expression des sujets en français saguenéen: Étude variationniste. *Hétérogénéité et Homogénéité dans les pratiques langagieres: Mélanges offerts à Denise Deshaies. Quebec: Presses de l'Université Laval*, 67–86.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Branca-Rosoff, S., Fleury, S., Lefeuvre, F., & Pires, M. (2012). Discours sur la ville. Présentation du Corpus de Français Parlé Parisien des années 2000 (CFPP2000). *article en ligne,* `http://cfpp2000.univ-paris3.fr/Articles.html`.

Campion, E. (1984). *Left dislocation in Montréal French*. Unpublished doctoral dissertation, University of Pennsylvania.

Creider, C. A., & Creider, J. T. (1983). Topic: Comment relations in a verb-initial language. *Journal of African Languages and Linguistics*, *5*(1), 1–15.

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

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

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*(2), 193–210. doi: 10.1016/j.cognition.2008.07.008

Demberg, V., Sayeed, A. B., Gorinski, P. J., & Engonopoulos, N. (2012). Syntactic surprisal affects spoken word duration in conversational contexts. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning* (pp. 356–367). Jeju Island, Korea..

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

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, *140*, 1–11. doi: 10.1016/j.bandl.2014.10.006

Futrell, R., Gibson, E., & Levy, R. P. (2020). Lossy-Context Surprisal: An Information-Theoretic Model of Memory Effects in Sentence Processing. *Cognitive Science*, *44*(3). doi: 10.1111/cogs.12814

Gadet, F. (Ed.). (2017). *Les parlers jeunes dans l'île-de-France multiculturelle*. Paris and Gap: Ophrys.

Gadet, F., & Guerin, E. (2016). Construire un corpus pour des façons de parler non standard : Multicultural Paris French. *Corpus*, *15*. doi: 10.4000/corpus.3049

Giulianelli, M., Sinclair, A., & Fernández, R. (2021). Is Information Density Uniform in Task-Oriented Dialogues? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 8271–8283). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Grobol, L., & Crabbé, B. (2021). Analyse en dépendances du français avec des plongements contextualisés (French dependency parsing with contextualized embeddings). In *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale* (pp. 106–114). Lille, France: ATALA.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies* (pp. 1–8). USA: Association for Computational Linguistics. doi: 10.3115/1073336.1073357

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, 23-62.

Jaeger, T. F. (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.

Kayne, R. S. (1975). *French syntax: The transformational cycle* (Vol. 30). MIT press Cambridge, MA.

Komagata, N. (2003). Contextual Effects on Word Order: Information Structure and Information Theory. In P. Blackburn, C. Ghidini, R. M. Turner, & F. Giunchiglia (Eds.), *Modeling and Using Context* (Vol. 2680, pp. 190–203). Berlin, Heidelberg: Springer Berlin Heidelberg. doi: 10.1007/3-540-44958-2_16

Kuribayashi, T., Oseki, Y., Ito, T., Yoshida, R., Asahara, M., & Inui, K. (2021). Lower Perplexity is Not Always Human-Like. In *ACL*. arXiv.

Lemke, R. (2021). *Experimental investigations on the syntax and usage of fragments*. Language Science Press.

Lemke, R., Horch, E., & Reich, I. (2017). Optimal encoding! - Information Theory constrains article omission in newspaper headlines. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers* (pp. 131–135). Valencia, Spain: Association for Computational Linguistics.

Lemke, R., Reich, I., Schäfer, L., & Drenhaus, H. (2021). Predictable Words Are More Likely to Be Omitted in Fragments–Evidence From Production Data. *Frontiers in Psychology*, *12*.

Levy, R., & Jaeger, T. (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* (Vol. 19, pp. 849–856). MIT Press.

Liang, Y., Donati, C., & Burnett, H. (2023). *French Subject Doubling: A Third Path* [Oral Presentation]. Paper presented at 53rd Linguistic Symposium on Romance Languages, Paris, France. Paris, France. Retrieved from https://lsrl53.sciencesconf.org/475601

Meister, C., Pimentel, T., Haller, P., Jäger, L., Cotterell, R., & Levy, R. (2021). Revisiting the Uniform Information Density Hypothesis. *arXiv:2109.11635 [cs]*.

Merkx, D., & Frank, S. L. (2021). Human Sentence Processing: Recurrence or Attention? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics* (pp. 12–22). Online: Association for Computational Linguistics. doi: 10.18653/v1/2021.cmcl-1.2

Mithun, M. (1995). Morphological and prosodic forces shaping word order. *Word order in discourse*, *30*, 387.

Multicultural Parisian French [corpus]. (2019). *Multicultural Parisian French [corpus], version 3.* Retrieved from https://hdl.handle.net/11403/mpf/v3 (ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr)

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

Oh, B.-D., Clark, C., & Schuler, W. (2022). Comparison of Structural Parsers and Neural Language Models as Surprisal Estimators. *Frontiers in Artificial Intelligence*, *5*.

Oh, B.-D., & Schuler, W. (2022). Entropy- and Distance-Based Predictors From GPT-2 Attention Patterns Predict Reading Times Over and Above GPT-2 Surprisal. In *EMNLP*. arXiv. (http://arxiv.org/abs/2212.11185)

Oh, B.-D., & Schuler, W. (2023). Why Does Surprisal From Larger Transformer-Based Language Models Provide a Poorer Fit to Human Reading Times? *Transactions of the Association for Computational Linguistics*, *11*, 336–350. doi: 10.1162/tacl_a_00548

Pabst, K., Konnelly, L., Wilson, F., & Nagy, N. (2020). Variation in subject doubling in Homeland and Heritage Faetar. *Toronto Working Papers in Linguistics*, *42*.

Pimentel, T., Meister, C., Salesky, E., Teufel, S., Blasi, D., & Cotterell, R. (2021). A surprisal–duration trade-off across and within the world's languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 949–962). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.73

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. *arXiv:2003.07082 [cs]*.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2019). *Improving Language Understanding by Generative Pre-Training*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language Models are Unsupervised Multitask Learners.* Open AI Technical Report.

Reich, I. (2007). Toward a uniform analysis of short answers and gapping. *On information structure, meaning and form*, 467–484.

Rizzi, L. (1986). Null Objects in Italian and the Theory of pro. *Linguistic Inquiry*, *17*(3), 501–557.

Roberge, Y. (1990). *Syntactic Recoverability of Null Arguments*. McGill-Queen's University Press.

Sankoff, G. (1982). Usage linguistique et grammaticalisation: Les clitiques sujets en français. *La sociolinguistique dans les pays de langue romane. Tübingen: Gunter Narr Verlag*, 81–85.

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

Simoulin, A., & Crabbé, B. (2021). Un modèle Transformer Génératif Pré-entrainé pour le_____ français (Generative Pre-trained Transformer in_____ (French) We introduce a French adaptation from the well-known GPT model). In *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale* (pp. 246–255). Lille, France: ATALA.

Team, R. C. (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Torabi Asr, F., & Demberg, V. (2015). Uniform Surprisal at the Level of Discourse Relations: Negation Markers and Discourse Connective Omission. In *Proceedings of the 11th International Conference on Computational Semantics* (pp. 118–128). London, UK: Association for Computational Linguistics.

Weir, A. (2014). Fragment answers and the Question under Discussion. In *Proceedings of NELS* (Vol. 44, pp. 255–266).

Wilcox, E. G., Gauthier, J., Hu, J., Qian, P., & Levy, R. (2020). On the predictive power of neural language models for human real-time comprehension behavior. In *CogSci.* Retrieved from https://arxiv.org/abs/2006.01912 (https://arxiv.org/abs/2006.01912)

Zahler, S. (2014). Variable subject doubling in spoken Parisian French. *University of Pennsylvania Working Papers in Linguistics*, *20*(1), 38.