Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Uniformité de la densité informationnelle: le cas du redoublement du sujet

Yiming Liang[1] Pascal Amsili[2] Heather Burnett[1]
(1) Laboratoire de linguistique formelle, Université Paris Cité/CNRS, Paris, France
(2) Laboratoire Lattice, CNRS/PSL-ENS/Sorbonne Nouvelle, Paris, France
yiming.liang@etu.u-paris.fr, pascal.amsili@ens.fr
heather.susan.burnett@gmail.com

RÉSUMÉ

Nous présentons les résultats d'une expérience visant à savoir si la densité d'information (ou de surprise) affecte le redoublement du sujet dans des conversations spontanées. En utilisant la version française de GPT, nous estimons la surprise lexicale du sujet NP étant donné un contexte précédent et vérifions si la surprise du sujet affecte son redoublement. L'analyse de régression à effet mixte montre que, en plus des facteurs qui ont été montrés comme affectant le redoublement du sujet dans la littérature, la prévisibilité du sujet nominal est un prédicteur important du non-redoublement. Les sujets nominaux moins prédictibles tendent à être redoublés par rapport à ceux qui sont plus prédictibles. Notre travail confirme l'intérêt de l'hypothèse de l'Uniformité de la densité informationnelle (UID) pour le français et illustre l'opérationalisation de la densité informationnelle à l'aide de grands modèles neuronaux de langage.

ABSTRACT

The Uniform Information Density Hypothesis : the case of subjet-doubling in French

We present the results of an experiment investigating whether information density affects subject doubling in conversations in French. Using the French version of GPT, we estimate the lexical surprisal of the subject NP subject given a certain left context and verify whether the surprisal of the subject affects its doubling. A Mixed effect regression analysis shows that, in addition to factors that have been shown to affect subject doubling in the literature, the predictability of the NP is an important predictor of subject doubling. Less predictable NPs tend to be more often doubled with a clitic than more predictable ones. Our work thus provides additional support to the Uniform Information Density (UID) hypothesis in French and points to a way to the operationalization of information density with the help of large neural language models.

MOTS-CLÉS : uniformité de la densité informationnelle, redoublement du sujet, surprise, français oral, modèle Tranformer Génératif Pré-entraîné (GPT).

KEYWORDS: Uniform Information Density, subject doubling, surprisal, spoken French, Generative Pre-trained Transformer (GPT).

1 Introduction

Le redoublement du sujet à l'oral (1) est un phénomène fréquent en français, en particulier à l'oral, et il est conditionné par de nombreux paramètres linguistiques, qui vont de la phonologie à la structure

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : travaux de recherche originaux – articles longs, pages 164–176.

informationnelle.

(1) Marie$_i$ elle$_i$ m'a prêté son vélo.

Ce phénomène se caractérise aussi par le fait qu'il n'apporte en général pas d'information nouvelle, les locuteurs ont donc le choix entre deux constructions sémantiquement comparables. Ce genre de phénomène d'alternance, que l'on peut aussi observer à propos de l'ordre des mots par exemple (2-a) *vs.* (2-b), est un terrain privilégié pour étudier l'influence d'un principe cognitif qui s'est révélé pertinent dans de nombreuses études (Maurits *et al.*, 2010; Cuskley *et al.*, 2021, parmi d'autres), le principe selon lequel les locuteurs choisissent la version qui uniformise la densité informationnelle de l'énoncé (hypothèse UID, définie et illustrée à la section 2).

(2) a. J'ai donné ce livre à Jade.
b. J'ai donné à Jade ce livre.

Le travail que nous présentons dans cet article s'inscrit dans cette lignée, et il vise à étudier la façon dont ce principe peut contribuer à expliquer les choix des locuteurs. À partir de corpus oraux récents annotés, nous proposons un modèle de régression logistique intégrant tous les facteurs connus pour avoir une influence sur le redoublement du sujet. Nous montrons que prendre en compte la dimension d'uniformité de la densité informationnelle augmente la capacité prédictive du modèle, ce qui conforte l'intérêt de cette hypothèse pour expliquer le phénomène.

Nous présentons à la section 2 l'hypothèse elle-même, et évoquons quelques travaux qui ont proposé de la prendre en compte. Le phénomène du redoublement du sujet fait l'objet de la section 3, où nous détaillons les variables dont l'influence a été établie, et proposons un premier modèle de régression logistique. Nous montrons à la section 4 qu'en formulant la densité informationnelle au moyen d'une probabilité obtenue par un modèle de langue, et en intégrant cette variable dans le modèle, on prédit mieux le redoublement du sujet. Les limites et perspectives ouvertes par cette étude sont présentées à la section 5.

2 L'hypothèse de l'uniformité de la densité informationnelle

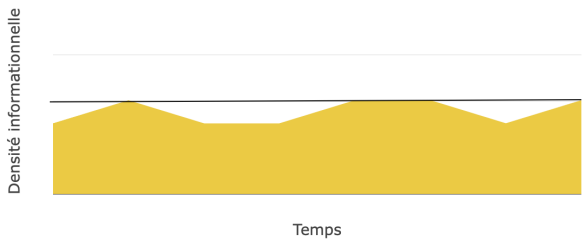

FIGURE 1 – Bonne utilisation du canal.

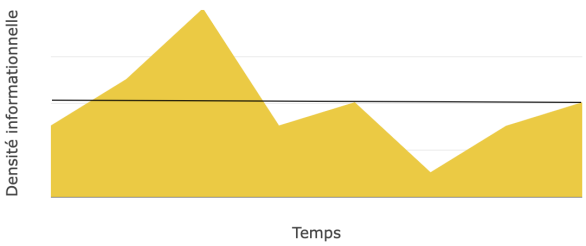

FIGURE 2 – Mauvaise utilisation du canal.

L'hypothèse de l'uniformité de la densité informationnelle (*Uniform Information Density*, UID) (Levy & Jaeger, 2007; Jaeger, 2010) pose que, dans la mesure où la grammaire le permet, les locuteurs distribuent l'information de la façon la plus homogène possible au sein d'une phrase. L'intuition

qui sous-tend cette hypothèse est double : d'une part, la communication peut être vue comme une transmission d'information à travers un canal bruité et limité par nos ressources cognitives, ce qui peut donner lieu à des erreurs, voire à l'échec de la transmission du message initial. Pour augmenter les chances de succès de la communication, il est préférable que l'information transmise à chaque unité de temps ne dépasse pas la limite de la capacité du canal (illustrée par une ligne noire dans les figures 1 et 2). D'autre part, une communication efficace suppose que le canal de communication ne soit pas sous-employé, et par conséquent qu'un niveau minimal d'informativité soit garanti. La figure 1 représente l'évolution de la densité informationnelle au cours du temps et illustre une communication proche de l'idéal selon UID, par constraste avec une utilisation sous-optimale illustrée à la figure 2.

On se place ici dans le cadre de la théorie de l'information (Shannon, 1948), qui permet de définir la densité informationnelle au moyen de la surprise pour chaque mot w_i, définie comme le logarithme négatif de la probabilité de l'apparition d'un mot étant donné le contexte (équation 1). Plusieurs études psycholinguistiques ont montré que les unités linguistiques présentant un niveau de surprise plus élevé sont associées à des difficultés de compréhension et nécessitent davantage de temps de lecture (Demberg & Keller, 2008; Wilcox *et al.*, 2020; Futrell *et al.*, 2020; Frank *et al.*, 2015).

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

Dans ce qui suit, nous parlerons de façon interchangeable de *surprise*, d'*informativité* ou de *niveau d'information* : par hypothèse, une unité peu prédictible (ayant donc une probabilité faible), est associée à un niveau élevé de surprise, et est donc considérée comme informative.

Diverses études portant sur des phénomènes de variation à différents niveaux linguistiques ont permis de conforter l'hypothèse UID. À titre d'illustration et de manière non exhaustive : niveau phonétique : un son ou un mot ayant une surprise plus élevée est prononcé plus lentement au sein et à travers les langues (Demberg *et al.*, 2012; Pimentel *et al.*, 2021) ; niveau syntaxique : les mots fonctionnels tendent à être omis lorsque la structure qu'ils introduisent est plus prédictible (par exemple, omission de *that* dans les complétives (Jaeger, 2010) et les relatives (Jaeger, 2011) en anglais, omission d'article dans les titres de journaux en allemand (Lemke *et al.*, 2017)) ; niveau discursif : l'omission des connecteurs est plus fréquente lorsque la relation discursive est moins surprenante (Torabi Asr & Demberg, 2015) ; entre autres.

Ces travaux montrent de façon convaincante l'importance de la densité informationnelle, mais la question de l'opérationalisation de l'hypothèse reste aujourd'hui encore délicate. Deux aspects sont particulièrement discutés dans la littérature. D'une part, il faut se demander si le calcul se fait localement ou globalement (voir (Meister *et al.*, 2021) pour une discussion approfondie). D'autre part, et de façon plus cruciale encore, il s'agit de savoir par quel moyen on va estimer le niveau d'informativité d'une unité (typiquement un token). Les études qui adoptent une approche de corpus utilisent souvent un modèle bigramme basé sur des fréquences (Jaeger, 2010, 2011; Lemke *et al.*, 2017). D'autres études se basent sur des indices plus linguistiquement motivés, mais plus spécifiques. Par exemple, Torabi Asr & Demberg (2015) s'appuient sur la présence d'une négation dans la première phrase en tant qu'indicateur approximatif de la prédictibilité d'une relation causale avec la deuxième phrase. Ces mesures prennent en considération un contexte assez limité et ne sont pas faciles à généraliser à d'autres phénomènes. Ainsi, dans l'article de Jaeger (2010) sur l'omission de *that*, la surprise d'une complétive est estimée par la fréquence de sous-catégorisation du verbe. Par exemple, *think* « penser » est souvent suivi d'une complétive et donc l'apparition d'une complétive est peu surprenante après ce verbe. Cependant, même pour un verbe donné, la probabilité de la

complétive peut varier selon les contextes. Par exemple, l'apparition d'une complétive après *think* est plus probable quand *I think* se trouve en position initiale de la phrase (3-a) que quand il se trouve en position non initiale (3-b).

(3) a. I think he is crazy.
b. He is crazy, I think.

Les modèles de langues sont largement utilisés dans des travaux de modélisation cognitive ou de psycholinguistique pour fournir des probabilités conditionnelles de mots étant donné le contexte gauche, afin de mesurer la corrélation entre la surprise des modèles et diverses mesures comportementales (temps de lecture, présence de difficultés de compréhension, mouvements occulaires, etc.). Des corrélations ont été observées dans de nombreuses études utilisant différents types de modèles tels que n-grammes, GRU (*Gated Recurrent Unit*), LSTM (*Long Short-Term Memory*) et plus récemment, GPT (*Generative Pre-trained Transformer*) (Goodkind & Bicknell, 2018; Wilcox *et al.*, 2020; Merkx & Frank, 2021; Kuribayashi *et al.*, 2022). Cependant, ils sont rarement utilisés pour explorer le rôle de la densité informationnelle dans la variation syntaxique. De plus, les phénomènes de variation restent peu étudiés en français sous cette optique (mais voir (Liang *et al.*, 2021) sur l'omission de *que* en français montréalais). Nous proposons d'utiliser un modèle de langue génératif, en l'occurrence GPT, pour estimer la probabilité (et donc l'information) des mots, afin de contribuer à cette ligne de recherche en français, sur un phénomène de redondance syntaxique, le redoublement du sujet.

3 Le phénomène

3.1 Redoublement du sujet

Le redoublement du sujet, où un sujet lexical et un clitique coréférentiel apparaissent en même temps dans une phrase ((4-a), à comparer à la version non-redoublée (4-b)), est un phénomène largement répandu en français oral [1].

(4) a. Mon père$_i$ il$_i$ est venu.
b. Mon père est venu.

De nombreux facteurs ont été proposés pour expliquer la variation entre les deux constructions (doublée *vs.* non-redoublée) : par exemple, le type de sujet nominal (Nadasdi, 1995; Auger, 1998; Auger & Villeneuve, 2010), le type de proposition (Auger & Villeneuve, 2010), la présence d'éléments intervenant entre le sujet et le verbe (Zahler, 2014), le statut informationnel (Zahler, 2014), entre autres. Malgré cette littérature riche sur le rôle des facteurs grammaticaux pertinents, peu d'études ont examiné ce phénomène sous l'angle de la théorie de l'information.

Pourtant, le redoublement peut être considéré comme un exemple de redondance syntaxique, un cas où l'hypothèse UID pourrait contribuer à expliquer le choix des locuteurs, dans la mesure où la présence du sujet clitique n'apporte pas d'information nouvelle à la phrase. On pourrait proposer le raisonnement suivant : le pronom redoublé est par essence peu informatif, et par conséquent introduire

1. Il est aussi possible de redoubler un sujet nominal postverbal par un sujet clitique (*Il$_i$ est venu mon père$_i$*). Dans la présente étude, nous nous limitons aux cas de redoublement de sujet préverbal.

un pronom conduit à une baisse locale du niveau de surprise. L'hypothèse UID prédit que l'on aura plus tendance à introduire ce pronom quand la surprise associée au sujet lexical est élevée, pour lisser le niveau de surprise, alors qu'au contraire avec un sujet lexical très prédictible, et donc de surprise peu élevée, l'ajout d'un pronom lui-même peu informatif n'a pas d'impact pertinent sur l'uniformité de la densité informationnelle.

Afin d'examiner la pertinence de cette proposition, nous avons mené une étude du redoublement du sujet dans un corpus de français oral, *Multicultural Parisian French* (MPF) (Gadet & Guerin, 2016; Gadet, 2017). Avant de procéder à l'opérationalisation de UID dans ce nouveau cas, il est important d'identifier et de contrôler les autres facteurs potentiels.

3.2 Une première modélisation

3.2.1 Extraction des données

Le corpus MPF se compose de 66 entretiens pour un total de près de 790 000 mots transcrits à ce jour, et vise à documenter le langage oral de jeunes âgés de 12 à 37 ans, issus d'un milieu familial multiculturel et résidant dans les banlieues parisiennes [2]. Il s'agit de conversations face à face entre amis ou connaissances portant sur des thèmes variés tels que la famille, la vie quotidienne, l'évolution des langues, entre autres. Le corpus n'étant pas muni d'annotations morpho-syntaxiques, nous l'avons d'abord segmenté et annoté en parties du discours à l'aide de *Stanza* (Qi *et al.*, 2020), puis utilisé l'*analyseur HOPS* (Grobol & Crabbé, 2021) pour en obtenir une analyse en dépendance. Nous avons ensuite procédé à l'extraction de toutes les phrases qui contiennent un sujet nominal (par exemple *mon père*, *un garçon*, *certains*, *tout le monde*, *Marie*...) de l'ensemble du corpus et indiqué si le sujet nominal est redoublé par un clitique (comme *il(s), elle(s), ce, ça*). Seuls les sujets préverbaux de troisième personne ont été pris en compte. Les pronoms forts, tels que *lui* et *eux*, ont été exclus. L'extraction a été faite à l'aide d'un script Python analysant les annotations morpho-syntaxiques et syntaxiques du corpus, complété par une vérification manuelle. Pour ce modèle comme pour le suivant, nous avons limité notre analyse aux cas où la tête du sujet est réduite à un seul token suite à la prétokenization par *Stanza* pour qu'il soit possible d'établir une valeur de surprise pour ce token. Ce processus a abouti à un jeu de données de 4 136 occurrences, dont le taux de redoublement est de 75,5 %.

3.2.2 Facteurs de contrôle

Un second script Python a été utilisé pour annoter de façon semi-automatique toutes les occurrences extraites pour les facteurs listés ci-dessous. L'objectif était de relever les facteurs qui conditionnent le redoublement du sujet, afin de les contrôler lors de l'exploration de l'effet de la densité informationnelle décrite dans la section suivante :

— Polarité de la phrase : affirmative, négative avec *ne*, négative sans *ne* ;
— Type de sujet : sujet quantifié universellement (par ex. *tout le monde*, *rien*), groupe nominal indéfini (par ex. *un garçon*), groupe nominal défini (par ex. *le garçon*, *Daniel*) ;
— Type de proposition : proposition principale, proposition subordonnée autre que les relatives, proposition relative ;

2. Les entretiens et les transcriptions sont en accès libre sur le site du corpus : https://www.ortolang.fr/market/corpora/mpf/v3 (mpf, 2019).

— Fréquence du verbe mesurée dans le même corpus ;
— Distance en nombre de mots entre la tête du sujet et le verbe.

3.2.3 Modélisation statistique

Une modélisation au moyen d'un modèle de régression logistique à effets mixtes a été réalisée avec le logiciel R Studio (Team, 2022), implémentée grâce à la fonction `glmer()` du package lme4 (Bates *et al.*, 2015). Les facteurs numériques sont transformés en logarithme et centrés. Pour les facteurs catégoriels, le codage de *Backward Difference* est employé, comparant les niveaux adjacents dans l'échelle définie ci-dessous, le niveau supérieur est comparé avec le précédent. En plus des effets fixes, le modèle prend en compte trois effets aléatoires : locuteur, lemme du verbe, lemme de la tête du sujet. Le modèle de référence est ainsi défini (redoublement = 1, non redoublement = 0) :

MODÈLE DE RÉFÉRENCE : Redoublement ~ polarité + type_sujet + type_proposition + fréquence_verbe + distance + (1 | lemma_verbe) + (1 | lemma_sujet) + (1 | locuteur)

La table 1 (première colonne, « modèle de référence ») récapitule les facteurs qui jouent un rôle significatif sur le redoublement du sujet dans le modèle de référence. Tous les facteurs de contrôle pris en compte sont significatifs.

	Modèle de référence				**Modèle principal**			
	Coef.	z	p	Sig.	Coef.	z	p	Sig.
(Intercept)	-2.96	-6.68	2.47e-11	***	-2.99	-6.76	1.35e-11	***
polarité (*ne* vs. aff.)	-6.43	-6.06	1.33e-09	***	-6.43	-6.06	1.36e-09	***
polarité (sans *ne* vs. *ne*)	6.87	6.39	1.62e-10	***	6.88	1.08	1.65e-10	***
sujet (ind. vs. uni.)	1.61	2.74	0.00608	**	1.62	2.80	0.00517	**
sujet (defini vs. ind.)	2.47	5.28	1.30e-07	***	2.44	5.28	1.26e-07	***
prop. (sub. vs. prin.)	1.43	4.42	9.95e-06	***	1.45	4.48	7.37e-06	***
prop. (rel. vs. sub.)	0.97	7.86	3.78e-15	***	0.95	7.64	2.19e-14	***
fréquence verbe	0.27	3.18	0.00148	**	0.26	3.10	0.00193	**
distance tête suj. - verbe	0.26	4.46	8.21e-06	***	0.26	4.51	6.53e-06	***
surprise (cf. § 4.2)	-	-	-	-	**0.15**	**2.23**	**0.02566**	*

TABLE 1 – Modèle de régression logistique à effets mixtes du redoublement du sujet en fonction des facteurs de contrôle (modèle de référence) et de la surprise du sujet (modèle principal), avec les coefficients de pente, les valeurs *z*, les valeurs *p* ainsi que les niveaux de significativité des effets fixes. Le locuteur, les lemmes du sujet et du verbe ont été considérés comme des intercepts aléatoires.

Les effets des différents facteurs de contrôle dans le modèle de référence (table 1) peuvent être résumés brièvement :

— Polarité de la phrase : les phrases négatives contenant « ne » défavorisent le redoublement du sujet en comparaison avec les phrases négatives sans « ne » et les phrases affirmatives.
— Type de sujet : les sujets quantifiés universellement sont redoublés le moins souvent, tandis que les sujets définis sont redoublés le plus souvent. Les sujets indéfinis se situent entre les deux.
— Type de proposition : les propositions principales sont corrélées avec un taux d'omission le plus élevé, tandis que les subordonnées défavorisent le redoublement. Parmi les subordonnées,

les propositions relatives sont celles qui défavorisent le plus le redoublement et les autres types de subordonnées se situent entre les deux.
— Fréquence du verbe : plus un verbe est fréquent dans le corpus, plus il y a de chances que le sujet soit redoublé.
— Distance en nombre de mots entre la tête du sujet et le verbe : plus le verbe est distant de la tête du sujet, plus il est probable que le sujet soit redoublé.

4 Prise en compte de la densité informationnelle

4.1 Utilisation d'un modèle de langue génératif

Afin d'estimer la surprise du sujet nominal, nous avons utilisé un modèle GPT (transformer génératif pré-entraîné). Basés sur les transformers (Vaswani *et al.*, 2017), et utilisant des blocs de type décodeur (i.e. avec auto-attention masquée), les modèles GPT sont capables de générer un token à la position t_i à partir d'un préfixe $t_0 ... t_{i-1}$ qui peut être très long (1 024 tokens au maximum). Puisque le modèle GPT est capable de prendre en considération un contexte plus étendu en comparaison avec d'autres types de modèles tels que n-gramme et LSTM, nous l'avons choisi dans la présente étude. Toutefois, il a été constaté que les valeurs de surprise estimées par un modèle GPT ayant plus de paramètres sont moins efficaces pour prédire les temps de lecture de mots, en comparaison avec ses versions plus petites (Oh *et al.*, 2022; Oh & Schuler, 2023). De ce fait, notre étude est basée sur le modèle GPT le plus petit disponible pour le français, avec 124 millions de paramètres (Simoulin & Crabbé, 2021), GPT$_{fr}$-124M, lui-même adapté des modèles OpenAI GPT et GPT-2 (Radford *et al.*, 2019a,b).

4.1.1 Ajustement du modèle

Comme GPT$_{fr}$-124M a été entraîné principalement sur les documents écrits, il est moins capable de prendre en considération les caractéristiques des textes oraux (hésitations, répétitions, reformulations, registre lexical familier...). Par conséquent, nous avons procédé à un ajustement du modèle (*fine-tuning*) avec un autre corpus du français parisien parlé, le Corpus de Français Parlé Parisien des années 2000 (CFPP2000) (Branca-Rosoff *et al.*, 2012). Collecté à partir de 2005-2006, il se compose d'un ensemble d'entretiens réalisés dans différents quartiers de Paris et de la proche banlieue. Actuellement, 51 entretiens transcrits contenant un total de 750 000 tokens (d'après la prétokenization par *Stanza*) sont disponibles sur le site [3], à partir desquels le corpus d'entraînement et d'évaluation a été construit pour l'ajustement du modèle. La tâche est de prédire le prochain mot vu le contexte précédent. Afin de ne pas dégrader la généralité du modèle, pendant l'ajustement, nous avons utilisé la même tokenization définie par le modèle GPT$_{fr}$-124M (Simoulin & Crabbé, 2021) et construit le corpus d'entraînement d'une façon similaire.

Chaque entretien est divisé en tours de parole. Un exemple d'apprentissage est construit à partir de tours de parole successifs concaténés dans la limite de 1 024 tokens. Une fois le nombre atteint, un nouvel exemple est construit à partir du tour de parole suivant, et aucun exemple ne franchit la frontière des entretiens (le dernier exemple est complété (*padding*)). Nous avons divisé les entretiens en corpus d'entraînement et de test, contenant respectivement 583 (46 entretiens) et 97 exemples (5 entretiens). L'ajustement s'est déroulé pendant 30 époques avec les paramètres par défaut.

3. http://cfpp2000.univ-paris3.fr/

4.1.2 Évaluation du modèle ajusté

Le modèle ajusté est évalué sur le corpus de test du CFPP2000 (10% des entretiens) et sur 50% des entretiens de MPF respectivement. Nous avons évalué la perplexité sur la base de la tokenization qui est identique pour le modèle original (GPT$_{fr}$-124M) et le modèle ajusté. Comme on le voit dans la table 2, l'ajustement a permis une réduction significative dans la perplexité évaluée sur les exemples de test dans les deux corpus, suggérant que le modèle ajusté s'adapte mieux aux données orales que le modèle original.

	CFPP2000	MPF
nombre d'exemples de test	97	379
perplexité moyenne du modèle original	42,25	42,61
perplexité moyenne du modèle ajusté	28,93	40,97
p valeur	$< 2.2e-16$	0,043

TABLE 2 – Comparaison de la perplexité par les deux modèles dans les deux corpus oraux. Des t-tests ont été employés pour tester la significativité de la différence.

4.1.3 Estimation de l'information du sujet

Grâce à ce modèle ajusté, il est possible d'estimer la probabilité (et donc la surprise) du sujet nominal dans toutes les occurrences extraites du corpus. Afin de faciliter le calcul, on calcule la probabilité seulement pour la tête du syntagme nominal (dont on s'est assuré qu'elle était mono-lexicale, voir plus haut). Il reste à établir la longueur du contexte gauche que nous devons fournir au modèle de langue pour produire cette probabilité. Comme le sujet d'une phrase est souvent situé en position initiale d'une phrase, ce qui rend l'estimation de la probabilité peu fiable, il convient d'élargir le contexte considéré. Généralement, ce genre de modèle génératif est d'autant plus performant que le contexte gauche est étendu. Cependant, ce que nous cherchons à savoir est plutôt quelle taille du contexte à gauche serait la plus appropriée pour simuler le comportement humain. Dans une étude récente, Kuribayashi *et al.* (2022) montrent que plus le contexte précédent est limité, plus la probabilité du mot estimée par le modèle GPT est corrélée avec le temps de lecture moyen de ce mot mesuré par oculométrie en anglais et en japonais. C'est la raison pour laquelle, dans la présente étude, nous avons pris comme contexte gauche un seul tour de parole avant la cible, indépendamment du nombre de mots (voir une illustration à la table 3). Nous revenons en conclusion sur la question de la taille du contexte gauche. Un tour de parole compte en moyenne 10 tokens. La tâche consiste à prédire le mot à la position de la tête du sujet en position t_i (en gras). La probabilité logarithmique négative du sujet sera considérée comme la surprise du sujet.

Contexte ($t_0 ... t_{i-1}$)		Tête du sujet (t_i)	$P(t_i\|\text{contexte})$	$-\log P(t_i\|\text{contexte})$
Tour précédant la cible	*Cible*			
Oui c'est mes origines.	Ben c'est moi quoi chez moi euh on ma	**mère**	0,0125	4,3782

TABLE 3 – Exemple de l'estimation de la surprise du sujet en prenant en compte un tour de parole.

Puisque le modèle GPT utilise un vocabulaire de type *bytepair encoding* (BPE), il est possible que la tête du sujet soit décomposée en sous-mots. Dans la présente étude, 21 % des tokens du sujet ont été décomposés en 2 à 6 sous-mots. Il s'agit souvent de noms propres, de mots contenant un symbole

(comme *quelqu'un*, *grand-mère*) ou de mots familiers (verlan et emprunts). Puisque la fréquence des sous-mots est plus élevée que celle du mot en entier, la moyenne des probabilités sous-estimerait la surprise de l'ensemble du mot. Par conséquent, nous définissons la probabilité d'un mot comme la probabilité conjointe des sous-mots qui le composent. La surprise de la tête du sujet est donc définie comme la somme des log probabilités de ses sous-mots, ce qui est conforme aux pratiques de plusieurs études analysant la plausibilité cognitive des modèles de langue neuronaux (Wilcox *et al.*, 2020; Kuribayashi *et al.*, 2021; Oh & Schuler, 2022).

Nous utilisons le même modèle de régression logistique que celui de la section précédente, auquel est ajoutée la variable correspondant à la surprise de la tête du sujet telle qu'elle est estimée par le modèle GPT.

4.2 Résultats

Les résultats de ce nouveau modèle de régression logistique sont présentés en colonne de droite de la table 1 (déjà donnée à la fin de la section 3). Ils montrent que, en plus des effets identifiés précédemment, la surprise de la tête du sujet a un effet sur le taux de redoublement. Pour étudier plus finement la relation entre surprise et probabilité de redoublement, nous représentons à la figure 3 toutes les observations regroupées par valeur de surprise en groupes de 30 contre la proportion de redoublement. On voit que moins un sujet est prédictible (plus il est informatif, donc), plus il a tendance à être redoublé par un clitique. Un test anova confirme que l'inclusion de la surprise de la tête du sujet dans le modèle contribue à une amélioration de performance significative ($p < 0.05$).

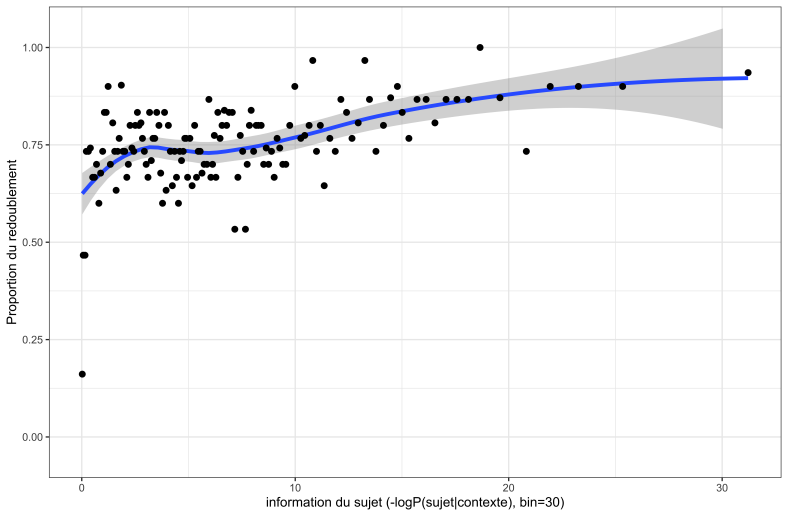

FIGURE 3 – Taux de redoublement du sujet en fonction de la surprise de la tête du sujet nominal. Chaque point représente un groupe de 30 observations regroupées par niveau de surprise. La courbe est générée par la méthode *loess*.

5 Conclusion

5.1 Nouveau champ d'aplication de UID

Le résultat montre une relation entre le taux de redoublement et la surprise du sujet nominal, ce qui suggère que l'hypothèse UID contribue à expliquer le phénomène du redoublement du sujet étudié ici. On peut en effet supposer que, comme la densité informationnelle au niveau du sujet risque de dépasser la capacité du canal quand le nom est moins prédictible, les locuteurs préfèrent utiliser un clitique qui n'apporte rien sur l'information du sujet, afin de réduire la densité informationnelle. En revanche, lorsque le nom en position de sujet est très prédictible, l'ajout d'un clitique conduirait à une densité plus basse, et explique donc que la version sans clitique soit préférable. On peut noter cependant que le taux de redoublement est très élevé en général (75,5 %), ce qui suggère que le sujet clitique pourrait être en voie d'être grammaticalisé comme un marqueur d'accord faisant partie de la flexion du verbe en français informel (Auger, 1995; Culbertson, 2010).

5.2 Limites et perspectives

Bien que nos résultats confirment l'intérêt de l'UID pour le français, il convient de remarquer que l'effet de la densité informationnelle du sujet est faible, son coefficient étant le plus petit parmi les facteurs testés (table 1). De plus, cet effet est sensible aux effets aléatoires. En effet, lorsque le lemme du sujet est exclu de l'analyse, la corrélation entre la surprise et le redoublement devient plus forte dans le modèle mixte (coefficient : 0.18, $p < 0.001$). Il est probable que cela tienne en partie au fait que dans nos données, 60,6 % des noms (lemmatisés) n'apparaissent qu'une fois comme sujet et 91,7 % des noms (lemmatisés) apparaissent moins de 5 fois comme sujet (au total 1 178 différents lemmes du sujet). Pour ces cas-là, il n'est pas surprenant que l'intercept aléatoire du lemme du sujet explique mieux la variance du sujet en ce qui concerne le redoublement et diminue donc la force d'explication de l'effet de la surprise du sujet en général.

Par ailleurs, il s'avère que les résultats sont sensibles à la taille de fenêtre du contexte considéré. En effet, la performance du modèle GPT en termes de prédiction de sujet s'améliore si on accroit la taille du contexte, jusqu'à une exactitude maximale de 28,7% avec le contexte le plus large testé. Toutefois, il a aussi été observé que l'effet de la surprise du sujet diminue progressivement à mesure que le modèle GPT a accès à un contexte plus large lors de l'estimation de la probabilité du sujet. Cela pourrait s'expliquer par le fait que plus le contexte est large, plus le modèle est certain du choix lexical du sujet, ce qui rend la distribution de surprises moins équilibrée. Par ailleurs, Kuribayashi *et al.* (2022) mettent en évidence l'existence d'un écart entre la capacité d'accès au contexte des modèles et des humains, et montrent qu'une limitation du contexte rend le comportement des modèles de langue plus similaires à ceux des humains. Dans une autre étude, Kuribayashi *et al.* (2021) montrent aussi qu'un modèle de langue avec une perplexité plus basse (donc plus performant) reflète mieux le comportement humain pour l'anglais, mais pas pour le japonais. Des investigations supplémentaires sont nécessaires afin d'explorer l'impact de la taille du contexte sur le temps de lecture en français, en vue d'identifier la taille de fenêtre la plus appropriée pour l'estimation de la surprise.

Notre travail met donc en évidence l'intérêt que peut présenter l'utilisation d'un modèle de langue pour estimer le niveau de surprise des mots dans une phrase. Nous pensons que ce type de modèle présente l'intérêt d'être très général, d'incorporer des connaissances linguistiques grâce au pré-apprentissage, et de permettre la prise en compte d'un contexte gauche assez large, et devrait donc être privilégié par

rapport à différentes opérationalisations de l'hypothèse UID qui ont été proposées et qui n'ont pas ces qualités. [4]

Remerciements

Nous remercions Marie Candito pour sa relecture d'une version précédente de ce papier et ses conseils. Nous tenons également à remercier Vera Demberg pour une discussion sur l'analyse statistique, ainsi que Benoît Crabbé pour son aide dans l'analyse en relations de dépendance du corpus étudié. Ces travaux ont bénéficié du financement de l'ERC dans le cadre du programme de recherche et d'innovation Horizon 2020 de l'Union européenne (N°850539).

Références

(2019). MPF. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

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*, p. 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`.

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.

CUSKLEY C., BAILES R. & WALLENBERG J. (2021). Noise resistance in communication : Quantifying uniformity and optimality. *Cognition*, **214**, 104754. DOI : 10.1016/j.cognition.2021.104754.

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*, p. 356–367, Jeju Island, Korea.

4. Nous avons cependant testé, en manière de contrôle, une mesure beaucoup plus grossière pour estimer la surprise du sujet : la fréquence de la tête du sujet (lemmatisé) dans le corpus MPF. En remplaçant la surprise calculée par GPT par cette mesure dans le modèle, nous avons observé une corrélation négative entre la fréquence du sujet et le redoublement, ce qui est de nouveau compatible avec l'UID, car un sujet plus fréquent est plus prédictible et donc moins informatif en général. Ce résultat n'est pas surprenant puisque, comme nous l'avons vérifié, la fréquence du lemme et la probabilité donnée par GPT sont corrélés (test de Pearson : cor : -0.58, $p < 2.2e-16$).

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., Éd. (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.

GOODKIND A. & BICKNELL K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, p. 10–18, Salt Lake City, Utah : Association for Computational Linguistics. DOI : 10.18653/v1/W18-0102.

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*, p. 106–114, Lille, France : ATALA.

JAEGER T. F. (2010). Redundancy and reduction : Speakers manage syntactic information density. *Cognitive Psychology*, p. 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*, p. 161–198.

KURIBAYASHI T., OSEKI Y., BRASSARD A. & INUI K. (2022). Context Limitations Make Neural Language Models More Human-Like. In *EMNLP* : arXiv. http://arxiv.org/abs/2205.11463.

KURIBAYASHI T., OSEKI Y., ITO T., YOSHIDA R., ASAHARA M. & INUI K. (2021). Lower Perplexity is Not Always Human-Like. In *ACL* : arXiv. http://arxiv.org/abs/2106.01229.

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*, p. 131–135, Valencia, Spain : Association for Computational Linguistics.

LEVY R. & JAEGER T. (2007). Speakers optimize information density through syntactic reduction. In B. SCHÖLKOPF, J. PLATT & T. HOFMANN, Éds., *Advances in Neural Information Processing Systems 19 : Proceedings of the 2006 Conference*, volume 19, p. 849–856 : MIT Press.

LIANG Y., AMSILI P. & BURNETT H. (2021). New ways of analyzing complementizer drop in Montréal French : Exploration of cognitive factors. *Language Variation and Change*, **33**(3), 359–385. DOI : 10.1017/S0954394521000223.

MAURITS L., NAVARRO D. & PERFORS A. (2010). Why are some word orders more common than others ? A uniform information density account. In *Advances in Neural Information Processing Systems*, volume 23 : Curran Associates, Inc.

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*, p. 12–22, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.cmcl-1.2.

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.

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*, p. 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. (2019a). Improving Language Understanding by Generative Pre-Training.

RADFORD A., WU J., CHILD R., LUAN D., AMODEI D. & SUTSKEVER I. (2019b). Language Models are Unsupervised Multitask Learners. Open AI Technical Report.

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). In *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, p. 246–255, Lille, France : ATALA.

TEAM R. C. (2022). R : A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.

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*, p. 118–128, London, UK : Association for Computational Linguistics.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. & POLOSUKHIN I. (2017). Attention Is All You Need. In *31st Conference on Neural Information Processing Systems*, Long Beach, CA, USA.

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*. 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.