Application of the UID Hypothesis to Alternation Phenomena in French Corpora

Pascal Amsili

Lattice (CNRS, ENS, Université Sorbonne Nouvelle)

Joint work with Yiming Liang (UGent)

and others

May 28, 2025

`https://tinyurl.com/6yecd8fe`

Abstract

Amsili examines the impact of the Uniform Information Density (UID) hypothesis on various linguistic phenomena in French, such as subject repetition, the alternation between futur simple and futur proche, and the omission of que in Quebec French. This hypothesis posits that a speaker with multiple expressive options tends to choose the one that results in the least variation in information density across words (Jaeger, 2010; Levy & Jaeger, 2007). It was originally proposed to explain, in addition to lexical, semantic, and syntactic variables, the phenomenon of that-omission in English. The hypothesis thus appears to be relevant in certain French contexts, although its application involves significant methodological challenges.

Project

- Corpus studies
- with a quantitative approach
- on phenomena that involve an alternation
- where speakers' choice is conditionned
- by a variety of variables
- among which some have a cognitive origin
- especially those which implement the uniformity of information density hypothesis.

Alternations

(1) a. You must give this man a good book.
b. You must give a good book to this man.

(Bresnan et al. *, 2007)*

(2) a. Luc dit qu'il pense il va partir.
b. Luc dit qu'il pense qu'il va partir.

(3) a. Le porte elle est fermée.
b. La porte est fermée.

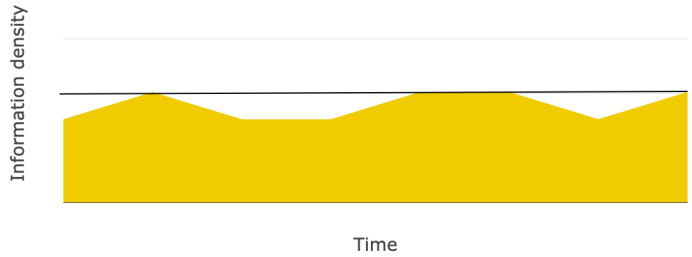
(4) a. La conférence va être organisée à Toulouse.
b. La conférence sera organisée à Toulouse.

(5) a. Je suis sûr que ça ne va pas marcher.
b. Je suis sûr que ça ne marchera pas.

Overview

UID hypothesis (Jaeger, 2010; Levy & Jaeger, 2007)

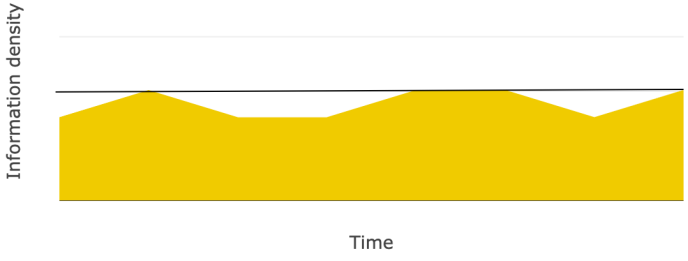
If grammar permits, speakers prefer utterances that distribute information uniformly across a sentence, so as to avoid peaks and troughs in information density.



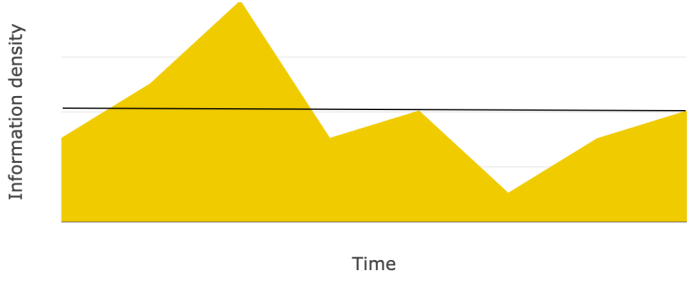
Good distribution.

UID hypothesis (Jaeger, 2010; Levy & Jaeger, 2007)

If grammar permits, speakers prefer utterances that distribute information uniformly across a sentence, so as to avoid peaks and troughs in information density.



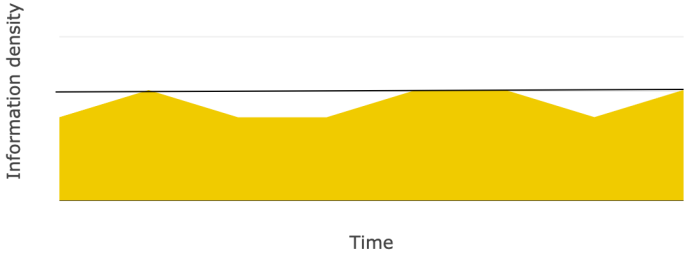
Good distribution.



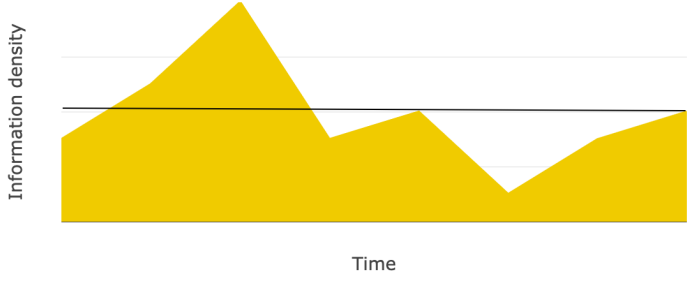
Bad distribution.

UID hypothesis (Jaeger, 2010; Levy & Jaeger, 2007)

If grammar permits, speakers prefer utterances that distribute information uniformly across a sentence, so as to avoid peaks and troughs in information density.



Good distribution.

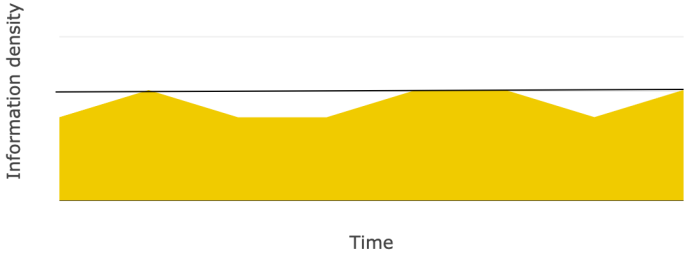


Bad distribution.

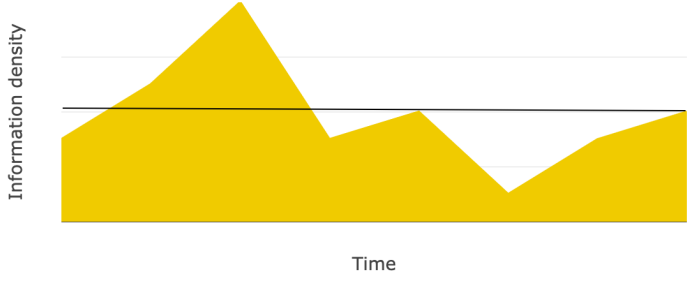
- **Information of a word :** surprisal

UID hypothesis (Jaeger, 2010; Levy & Jaeger, 2007)

If grammar permits, speakers prefer utterances that distribute information uniformly across a sentence, so as to avoid peaks and troughs in information density.



Good distribution.

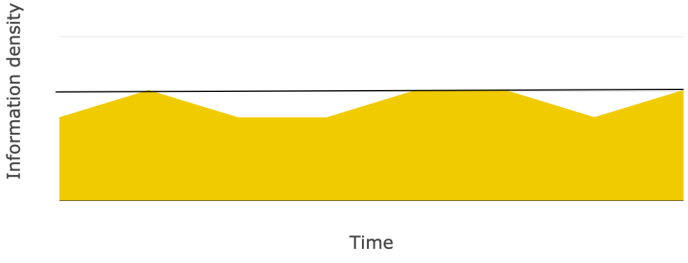


Bad distribution.

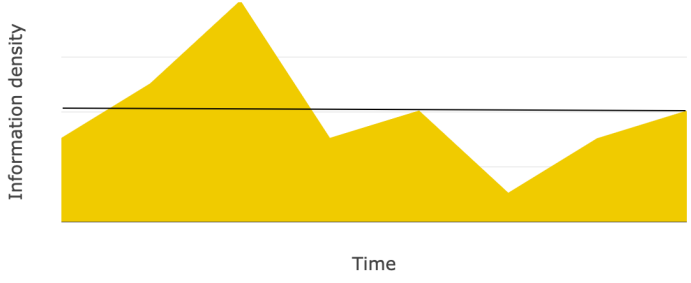
- **Information of a word :** surprisal $I(w_i) = -\log P(w_i|w_{<i})$

UID hypothesis (Jaeger, 2010; Levy & Jaeger, 2007)

If grammar permits, speakers prefer utterances that distribute information uniformly across a sentence, so as to avoid peaks and troughs in information density.



Good distribution.

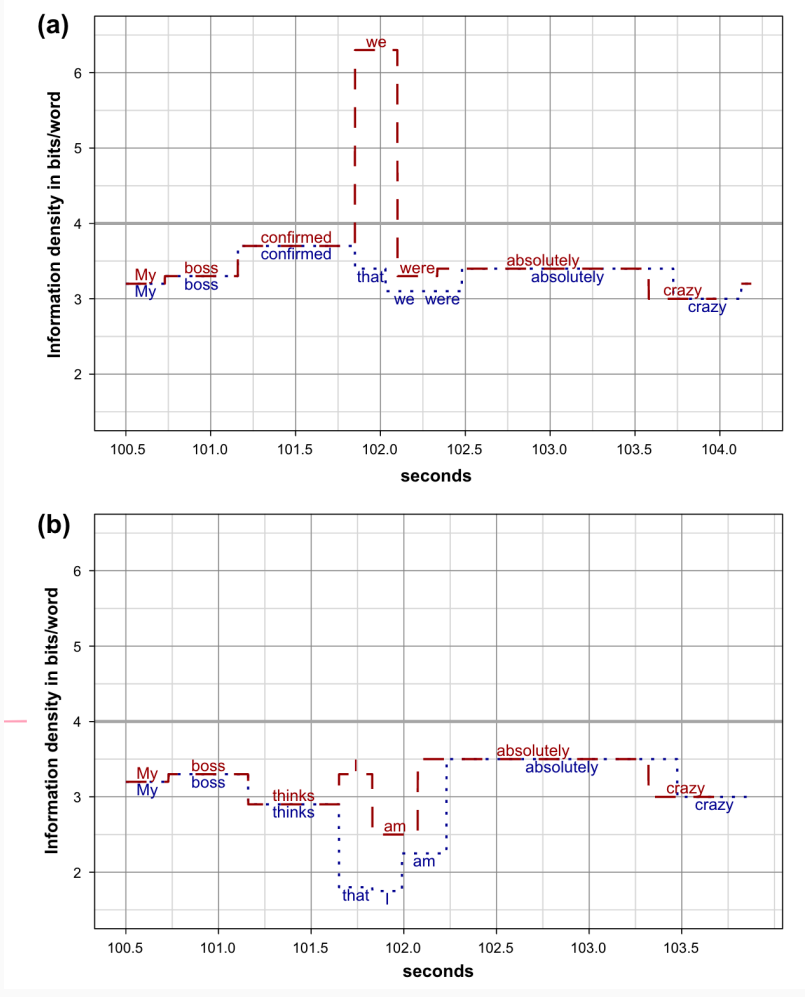


Bad distribution.

- **Information of a word :** surprisal $I(w_i) = -\log P(w_i|w_{<i})$
- **Prediction for syntactic redundancy :**
  - high surprisal $\rightarrow$ lengthened
  - low surprisal $\rightarrow$ shortened or omitted

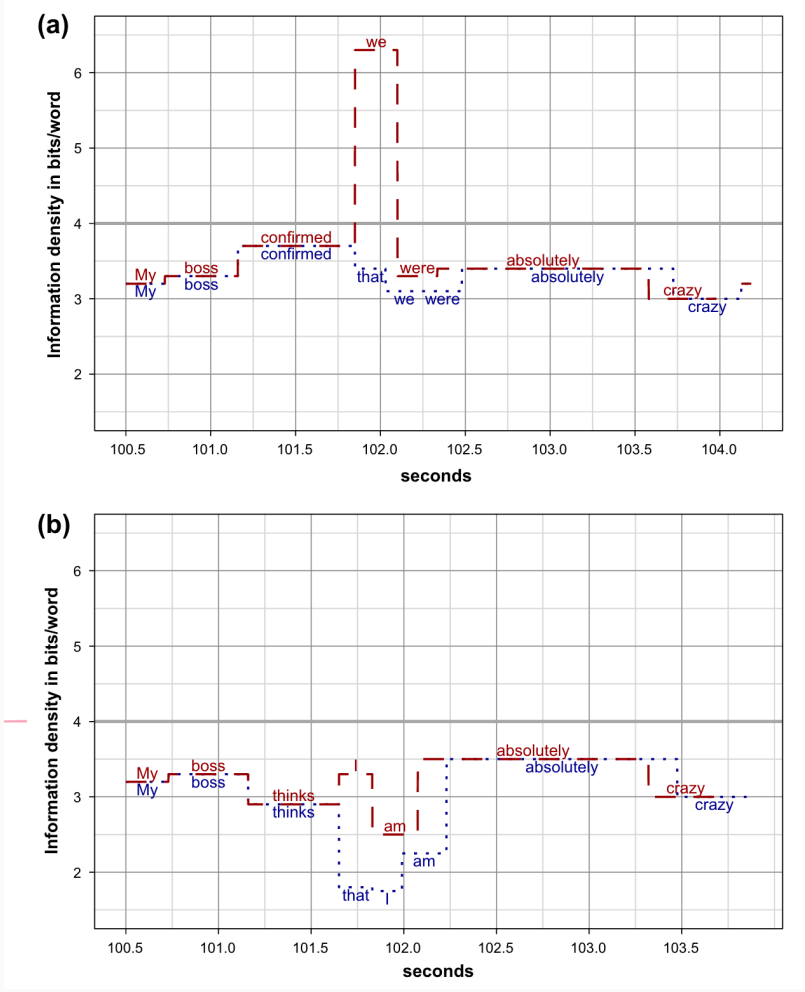
Evidence for UID at syntactic and discourse levels

Complementizer drop (Jaeger, 2010):



Evidence for UID at syntactic and discourse levels

Complementizer drop (Jaeger, 2010):



But also:
article drop in German (Lemke et al., 2017), discourse connector drop in English (Torabi Asr & Demberg, 2015)...

What about subject doubling in French?

Sorbonne Nouvelle

General Methodology

Prediction in an alternation case

- Design a multifactorial model
- Include all potential variables of influence
- Fit the model so as to maximize its accuracy :
  - collect data and compute for each variable the extent to which it contributes to explain the variance ;
  - remove non significant factors
- Implement a proxy to operationalize UID
- Determine how much the model is improved when the UID variable is inserted

Overview

Omission du complémenteur

(1) Français canadien :

a. Faut que je t'apporte quelque-chose.

b. Faut ∅ je t'apporte quelque-chose.

(Corpus du français parlé à Ottawa-Hull, Poplack, 1989)

(2) Anglais :

a. My boss think that you're absolutely right.

b. My boss think ∅ you're absolutely right.

(Corpus Switchboard, Godfrey et al., 1992)

Facteurs phonologiques ou syntaxiques ?

Contexte phonologique :

- sibilante ([s], [z], [ʃ] et [ʒ]) : le plus favorable à l'omission de *que*, pour le contexte précédent et suivant (Sankoff et al., 1971 ; Sankoff, 1980)
- Moins le contexte à droite est sonore, plus on observe d'omission (Martineau, 1985 ; Warren, 1994 ; Dion, 2003).

(3) a. Je pense ∅ c'est bien. ([s], obstruante)

b. Je pense ∅ ma mère est là. ([m], sonante)

c. Je pense ∅ il est bien. ([i], voyelle)

a > b > c

Facteurs phonologiques ou syntaxiques ?

Contexte syntaxique à droite :

- Une complétive dont le sujet est un pronom favorise l'omission alors qu'un syntagme nominal la défavorise. (Connors, 1975)

  (4) a. Je pense ∅ c'est bien. (ce, pronom)

  b. Je pense ∅ ma mère est là. (ma mère, syntagme nominal)

  a > b

- Connors (1975) : L'effet phonologique est un épiphénomène :
  - contextes droits sibilants ≈ sujets pronominaux *je*, *ce* et *ça*
- Cependant, pour Dion (2003), l'effet syntaxique n'est pas indépendant
  - Les pronoms commençant par une voyelle (e.g. *il*, *elle*) défavorisent clairement l'omission.

D'autres facteurs identifiés en français

Facteur lexical (Martineau, 1985 ; Warren, 1994 ; Dion, 2003) :

- Certains verbes (e.g., *rappeler*, *sembler*, *penser*) ou contextes (e.g., *je pense*) favorisent l'omission.
- effet difficile à interpréter

Facteurs sociaux

- Éducation (Warren, 1994) : les locuteurs ayant une éducation post-secondaire mais pas universitaire tendent à omettre *que*
- Niveau socio-économique du locuteur (Dion, 2003) : les ouvriers non-qualifiés omettent plus souvent le complémenteur que les autres

État de l'art en anglais

En plus des facteurs sociaux et linguistiques, les études en anglais ont interrogé des facteurs cognitifs sur le choix de l'omission :

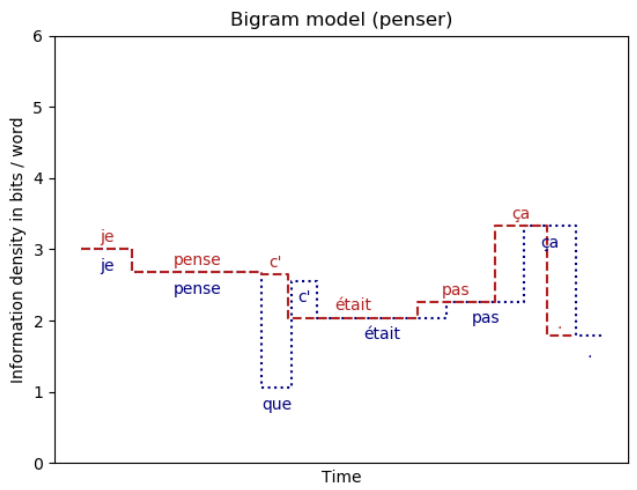
- difficultés de la production : présence d'une pause, disfluence, répétitions (Ferreira & Firato, 2002 ; Jaeger, 2005, parmi d'autres)
- fréquence du verbe de la matrice (Elsness, 1984 ; Roland et al., 2006)
- densité informationnelle → Hypothèse de la densité informationnelle (Jaeger, 2006, 2010 ; Levy & Jaeger, 2007)

La probabilité de la présence d'une complétive = la fréquence de l'apparition d'une complétive étant donné le contexte précédent (≈ le verbe de la matrice) :
e.g., la probabilité de l'apparition d'une complétive après le verbe *penser* :

$$P(completive|penser) = \frac{\#(penser + completive)}{\#penser}$$

Pour l'omission du complémenteur

Prédiction : plus la densité informationnelle au début d'une complétive est basse (≈ plus l'apparition d'une complétive est prédictible), plus on omet "que" qui est peu informatif (i.e. redondant)



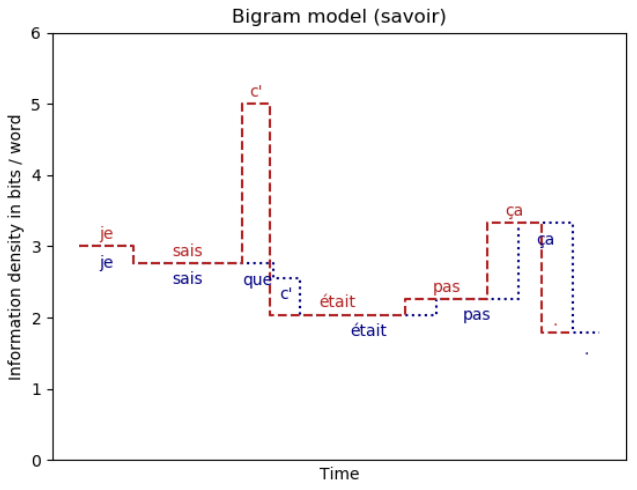


Figure 3: Je pense (que) c'était pas ça. **Figure 4:** Je sais (que) c'était pas ça.

Corpus Montréal 84 (Thibault & Vincent, 1990)

- entretiens traditionnels
- 72 locuteurs de Montréal, sexes et âges bien distribués ; informations sociales documentées
- corpus oral avec 1,6 millions de mots transcrits au total, non annoté
- traitement semi-automatique du corpus :
  - limite aux verbes apparaissant plus de 100 fois dans le corpus
  - 5818 complétives introduites par 17 verbes (e.g., *sembler, penser, croire, trouver*)

Modèle de régression logistique à effets mixtes

Effets fixes :

- Facteurs sociaux :
  - âge, sexe, niveau scolaire, profession
- Facteurs linguistiques :
  - contextes phonologique à gauche et à droite (obstruante, sonante, voyelle)
  - sujets de la matrice et de la complétive (*je_tu*, autre pronom, syntagme nominal)
- Facteurs cognitifs :
  - fréquence du verbe
  - prédictibilité de la présence d'une complétive (opérationnalisée comme la fréquence catégorielle du verbe comme une approximation)

(5) Je pense pas qu'en soixante-et-onze je travaillais là.
(locuteur 2, corpus Montréal 84)

Effet aléatoire : locuteur

Effets sociaux

Éducation ($p<0.01$) : ceux qui ont arrêté leurs études avant l'université $>$ diplômés universitaires

Profession ($p<0.05$) : techniciens, cols bleus et chômeurs $>$ patrons, diplômés universitaires, cols blancs

Le marché linguistique (Bourdieu & Boltanski, 1975) : marché symbolique où ont lieu les échanges linguistiques

- L'usage d'un langage approprié dans une situation donnée permet de gagner plus de profit.
- Le marché linguistique correspondant aux locuteurs ayant une éducation supérieure ou un métier tel que patron et col blancs valorise l'emploi d'un langage standard.
- Ces locuteurs subissent donc plus de pression pour omettre *que*.

Effet phonologique

Contexte phonologique à droite (p<0,001) :

- obstruantes > sonnantes > voyelles (omission dominée par l'échelle de sonorité)
- effet le plus dominant parmi les autres facteurs
- simplification des groupes consonantiques en attaque (Côté, 2004) en cas de violation du *Principe de la Sonorité Séquentielle* (SSP) (Clements, 1990) :

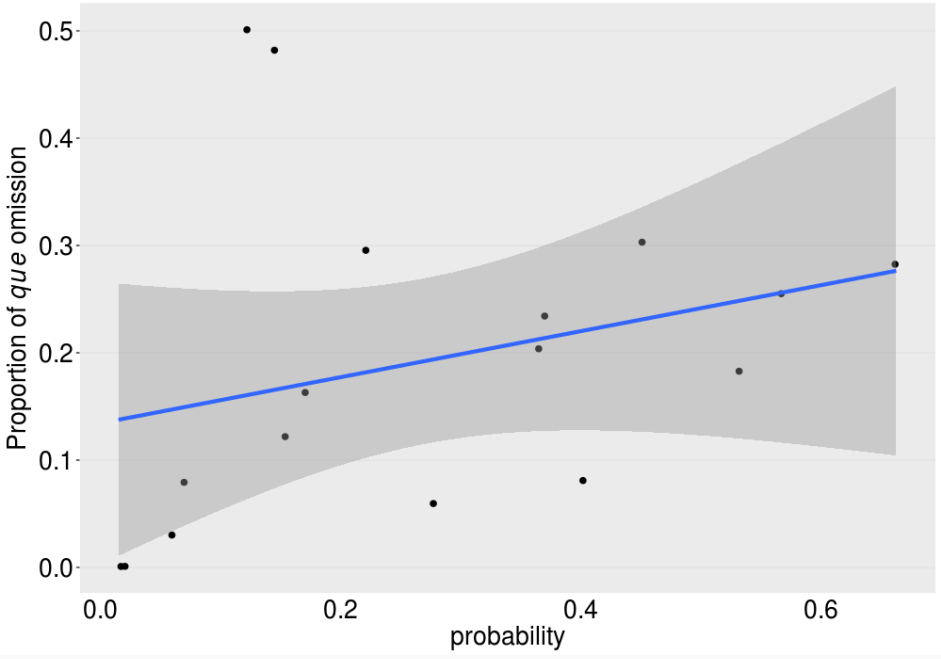
(6) a. Je pense que c'est bien. ([ks] → violation de SSP)

b. Je pense qu'il est bien. ([ki] → pas de violation)

Effet de la prédictibilité

Prédictibilité d'apparition d'une complétive (p<0,001) :

- Plus une complétive est attendue, plus le complémenteur sera susceptible d'être omis.
- conforme aux prédictions de l'hypothèse de la densité informationelle
- Mais l'importance de cet effet s'est classée en troisième.



Effet de la fréquence du verbe

Fréquence du verbe principal ($p<0{,}001$) :

- Un verbe fréquent ($\approx$ accessible) favorise l'omission de *que*.
- potentiellement soutien à l'hypothèse de la densité informationnelle : un verbe fréquent $\rightarrow$ plus attendu $\rightarrow$ moins informatif $\rightarrow$ l'omission du complémenteur

Taking stock

- As was show in Jaeger's seminal paper for *that* in English,
- we show that in Quebec French a similar pressure is playing,
- along with other factors, to explain speakers choices to drop or maintain the complementizer *que*.

	Variable	df	χ^2	*p* value	
1	Premier_phone_cc	2	436.3	2.2e-16	***
2	Fréquence du verbe matrice	1	72.673	5.2e-10	***
3	Probabilité de complétive	1	38.671	2.2e-16	***
4	Sujet_cc	2	22.217	1.498e-05	***
5	Groupe socio-professionnel	5	11.998	0.03481	*
6	Éducation	2	7.272	0.02636	*
7	Dernier_Phone_Avant_CC	2	3.8616	0.145	
8	Sujet_Matrice	2	1.5807	0.4537	
9	Âge	1	0.5779	0.4474	
10	Sexe	1	0	0.9958	

Overview

The phenomenon

Subject doubling: a lexical subject co-occurs with a co-indexed clitic:

(1) a. Marie veut une pomme.

b. Marie$_i$ elle$_i$ veut une pomme.

'Marie$_i$ (she$_i$) wants an apple'

The phenomenon

Subject doubling: a lexical subject co-occurs with a co-indexed clitic:

(2) a. Marie veut une pomme.

b. Marie$_i$ elle$_i$ veut une pomme.

'Marie$_i$ (she$_i$) wants an apple'

Two grammatical accounts:

- **Topicalization** analysis (Kayne, 1975; De Cat, 2005):
  [$_{Top}$ Marie [$_{TP}$ elle [$_{T}$ veut [$_{VP}$ une pomme]]]
  → lexical NP as the **topic**
- **Morphological** analysis (Roberge, 1990; Culbertson, 2010):
  [$_{TP}$ Marie [$_{T}$ elle-veut [$_{VP}$ une pomme]]]
  → lexical NP as the **subject**

Mutifactoral approach

(3) a. Marie veut une pomme.

b. Marie$_i$ elle$_i$ veut une pomme.

'Marie$_i$ (she$_i$) wants an apple'

Multifactoral approach to address linguistic variation:

- NP subject type (Nadasdi, 1995; Auger, 1995): definite > indefinite
- clause type (Auger & Villeneuve, 2010): main > subordinate
- information status: new > old (Barnes, 1985), but old > new (Zahler, 2014)
- However: cognitive factors have been barely studied

Subject doubling: syntactic redundancy
→ may be conditioned by more general accounts, like the *Uniform Information Density* (UID) Hypothesis?

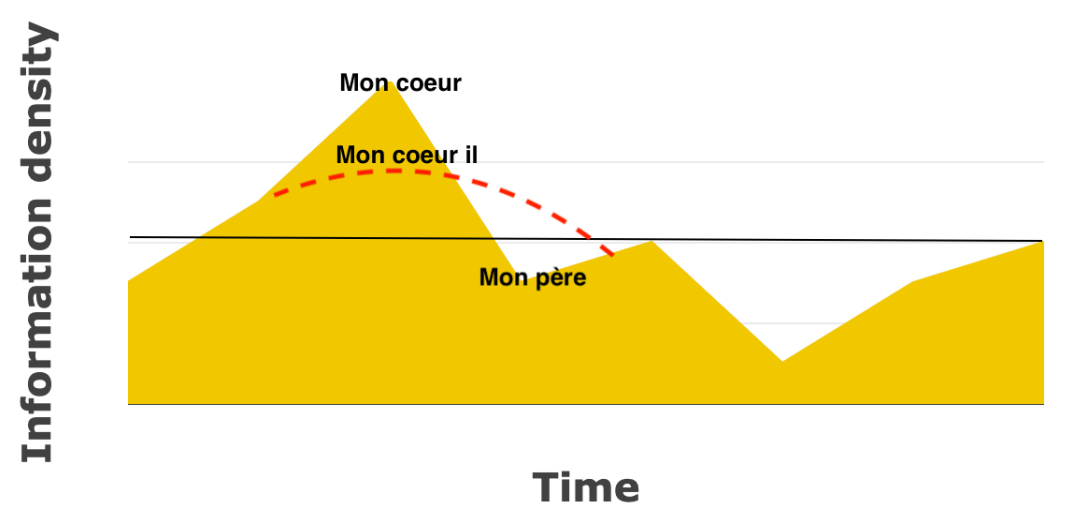
Predictions from UID for Subject doubling

(4) a. **Mon cœur** est blessé. "My heart is hurt."

b. **Mon cœur**$_i$ il$_i$ est blessé. "My heart it is hurt."

→ Subject doubling: syntactic redundancy

Predictions:



Variation context

Corpus: *Multicultural Parisian French* (MPF) (Gadet, 2017)

- Spoken French of young residents of Île-de-France (12-37 y.o.)
- conversations among friends or acquaintances
- 66 interviews, 790,000 words transcribed

Extraction: all tokens containing a **third-person** preverbal nominal subject (n= 4 057)

(5) **Definite NP:** Mon père (il) est parti. 'My father (he) left.'

(6) **Indefinite NP:** Un garçon (il) est venu. 'A boy (he) came.'

(7) **Universal NP:** Tout le monde (il) est là. 'Everyone (he) is here.'

Case study 4: Subject doubling as an optimal encoding strategy?

(25) **Marie** mange. "Marie eats"

(26) **Marie**$_i$ elle$_i$ mange. "Marie$_i$ she$_i$ eats."

Hypothesis: doubling subject to lower down information density

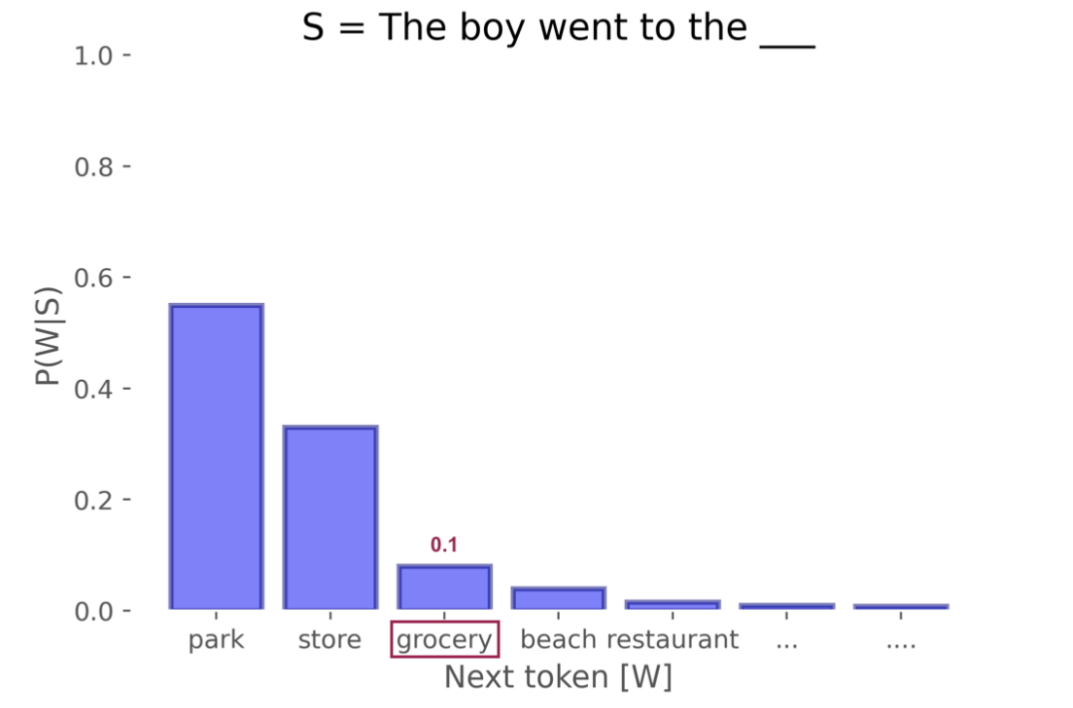
- doubling: DP subject with high surprisal
- non-doubling: DP subject with low surprisal

$$\begin{aligned} Surprisal(DP\ subject) &= -logP(DP\ subject|preceding\ context) \\ &\approx -logP(head|preceding\ context) \end{aligned} \quad (2)$$

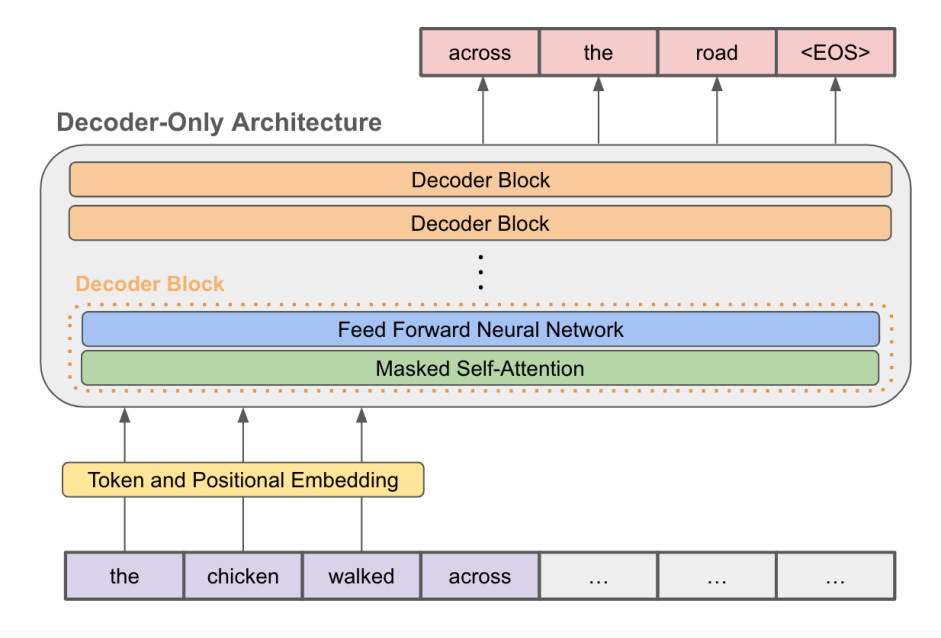
How to assess this probability?

Language model

Language model: probability distribution over words or word sequences



Generative Pretrained Transformer (GPT)



- incremental processing
- world knowledge and general linguistic competence
- incorporating broad contexts
- capture human reading time (Merkx & Frank, 2021; Wilcox et al., 2020) → more human-like

Use a French GPT model

French GPT (Simoulin & Crabbé, 2021):

- GPT-2 model with 124M parameters
- mainly trained on texts → less adapted to oral data

Fine-tuning:

- 90% of CFPP2000 (Branca-Rosoff et al., 2012) ≈ 650,000 tokens

	CFPP2000 (10%)	MPF (50%)
GPTfr-124M original	42.25	42.61
GPTfr-124M finetuned	**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.

→ better speaker of oral Parisian French

Estimation of NP subject's surprisal

(8) Ben parce qu'ils pensaient que **la vie** ... "Well, because they think that **the life**..."

$\Rightarrow Surprisal(NP\ subject) = -logP(NP\ subject|preceding\ context)$

Prediction of the upcoming subject: relies on a longer context → incorporating previous speaking turns into the context

Context ($t_0 ... t_{i-1}$)		NP subject (t_i)	$P(t_i\|\text{context})$	$-\log P(t_i\|\text{context})$
Previous turn	*Target*			
Pourquoi ils sont partis tes parents?	Ben parce qu'ils pensaient que	**la vie**	0.00013	8.93

Statistical modelling

Mixed-effect logistic model (n= 4 057, doubling = 1, non-doubling = 0)

Predictors:

- **polarity:** affirmative, neg. with *ne*, neg. w/o *ne* (Zahler, 2014)
- **subject type:** universal, indefinite, definite (Nadasdi, 1995)
- **clause type:** main, other subordinate, relatives (Zahler, 2014)
- **verb frequency** in the corpus (log-transformed)
- **distance** in words between subject head and verb (Zahler, 2014)
- **NP subject length**
- **NP surprisal**: estimated by French GPT

Random effects: speaker, verb lemma

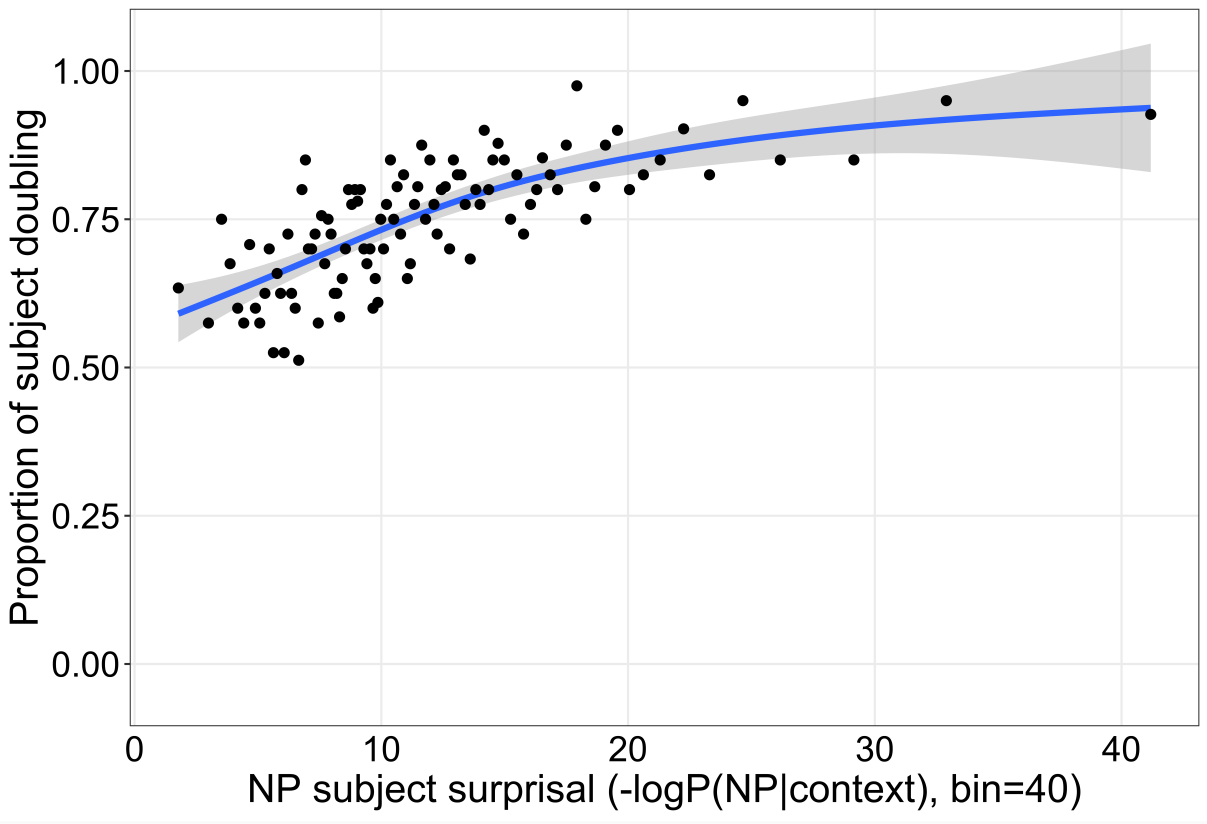
Overall rate of doubling: 74.8%

Accuracy of the statistical model: 86.7%

Results

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 (w/o *ne* vs. *ne*)	6.215579	1.066526	5.828	5.61e-09	***
subject (indef. vs. uni.)	2.289983	0.402894	5.684	1.32e-08	***
subject (def. vs. indef.)	2.436020	0.330875	7.362	1.81e-13	***
clause (sub. vs. rel,)	1.522349	0.302683	5.030	4.92e-07	***
clause (main vs. sub.)	0.961943	0.112814	8.527	$<$ 2e-16	***
verb frequency	0.270007	0.079023	3.417	0.000634	***
distance	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	***

Effect of NP subject surprisal



Subject doubling is conditioned by UID:

NP subjects with higher surprisal → more doubling ($p < 0.001$)

Discussion: UID explains subject doubling

- **Robustness** of the effect: estimate NP surprisal using different lengths for the preceding context (1-11 speech turns) → all significant
- We also found evidence for both grammatical analyses:
  - Topicalization analysis (Kayne, 1975; De Cat, 2005): NP subject type, clause type
  - Morphological analysis (Roberge, 1990; Culbertson, 2010): high doubling rate (75%)
  - But these grammatical accounts are very specific, only applied to doubling phenomena
- UID provides a more general cognitive/communication explanation, applicable to other types of linguistic phenomena
- Factors related to information structure, can be covered by information theory (Komagata, 2003; Lemke, 2021) → future studies for subject doubling

Overview

Case study 5: FTR as an optimal encoding strategy?

(27) a. Je mangerai. (Synthetic Future)

b. Je vais manger. (Periphrastic Future)

"I will eat."

Hypothesis: Synthetic future as a less redundant encoding strategy

- Synthetic future: used with future events with low surprisal
- Periphrastic future: used with future events with high surprisal

$$\begin{aligned} Surprisal(future\ event) &= -logP(future\ event|preceding\ context) \\ &\approx -logP(verb|preceding\ context) \end{aligned} \tag{5}$$

Incorporate event surprisal in the statistical model

Baseline model:

$$\begin{aligned} &\textit{Usage of SF constrasted with PF is dependent on:} \\ &\textit{sentential polarity} + \textit{subject type} \\ &+ \textit{verb type} + \textit{verb irregularity} \\ &+ (1 \mid \textit{speaker}) \end{aligned} \tag{6}$$

Model with event surprisal:

$$\text{Baseline model} + \textit{event surprisal} \tag{7}$$

Incorporate event surprisal in the statistical model

Baseline model:

$$\begin{aligned} &\textit{Usage of SF constrasted with PF is dependent on:} \\ &\textit{sentential polarity} + \textit{subject type} \\ &+ \textit{verb type} + \textit{verb irregularity} \\ &+ (1 \mid \textit{speaker}) \end{aligned} \tag{6}$$

Model with event surprisal:

$$\text{Baseline model} + \textit{event surprisal} \tag{7}$$

$\rightarrow \chi^2 = 0.0261$, $p > 0.05$

$\rightarrow$ failed to reveal an effect of event surprisal

Possible explanation

Future temporal reference:

1. not redundancy
2. but word order variation
   - Synthetic future mangerai: verb + future suffix
   - Periphrastic future vais manger: future + verb
3. other processing accounts more adequate

Overview

Conclusion

- It is well known that variation in language is a multifactorial phenomenon
- It has also been shown that variables conditioning variation in general can belong to many linguistic and social realms,
- Cognitive factors have been suggested to play also a role,
- among which the UID hypothesis can be shown to have an impact,
- however not necessarily on all variation phenomena

Conclusion

- It is well known that variation in language is a multifactorial phenomenon
- It has also been shown that variables conditioning variation in general can belong to many linguistic and social realms,
- Cognitive factors have been suggested to play also a role,
- among which the UID hypothesis can be shown to have an impact,
- however not necessarily on all variation phenomena

Perspectives

- Operationalisation of UID : local *vs.* local approaches
- Better collection and curation of oral, spontanuous, corpora
- Interaction with other cognitive factors

Remerciements

Mes remerciements les plus chaleureux à
Yiming Liang, Heather Burnett, Vera Demberg,
Aurélien Said-Housseini...



The work was funded by

- Heather Burnett's ERC SMIC project (grant agreement n°850539),
- Labex EFL (ANR-10-LABX-0083),
- the Deutsche Forschungsgemeinschaft (232722074-SFB1102).

References

Bresnan, Joan, Cueni, Anna, Nikitina, Tatiana, Baayen, R Harald, *et al.* . 2007. Predicting the dative alternation. *Cognitive foundations of interpretation*, 69–94.

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

Levy, Roger, & Jaeger, T. Florian. 2007. Speakers optimize information density through syntactic reduction. *In :* Schölkopf, Bernard, Platt, John, & Hoffman, Thomas (eds), *Advances in Neural Information Processing Systems 19 : Proceedings of the 2006 Conference*, vol. 19. MIT Press.

Liang, Yiming, Amsili, Pascal, & Burnett, Heather. 2019 (Oct.). Faut je t'apporte quelque chose : Complementizer drop in Quebec French. *In : Poster presented at CSSP 2019 (Colloque de Syntaxe et Sémantique à Paris)*.

Liang, Yiming, Amsili, Pascal, & Burnett, Heather. 2021a (Sept.). Exploring the polarity effect on future variation in Parisian French. *In : Paper presented at the 13th biennal conference "UK Language Variation and Change"*. University of Glasgow.

Liang, Yiming, Amsili, Pascal, & Burnett, Heather. 2021b. New ways of analyzing complementizer drop in Montréal French : Exploration of cognitive factors. *Language Variation and Change*, **33**(3), 359–385.

Liang, Yiming, Amsili, Pascal, & Burnett, Heather. 2021c (June). Revisiter l'omission du complémenteur "que" en français montréalais : exploration de facteurs cognitifs. *In : Communication présentée au congrès de l'Association Canadienne de Linguistique*.

Liang, Yiming, Amsili, Pascal, & Burnett, Heather. 2022 (Sept.). Revisiter l'omission du complémenteur que en français montréalais : exploration de facteurs cognitifs. *In :* Blondeau, Hélène, Laforest, Marty, & Remysen, Wim (eds), *Colloque international «1971-2021 : 50 ans de corpus montréalais». Résumés des communications*.

Liang, Yiming, Amsili, Pascal, & Burnett, Heather. 2023. Uniformité de la densité informationnelle : le cas du redoublement du sujet. *Pages 164–176 of :* Servan, Christophe, & Vilnat, Anne (eds), *Actes de CORIA-TALN 2023*. Paris, France : ATALA.

Liang, Yiming, Amsili, Pascal, & Burnett, Heather. 2024a. The Polarity Effect in Future Temporal Reference : new insights from Parisian French. *Revue canadienne de linguistique/Canadian Journal of Linguistics*. Squib accepté pour publication.

Liang, Yiming, Amsili, Pascal, Burnett, Heather, & Demberg, Vera. 2024b (july). Uniform information density explains subject doubling in French. *In : Proceeding of The 46th Annual Meeting of the Cognitive Science Society*.

Future temporal reference

Future events may be expressed by using a Synthetic Future (SF) or a Periphrastic Future (PF)[1]:

(7) a. Je le dir-ai. (Synthetic future or *futur simple*)

b. Je vais le dire. (Periphrastic future or *futur proche*)

"I will say it."

[1] We will set aside the Futurate Present for this study.

Polarity: strong but puzzling

Laurentian French (Poplack & Dion, 2009; Wagner & Sankoff, 2011):

- affirmative contexts: periphrastic form $>$ synthetic form
- negative contexts: synthetic form ($\geqslant$ 97%)

Metropolitan French (except Vimeu French) (Roberts, 2012; Tristram, 2020):

- variability in negative context, with a small preference for synthetic forms

Vimeu French (Villeneuve & Comeau, 2016):

- no effect, but reliable effect of temporal distance

$\rightarrow$ **Parisian French??**

(8) a. Je vais le dire. affirmative, PF

b. Je le dirai pas. negative, SF
"I will say it/I will not say it."

Corpus

Multicultural Paris French (MPF) (Gadet & Guerin, 2016; Gadet, 2017)

- young residents of Île-de-France (12-37 y.o.)
- with a multicultural background
- 66 interviews, around 790 000 words transcribed
- conversations among friends or acquaintances

Statistical analysis

Mixed-effect logistic regression in R (n=3,622)

Fixed effects:

- 4 social factors: sex, age, education level, profession
- 4 linguistic factors:
  - Sentential polarity: **affirmative**, **negation with *pas***, **negation with other negative words** (e.g., *jamais* "never", *plus* "no longer", *rien* "nothing"...)
  - Verb's frequency
  - verb's type (regular or irregular)
  - subject type

Random effects:

- speaker (n=93), verb lemma (n=428)

Results

Mixed-effect logistic regression in R (n=3,622)

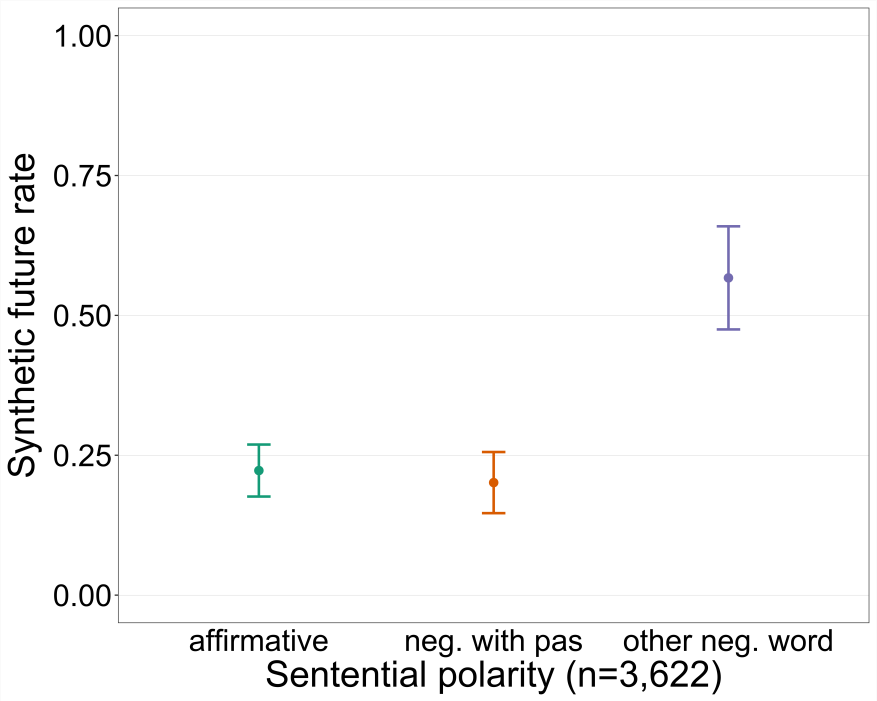
Fixed effects:

- 4 social factors: sex, age, education level, profession
- 4 linguistic factors:
  - sentential polarity: **affirmative**, **negation with *pas***, **negation with other negative words** (e.g., *jamais* "never", *plus* "no longer", *rien* "nothing"...)
  - verb's frequency
  - verb's type (regular or irregular)
  - subject type

Random effects:

- speaker (n=93), verb lemma (n=449)

Polarity effect



(9) a. Tu verras. "You will see"

b. Tu verras pas. "You will not see"

c. Non mais tu verras rien. "No but you will see nothing"

Interim summary

Laurentian French (Poplack & Dion, 2009; Wagner & Sankoff, 2011):
SF used nearly categorically in all negative contexts ($\geqslant$ 97%)

Parisian French:

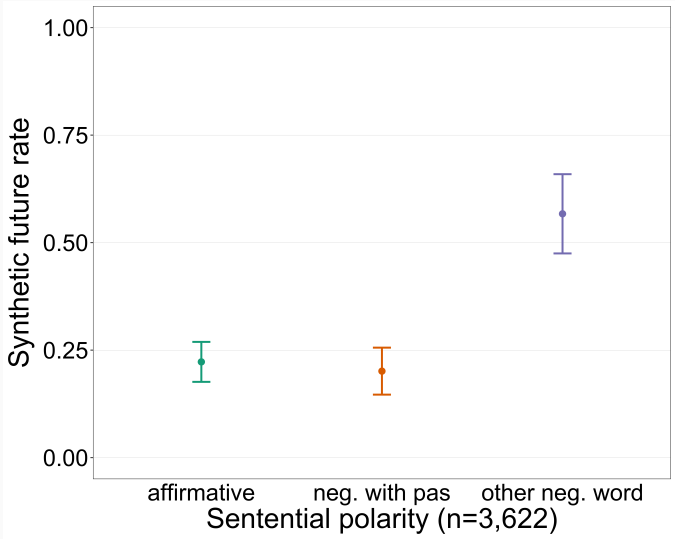


Figure 1: MPF corpus.

Interim summary

Laurentian French (Poplack & Dion, 2009; Wagner & Sankoff, 2011):
SF used nearly categorically in all negative contexts ($\geqslant$ 97%)

Parisian French:

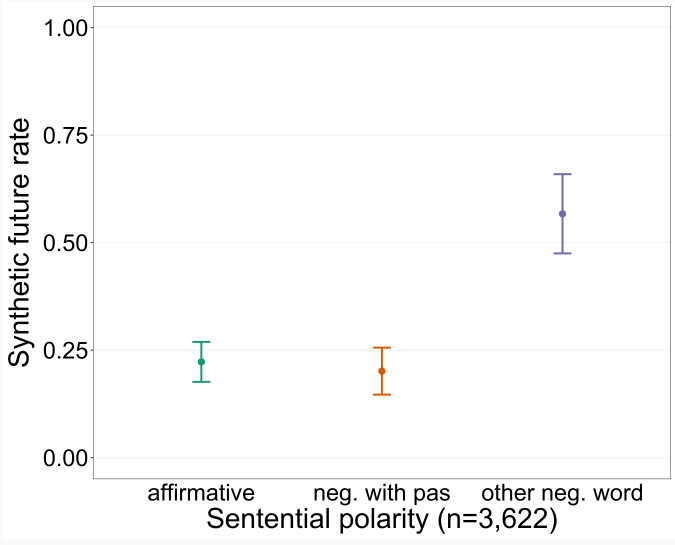


Figure 1: MPF corpus.

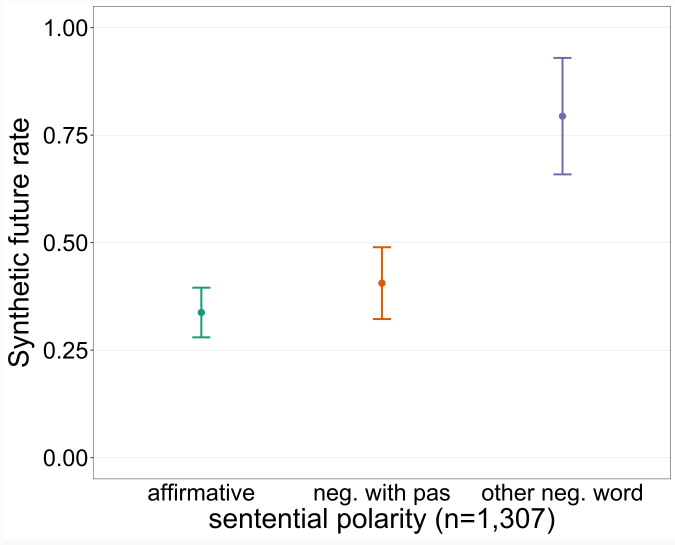


Figure 2: CFPP2000 corpus.

Interim summary

Laurentian French (Poplack & Dion, 2009; Wagner & Sankoff, 2011):
SF used nearly categorically in all negative contexts ($\geqslant$ 97%)

Parisian French:

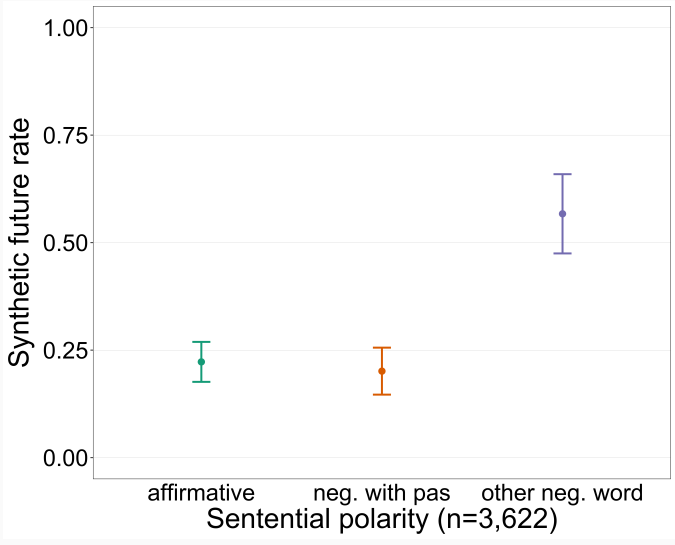


Figure 1: MPF corpus.

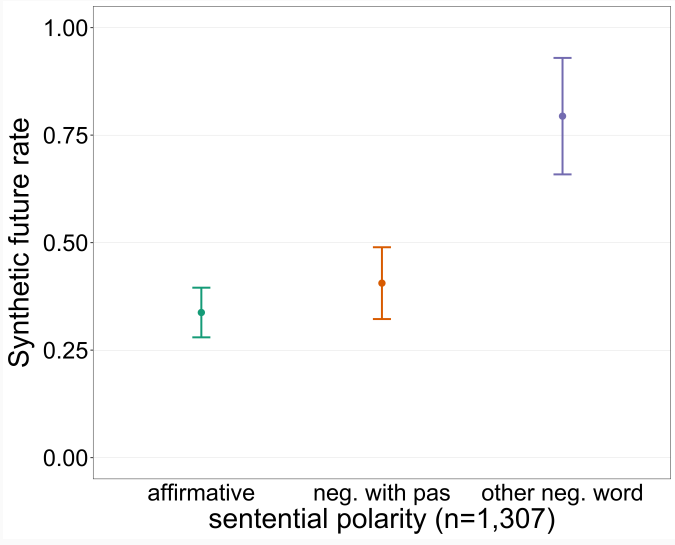


Figure 2: CFPP2000 corpus.

$\rightarrow$ Explanation of polarity effect??