

# **CLASSIFICATION AUTOMATIQUE DE QUESTIONS SPONTANÉES VS. PRÉPARÉES DANS DES TRANSCRIPTIONS DE L'ORAL**

Iris Eshkol-Taravella, Angèle Barbedette, Xingyu Liu, Valentin-Gabriel Soumah




# INTRODUCTION

La communication avec les machines est essentielle dans de nombreuses applications de Traitement Automatique du Langage (TAL), comme les dialogues homme-machine et les chatbots. L'un des principaux défis de ces applications est de rendre les réponses générées par les machines plus proches du langage humain.

Pour améliorer la qualité et l'expressivité des systèmes de synthèse de la parole, des techniques telles que l'ajout automatique de disfluences et de pauses remplies ont été proposées. Cependant, ces éléments ne suffisent pas à imiter parfaitement le discours humain spontané.


La recherche abordée dans cet article se penche sur la distinction entre le discours spontané et préparé en utilisant des indices linguistiques, dans le but d'améliorer la performance de ces systèmes de communication homme-machine.





Données : Les transcriptions d'enregistrements provenant des corpus ESLO2(enquête sociolinguistique à Orléans 2) et ACSYNT(un corpus oral du français contemporain)

Objectif : Développer un modèle linguistique pour classer automatiquement les questions en “spontané” ou “préparé” et évaluer les performances de différentes méthodes



# QUESTIONS

## SPONTANÉ

Les questions spontanées sont caractéristiques de l'oralité, elles sont construites en direct au cours de la conversation. Parmi les différents marqueurs des questions spontanées, on retrouve les disfluences (hésitations, répétitions de parties de mots, insertion de mots non significatifs à la conversation) une influence du contexte immédiat (la question spontanée peut par exemple servir à obtenir des précisions ou des informations supplémentaires)

## PRÉPARÉ

Les questions préparées sont en général planifiées à l'avance, en amont de la production du discours. Par exemple, dans un entretien sociologique, les questions préparées sont écrites sur une trame et auront donc souvent une syntaxe et une formulation plus claires que les questions spontanées. Pour ce qui est de leurs caractéristiques récurrentes, elles présentent en général moins de disfluences, une répétition d'entités nommées, un mot interrogatif en position initiale. Elles peuvent aussi être annoncées ("une dernière question")

## NON-QUESTIONS

Les énoncés se terminant par un point d'interrogation qui ne relèvent pas de la recherche d'information. Les non-questions peuvent être par exemple des demandes de clarification, de répétition, d'action, elles servent parfois au bon déroulement de la conversation, par exemple pour vérifier que l'interlocuteur écoute et comprend ce qui lui est dit. Elles peuvent aussi être injonctives (demande d'action) ou servir à contrôler le dialogue.

# GUIDE D'ANNOTATION (NON-QUESTION)

Tag questions : un moyen de vérifier que l'interlocuteur a compris et accepté ce qui est dit

- You're coming to the party, aren't you?
- That was a great movie, wasn't it?
- You like chocolate, don't you?

Demandes de répétition : de forme interrogative, elles sont associées à des demandes de Clarification

- What did you say? I didn't catch that.
- Could you repeat that for me, please?
- Sorry, can you say that again?

Vérifications de compréhension : Ce sont des demandes de clarification d'après

- So, you're saying we should meet at 3 PM, right?
- If I understand correctly, you want me to email the report by tomorrow?
- Just to clarify, you mean we should start at 9 AM?

Questions rhétoriques : malgré une syntaxe semblable à celle d'une question, elles ne nécessitent pas de réponse et ne correspondent pas à une recherche d'information

- Can you believe this weather?
- Who could resist such a delicious meal?
- Do I look like I care?

# GUIDE D'ANNOTATION (NON-QUESTION)

Questions injonctives : sont des demandes d' action

- Can you pass the salt?
- Could you lend me a hand with this?
- Help me with these groceries, will you?

Questions d' obligation sociale : ces actes correspondent à la gestion des obligations sociales, qui sont des actes de contrôle du dialogue.

- Should I RSVP to the event?
- should I send them a thank-you card.
- Do you think I ought to attend the meeting?

Questions non annotables : questions coupées, questions trop larges, discours rapporté ou questions non interprétables

- Why did you... What's the point?
- Can you pass me the... Never mind.
- How could you... I don't understand.



# GUIDE D'ANNOTATION (NON-QUESTION)

Questions ouvertes : contiennent un mot interrogatifs et dont la réponse n' est pas "oui", "si" ou "non"

- What is your favorite book?
- How do you feel about the new policy?
- Where did you spend your last vacation?

Questions fermées : supposent une réponse par "oui", "si" ou "non"

- Did you finish your homework?
- Is the meeting at 3 PM?
- Have you ever been to Paris?

Questions alternatives ou alt-questions : les réponses possibles attendues sont contenues dans la question posée.

- Do you prefer coffee or tea?
- Is your favorite color blue or green?
- Would you like to go to the movies or stay in tonight?

# GUIDE D'ANNOTATION (SPONTANÉ)



## Anaphores

- Speaker 1: "In my current role, I manage divisional controls."
- Speaker 4: "And what does that entail?"

## Demandes de clarification ou d'informations supplémentaires

- Speaker 1: "I was thinking, wouldn't it be interesting if children could learn foreign languages from a very young age?"
- Speaker 4: "Why do you believe that's important?"

## Thème précédant le rhème

- Speaker 1: "So, you've been in London for quite a while. How long has it been now?"

## Mot interrogatif en position finale

- Speaker 3: "This team seems impressive, but how is it constituted?"
- Speaker 2: "You've mentioned a new initiative. What inspired you to start it?"

## Disfluences

- Speaker 1: "He's my super... um, superior. How does the hierarchy work in your workplace?"
- Speaker 4: "Could you, uh, provide more details about your role?"



# GUIDE D'ANNOTATION (PRÉPARÉ)

## Répétition d' entité nommée

- Speaker 1: "How many years have you been practicing yoga?"
- Speaker 3: "Oh, it's been a decade since two thousand and ten."
- Speaker 1: "Do you find it beneficial for your well-being?"

## Rupture d' isotopie

- Speaker 2: "You mentioned your passion for hiking. Do you have a favorite trail?"
- Speaker 4: "Yes, I love exploring new paths in the mountains."
- Speaker 2: "Can you share an unforgettable hiking experience?"

## Rhème précédant le thème

- Speaker 1: "Tell me, what motivated you to pursue a career in journalism?"
- Speaker 4: "I find the power of storytelling truly inspiring"

## Mot interrogatif en position initiale

- Speaker 2: "How do you manage stress in your daily life?"
- Speaker 1: "Any particular strategies or routines you follow?"

## Annonce de la question

- Speaker 4: "Now, I'd like to delve into some questions that might touch on more personal aspects of your life."

## APPRENTISSAGE SUPERVISÉ DE CES CATÉGORIES



- Q Les expériences de classification automatique sont réalisées avec deux classifications, une multiclasse avec les trois catégories “spontané” “préparé” et “non-question” ainsi qu’une classification binaire des questions “spontané” et “préparé”.
- Q Les données (questions avec leurs annotations) sont divisées en deux ensembles, un ensemble d’entraînement (75% des données) ainsi qu’un ensemble de test (25% des données), cela permet de tester le modèle sur un ensemble indépendant pour évaluer sa performance.
- Q Les questions ont été représentées en utilisant une vectorisation du corpus basée sur le modèle CBOW (200 dimensions) du corpus FrWaC, avec une normalisation utilisant des poids TF-IDF pour tenir compte de l’importance, de la rareté et de la fonction discriminante des mots du corpus. Le modèle Skip-Gram a également été testé, mais il a généralement donné des résultats inférieurs.

## APPRENTISSAGE SUPERVISÉ DE CES CATÉGORIES

- Q Pour la classification, plusieurs critères linguistiques ont été sélectionnés, tels que la longueur de la question, la présence de disfluences, l'annonce d'une question, l'inversion du sujet, la répétition d'une entité nommée, la distance vectorielle avec des vecteurs Word2Vec, la position du mot interrogatif, la présence de la forme interrogative "est-ce que" et du marqueur de l'inversion du sujet "t-il".
- Q Les résultats montrent que les meilleures performances ont été obtenues avec un score de 0,74 pour la classification binaire et un score de 0,66 pour la classification multiclasse, en utilisant une régression logistique et des critères linguistiques uniquement. La sélection des critères pertinents pour la classification s'est basée sur des critères linguistiques mentionnés et des observations du corpus.
- Q En résumé, Les résultats montrent en général de meilleures performances lorsqu'on applique une classification binaire, surtout pour les questions spontanées, en raison de similitudes syntaxiques avec les non-questions. De plus, l'étude a confirmé l'importance des disfluences et de la position du mot interrogatif pour distinguer les catégories de questions.

## CONSTITUTION D'UN CORPUS ANGLOPHONE SIMILAIRE

Problème: difficulté de trouver un corpus similaire

- Transcription de l'oral
- Oral de type “interview” avec des questions spontanées et préparées
- Disponible librement sur internet

Premier essai avec des podcasts, émission radio et télévision

Problème: les formats des transcriptions diffèrent, absence d'information importantes pour l'annotation (qui parle)

<https://github.com/chasche/annotation-question-spon-prep>

# CORPUS ANGLOPHONE SIMILAIRE

## Family Support of Third-Grade Reading Skills, Motivation, and Habits

openICPSR (Inter-university Consortium for Political and Social Research,  
une archive de données de sciences sociales)

DOI: 10.1177/2332858417714457 • Corpus ID: 149415011

### Family Support of Third-Grade Reading Skills, Motivation, and Habits

[Lauren Capotosto](#), [James S. Kim](#), [Mary A. Burkhauser](#), [Soojin Oh Park](#), [B. Mulimbi](#), [Maleka Donaldson](#), [H. Kingston](#) [less](#) •

Published 1 June 2017 • Psychology, Education • AERA Open

This qualitative study investigated the ways in which 84 parents from predominantly low-income communities described supporting their third graders' reading skills, motivation, and habits. Thematic analysis of open-ended parent interviews indicated that parents actively and deliberately scaffolded their children's progress toward developing independent reading skills. Parents explicitly communicated the value of reading in everyday conversations; actively listened to their children read, even if they did not understand the language in which the text was written; asked reading comprehension questions; created a home environment conducive to sustained reading; promoted reader autonomy through encouragement of strategy use; and incorporated reading practices into daily routines. Parents often described their own efforts as responsive to their children's level of reading motivation and reading performance, thus highlighting the reciprocal nature of parent-child reading interactions. Findings reveal a variety of ways in which families support their children's reading skills, motivation, and habits. [Collapse](#)





# DIFFÉRENCES ENTRE CORPUS

- thème et trame de questions plus spécifiques
- différence de taille (84 interview)
- corpus final: 200 questions annotées (contre environ 1900)

## Interview protocol

We used a qualitative interview design to address our research question. Open-ended interviews are particularly useful when investigating under-explored topics ([Johnson & Onwiegbufie, 2004](#)), such as the role parents play in supporting their children's reading development in middle childhood. Our interview protocol consisted of six open-ended questions asked of all families:

1. Tell me about a typical day for your child from morning to bedtime as well as you can remember.
2. What are some things that you do to help your child become a good reader?
3. What, if anything, do you do when your child has a hard time with a book?
4. What are some questions that you ask your child when you talk about books that s/he has read?
5. What, if anything, do you do to motivate your child to read?
6. Where does your child get most of his or her books from?

## THEME LANGUAGE

- S51 : Est-ce qu'il y a une façon de parler propre à Orléans ?
- S52 : Est-ce qu'on parle bien à Orléans ?
- S53 : Langue des jeunes.
- S54 : Est-ce qu'il y a des choses qui vous agacent ou qui vous amusent dans la façon de parler de certaines personnes ?
- S55 : Pratiques plurilingues (les autres langues entendues par le témoin dans Orléans).
- S56 : Question sur les autres langues parlées par le témoin (et/ou entourage)
- Si oui questions sur les pratiques, transmission,...

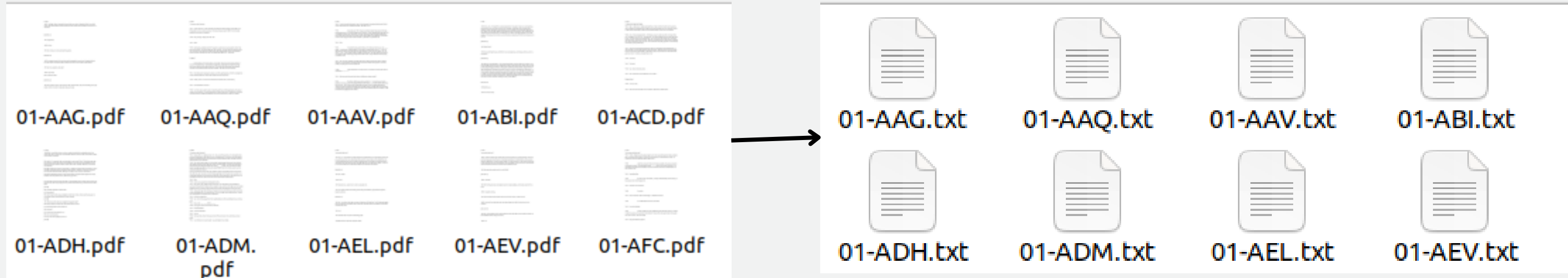
## EXTRA

- S61 : La recette de l'omelette.



# CONSTITUTION D'UN CORPUS ANGLOPHONE SIMILAIRE

Utilisation de PyPDF pour convertir le corpus en format txt et fusionner les fichier txt entre eux



12194 improved, that he was in G level and now he had risen to level M.  
12195 HV1: What are those levels?  
12196 PAR: Like in the alphabet, I guess it is.  
12197 HV1: Ah, okay.  
12198 PAR: Like, G... E, F, G, H, I, J, K, M. So he increased 9 levels from what he had.  
12199 HV1: And that was... This last month, you say.  
12200 PAR: Uh-huh. Around a month and a half. He was in that level. G. And he passed to level M. He has  
12201 improved quite a bit in reading.  
12202 HV1: What goals do you have for Child J's reading this summer?  
12203 PAR: To be honest... Reading is very, very important, something that I hadn't paid enough attention to  
12204 before, and I want him to improve his level. He has to be on M level, that is his grade. I want him to be  
12205 on his right level, not behind on the reading level.  
12206 HV1: Okay. This ended the second part of the conversation.  
12207

# EXCTRACTION DES QUESTIONS

```
pattern = r'(HV1|PAR)(.*?)\?'
```

limite: Les énoncés qui relèvent de la recherche d’information mais ne se terminent pas par un ? ne sont pas extraits.

→ 4 Please tell me what Child S does from the moment she wakes up until she goes to bed.  
5 PAR: Like what?  
6 HV1: Do you want me to repeat the question?  
7 PAR: Yes, please, I didn't understand.  
8 [0:09:00]  
9 HV1: We are interested in knowing what you and Child S do in your daily routine. Please tell me about a  
10 typical day of Child S from the moment she wakes up until she goes to bed.  
11 PAR: Like what she does?  
12 HV1: Uh-huh. Like a typical day. Her routine. Uh-huh.

2	PAR: Like what?
3	HV1: Do you want me to repeat the question?
4	PAR: Like what she does?
5	HV1: Okay. From any other place?
6	HV1: Okay. What are some of the things you do to make Child S
7	HV1: Okay. Do you do something special to help her learn new v
8	HV1: Do you do something special to help her find time to read?
9	HV1: Um, so that is one of the ways you motivate her?

# ANNOTATION

On annotate 200 questions

1	questions	qi	charlotte
2	PAR: Like what?	non-question	non-question
3	HV1: Do you want me to repeat the question?	non-question	non-question
4	PAR: Like what she does?	non-question	non-question
5	HV1: Okay. From any other place?	non-question	spontané
6	HV1: Okay. What are some of the things you do to make Child S a great reader?	préparé	préparé
7	HV1: Okay. Do you do something special to help her learn new words?	spontané	spontané
8	HV1: Do you do something special to help her find time to read?	spontané	spontané
9	HV1: Um, so that is one of the ways you motivate her?	non-question	spontané
10	HV1: What are some of the questions you ask Child S about the books she reads?	préparé	préparé
11	HV1: What do you do, if anything, when Child S has difficulties reading a book?	préparé	préparé
12	HV1: Is there anyone here at home that can help her read?	spontané	spontané

Accuracy: 0.62				
	precision	recall	f1-score	support
non-question	0.67	0.13	0.22	15
préparé	0.79	0.79	0.79	14
spontané	0.55	0.86	0.67	21
accuracy			0.62	50
macro avg	0.67	0.59	0.56	50
weighted avg	0.65	0.62	0.57	50

Les questions préparées sont pré annotées en se servant de la trame d’entretien

	Annotateur 2 préparé	Annotateur 2 spontané	Annotateur 2 non-question
Annotateur 1 préparé	63	0	0
Annotateur 1 spontané	0	73	6
Annotateur 1 non-question	0	16	42

	Annotateur 2 spontané	Annotateur 2 non question
Annotateur 1 spontané	73	6
Annotateur 1 non-question	16	42

cohen de kappa: 0,625

Nous avons ensuite fait une version de consensus pour la regression logistique



## AUTOCRITIQUE

### Principales difficultés

- fausse piste au début pour trouver un corpus comparable, perte de temps
- Le corpus est très spécifique
- Enoncés interrogatifs sans “?” non relevés
- Difficile de différencier les questions spontanées des non questions



## BIBLIOGRAPHIE

- Capotosto, L., Kim, J. S., Burkhauser, M. A., Oh Park, S., Mulimbi, B., Donaldson, M., & Kingston, H. C. (2017). Family Support of Third-Grade Reading Skills, Motivation, and Habits. AERA Open, 3(3). <https://doi.org/10.1177/2332858417714457>
- Iris Eshkol-Taravella, Angèle Barbedette, Xingyu Liu, Valentin-Gabriel Soumah. Classification automatique de questions spontanées vs. préparées dans des transcriptions de l’oral. Traitement Automatique des Langues Naturelles, 2022, Avignon, France. pp.305-314. ffhal-03701483f

# MERCI

