*Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles*
*Avignon, France, 27 juin au 1er juillet 2022*

# Classification automatique de questions spontanées vs. préparées dans des transcriptions de l'oral

Iris Eshkol-Taravella[1, 2] Angèle Barbedette[3] Xingyu Liu[2] Valentin-Gabriel Soumah[2]

(1) MoDyCo UMR7114, 200, avenue de la Republique, 92001, Nanterre, France
(2) Université Paris Nanterre, 200, avenue de la Republique, 92001, Nanterre, France
(3) Université Sorbonne Nouvelle - Paris 3, 19, rue des Bernardins, 75005, Paris, France
ieshkolt@parisnanterre.fr, angele.barbedette@gmail.com,
xingyu.liu@univ-grenoble-alpes.fr, soumahvg@gmail.com

## RÉSUMÉ

Ce travail a pour objectif de développer un modèle linguistique pour classifier automatiquement des questions issues de transcriptions d'enregistrements provenant des corpus ESLO2 et ACSYNT en deux catégories "spontané" et "préparé". Avant de procéder au traitement automatique, nous proposons une liste de critères définitoires et discriminants permettant de distinguer les questions parmi d'autres énoncés. Les expériences basées sur des méthodes d'apprentissage supervisé sont réalisées selon une classification multiclasse comprenant les catégories "spontané", "préparé" et "non-question" et selon une classification binaire incluant les catégories "spontané" et "préparé" uniquement. Les meilleurs résultats pour les méthodes traditionnelles d'apprentissage automatique sont obtenus avec une régression logistique combinée aux critères linguistiques significatifs uniquement (F-score de 0.75). Pour finir, nous mettons en parallèle ces résultats avec ceux obtenus en utilisant des techniques d'apprentissage profond.

## ABSTRACT

**Automatic Classification of Spontaneous vs. Prepared Questions in Speech Transcriptions**

This work aims at developing a linguistic model to automatically classify questions from speech transcriptions of ESLO2 and ACSYNT corpora into two categories "spontaneous" and "prepared". Before proceeding with the automatic processing, we provide a list of defining and discriminating criteria to distinguish questions from other statements. Experiments based on supervised machine learning methods are conducted using a multiclass classification including "spontaneous", "prepared" and "non-question" categories and a binary classification including "spontaneous" and "prepared" categories only. The best results for traditional machine learning methods are obtained with a logistic regression combined with significant linguistic criteria only (F-score of 0.75). Finally, we compare these results with those obtained using deep learning techniques.

MOTS-CLÉS : classification de questions, discours spontané, discours préparé, corpus oral, apprentissage supervisé.

KEYWORDS: questions classification, spontaneous speech, prepared speech, oral corpus, supervised machine learning.

Volume 1 : conférence principale, pages 305–314.

# 1 Introduction

La communication avec la machine est un objectif de plusieurs tâches et applications en TAL telles que le dialogue homme-machine ou les chatbots. Leur défi majeur est de rendre les énoncés générés par la machine les plus proches du langage humain. Ainsi, l'ajout automatique de disfluences ou de pauses remplies pour imiter la parole spontanée et améliorer par conséquence la qualité et l'expressivité des systèmes de synthèse de la parole a été proposé par exemple dans les systèmes de dialogue homme-machine (Sundaram & Narayanan, 2003; Qader *et al.*, 2017). Cependant, les disfluences et les pauses ne sont pas suffisantes pour imiter le discours humain spontané. D'autres caractéristiques linguistiques peuvent être exploitées par le TAL afin de rendre ces systèmes plus performants. C'est l'un des objectifs de la recherche présentée dans cet article qui s'intéresse au discours spontané vs. préparé et à leur distinction automatique fondée sur les indices linguistiques.

La distinction entre discours spontané et discours préparé ou élaboré est souvent associée dans les travaux en linguistique à la distinction entre langage oral et écrit. Le discours oral est comparé par (Blanche-Benveniste *et al.*, 1990) aux brouillons qui précèdent la version finale des écrits car on assiste aux étapes de son élaboration. Le discours oral étudié dans les travaux de Blanche-Benveniste est un discours spontané. Or, il existe différents types de langages parlés qui ne se limitent pas seulement aux conversations spontanées, tels que les interviews ou conférences par exemple. C'est la situation ou le contexte de production qui va influencer les aspects de la communication. Le discours oral est un continuum qui s'étend de la parole lue à de la parole enregistrée dans des conditions dites "naturelles" (au cours d'un repas, dans un environnement familier par exemple), en passant par la parole utilisée par des professionnels (journalistes par exemple) dans le cadre de l'exercice de leurs fonctions (Delgado-Martins & Freitas, 1991). Nous considérons pour ce travail que l'oral spontané correspond à un discours en construction directe et continue tandis que l'oral preparé s'appuie sur une élaboration en amont du discours (Dutrey *et al.*, 2014; Jousse *et al.*, 2008) et implique une séparation spatio-temporelle entre l'élaboration du discours et sa production (Guerin, 2015).

Plusieurs indices ont été proposés pour caractériser le discours spontané. Dans (Jousse *et al.*, 2008; Shriberg, 1999), les auteurs considèrent que le degré de spontanéité du discours peut en partie être établi d'après des critères prosodiques tels que la durée des voyelles ou l'allongement des syllabes en fin de mots et qu'il existe donc un lien entre prosodie et discours spontané. Le discours spontané est également souvent caractérisé par la présence de disfluences, c'est à dire d'interruptions du flux de la parole qui n'apportent pas d'informations supplémentaires à l'interaction, de contenu propositionnel (Blanche-Benveniste *et al.*, 1990; Tree, 1995). Cependant, d'autres éléments peuvent indiquer la nature spontanée d'un énoncé.

Afin d'étudier et de repérer le discours spontané, nous nous sommes concentrés sur les énoncés interrogatifs issus de transcriptions de l'oral [1]. Les informations prosodiques ont été laissées de côté. Les résultats obtenus peuvent avoir un impact important dans différentes disciplines. Par exemple, il est intéressant de pouvoir distinguer une question spontanée d'une question préparée dans le cadre de l'analyse de discours politiques. Les traits linguistiques caractéristiques relevés dans ce travail peuvent aussi être exploités dans les systèmes de génération automatique de texte pour améliorer certains aspects du dialogue homme-machine ou encore des chatbots en rendant le discours plus naturel, la spontanéité étant une caractéristique indéniable du langage humain.

Les données traitées proviennent des corpus ESLO2 et ACSYNT qui comprennent entre autres des

1. Ces énoncés ont déjà fait l'objet d'une étude dans (Barbedette & Eshkol-Taravella, 2021)

interviews realisées selon une trame de questions prédéfinies qu'on considère dans ce travail comme des questions préparées. Les données collectées à partir d'Ortolang (www.ortolang.fr) sont :

— les modules *Entretiens*, *Interviews Personnalités* (et leurs trames), *Cinéma* et *Repas* d'ESLO2 (Baude & Dugua, 2011; Eshkol-Taravella *et al.*, 2011) ;
— les entretiens guidés (et leurs trames) du corpus ACSYNT (CLLE-ERSS, 2013).

Les questions ont été extraites automatiquement en utilisant le point d'interrogation transcrit (les énoncés interrogatifs ne passant pas par l'utilisation du point d'interrogation n'ont donc pas été utilisés), ainsi que les cinq tours de parole précédant et suivant chaque question. Ce seuil a été considéré comme suffisant pour permettre une bonne compréhension de la question par les annotateurs, repérer de potentielles informations pertinentes et donc faciliter l'annotation manuelle des questions. De plus, il a permis par la suite de dégager des critères linguistiques basés sur le contexte pour la tâche de classification automatique. Le corpus final est composé de 1 298 échantillons annotés pour ESLO2 et de 588 pour ACSYNT.

## 2 Annotation manuelle

La distinction entre une question spontanée et une question préparée ne fait sens qu'une fois qu'il a été établi que la production est bien une question. La DIDT (Dynamic Interpretation and Dialogue Theory) (Bunt, 1999) et le DAMSL (Dialog Act Markup in Several Layers) (Jurafsky, 1997) définissent une question comme un acte de dialogue consistant en une recherche d'information. D'après (Ginzburg & Sag, 2000), un énoncé interrogatif est composé de propositions qui constituent une question. Une question déclarative n'est donc pas syntaxiquement une question, mais sa signification est conforme à une demande d'information. Au contraire, une question rhétorique est syntaxiquement une question, mais aucune demande d'information n'est formulée. Du point de vue du *commitment* (Beyssade & Marandin, 2009), une question est un acte de langage qui engage à la fois le locuteur comme "désireux d'obtenir une information" et l'interlocuteur à accepter cet acte comme un désir d'obtenir une information.

Le corpus a été réparti entre sept annotateurs et des consignes ont été établies avant la tâche pour déterminer les critères sur lesquels s'appuyer et les étiquettes à utiliser. La première étape de l'annotation consiste à déterminer si l'énoncé à annoter est une question, une non-question ou une question non annotable :

— Non-questions :
  — Tag questions : sont définies par la DIDT (Bunt, 1999) comme un moyen de vérifier que l'interlocuteur a compris et accepté ce qui est dit ;
  — Demandes de répétition : de forme interrogative, elles sont associées à des demandes de clarification (Ginzburg, 2012; Purver *et al.*, 2003; Boritchev, 2021) ;
  — Vérifications de compréhension : sont similaires aux demandes de répétition et aux tag questions (Bunt, 1999), le locuteur s'assure qu'il a bien compris le tour de parole précédent en en répétant une partie ou le tout. Ce sont des demandes de clarification d'après (Ginzburg, 2012; Purver *et al.*, 2003; Boritchev, 2021) ;
  — 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 (Boritchev, 2021) ;
  — Questions injonctives : sont des demandes d'action ;
  — Questions d'obligation sociale : pour (Bunt, 1999), ces actes correspondent à la gestion

des obligations sociales, qui sont des actes de contrôle du dialogue. Il en est de même pour (Jurafsky, 1997), qui les considère comme faisant partie de la gestion de la communication ;
- Questions non annotables (inutilisables en raison d'un problème de transcription préexistant) : questions coupées, questions trop larges, discours rapporté ou questions non interprétables ;
- Questions (demandes d'information qui apportent une valeur ajoutée au dialogue en cours) :
  - Questions ouvertes : contiennent un mot interrogatifs et dont la réponse n'est pas "oui", "si" ou "non" (Bunt, 1999; Jurafsky, 1997; Ginzburg & Sag, 2000; Boritchev, 2021) ;
  - Questions fermées : supposent une réponse par "oui", "si" ou "non ;
  - Questions alternatives ou alt-questions (Bunt, 1999; Aarts *et al.*, 2018; Boritchev, 2021) : les réponses possibles attendues sont contenues dans la question posée.

Une vraie question peut alors être classée comme une question spontanée ou préparée. Les critères énumérés dans le tableau 1 aident à décider de la catégorie mais ne sont pas décisifs : ils sont combinés avec la compréhension globale de la question et son contexte de production. Les trames d'entretiens existantes ont également été prises en compte lors de l'annotation.

| | | |
|---|---|---|
| Spontané | Anaphores | spk1 : je suis contrôleur divisionnaire aux PTT |
| | | spk4 : et en quoi *ça* consiste ? |
| | Demandes de clarification ou d'informations supplémentaires (Purver, 2004) | spk1 : mais // ce qui serait intéressant c'est que justement // on puisse euh // les enfants puissent apprendre de très bonne heure // euh très jeunes // une ou deux langues étrangères une au moins |
| | | spk4 : *et pourquoi ?* |
| | Thème précédant le rhème [2] | spk1 : mais alors là vous êtes en France pour combien de temps maintenant ? |
| | Mot interrogatif en position finale | spk3 : oui non mais cette équipe elle est constituée *comment* ? |
| | Disfluences | spk1 : il m'est supérieur *euh* comment ça dans le travail ou ? |
| Préparé | Répétition d'entité nommée | spk4 : depuis combien de temps habitez-vous Orléans ? |
| | | spk1 : oh ça fait neuf ans depuis dix neuf cent soixante |
| | | spk4 : vous vous plaisez à *Orléans* ? |
| | Rupture d'isotopie | spk1 : oui oui // et vous comptez rester ? |
| | | spk2 : oui enfin // tant que je serai célibataire |
| | | spk1 : ah oui |
| | | spk2 : après on en sait rien de ça |
| | | spk1 : *alors est-ce qu'on pourrait parler un peu de votre travail ?* |
| | Rhème précédant le thème [3] | spk4 : qu'est-ce qui vous plaît dans votre travail ? |
| | Inversion du sujet | spk4 : *faites-vous* un brouillon ? |
| | Mot interrogatif en position initiale | spk4 : et *qu*'est-ce que vous pensez du latin à l'école ? |
| | Annonce de la question | spk4 : alors maintenant je vais vous poser des questions euh // peut-être un peu // un peu plus pers- personnelle mais |

TABLE 1 – Indicateurs et exemples issus du corpus pour définir les questions spontanées et préparées

Enfin, un accord inter-annotateur a été calculé avec un Kappa de Cohen (Cohen, 1960) sur l'annotation

2. Les concepts de thème et rhème sont liés à la distinction entre information existante et nouvelle (Gundel, 1988). Ici, la nouvelle information liée au thème "mais alors là vous êtes en France", c'est-à-dire le rhème, est la question "pour combien de temps maintenant" qui suit le thème.
3. Ici, le thème "dans votre travail" suit le rhème qui est véhiculé par la question "qu'est-ce qui vous plaît".

manuelle de 200 questions, produite par deux annotateurs experts qui sont les deux premiers auteurs de ce travail. La valeur de Kappa obtenue est de 0.75, ce qui est considéré comme un accord considérable et satisfaisant d'après (Landis & Koch, 1977).

# 3 Expériences de classification automatique

Les expériences basées sur des méthodes d'apprentissage automatique supervisé ont été réalisées selon une classification multiclasse comprenant les catégories "spontané", "préparé" et "non-question" et selon une classification binaire incluant les catégories "spontané" et "préparé" uniquement, sur un jeu de données annotées composé de 731 questions spontanées, 478 questions préparées et 284 non-questions des corpus ESLO2 et ACSYNT qui ont été divisées en 0.75 et 0.25 pour les données d'entraînement et de test respectivement. Les questions ont été représentées en utilisant une vectorisation du corpus obtenue avec le modèle CBOW basé sur le corpus FrWaC (Fauconnier, 2015) (200 dimensions) et en utilisant une normalisation avec des poids TF-IDF pour prendre en compte l'importance, la rareté et la fonction discriminante des mots du corpus. Le modèle Skip-Gram a également étét testé pour les expériences mais a donné globalement des résultats inférieurs. Notre sélection de critères pertinents pour la classification se base sur les critères linguistiques mentionnés dans (Blanche-Benveniste & Bilger, 1999), ainsi que sur les observations faites à partir d'un extrait du corpus. Les critères ont été extraits à partir des librairies Python suivantes : SpaCy (Honnibal & Montani, 2017) pour le repérage d'entités nommées et le POS tagging notamment, NLTK (Bird *et al.*, 2009) pour la tokenisation et la lemmatisation, Gensim (Rehurek & Sojka, 2011) et Sk-Learn (Pedregosa *et al.*, 2011) :

1. Longueur de la question (nombre de tokens) ;
2. Présence de disfluences (ratio entre le nombre de disfluences et le nombre total de mots) ;
3. Présence d'une annonce d'une question (critère binaire) ;
4. Présence d'une inversion du sujet (critère binaire) ;
5. Présence de la répétition d'une entité nommée (critère binaire) ;
6. Distance vectorielle mesurée avec des vecteurs Word2Vec (Fauconnier, 2015) entre les questions et le contexte précédant ;
7. Position du mot interrogatif (trois critères binaires initiale (a), intermédiaire (b) ou finale (c)) ;
8. Présence de la forme interrogative "est-ce que" (critère binaire) ;
9. Présence du marqueur de l'inversion du sujet "t-il" (critère binaire).

Les résultats des expériences sont résumés dans le tableau 2. Ils ont été obtenus sans équilibrer les données afin de conserver leur distribution naturelle dans chaque catégorie. La proportion de la classe minoritaire a été considérée comme suffisante pour la tâche, car elle représente 19% des données utilisées pour la classification multiclasse et 39,5% des données utilisées pour la classification binaire. Les meilleurs résultats sont obtenus avec un score de 0.74 pour la classification binaire et avec un score de 0.66 pour la classification multiclasse avec une régression logistique et des critères linguistiques uniquement.

Après avoir vérifié la pertinence des critères linguistiques à l'aide d'un modèle de regression logistique, des valeurs de significativité p pour chacun des critères évalués et pour chacune des deux catégories "spontané" et "préparé" ont été obtenues avec la librairie Python statsmodel (Seabold & Perktold,

| | W2V et TF-IDF | | | Critères linguistiques | | | Critères combinés | | |
|---|---|---|---|---|---|---|---|---|---|
| | Micro | Macro | Pond. | Micro | Macro | Pond. | Micro | Macro | Pond. |
| RF (multi) | 0.55 | 0.51 | 0.53 | 0.58 | 0.55 | 0.58 | 0.62 | 0.56 | 0.6 |
| k-NN (multi) | 0.51 | 0.5 | 0.51 | 0.55 | 0.53 | 0.55 | 0.49 | 0.48 | 0.49 |
| SVM (multi) | 0.6 | 0.56 | 0.59 | 0.61 | 0.56 | 0.59 | 0.61 | 0.57 | 0.6 |
| LR (multi) | 0.55 | 0.53 | 0.55 | 0.67 | 0.63 | **0.66** | 0.6 | 0.58 | 0.6 |
| RF (bin) | 0.64 | 0.57 | 0.6 | 0.68 | 0.67 | 0.68 | 0.63 | 0.57 | 0.59 |
| k-NN (bin) | 0.59 | 0.59 | 0.6 | 0.7 | 0.68 | 0.69 | 0.61 | 0.6 | 0.61 |
| SVM (bin) | 0.65 | 0.58 | 0.61 | 0.59 | 0.4 | 0.45 | 0.66 | 0.59 | 0.62 |
| LR (bin) | 0.63 | 0.62 | 0.63 | 0.75 | 0.73 | **0.74** | 0.69 | 0.67 | 0.68 |

TABLE 2 – Résumé des moyennes micro, macro et pondérées de F-mesures obtenues en combinant les algorithmes et des critères sélectionnés (représentations Word2Vec et TD-IDF, critères linguistiques ou les deux), en faisant une classification multiclasse (multi) ou binaire (bin) sur des classes non équilibrées

2010). Les critères semblant significatifs (valeur de p inférieure au seuil standard de 0.05) pour la classification (nombre de tokens, rapport entre les disfluences et le nombre total de mots, inversion du sujet, position initiale du mot interrogatif, présence de "est-ce que" et "-t-il") ont été sélectionnés pour réexécuter l'algorithme de régression logistique pour la classification binaire et les scores ont été légèrement améliorés (tableau 3).

| | Précision | Rappel | F-mesure |
|---|---|---|---|
| Spontané | 0.76 | 0.83 | 0.8 |
| Préparé | 0.74 | 0.64 | 0.69 |
| Micro | | | 0.75 |
| Macro | 0.75 | 0.74 | 0.74 |
| Moyenne pondérée | 0.75 | 0.75 | 0.75 |

TABLE 3 – Résultats pour la classification binaire "spontané" vs. "préparé" en utilisant l'algorithme de régression logistique avec les critères linguistiques significatifs seulement ($p<0.05$)

Les résultats présentés dans le tableau 2 montrent en général de meilleures performances lorsqu'on applique une classification binaire. D'après les observations faites au cours des expériences, la classification multiclasse a de meilleurs précision et rappel pour les questions spontanées que pour les questions préparées, ce qui peut être lié au fait que les non-questions, troisième catégorie pour la classification multiclasse, peuvent partager les mêmes caractéristiques syntaxiques que les questions spontanées. Cette difficulté peut être accentuée par le fait que les questions spontanées représentent la classe majoritaire dans les données utilisées. Pour tester l'hypothèse d'un effet sur les résultats, le corpus a été équilibré en réduisant les données et les algorithmes ont été réexécutés. Les résultats sont globalement similaires aux premiers. Les scores pour les questions spontanées sont généralement plus faibles et le rappel est plus élevé pour les non-questions, mais avec une précision inférieure à 0.6. En ce qui concerne la sélection des seuls critères linguistiques pertinents, l'amélioration est mineure mais les scores sont restés stables, ce qui signifie que les critères retenus après l'analyse de signification sont pertinents et suffisants pour obtenir le score satisfaisant de 75% de bonnes prédictions. De plus, les traits pertinents mis en évidence par l'analyse statistique montrent une cohérence avec l'analyse manuelle du corpus qui a montré notamment que les disfluences étaient représentatives des questions spontanées et que la présence d'un mot interrogatif en début de question était typique des questions préparées.

# 4 Discussion et conclusion

En ayant pour but de classifier automatiquement les questions spontanées et préparées, cette étude a également permis de dégager des traits permettant de distinguer les deux catégories à partir de transcriptions de la parole, sans se baser sur des informations audio comme cela a déjà été fait dans (Shriberg *et al.*, 2009). Les meilleurs résultats pour les méthodes traditionnelles d'apprentissage automatique sont obtenus avec une régression logistique combinée aux critères linguistiques significatifs uniquement (F-score de 0.75).

Des développements possibles de ce travail consisteraient à tester des méthodes d'apprentissage profond sur le corpus avec les mêmes caractéristiques linguistiques pertinentes et également à détecter les questions et non-questions. Des résultats préliminaires ont été obtenus d'abord en réutilisant les critères définis pour classer les questions spontanées et préparées, puis en utilisant un modèle de langage pré-entraîné, sans critère supplémentaire, pour prédire les questions et les non-questions :

— Réseau de neurones à propagation avant : les critères linguistiques significatifs ont été utilisés et le réseau construit avec la bibliothèque python Keras (Chollet *et al.*, 2015). Une méthode heuristique a été appliquée pour ajuster l'algorithme jusqu'à obtenir les meilleurs résultats possibles. L'architecture finale du modèle utilisé est simple (composée de trois couches denses séparées par des couches de dropout) et ce dernier a été entraîné pendant 200 époques. Le corpus a été équilibré en réduisant les données pour éviter un écart important entre les scores des questions spontanées et ceux des questions préparées. L'ensemble d'entraînement contient donc 717 questions dont 142 font partie du jeu de validation, tandis que l'ensemble de test est composé de 239 questions. Une F-mesure moyenne pondérée de 0.7 a finalement été obtenue (tableau 4).

— CamemBERT (Martin *et al.*, 2019) : la librairie Simple Transformers (Rajapakse, 2019) a été utilisée (paramètres par défaut, 5 époques) avec un jeu de données équilibré composé de 956 questions. Une F-mesure moyenne pondérée de 0.73 a été obtenue pour la classification en question spontanée / préparée (tableau 4), valeur proche des résultats obtenus avec la régression logistique et les critères significatifs, avec des résultats légèrement améliorés pour les questions préparées, et de 0.84 pour la classification en question / non-question (tableau 5), score qui pourrait être amélioré en déterminant des traits pertinents pour distinguer ces deux catégories.

| | NN à propagation avant | | | CamemBERT | | |
|---|---|---|---|---|---|---|
| | Précision | Rappel | F-mesure | Précision | Rappel | F-mesure |
| Spontané | 0.71 | 0.74 | 0.73 | 0.77 | 0.67 | 0.72 |
| Préparé | 0.68 | 0.65 | 0.67 | 0.71 | 0.8 | 0.75 |
| Moyenne pondérée | 0.7 | 0.7 | 0.7 | 0.74 | 0.73 | 0.73 |

TABLE 4 – Résultats pour la classification binaire "spontané" vs. "préparé" en utilisant un réseau de neurones à propagation avant et les critères significatifs seulement ($p<0.05$), et en utilisant le modèle pré-entraîné de CamemBERT

| | Précision | Rappel | F-mesure |
|---|---|---|---|
| Question | 0.85 | 0.84 | 0.84 |
| Non-question | 0.84 | 0.85 | 0.85 |
| Moyenne pondérée | 0.84 | 0.84 | 0.84 |

TABLE 5 – Résultats pour la classification binaire "question" vs. "non-question" en utilisant le modèle pré-entraîné de CamemBERT

D'après l'ensemble des résultats, nos critères ne semblent pas suffisamment pertinents pour la détection des non-questions et il semble y avoir des caractéristiques communes entre les non-questions et les questions spontanées qui complexifient la tâche. La classification des questions et des non-questions avec CamemBERT a donné 84% de bonnes prédictions. Il serait intéressant d'affiner les caractéristiques linguistiques pour être capable de discriminer les questions spontanées, les questions préparées et les non-questions, mais aussi d'étendre ce travail à d'autres types d'énoncés interrogatifs comme les questions injonctives ou d'obligation sociale par exemple.

# Remerciements

Nous souhaitons remercier Marina Baidina et Martin Anthony Salud pour leur participation dans toutes les étapes qui ont mené à une première version de ce travail : de la définition de la typologie des questions spontanées vs. préparées, jusqu'aux expériences de classification automatique, en passant par l'annotation manuelle des questions.

# Références

AARTS B., CHALKER S., WEINER E. & PRESS O. (2018). The oxford dictionary of english grammar .(2014). *URL https ://en. oxforddictionaries. com/definition/heterogeneous. Accessed*.

BARBEDETTE A. & ESHKOL-TARAVELLA I. (2021). Quand les questions en disent plus que les réponses : classification automatique des intentions dans les questions. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (28).

BAUDE O. & DUGUA C. (2011). (re) faire le corpus d'orléans quarante ans après : quoi de neuf, linguiste ? *Corpus*, (10), 99–118.

BEYSSADE C. & MARANDIN J.-M. (2009). Commitment : une attitude dialogique. *Langue française*, (2), 89–107.

BIRD S., KLEIN E. & LOPER E. (2009). *Natural language processing with Python : analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

BLANCHE-BENVENISTE C. & BILGER M. (1999). Français parlé-oral spontané. quelques réflexions. *Revue française de linguistique appliquée*, **4**(2), 21–30.

BLANCHE-BENVENISTE C., BILGER M., ROUGET C., VAN DEN EYNDE K., MERTENS P. & WILLEMS D. (1990). Le français parlé(études grammaticales). *Sciences du langage*.

BORITCHEV M. (2021). *Modélisation dynamique des dialogues*. Thèse de doctorat, Université de Lorraine.

BUNT H. (1999). Dynamic interpretation and dialogue theory. *The structure of multimodal dialogue*, **2**, 139–166.

CHOLLET F. *et al.* (2015). keras.

CLLE-ERSS (2013). Acsynt. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

COHEN J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, **20**(1), 37–46.

DELGADO-MARTINS M. R. & FREITAS M. J. (1991). Temporal structures of speech :" reading news on tv". In *Phonetics and Phonology of Speaking Styles*.

DUTREY C., ROSSET S., ADDA-DECKER M., CLAVEL C. & VASILESCU I. (2014). Disfluences dans la parole spontanée conversationnelle : détection automatique utilisant des indices lexicaux et acoustiques. *XXXe Journées d'Étude sur la Parole (JEP'14)*, p. 366–373.

ESHKOL-TARAVELLA I., BAUDE O., MAUREL D., HRIBA L., DUGUA C. & TELLIER I. (2011). Un grand corpus oral «disponible» : le corpus d'orléans 1 1968-2012.

FAUCONNIER J.-P. (2015). French word embeddings.

GINZBURG J. (2012). *The interactive stance*. Oxford University Press.

GINZBURG J. & SAG I. (2000). *Interrogative investigations*. Stanford : CSLI publications.

GUERIN E. (2015). *Observer, décrire,... enseigner, le français" langue vivante"*. Thèse de doctorat, Univ. Poitiers.

GUNDEL J. K. (1988). Universals of topic-comment structure. *Studies in syntactic typology*, **17**(1), 209–239.

HONNIBAL M. & MONTANI I. (2017). spacy 2 : Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, **7**(1), 411–420.

JOUSSE V., ESTEVE Y., BÉCHET F., BAZILLON T. & LINARES G. (2008). Caractérisation et détection de parole spontanée dans de larges collections de documents audio. *JEP*, **2008**, 9–13.

JURAFSKY D. (1997). Switchboard swbd-damsl shallow-discourse-function annotation coders manual. *Institute of Cognitive Science Technical Report*.

LANDIS J. R. & KOCH G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, p. 159–174.

MARTIN L., MULLER B., SUÁREZ P. J. O., DUPONT Y., ROMARY L., DE LA CLERGERIE É. V., SEDDAH D. & SAGOT B. (2019). Camembert : a tasty french language model. *arXiv preprint arXiv :1911.03894*.

PEDREGOSA F., VAROQUAUX G., GRAMFORT A., MICHEL V., THIRION B., GRISEL O., BLONDEL M., PRETTENHOFER P., WEISS R., DUBOURG V. *et al.* (2011). Scikit-learn : Machine learning in python. *the Journal of machine Learning research*, **12**, 2825–2830.

PURVER M., GINZBURG J. & HEALEY P. (2003). On the means for clarification in dialogue. In *Current and new directions in discourse and dialogue*, p. 235–255. Springer.

PURVER M. R. J. (2004). *The theory and use of clarification requests in dialogue*. Thèse de doctorat, Citeseer.

QADER R., LECORVÉ G., LOLIVE D. & SÉBILLOT P. (2017). Ajout automatique de disfluences pour la synthèse de la parole spontanée : formalisation et preuve de concept. In *Traitement automatique du langage naturel (TALN)*.

RAJAPAKSE T. C. (2019). Simple transformers. https://github.com/ThilinaRajapakse/simpletransformers.

REHUREK R. & SOJKA P. (2011). Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, **3**(2).

SEABOLD S. & PERKTOLD J. (2010). Statsmodels : Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference*, volume 57, p.61 : Austin, TX.

SHRIBERG E., FAVRE B., FUNG J., HAKKANI-TUR D. & CUENDET S. (2009). Prosodic similarities of dialog act boundaries across speaking styles. *Linguistic Patterns in Spontaneous Speech*, (A25), 213–239.

SHRIBERG E. E. (1999). *Phonetic consequences of speech disfluency*. Rapport interne, SRI INTERNATIONAL MENLO PARK CA.

SUNDARAM S. & NARAYANAN S. (2003). An empirical text transformation method for spontaneous speech synthesizers. In *Eighth European Conference on Speech Communication and Technology*.

TREE J. E. F. (1995). The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of memory and language*, **34**(6), 709–738.