Est-ce que l'extraction des interrogatives du français peut-elle être automatisée ?

Valentin D. Richard

Actes des journées LIFT 2023
Nancy, France, 20-21 novembre 2023

Est-ce que l'extraction des interrogatives du français peut-elle être automatisée ?

Valentin D. Richard[1]
(1) LORIA, Université de Lorraine, 615 Rue du Jardin-Botanique, 54506 Vandœuvre-lès-Nancy, France
valentin.richard@loria.fr

RÉSUMÉ

La quasi totalité des études linguistiques sur les interrogatives du français se contente d'extraire ces dernières d'un corpus à la main ou grâce à de simples heuristiques basées sur du texte brut (mots interrogatifs, point d'interrogation,...). Dans ce papier, je présente FUDIA (French UD Interrogative Annotator), un programme qui permet de détecter les interrogatives du français d'un corpus annoté en dépendances universelles (UD). FUDIA est un système de réécriture de graphe par règles, basé sur Grew. Je liste les obstacles à une telle tâche d'identification automatique des interrogatives et j'explique comment FUDIA en résout la plupart. Je montre que, couplé à un parseur affiné sur des données similaires, FUDIA obtient de bons résultats sur du texte brut (écrit et transcription de l'oral).

ABSTRACT

Can French Interrogative Retrieval be Fully Machine-Based ?

The vast majority of linguistic corpus studies on French interrogatives retrieve the researched patterns by hand or only based on simple heuristics on raw text (e.g. interrogative words, question marks). In this paper, I present FUDIA (French UD Interrogative Annotator), a program able to detect French interrogatives from a corpus annotated in Universal Dependencies (UD). FUDIA is a rule-based graph rewriting system based on Grew. I inventory the obstacles to such an interrogative identification task and I explain how FUDIA solves most of them. I show that, coupled with a parser fine-tuned on similar data, FUDIA obtains good results on raw text (written and speech transcription).

MOTS-CLÉS : interrogatives, français, Universal Dependencies, par règles, réécriture de graphe.

KEYWORDS: interrogatives, French, Universal Dependencies, rule-based, graph rewriting.

1 Introduction

Parmi les études de corpus récentes portant sur les interrogatives du français, la plupart récupèrent leurs données en les extrayant à la main (Reinhardt, 2019a; Bally, 2022). Quelques

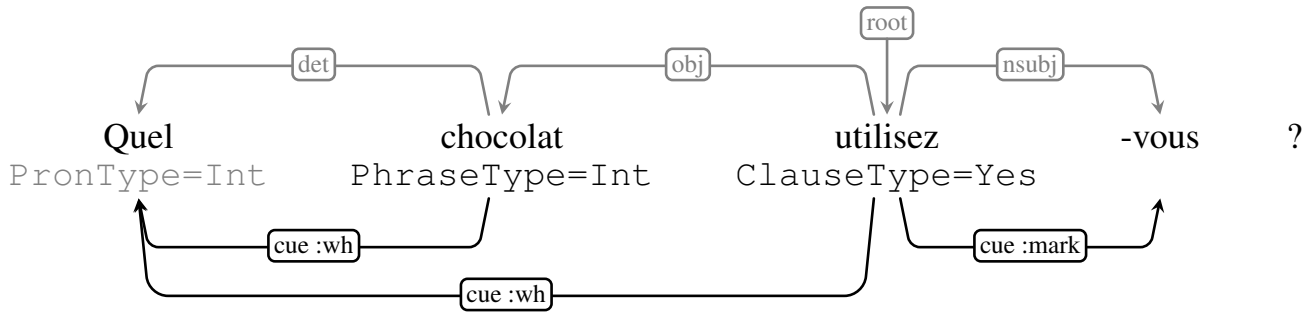

FIGURE 1 – Annotation d'une question de GSD par FUDIA. Les arcs et traits d'origine sont en gris. Les arcs `cue` peuvent être retirés pour conserver la structure d'arbre initiale.

méthodes semi-automatiques sont employées, comme l'utilisation d'un concordancier (Rossi-Gensane *et al.*, 2021; Benzitoun, 2022; Gillet, 2022), d'expressions régulières (Lefeuvre & Rossi-Gensane, 2017) ou d'heuristiques simples (Reinhardt, 2019b; Eshkol-Taravella *et al.*, 2022) (mots interrogatifs et points d'interrogations). Seule Lefeuvre (2021) profite d'un corpus (LM10 (Habert, 2005) [1]) préalablement annoté syntaxiquement grâce à l'outil Syntex (Bourigault *et al.*, 2005) pour identifier certains motifs.

2 Proposition

FUDIA Le programme FUDIA (French UD Interrogative Annotator) s'appuie sur la présence de corpus déjà annotés en syntaxe, notamment les corpus francophones du projet Universal Dependencies (UD) (de Marneffe *et al.*, 2021) : FQB (Seddah & Candito, 2016), GSD (Guillaume *et al.*, 2019), ParisStories (Kahane *et al.*, 2021), ParTUT (Bosco & Sanguinetti, 2014), PUD (McDonald *et al.*, 2013), Rhapsodie (Lacheret *et al.*, 2014) et Sequoia (Candito & Seddah, 2012). FUDIA contient des règles de réécriture de graphe basées sur Grew [2] (Bonfante *et al.*, 2018). Ces règles listent les structures attestées d'interrogatives en français [3].

Annotations et FIB À partir de la représentation UD d'une phrase, FUDIA rajoute le trait `ClauseType=Int` indiquant sur la tête d'une proposition (finie, infinitive ou averbale) qu'elle est interrogative. Les mots interrogatifs et marquages morphosyntaxiques (inversion du clitique sujet, *est-ce que* / *si*,...) sont aussi identifiés par des dépendances spécifiques, appelée arcs `cue` (voir Fig. 1).

Les interrogatives des corpus francophones UD ont été extraites en un corpus appelé French

1. Voir aussi `http://redac.univ-tlse2.fr/voisinsdelemonde/index.jsp`
2. `https://grew.fr/`
3. On s'intéresse ici aux interrogatives d'un point de vue syntaxique, comme définies dans la Grande Grammaire du Français (GGF) (Delaveau *et al.*, 2021).

Interrogative Bank (FIB). À l'aide d'un script fourni, il est possible d'y calculer automatiquement la proportion d'un certain type d'interrogative, selon la classification de Coveney (2011). Le FIB étend le FQB (French Question Bank) en y apportant une plus grande diversité de structures syntaxiques (ex. des phrase de la forme : *Est-ce que* + sujet + verbe) et des interrogatives enchâssées.

Le code source de FUDIA [4] ainsi que le French Interrogative Bank [5] sont librement disponible en ligne.

3 Difficultés

Obstacles Les difficultés rencontrées pour correctement identifier les interrogatives sont nombreuses. On recense des obstacles linguistiques, notamment la grande variabilité de ces constructions en français, et les formes proches en apparence (ex. inversion stylistique pour rapporter des paroles). Du fait de son format, UD a quelques limites. Par exemple, rien ne différencie lexicalement les *si* interrogatifs des *si* conditionnels. Mais un bon nombre de difficultés est attribuable à la variabilité et la non-uniformité des annotations UD. Par exemple, on a retrouvé 6 annotations différentes de *est-ce que* sur 36 occurrences, dont aucune avec la relation `fixed`, censée être employée pour les expression figées.

Solutions apportées Le développement de FUDIA a cherché à prendre en compte toutes les formes attestées d'interrogatives dans la littérature scientifique, mêmes celles non standards, comme le titre de ce papier (*est-ce que* + inversion du clitique). De plus, le programme évite le plus possible de dépendre de listes de mots de classe ouverte. Notamment il ne présume pas des verbes pouvant enchâsser une interrogative. Le faire risquerait de louper des occurrences non envisagées. Cette stratégie s'est avérée payante car elle a permis de "découvrir" 6 verbes introducteurs d'interrogative listés ni dans la GGF (Delaveau *et al.*, 2021, Tab. XXII-7 p.1414) ni par Defrancq (2005, chap. 1 n.b.p. 11). Ces verbes sont : *connaitre*, *enseigner*, *interroger*, *mesurer*, *souligner* et *tester*.

Les obstacles linguistiques et les limitations d'UD sont en partie résolues grâce à quelques heuristiques sur l'environnement syntaxique des structures recherchées, par exemple la forme du verbe principal ou la distinction complément / ajout. La présence de nombreux cas et exceptions a pu être traitée grâce à un schéma de disjonction de motifs écrit en python. Quelques expressions clés (ex. *est-ce que* et *qu'est-ce que*) sont réannotées de manière uniforme en suivant les théories de la GGF (Delaveau *et al.*, 2021) et les choix d'annotation du corpus CEFC/Orféo [6] (Benzitoun *et al.*, 2016).

4. https://github.com/Valentin-D-Richard/FUDIA
5. https://github.com/Valentin-D-Richard/UD_French-FIB
6. https://repository.ortolang.fr/api/content/cefc-orfeo/11/documentation/site-orfeo/guide-dannotation-syntaxique-du-corpus-orfeo/

	exactitude	précision	rappel	F1
FUDIA	0,905	0,966	0,770	0,857
sur l'écrit seul	0,860	0,917	0,647	0,759
sur l'oral seul	0,950	1,00	0,875	0,933

TABLE 1 – Score de FUDIA sur la détection d'interrogatives du français

4 Évaluation

Méthode Pour évaluer FUDIA, un corpus de 200 phrases a été constitué. Il contient des phrases tirées aléatoirement de corpus écrit (Annodis (Péry-Woodley *et al.*, 2011)), oral (OFROM (Avanzi *et al.*, 2012)), de questions (Maya (Reinhardt, 2016), TenNovels (Reinhardt, 2019b)) et d'interrogatives enchâssées (Defrancq, 2005). Une tâche d'annotation regroupant 12 participant·es les a annoté selon si elles contiennent au moins une interrogative ou pas d'interrogative du tout. On utilise ces étiquettes de référence pour calculer le score de FUDIA.

En premier lieu, les phrases sont annotées en UD par un parseur pré-entrainé grâce à ArboratorGrew [7]. La moitié du corpus d'évaluation issu de corpus écrits est parsée en affinant le parseur avec GSD (1476 phrases, LAS = 0,922). L'autre moitié, issue de corpus oraux, en affinant sur Rhapsodie et ParisStories (total : 2675 phrases, LAS = 0,818). Puis FUDIA est exécuté sur la sortie.

Résultats La tâche d'annotation obtient un bon score inter-annotateur·rice : Cohen κ minimum = 0,613, moyenne = 0,781, maximum = 0,924. Cependant, quelques phrases ont généré plus de désaccord. On compte 11 phrases sur 200 avec un écart-type élevé (supérieur à 0,47, c.à.d. au moins un tiers de désaccord).

Les résultats de l'évaluation de FUDIA, sur l'ensemble du corpus d'évaluation ainsi que sur chacune des parties écrites et orales de celui-ci, sont affichés en Table 1.

5 Interprétation

Tâche d'annotation Parmi les phrases engendrant le plus de désaccord, on trouve certaines interrogatives enchâssées, ex. (1-a) (crochets rajoutés pour délimiter l'interrogative). Ces structures apparaissent donc comme plus difficiles à détecter pour les humains. L'autre type majeur de débat concerne les marqueurs du discours insistant sur l'attente d'une réponse, tels

index.html

7. https://arboratorgrew.elizia.net/

hein ? ou *non ?* (1-b). Les consignes demandaient d'étiqueter les déclaratives questionnantes, qui auraient une intonation montante à l'oral, comme non-interrogatives. Mais le statut de ces marqueurs n'y était pas suffisamment précisé.

(1) a. À travers un récit largement autobiographique, le comique français a décrit avec humour [comment nombre des « gauchistes » d'alors ont troqué le manteau afghan et les sabots hollandais pour la veste et l'attaché-case d'aujourd'hui]. (Defrancq écrit)
b. Tu me fais confiance, non ? (TenNovels)

Scores de FUDIA La précision de FUDIA est élevée. Nous attribuons ça aux nombreuses heuristiques qui permettent d'éliminer les structures qui ressemblent à des interrogatives mais n'en sont pas.

Le point faible de FUDIA est son rappel. En tout, 17 interrogatives ne sont pas correctement identifiées (12 de la partie écrite du corpus dévaluation, 5 de la partie orale). Parmi ces faux négatifs, on estime que la quasi totalité sont dus au parseur en amont. L'erreur la plus fréquente concerne les mots QU annotés comme des mots relatifs (ex. (2-a)) ou des conjonctions de subordination, au lieu de comme des mots interrogatifs. Par exemple, dans (2-b) , le parseur annote les deux *comment* en tant que conjonction de subordination. C'est erroné et surprenant, car *comment* est toujours un adverbe en français, et toutes les occurrences de *comment* dans les données d'affinage sont annotées comme des adverbes.

(2) a. mais il fallait répondre à qui (Defrancq oral)
b. Nous ne voulons pas savoir comment les faits se sont déroulés, mais comment ils auraient pu ou dû arriver. (Defrancq écrit)
c. [...] l'AIEA n'était pas en mesure de vérifier s'il y a eu détournement ou non de matériel nucléaire [...] (Defrancq écrit)
d. ce qui me préoccupe c'est que va-t-on faire avec XXX (Defrancq oral)

Un autre type d'erreur concerne la confusion entre circonstancielle conditionnelle et interrogative complément. Par exemple, dans (2-c), le parseur a annoté le syntagme "*s'il y a eu détournement ou non de matériel nucléaire*" comme une proposition ajout au lieu d'une proposition complément de *vérifier*. De ce fait, FUDIA la classifie comme une conditionnelle, sans détecter la séquence *ou pas*.

Le dernier type de faux négatif est lié à la tokénisation. J'ai utilisé `spaCy` pour tokéniser le corpus d'évaluation. Mais le traitement du *-t-* euphonique dans l'inversion sujet-verbe n'y est pas le même que dans les données d'affinage. Ainsi, dans la séquence tokénisé *va -t -on* de (2-d), le parseur a annoté *-t* comme sujet explétif de *va* et *-on* comme son objet indirect. FUDIA ne détecte donc pas d'inversion avec un sujet appartenant à la liste des pronoms

clitiques.

Enfin, il est aussi étonnant à première vue de noter que FUDIA performe moins bien sur l'écrit que sur l'oral, et ce d'autant plus que le score LAS [8] du parseur affiné sur de l'écrit était supérieur à celui affiné sur de l'oral. Au vu des phrases et des types d'erreurs discutés ci-dessus, cela s'explique surement par la plus grande complexité en moyennes des phrases du sous-corpus écrit (longueur, structures enchâssées, relations à longue distance, etc.).

6 Conclusion

Je répondrais à la question du titre par l'affirmative. Les corpus arborés nous permettent de produire des programmes, comme FUDIA, qui détectent les interrogatives du français avec un bon score. Cependant, la tâche reste difficile, et FUDIA peut être moins performant sur des données bruitées ou très variables (ex. productions d'enfants (Gillet, 2022)). De plus, les performances (dont celles des parseurs) dépendent beaucoup de la qualité des annotations en entrée. C'est pourquoi la pérennité, le maintien et la révision constante des corpus arborés me semblent encore une des tâches essentielles de notre discipline.

Remerciements

Je remercie très fortement Bruno Guillaume, †Guy Perrier et Sylvain Kahane pour leur aide et leurs commentaires sur mon travail.

Références

AVANZI M., BÉGUELIN M.-J., CORMINBOEUF G., DIÉMOZ F. & JOHNSEN L. A. (2012). OFROM – corpus oral de français de Suisse romande.

BALLY A.-S. (2022). Les interrogatives totales en français québécois dans l'écrit SMS : à la croisée de l'oral et de l'écrit. In F. NEVEU, P. PRÉVOST, A. STEUCKARDT, G. BERGOUNIOUX & B. HAMMA, Éds., *8e Congrès Mondial de Linguistique Française*, volume 138, p. 12006, Orléans : SHS Web of Conferences. DOI : 10.1051/shsconf/202213812006.

BENZITOUN C. (2022). Évolution des interrogatives partielles directes en français : le cas de combien. In *8e Congrès Mondial de Linguistique Française*, volume 138 de *Syntaxe*, p. 13003, Orléans : SHS Web of Conferences. DOI : 10.1051/shsconf/202213813003.

BENZITOUN C., DEBAISIEUX J.-M. & DEULOFEU H.-J. (2016). Le projet ORFÉO : un corpus d'étude pour le français contemporain. *Corpus*, **15**(15). DOI : 10.4000/corpus.2936.

8. *labeled attachment score* : pourcentage de tokens ayant la bonne étiquette et le bon gouverneur.

BONFANTE G., GUILLAUME B. & PERRIER G. (2018). *Application of Graph Rewriting to Natural Language Processing*, volume 1. ISTE Wiley.

BOSCO C. & SANGUINETTI M. (2014). Towards a Universal Stanford Dependencies parallel treebank. In V. HENRICH, E. HINRICHS, D. DE KOK, P. OSENOVA & A. PRZEPIÓRKOWSKI, Éds., *Proceedings of the 13th Workshop on Treebanks and Linguistic Theories (TLT-13)*, p. 14–25, Tübingen (Germany).

BOURIGAULT D., FABRE C., FRÉROT C., JACQUES M.-P. & OZDOWSKA S. (2005). Syntex, analyseur syntaxique de corpus. In *Actes Des 12èmes Journées Sur Le Traitement Automatique Des Langues Naturelles*, Dourdan, France : Association pour le Traitement Automatique des Langues.

CANDITO M. & SEDDAH D. (2012). Le corpus Sequoia : Annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *Actes de La 19e Conférence Sur Le Traitement Automatique Des Langues Naturelles*, p. 321–334, Grenoble, France : Association pour le Traitement Automatique des Langues.

COVENEY A. (2011). L'interrogation directe. *Travaux de linguistique*, **63**(2), 112–145.

DE MARNEFFE M.-C., MANNING C. D., NIVRE J. & ZEMAN D. (2021). Universal Dependencies. *Computational Linguistics*, **47**(2), 255–308. DOI : 10.1162/coli_a_00402.

DEFRANCQ B. (2005). *L'interrogative enchâssée. Structure et interprétation*. Champs linguistiques. Louvain-la-Neuve : De Boeck Supérieur.

DELAVEAU A., CAPPEAU P. & DAGNAC A. (2021). Les phrases interrogatives. In A. ABEILLÉ & D. GODARD, Éds., *La Grande Grammaire du Français*, volume 2, p. 1402–1437. Arles : Actes Sud/Imprimeries nationales Éditions, 1 édition.

ESHKOL-TARAVELLA I., BARBEDETTE A., LIU X. & SOUMAH V.-G. (2022). Classification automatique de questions spontanées vs. préparées dans des transcriptions de l'oral. In Y. ESTÈVE, T. JIMÉNEZ, T. PARCOLLET & M. ZANON BOITO, Éds., *Traitement Automatique Des Langues Naturelles*, p. 305–314, Avignon, France : ATALA.

GILLET P. (2022). Développement du langage de l'enfant : L'exemple des interrogatives partielles. In *8e Congrès Mondial de Linguistique Française*, volume 138, p. 13002, Orléans : SHS Web of Conferences. DOI : 10.1051/shsconf/202213813002.

GUILLAUME B., DE MARNEFFE M.-C. & PERRIER G. (2019). Conversion et améliorations de corpus du français annotés en Universal Dependencies. *Revue TAL*, **60**(2), 71.

HABERT B. (2005). Text corpus of "Le Monde". DOI : https ://www.islrn.org/resources/421-401-527-366-2/.

KAHANE S., CARON B., STRICKLAND E. & GERDES K. (2021). Annotation guidelines of UD and SUD treebanks for spoken corpora : A proposal. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, p. 35–47, Sofia, Bulgaria : Association for Computational Linguistics.

LACHERET A., KAHANE S., BELIAO J., DISTER A., GERDES K., GOLDMAN J.-P., OBIN N., PIETRANDREA P. & TCHOBANOV A. (2014). Rhapsodie : un Treebank annoté pour l'étude de l'interface syntaxe-prosodie en français parlé. In *4e Congrès Mondial de Linguistique Française*, volume 8, p. 2675–2689, Berlin, Germany : SHS Web of Conferences. DOI : 10.1051/shsconf/20140801305.

LEFEUVRE F. (2021). Les interrogatives averbales dans la presse, stratégies discursives récurrentes ? *Langue francaise*, **212**(4), 107–122.

LEFEUVRE F. & ROSSI-GENSANE N. (2017). Les interrogatives indirectes en discours informel oral. *Langue française*, **196**(4), 51–74. DOI : 10.3917/lf.196.0051.

MCDONALD R., NIVRE J., QUIRMBACH-BRUNDAGE Y., GOLDBERG Y., DAS D., GANCHEV K., HALL K., PETROV S., ZHANG H., TÄCKSTRÖM O., BEDINI C., BERTOMEU CASTELLÓ N. & LEE J. (2013). Universal Dependency Annotation for Multilingual Parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, p. 92–97, Sofia, Bulgaria : Association for Computational Linguistics.

PÉRY-WOODLEY M.-P., AFANTENOS S., HO-DAC L.-M. & ASHER N. (2011). La ressource ANNODIS, un corpus enrichi d'annotations discursives. *Revue TAL*, **52**(3), 71.

REINHARDT J. (2016). Établir un corpus oral de questions : L'analyse semi-automatisée avec Praat et Perl à l'exemple de cinq épisodes de Maya l'Abeille. In *5e Congrès Mondial de Linguistique Française*, volume 27, p. 11007, Tours : SHS Web of Conferences. DOI : 10.1051/shsconf/20162711007.

REINHARDT J. (2019a). La transmission du manque d'information dans la téléréalité française. In *Transmission, oubli et mémoire dans les sciences du langage, JC2017 - 20èmes Rencontres des jeunes chercheurs en Sciences du Langage*, Paris, France.

REINHARDT J. (2019b). Les interrogatives directes tirées de dix romans policier. DOI : https ://hdl.handle.net/11403/interrogatives-in-novels/v1.

ROSSI-GENSANE N., CÓRDOBA L. F. A., URSI B. & LAMBERT M. (2021). Les structures interrogatives directes partielles fondées sur *où* dans les dialogues de romans français du XXe siècle. *Journal of French Language Studies*, **31**(2), 169–191. DOI : 10.1017/S0959269520000253.

SEDDAH D. & CANDITO M. (2016). Hard Time Parsing Questions : Building a QuestionBank for French. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, p. 2366–2370, Portorož, Slovenia : European Language Resources Association (ELRA).