Évaluation d'un corpus annoté en anaphores : le cas des chaines contenant un mot interrogatif

Valentin D. Richard
Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
ILLC, Universiteit van Amsterdam, Amsterdam, The Netherlands
valentin.richard@loria.fr

RÉSUMÉ

Ce papier évalue l'annotation des anaphores dans le corpus de français parlé ANCOR en se concentrant sur celles contenant un mot interrogatif. À l'aide d'une méthode semi-automatique, nous montrons que beaucoup de chaines anaphoriques incluant un syntagme interrogatif sont manquantes. L'annotation des questions en *quel(le)(s)* $\bar{N}$ est homogène et régulière, mais les autres mots interrogatifs sont absents ou annotés avec peu de cohérence.

ABSTRACT

Evaluating Chains Containing an Interrogative Word in an Anaphorically Annotated Corpus

This paper evaluates the annotation of anaphora in the spoken French corpus ANCOR, focussing on anaphora containing an interrogative word. Using a semi-automatic method, we show that many anaphoric chains that include an interrogative phrase are missing. The annotation of questions with *quel(le)(s)* $\bar{N}$ ('*which* $\bar{N}$') is homogeneous and regular, but the other interrogative words are absent or annotated with little consistency.

MOTS-CLÉS : anaphore, coréférence, mot interrogatif, annotation.

KEYWORDS: anaphora, co-reference, interrogative word, annotation.

Introduction L'anaphore est une relation entre plusieurs expressions dont l'interprétation sémantique dépend d'un même référent (Partee, 2014). Ces expressions sont appelées mentions et leur relation une chaine. Les mots interrogatifs peuvent faire partie d'une chaine anaphorique. Leur capacité à introduire un nouveau référent du discours a été mise en avant dans de nombreux travaux formels (van Rooij, 1998; Groenendijk, 1998; Haida, 2007; Li, 2020; Roelofsen & Dotlačil, 2023; Richard, 2024). Par exemple, dans le discours (1), *il* fait référence à l'élève introduit par *quel élève* dans la question précédente.

(1) A : [Quel élève]^k était assis là ? Il$_k$ a oublié son sac.

ANCOR (Muzerelle *et al.*, 2014) (488.000 mots) est un corpus de conversations en français

oral annoté en anaphores. Même si les mots interrogatifs ne sont pas explicitement mentionnés dans le guide d'annotation, le corpus comporte des chaines incluant un mot interrogatif. L'objectif de cet article est d'évaluer ces annotations quantitativement et qualitativement.

Méthode Dans un premier temps, nous avons extrait automatiquement un sous-corpus d'ANCOR composé des segments de discours contenant un mot interrogatif. Pour cela, nous avons parsé le corpus en syntaxe (schéma Universal Dependencies) grâce à ArboratorGrew (Guibon *et al.*, 2020) (modèle neuronal affiné sur des corpus oraux, LAS $= 0,8180$) et FUDIA (Richard, 2023), en préservant la segmentation et la tokénisation données par ANCOR. [1] Nous nous intéressons aux anaphores dont le référent est un individu ou un lieu. Parmi les 2580 mots interrogatifs détectés, nous ne retenons donc que les 745 ayant pour lemme *qui*, *quoi*, *où*, *quel* ou *lequel*. [2] Sur ces 745 mots, seulement 502 sont annotés par ANCOR comme faisant partie d'une mentions, et 347 de ces derniers appartiennent à (au moins) une chaine non triviale selon ANCOR.

Pour savoir combien de chaines réelles contenant un interrogatif sont manquantes dans ANCOR, nous avons annoté à la main les extraits contenant les 243 occurrences détectées ne faisant pas partie d'une mention et les 155 occurrences annotés en mention mais ne faisant pas partie d'une chaine ou étant dans une chaine triviale (c.à.d. une chaine ayant une seule mention). Dans la suite, nous rapportons les résultats en excluant les faux positifs (erreurs d'extraction, ex. usage non interrogatif).

Anaphores avec un interrogatif dans ANCOR ANCOR contient deux types de chaines incluant un mot interrogatif. Il y a des co-références, c.à.d. des chaines où toutes les mentions dénotent le même individu, ainsi que des anaphores associatives, où les mentions dénotent des individus dont l'interprétation est liée (ex. relation ensemble/élément, comme *[Les enfants]^j sont dans la cour. [Jean]$_{j'}$ joue au ballon*). [3] Ce sont principalement les syntagme nominaux (SN) en *[quel $\bar{N}$]* qui sont annotés. Dans la majorité des cas, ce SN fait partie d'une question qui est répondue par une phrase de la forme *c'est* X. Le démonstratif *ce* est annoté en co-référence avec *[quel $\bar{N}$]* et le SN X en anaphore associative avec *[quel $\bar{N}$]*, comme dans (2).

1. Pour plus d'informations sur les processus mentionnés ici, voir le répertoire https://github.com/Valentin-D-Richard/ANCOR_eval. Il contient les scripts, une description plus détaillée des annotations, ainsi des échantillons des chaines manquantes.

2. Aucune occurrence de *que* ou *qu'est-ce que* n'a été annotée en mention dans ANCOR dans le sous-corpus extrait. De plus, au vu du trop grand nombre de ces occurrence, nous avons choisi de les exclure de cette étude.

3. Les indices j indiquent les annotations issues d'ANCOR, les k sont nos propre annotations. On note l'indice en exposant pour l'antécédent (trait NEW=YES), et les anaphores associatives avec un prime (ex. $\textsc{assoc}(j, j')$).

Lemme	*quel*	*qui*	*quoi*	*où*	*lequel*	Total
Occurrences en chaine non triviale	319	5	0	0	4	328
Occurrences en mention seule	101	39	5	0	0	145
dont manquantes	38	36	3	0	0	77
Occurrences pas annotées en mention	98	9	38	40	0	185
dont manquantes	51	7	21	30	0	109
Total des occurrences interrogatives	518	53	43	40	4	658
dont vraiment impliquées dans une chaine	408	48	24	30	4	514
dont manquantes	89	43	24	30	0	186

TABLE 1 – Nombre d'occurrence des mots interrogatifs selon les annotations d'ANCOR.

(2) a. SPK1 : madame lorsque vous étiez encore à l'école dans [quelle matière]^{j} étiez vous le plus fort ? (ANCOR:021_C-6)
b. SPK2 : ah c$_{j}$'était littérature$_{j'}$.

Annotations manquantes Nous comptabilisons un grand nombre de mots interrogatifs non annotées par ANCOR pourtant liés anaphoriquement à d'autres expressions dans leur contexte droit (fenêtre de 20 segments). Par exemple, dans (3), le premier *qui* est annoté comme une mention mais aucune chaine d'ANCOR ne relie *qui* à *c'* ou à sa réponse *moi*. C'est un exemple de mention seule *manquante*.

(3) a. SPK1 : et qui^{k} est-ce qui remplissait les les f- les papiers administratifs
b. SPK2 : ah oui c$_{k}$'est moi$_{k'}$ (ANCOR:021_C-6)

Le tableau 1 recense les occurrences de mots interrogatifs selon leur annotation. Au total, environ 36 % des mots interrogatifs présents dans une anaphore sont manquants. Mais cela cache une grande disparité selon les lemmes. Par exemple, parmi les 51 occurrences de *qui* véritablement impliquées dans une chaine anaphorique, seulement 10 % sont présentes dans ANCOR, alors que pour *quel(le)(s)*, ce taux est de 78 %. Au final, seulement 22 % des mots interrogatifs considérés ne sont pas référés par un individu ou un lieu dans la suite du discours.

Problèmes qualitatifs et possibles explications Les chaines avec *quel* et *lequel* sont plus régulièrement identifiées dans ANCOR. Cela s'explique surement par le fait que ces deux mots dont D-liés (Pesetsky, 1987). Notamment une question en *quel* $\bar{N}$ ou *lequel* a typiquement une présupposition d'existentialité. Cette question demande l'identification d'un individu spécifique, dont le locuteur sait qu'il existe. Historiquement, l'annotation des anaphores s'est concentré sur les mentions spécifiques. La meilleure identification des syntagmes en *quel* ou *lequel* comme référentiels résulte donc surement de cette tradition d'annotation. Les

occurrences de *quel* manquantes sont pour la plupart des cas de *quel* adjectival ou *quel genre de* $\bar{N}$.

La grande proportion de chaines manquantes avec *qui*, *quoi* ou *où* suggère que les mots interrogatifs, dans leur ensemble, n'ont pas reçu une grande attention lors de l'annotation. Certes, certains occurrences participent à des relations moins communes et moins connues, telles les réponses non exhaustives, comme en (4). D'autres cas complexes incluent les interrogatives enchâssées ou les mentions dans la portée d'un opérateur modal ou d'une conditionnelle, par exemple en (5). Pourtant, on y retrouve le schéma typique de la réponse (potentielle) *(c'est)* X, bien annoté dans l'exemple (2). De plus, de telles relations apparaissent aussi avec *quel* et sont annotées par ANCOR comme ce que nous proposons dans (4) et (5).

(4) a. SPK2: il y a quoi^k dedans (ANCOR:1AP0073)
b. SPK1: c'est un guide d'informations sur la ville donc [ce qu'on peut trouver dans la ville]$_{k'}$ [les musées]$_{k''}$ [les choses comme ça]$_{k'''}$

(5) tout dépend où^k se trouve euh enfin profèsse le commerçant si c$_k$'est dans [à l'intérieur de la ville]$_{k'}$ [au centre de la ville]$_{k'}$ disons que sa mentalité diffère peu d'un d'une employée de bureau par exemple si c'est un commerçant qui est [à l'extérieur de la ville]$_{k''}$ qui se trouve [dans les quartiers ouvriers]$_{k'''}$ alors là euh son sa mentalité diffère (ANCOR:542_C-2)

De plus, certaines annotations d'ANCOR avec *qui* présentent des divergences. Parfois, la coréférence concerne le domaine de quantification plutôt que le référent, comme dans (6). Dans d'autres cas, le SN réponse est annoté en coréférence avec le syntagme interrogatif, et non pas en anaphore associative, ex. (7).

(6) il y a [des gens]^j que vous ne connaissez pas et que vous entendez parler parmi ceux-ci euh qui$_j$ est -ce qui parle le meilleur français ? (ANCOR:005_C-2)

(7) a. SPK1: qui^j est-ce qui remplit les papiers administratifs les feuilles d'impôts ?
b. SPK2: euh [une personne extérieure]$_j$ (ANCOR:079_C-2)

Conclusion En somme, bien qu'ANCOR ait établit un solide cadre d'annotation de chaines anaphoriques avec *quel*, les annotations avec les autres syntagmes interrogatifs à référent individuel ou de lieu sont manquantes pour la plupart ou présentent des incohérences. Établir un guide unifié pour les mots interrogatifs permettrait de combler ces lacunes.

Références

GROENENDIJK J. (1998). Questions in update semantics. In *Formal Semantics and Pragmatics of Dialogue : Proceedings of the Thirteenth Twente Workshop on Language Technology (Twendial '98)*, p. 125–137, Twente : Universiteit Twente, Faculteit Informatica.

GUIBON G., COURTIN M., GERDES K. & GUILLAUME B. (2020). When Collaborative Treebank Curation Meets Graph Grammars. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 5291–5300, Marseille, France : European Language Resources Association.

HAIDA A. (2007). *The Indefiniteness and Focusing of Question Words*. Thèse de doctorat, Humboldt University, Berlin.

LI H. (2020). *A Dynamic Semantics for Wh-Questions*. Thèse de doctorat, New York University.

MUZERELLE J., LEFEUVRE A., SCHANG E., ANTOINE J.-Y., PELLETIER A., MAUREL D., ESHKOL I. & VILLANEAU J. (2014). ANCOR_Centre, a large free spoken French coreference corpus : Description of the resource and reliability measures. In N. CALZOLARI, K. CHOUKRI, T. DECLERCK, H. LOFTSSON, B. MAEGAARD, J. MARIANI, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, p. 843–847, Reykjavik, Iceland : European Language Resources Association (ELRA).

PARTEE B. (2014). Formal Semantics and Typology of Anaphora.

PESETSKY D. (1987). Wh-In-Situ : Movement and Unselective Binding. In E. REULAND & A. TER MEULEN, Éds., *The Representation of (In)Definiteness*. Cambridge, England : MIT Press.

RICHARD V. D. (2023). Est-ce que l'extraction des interrogatives du français peut-elle être automatisée ? In K. FORT, C. GARDENT & Y. PARMENTIER, Éds., *5èmes journées du Groupement de Recherche CNRS " Linguistique Informatique, Formelle et de Terrain " (LIFT 2023)*, p. 69–76, Nancy, France : CNRS.

RICHARD V. D. (2024). Dynamic Effects of Modalized Questions. In F. CARCASSI, T. JOHNSON, S. BRINCK KNUDSTORP, S. DOMÍNGUEZ PARRADO, P. RIVAS-ROBLEDO & G. SBARDOLINI, Éds., *Proceedings of the 24th Amsterdam Colloquium*, p. 289–307, Amsterdam, The Netherlands.

ROELOFSEN F. & DOTLAČIL J. (2023). Wh-questions in dynamic inquisitive semantics. *Theoretical Linguistics*, **49**(1-2), 1–91. DOI : 10.1515/tl-2023-2001.

VAN ROOIJ R. (1998). Modal Subordination in Questions. In J. HULSTIJN & A. NIJHOLT, Éds., *Formal Semantics and Pragmatics of Dialogue. Proceedings of Twendial '98*, p. 237–247, Enschede : University of Twente.