Actes de CORIA-TALN 2026
Nantes, France, 29 juin au 3 juillet 2026

Quand le score F1 cache des métriques différentes : le cas de la détection de citations

Kirill Milintsevich[1] Agnès Saulnier[1]
(1) Institut national de l'audiovisuel, 4 avenue de l'Europe, 94360 Bry-sur-Marne, France
kmilintsevich@ina.fr, asaulnier@ina.fr

RÉSUMÉ

L'évaluation de la détection de citations repose souvent sur des scores agrégés tels que le score F1. Pourtant, des protocoles d'évaluation différents peuvent produire des scores similaires tout en mesurant des propriétés distinctes des prédictions, ou au contraire conduire à des scores différents pour un même ensemble de sorties. Cet article compare plusieurs métriques et protocoles d'évaluation utilisés dans la littérature à partir d'un cadre descriptif fondé sur quatre dimensions (unité évaluée, appariement, comparaison locale et agrégation). Nous analysons empiriquement l'impact de ces choix en appliquant différentes métriques à un même ensemble de prédictions, sur des configurations d'erreurs contrôlées et sur des données réelles issues du corpus FRACAS. Les résultats montrent que les scores obtenus peuvent varier sensiblement selon la métrique retenue. Ces enjeux deviennent particulièrement importants dans le cas des modèles génératifs.

ABSTRACT

When the F1 Score Conceals Different Metrics : The Case of Quotation Detection.

Evaluating quotation detection often relies on aggregated scores such as the F1 score. However, different evaluation protocols can yield similar scores while measuring distinct properties of the predictions, or conversely lead to different scores for the same set of outputs. This paper compares several metrics and evaluation protocols used in the literature within a descriptive framework based on four dimensions (evaluation unit, matching, local comparison, and aggregation). We empirically analyze the impact of these choices by applying different metrics to the same set of predictions, under controlled error configurations and on real data from the FRACAS corpus. The results show that the scores obtained can vary substantially depending on the metric chosen. These issues become particularly important in the case of generative models.

MOTS-CLÉS : Détection de citations, évaluation, score F1, modèles génératifs.

KEYWORDS: Quotation Detection, Evaluation, F1 Score, Generative Models.

1 Introduction

La détection automatique des citations occupe une place centrale dans de nombreux travaux en traitement automatique des langues, en analyse du discours et en sciences humaines et sociales. Les citations structurent les textes journalistiques, politiques et narratifs ; elles permettent d'identifier les prises de parole, les rapports d'autorité et les mécanismes de médiation discursive. Leur extraction automatique est ainsi mobilisée dans des tâches variées telles que l'attribution de propos, l'analyse des sources, la mesure de la visibilité médiatique ou encore l'étude des controverses.

Actes des 33ème Conférence sur le Traitement Automatique des Langues Naturelles.
Volume 1 : articles scientifiques originaux, pages 454–469.

Au cours des dernières années, les approches proposées pour la détection des citations se sont diversifiées : des systèmes à règles aux modèles discriminatifs de type BERT, jusqu'aux modèles génératifs récents (Pareti, 2016; Zhang & Liu, 2022; Richard *et al.*, 2024). Pourtant, si les méthodes évoluent rapidement, les pratiques d'évaluation demeurent hétérogènes et rarement discutées de manière explicite. Les performances sont le plus souvent résumées par un score unique – typiquement un F1 – sans que les choix sous-jacents d'unité d'évaluation, de protocole d'appariement ou de seuil de recouvrement soient systématiquement analysés.

S'agissant des grands modèles de langage (LLM), peu de travaux abordent l'extraction de citations et leur attribution dans leur définition stricte (Michel *et al.*, 2025). En revanche, plusieurs auteurs s'en servent pour ancrer les textes générés dans des informations factuelles sous forme de citations, notamment dans un contexte de génération augmentée par la recherche (RAG) ou de question–réponse fondée sur des preuves (Xiao *et al.*, 2024; Schreieder *et al.*, 2025; Bezerra & Weigang, 2025).

Une citation peut être définie, dans son acception la plus générale, comme la reprise d'un discours attribué à une source et intégré dans un nouveau discours. Elle correspond ainsi à une opération de médiation discursive : un locuteur rapporte les propos d'un autre, explicitement ou implicitement (Maingueneau, 2012). Dans les textes journalistiques et narratifs, cette opération se matérialise le plus souvent par une configuration minimale à trois composantes (Pareti, 2016) :

— un contenu rapporté (*quote*),
— une source à laquelle ce contenu est attribué,
— un marqueur d'introduction (*cue*) signalant l'acte de parole (par exemple un verbe déclaratif).

Du point de vue computationnel, évaluer la détection de citations consiste à comparer une prédiction produite par un système automatique à une annotation de référence. Cette évaluation peut porter sur plusieurs dimensions : l'identification des segments citationnels, la délimitation de leurs frontières, la segmentation de citations discontinues ou encore l'association entre le contenu rapporté et les entités qui lui sont liées (source, marqueur introducteur). Ces aspects relationnels ne sont toutefois pas l'objet du présent article, qui se concentre sur la détection des segments citationnels.

Par ailleurs, la citation constitue un objet linguistique structurellement instable. Selon les conventions d'annotation, une même séquence peut être analysée comme une citation unique ou comme plusieurs segments, et ses frontières varient en fonction de la ponctuation, des incises et des commentaires. Dès lors, les scores rapportés dans la littérature — souvent résumés par un F1 — peuvent refléter autant des différences de protocole d'évaluation que de véritables différences de performance des systèmes. L'évaluation dépend également de la définition opérationnelle de la citation adoptée dans les corpus annotés : des divergences peuvent ainsi refléter des différences de définition plutôt que de véritables erreurs de détection. Cette question dépasse toutefois le cadre du présent article (voir Saulnier (2026)).

Ce constat soulève une question méthodologique centrale : que mesure réellement un score rapporté pour la détection de citations, et dans quelle mesure ces scores sont-ils comparables d'un corpus à l'autre ?

Dans cet article, nous analysons les pratiques d'évaluation utilisées pour la détection de citations et leur impact sur l'interprétation des performances, en particulier lorsque celles-ci sont résumées par F1-score. Notre contribution est triple :

1. une analyse comparative des protocoles d'évaluation existants ;
2. la proposition d'un cadre descriptif pour caractériser ces protocoles ;
3. une analyse empirique de l'impact de ces choix d'évaluation sur un même ensemble de

Corpus	Langue	Domaine	Source	Cue	Quote	Types de citation
FRACAS (2024)	FR	presse	✓	✓	✓	D, I, M
PARC 3.0 (2016)	EN	presse	✓	✓	✓	D, I, M
NewsQuote (2023)	EN	presse	✓	✓	✓	D, I, M
DE-News (2024)	DE	presse	✓	✓	✓	D, I, IL
QuoteBank (2021)	EN	presse	✓	✗	✓	D
DirectQuote (2022)	EN	presse	✓	✗	✓	D
FI-News (2023)	FI	presse	✓	✗	✓	D, I
AADS French (2023)	FR	romans	✗	✗	✓	D

TABLE 1 – Vue d'ensemble des jeux de données existants dédiés à l'extraction et à l'attribution des citations. Pour les types de citation, **D** désigne « directe », **I** désigne « indirecte », **IL** désigne « indirecte libre », et **M** désigne « mixte ».

prédictions.

Ce travail ne propose ni nouveau système ni nouvelle métrique. Il vise à clarifier les dimensions constitutives de l'évaluation en détection de citations afin de favoriser une comparaison plus rigoureuse des approches existantes, notamment dans le contexte récent des modèles génératifs.

Le code et les données utilisés dans ce travail sont disponibles à `https://github.com/ina-foss/quoteval`.

2 État de l'art sur les datasets existants

Les corpus mobilisés pour l'étude automatique des citations constituent des ressources essentielles pour l'évaluation des systèmes. Ils se distinguent toutefois par des choix méthodologiques et théoriques hétérogènes, tant dans la définition de la citation que dans les schémas d'annotation retenus. Table 1 synthétise les principales ressources utilisées dans la littérature, en distinguant les corpus annotés manuellement, conçus comme vérités terrain, des ressources dérivées, issues d'extractions automatiques à grande échelle.

Des corpus tels que PARC (Pareti, 2016), FRACAS (Richard *et al.*, 2024) et DE-News (Petersen-Frey & Biemann, 2024) reposent sur des annotations explicites des citations et de leurs sources. Ils partagent un objectif commun d'attribution du discours rapporté, mais diffèrent sensiblement dans leur conception. PARC privilégie une modélisation des événements de discours (parole, pensée, écriture), les distinctions entre discours direct et indirect étant reconstruites a posteriori. FRACAS adopte une définition volontairement large de la parole rapportée, intégrant paraphrases et commentaires, tandis que DE-News propose un schéma particulièrement expressif, intégrant plusieurs rôles discursifs, au prix d'un coût annotatif plus élevé.

À l'opposé, certaines ressources se concentrent exclusivement sur le discours direct canonique. DirectQuote (Zhang & Liu, 2022) constitue un corpus gold annoté au niveau des jetons,[1] avec un alignement explicite des locuteurs, mais son périmètre est volontairement restreint. QuoteBank (Vaucher *et al.*, 2021), quant à lui, fournit une base de citations à très grande échelle extraite automatiquement à partir de guillemets et de règles d'attribution. En l'absence d'annotation manuelle

1. Jeton, aussi appelé *token* en anglais : unité élémentaire produite par la tokenisation.

et de modélisation des citations indirectes ou des marqueurs introducteurs, il ne peut être considéré comme une vérité terrain, mais plutôt comme une ressource d'exploitation ou de pré-entraînement.

D'autres corpus, comme le corpus finnois de Janicki *et al.* (2023) ou AADS French (Durandard *et al.*, 2023), illustrent des objectifs encore différents. Le premier vise principalement l'attribution des citations dans la presse, sans formaliser systématiquement les marqueurs introducteurs, tandis que le second s'inscrit dans une perspective littéraire, centrée sur la détection du discours direct et la séparation narration/parole, sans annotation fine de l'attribution.

Cette diversité de définitions, de schémas et de périmètres rend les comparaisons inter-corpus délicates. Elle souligne la nécessité d'expliciter les hypothèses théoriques et les choix d'annotation sous-jacents à chaque ressource, en particulier lors de l'évaluation et de la comparaison des systèmes de détection et d'attribution des citations.

Ces constats rejoignent les conclusions d'une étude multi-corpus récente (Zhong *et al.*, 2024), qui montre que les résultats obtenus sur un corpus donné ne peuvent être directement comparés ni transférés à d'autres, en raison de divergences structurelles entre les ressources.

3 État de l'art sur les mesures d'évaluation

3.1 Mesures d'évaluation utilisées pour les corpus de citations

La détection de citations ne dispose pas encore d'un protocole d'évaluation pleinement standardisé. La majorité des travaux s'appuie néanmoins sur des métriques issues des tâches d'extraction de segments textuels, en particulier les mesures de précision, de rappel et le score F1, appliquées à des correspondances exactes ou partielles des frontières des citations.

La plupart des travaux antérieurs utilisent une approche d'étiquetage séquentiel IOB pour extraire le segment de citation (Pareti, 2016; Zhang & Liu, 2022; Zhang *et al.*, 2023; Janicki *et al.*, 2023; Durandard *et al.*, 2023; Petersen-Frey & Biemann, 2024). Dans le schéma d'étiquetage IOB, chaque jeton de la séquence d'entrée reçoit l'une des étiquettes suivantes : *I* (Inside), *O* (Outside), ou *B* (Beginning). Ces étiquettes sont ensuite utilisées pour reconstruire les bornes des segments.

Puisque cette approche est répandue dans les systèmes de reconnaissance d'entités nommées (NER), les métriques d'évaluation suivent des conventions similaires. Pour évaluer la qualité d'extraction des segments, trois variantes de la métrique sont rapportées : stricte, partielle et souple. En appariement strict (ou exact), un segment est considéré comme correctement prédit uniquement si tous ses jetons correspondent à ceux d'un segment de référence. En appariement partiel, un segment prédit est évalué proportionnellement au nombre de jetons qui se recouvrent avec le segment de référence (par exemple, le taux de recouvrement au niveau des jetons). Enfin, en appariement souple, un segment est considéré comme correctement prédit s'il présente un recouvrement avec le segment de référence.

Les résultats d'appariement sont généralement résumés en termes de précision et de rappel, où la précision est la proportion de segments correctement prédits parmi tous les segments prédits, et le rappel la proportion de segments correctement prédits parmi tous les segments de référence. Pour construire le score final, la précision et le rappel peuvent être moyennés au niveau micro (*micro-averaging*) ou au niveau macro (*macro-averaging*).

En résumé, la plupart des travaux rapportent des scores F1 stricts (Pareti, 2016; Zhang *et al.*, 2023;

Durandard *et al.*, 2023; Richard *et al.*, 2024) et des scores F1 partiels (Pareti (2016) utilisent un moyennage micro, Zhang *et al.* (2023) un moyennage macro, et Petersen-Frey & Biemann (2024); Richard *et al.* (2024); Durandard *et al.* (2023) ne documentent pas explicitement la méthode de moyennage). Janicki *et al.* (2023); Zhang & Liu (2022) rapportent la précision, le rappel et le F1, mais ne documentent pas explicitement les méthodes de moyennage et d'agrégation utilisées.

Outre les scores fondés sur le recouvrement de jetons, Durandard *et al.* (2023) est le seul travail à rapporter une métrique supplémentaire, Zone Map Error (ZME) (Galibert *et al.*, 2014), qui distingue différentes catégories d'erreurs (omission, split, merge), offrant un diagnostic complémentaire au score global.

3.2 Mesures d'évaluation issues d'autres tâches de NLP

Des tâches proches rendent explicites certaines dimensions que l'évaluation des citations traite souvent implicitement. En diarisation de locuteurs, la métrique principale est le Diarization Error Rate (DER), qui repose sur une assignation optimale entre segments prédits et segments de référence et agrège les erreurs en un taux global (Bredin, 2017). Ce cadre explicite notamment la structure d'appariement adoptée.

D'autres mesures utilisées dans ce domaine, telles que *Coverage* et *Purity*, reposent au contraire sur une logique de recouvrement local maximal (*best overlap*), sans contrainte d'exclusivité globale. Plusieurs segments peuvent ainsi contribuer partiellement au score d'un même segment opposé, ce qui les rend structurellement plus tolérantes aux phénomènes de *split* et de *merge* que les assignations exclusives 1–1. En représentant une citation discontinue comme un même groupe de segments, ces métriques permettent de mesurer si les fragments annotés et prédits sont correctement regroupés. Ces métriques sont notamment implémentées dans la bibliothèque `pyannote.metrics` (Bredin, 2017). Ces cadres montrent que la tolérance aux splits, merges ou micro-décalages ne dépend pas du F1 en lui-même, mais du protocole d'appariement et de recouvrement qui le sous-tend.

L'essor des modèles génératifs introduit des configurations supplémentaires pour l'évaluation. Lorsque les sorties produites ne sont pas directement alignées sur les offsets du texte source, des métriques de similarité textuelle telles que ROUGE (Lin, 2004), METEOR (Banerjee & Lavie, 2005) ou BERTScore (Zhang *et al.*, 2019) peuvent être utilisées pour comparer les segments générés aux segments de référence.

Ces métriques évaluent une proximité lexicale ou sémantique entre séquences, mais ne permettent pas de vérifier la conformité à une annotation par segment (*span*) ni la cohérence structurelle entre entités (quote, source, cue). Elles répondent donc à une problématique différente de celle des protocoles d'évaluation structurés.

Des approches récentes explorent également l'usage de modèles de langage comme évaluateurs (*LLM-as-a-judge*), fondés sur un jugement sémantique global. Ces méthodes soulèvent toutefois des enjeux de reproductibilité et restent difficilement comparables aux protocoles d'évaluation structurés (Gu *et al.*, 2024).

Au-delà de la détection des citations, l'extraction de la source et de l'indice (*cue*) souffre des variations d'annotation. Par exemple, inclure l'apposition avec le nom de la personne ou omettre l'auxiliaire *avoir* au passé composé peut rester logiquement correct. On peut l'accommoder via une métrique de *head-match*, courante en coréférence, où un segment est correct si sa tête syntaxique correspond

à celle du segment de référence (Novák *et al.*, 2024). Cette approche requiert toutefois une analyse syntaxique, source d'erreurs et parfois indisponible pour les langues faiblement dotées.

4 Enjeux méthodologiques et limites des protocoles actuels

La plupart des travaux en détection automatique de citations rapportent leurs performances à l'aide des métriques classiques précision, rappel et F1. Cette homogénéité apparente masque toutefois des différences importantes : selon les corpus et protocoles, l'évaluation peut porter sur des segments citationnels, des sources, des marqueurs introducteurs ou sur les relations entre ces entités. Deux scores F1 rapportés dans la littérature peuvent ainsi reposer sur des protocoles d'évaluation différents et ne pas mesurer exactement le même phénomène. Un même score F1 peut alors agréger des phénomènes linguistiques distincts, tels que des erreurs de segmentation des citations, des divergences référentielles dans l'attribution des sources ou des variations morpho-syntaxiques dans les marqueurs introducteurs.

Ces différences sont également liées aux protocoles d'appariement adoptés (matching exact, recouvrement partiel, assignation 1–1 ou non exclusive), qui déterminent la sensibilité du score aux décalages de frontières, aux phénomènes de split ou de merge, ou encore aux variations de surface.

L'usage récent de modèles génératifs introduit en outre de nouvelles configurations : les sorties peuvent comporter des reformulations, ne pas fournir d'offsets explicites, ou produire des structures seulement partiellement alignables avec les annotations de référence. Les protocoles d'évaluation existants, conçus pour des sorties structurées directement alignées sur le texte source, ne sont pas toujours adaptés à ces formes de génération.

La question ne se limite donc pas au choix d'une métrique particulière, mais concerne la structure même du protocole d'évaluation : unité considérée, stratégie d'appariement, degré de tolérance aux divergences de surface et phénomène linguistique effectivement mesuré. Rendre explicites ces dimensions est nécessaire pour interpréter correctement les scores rapportés et comparer des protocoles d'évaluation qui utilisent des indicateurs similaires. C'est dans cette perspective que nous proposons, dans la section suivante, un cadre descriptif permettant de formaliser ces dimensions et de situer différents protocoles d'évaluation dans un espace commun.

5 Une décomposition analytique des protocoles d'évaluation

Les protocoles d'évaluation utilisés pour la détection automatique de citations reposent sur une diversité de choix méthodologiques souvent implicites. En conséquence, des scores rapportés à l'aide d'un même indicateur — notamment le F1-score — peuvent renvoyer à des réalités évaluatives sensiblement différentes. Les écarts observés entre systèmes reflètent ainsi non seulement des différences algorithmiques, mais également des divergences dans les protocoles d'évaluation adoptés.

Dans ce travail, nous ne proposons ni nouvelle métrique ni nouvelle procédure d'évaluation. Notre objectif est de fournir un cadre descriptif permettant (i) d'expliciter les choix d'évaluation existants, (ii) de situer différents protocoles dans un espace commun, et (iii) d'analyser empiriquement l'impact de ces choix à prédictions constantes.

Ressource / Système	D0	D1	D2	D3
FRACAS (2024)	Segment	1–1 exclusif	EM (segment)	Micro P/R/F1 (strict/boundaries)
Radar de Parité (2023)	Segment	1–1 exclusif	Recouvrement $\geq \theta$	Micro P/R/F1
PARC 3.0 (2016)	Segment	1–1 exclusif	EM (segment)	P/R/F1
DE-News (2024)	Segment	1–1 exclusif	Recouvrement (par jeton)	Micro P/R/F1
QuoteBank (2021)	Segment	1–1 exclusif	EM (segment)	P/R/F1 (subset)
DirectQuote (2022)	Jeton	–	EM (jeton)	P/R/F1
FI-News (2023)	Jeton	–	EM (jeton)	P/R/F1 + exactitude de source
AADS French (2023)	Jeton + segment	Non exclusif	Recouvrement (segment + jeton)	P/R/F1 + ZME
Diarisation (2017)	Segment/temps	Non exclusif	Recouvrement directionnel	Coverage/Purity

TABLE 2 – Positionnement des protocoles d'évaluation (D0–D3).

5.1 Cadre descriptif des protocoles d'évaluation

Nous proposons un cadre descriptif qui explicite ces choix, sans introduire de nouvelle métrique ni prescrire de standard.

Dans le cas des modèles génératifs, une étape préalable peut être nécessaire afin de réaligner les segments produits sur le texte source et de reconstruire leurs offsets. Ce pré-traitement conditionne la comparabilité des unités évaluées, tout en restant distinct du protocole d'évaluation à proprement parler.

L'évaluation peut alors être décrite selon quatre dimensions indépendantes :

— **D0 — Unité évaluée** : niveau auquel porte l'évaluation (jeton, segment, relation entre entités ou segmentation de texte) ;
— **D1 — Appariement** : mise en correspondance entre prédictions et références, soit 1—1 exclusif (chaque prédiction associée à au plus une référence), soit non exclusif (plusieurs correspondances partielles possibles), ou absent au niveau des jetons ;
— **D2 — Comparaison locale** : critère de correspondance entre une prédiction et une référence, fondé sur une correspondance exacte (*Exact Match, EM*), un recouvrement partiel (*overlap*) ou un recouvrement directionnel (évaluation séparée de la couverture d'un segment par l'autre) ;
— **D3 — Agrégation globale** : mode de combinaison des comparaisons locales en un score global (par exemple Précision / Rappel / F1).

Ce cadre (D0—D3) ne définit pas une nouvelle méthode ; il constitue un outil descriptif pour comparer des protocoles existants.

5.2 Positionnement des protocoles existants

Toute stratégie d'évaluation peut être représentée comme une combinaison particulière des dimensions D0–D3. Le Tableau 2 positionne plusieurs ressources et protocoles fréquemment utilisés dans cet

espace. Au-delà des dimensions D0–D3, ces ressources se distinguent par leur focalisation empirique : FRACAS met l'accent sur les relations et liens entre entités ; le Radar de Parité (aussi appelé F1@θ dans cet artcile) illustre une variante sensible au réglage du seuil de recouvrement sur les frontières ; PARC 3.0 cible l'attribution relationnelle (événements de discours) ; DE-News intègre une structure multi-rôles autour des citations ; QuoteBank s'inscrit dans un pipeline d'extraction à grande échelle (sous-ensemble évalué) ; DirectQuot combine détection du contenu de citation et alignement du locuteur ; FI-News évalue la citation et l'exactitude d'attribution ; AADS French privilégie l'analyse fine des structures d'erreurs de segmentation (via ZME) ; enfin, le cadre de diarisation se focalise sur la fragmentation (*split*/*merge*) et la couverture temporelle (Coverage/Purity).

Ce positionnement montre que des protocoles reposant tous sur un F1-score peuvent différer fortement selon l'unité évaluée, la stratégie d'appariement ou le critère de recouvrement. Les benchmarks de détection de citations privilégient majoritairement des appariements stricts 1–1, tandis que des approches non exclusives ou directionnelles — courantes dans des tâches voisines comme la diarisation du locuteur — restent peu exploitées.

Le tableau inclut également des métriques ne reposant pas sur un F1-score, telles que Coverage et Purity, qui mesurent le recouvrement entre segments prédits et segments de référence. Contrairement aux métriques fondées sur un comptage discret de TP/FP/FN, elles reposent sur des proportions de contenu couvert et ne nécessitent pas d'appariement explicite entre prédictions et références.

Les métriques telles que DE-News (Petersen-Frey & Biemann, 2024) produisent un score global de détection des citations considérées comme unités, tandis que le cadre AADS (Durandard *et al.*, 2023) propose une évaluation fondée sur le recouvrement du contenu citationnel et fournit des diagnostics détaillés des divergences de segmentation.

5.3 Evaluation empirique des protocoles

Afin d'examiner concrètement l'impact des choix d'évaluation présentés précédemment, nous menons une analyse empirique contrôlée de différents protocoles. Les prédictions restent strictement identiques pour l'ensemble des mesures comparées : seule la fonction d'évaluation varie. Les différences observées doivent donc être interprétées comme des effets de *régime évaluatif*, et non comme des variations de performance algorithmique.

L'analyse repose sur deux types de données. D'une part, des lots synthétiques conçus pour isoler des phénomènes spécifiques (erreurs de segmentation ou de formulation des citations) ; chaque lot regroupe quelques phrases (cinq exemples par configuration) construites de manière contrôlée. D'autre part, nous appliquons les mêmes protocoles d'évaluation à des données réelles, en utilisant le corpus FRACAS (Richard *et al.*, 2024).

Les lots synthétiques visent à couvrir des variations susceptibles d'influer sur l'évaluation (déplacements de frontières, chevauchements, *splits*, *merges*, variations textuelles), non pas pour simuler l'ensemble des sorties possibles d'un système, mais pour caractériser le comportement des métriques sous des proportions d'erreurs maîtrisées. Les lots synthétiques ne proviennent pas directement du corpus FRACAS, celui-ci ne pouvant être redistribué publiquement. Ils ont été construits manuellement à partir de formulations journalistiques et de structures de citations proches de celles observées dans FRACAS, afin de conserver une compatibilité méthodologique avec les expériences réalisées sur ce corpus, utilisé à la fois pour l'entraînement des modèles et les évaluations sur données réelles.

Les lots couvrent les phénomènes suivants (voir Annexe A.1 pour les exemples) :

1. **Perfect match** : les segments prédits correspondent exactement aux segments de référence ;
2. **No match** : aucune citation n'est correctement détectée ;
3. **Boundary shift** : la citation détectée recouvre largement la citation de référence mais présente un décalage de frontières (par ex. début ou fin décalés, suppression d'un déterminant, suppression d'un guillemet, ajout d'une virgule ou d'un mot) ;
4. **Containment (inclusion)** : la citation prédite est entièrement contenue dans la citation de référence, ou inversement ;
5. **Crossing overlap (chevauchement non inclusif)** : les segments se chevauchent partiellement sans relation d'inclusion (par ex. début correct mais fin qui dépasse, fin correcte mais début trop tard, recouvrement partiel) ;
6. **Split** : une citation de référence est divisée en plusieurs segments prédits ;
7. **Merge** : plusieurs citations de référence sont fusionnées en un seul segment prédit ;
8. **Nested** : une citation de référence est entièrement incluse dans une autre citation de référence, le système ne prédit qu'un seul des deux segments imbriqués ;
9. **Text Formatting** : variations purement typographiques à contenu inchangé (p. ex. capitalisation intégrale, changement de casse, présence/absence de diacritiques, ponctuation ou espaces), l'hypothèse restant textuellement équivalente à la référence ;
10. **Text Modification** : micro-modifications morpho-syntaxiques conservant l'information (p. ex. changement de temps/aspect ou de forme verbale, *sera appliquée* → *va être appliquée*), sans déplacement intentionnel des frontières ;
11. **Paraphrase** : variations lexicales de sens équivalent (*sera appliquée* → *sera mise en œuvre*, *dès* → *à compter de*), susceptibles d'affecter les mesures de similarité textuelle.

Nous appliquons ensuite un modèle génératif (Mixtral 8×7B [2], finetuné sur le corpus FRACAS) à la détection de citations sur le fichier de test du corpus FRACAS. Le choix s'est porté sur ce modèle en raison de la possibilité d'affiner localement le modèle génératif à l'aide des données FRACAS, sans les redistribuer. Ce choix permet également de conserver un cadre expérimental reproductible avec des ressources computationnelles académiques limitées. Les sorties générées sont d'abord réalignées sur le texte source afin de reconstruire leurs offsets. Les segments qui ne peuvent pas être alignés (hallucinations, sorties non *parsables*) sont exclus de l'évaluation et comptabilisés séparément (voir Annexe A.2).

Les métriques structurales (F1@θ, DE-News F1, AADS, Coverage/Purity) sont calculées sur les segments ainsi réalignés, afin d'évaluer les frontières des citations indépendamment des variations textuelles introduites par le modèle. Pour les prédictions qui n'ont pas pu être parfaitement réalignées sur le texte source, nous retenons comme prédiction finale la plus longue sous-chaîne commune entre la citation prédite et le texte d'origine. Dans ces cas, nous rapportons en outre le taux d'erreur moyen au niveau caractère (CER) entre cette plus longue sous-chaîne commune et la citation prédite, afin d'estimer le taux de corruption textuelle. L'Annexe A.3 présente les détails d'implémentation des métriques.

Le Tableau 3 présente les résultats des différentes métriques, tant sur les lots synthétiques que sur l'ensemble de test FRACAS. On observe que le choix du seuil θ peut entraîner des écarts marqués dans les cas *Boundary Shift* et *Split*. Les F1 à appariement exclusif 1–1 au niveau des segments (DE-News) et au niveau des jetons (AADS F1) présentent un comportement similaire sur la plupart

2. https://mistral.ai/news/mixtral-of-experts/

Dataset	F1@0.3	F1@0.8	DE-News	AADS F1	Coverage	Purity	CER
Perfect Match	1.0	1.0	1.0	1.0	1.0	1.0	0.0
No Match	0.0	0.0	0.0	0.0	0.80	0.70	0.0
Boundary Shift	1.0	0.4	0.73	0.73	0.75	0.78	0.0
Containment	1.0	0.8	0.8	0.77	0.81	0.77	0.0
Crossing Overlap	1.0	0.8	0.78	0.78	0.8	0.77	0.0
Split	0.67	0.0	0.5	0.92	0.73	1.0	0.0
Merge	0.67	0.67	0.5	0.92	1.0	0.73	0.0
Nested	0.67	0.67	0.5	0.92	0.91	0.91	0.0
Text Formatting	0.8	0.2	0.67	0.7	0.88	0.73	9.91
Text Modification	0.8	0.0	0.49	0.49	0.86	0.64	1.53
Paraphrase	1.0	0.2	0.69	0.7	0.87	0.79	1.37
FRACAS	0.81	0.79	0.8	0.84	0.94	0.86	0.005

TABLE 3 – Comparaison de protocoles d'évaluation sur des lots synthétiques illustrant des erreurs typiques.

des lots, à l'exception de *Split* et *Merge*. Cela s'explique par le fait que l'appariement exclusif ne retient qu'une partie d'une citation scindée (ou fusionne plusieurs citations en une correspondance unique), tandis que le F1 au niveau des jetons ne tient pas compte des frontières de segment. Enfin, les métriques Coverage et Purity fournissent une indication globale de la qualité des frontières : leurs valeurs diminuent lorsque les frontières sont incorrectes (*Boundary Shift*, *Containment*, *Crossing Overlap*), même lorsque le F1 reste élevé. Elles affichent toutefois des valeurs élevées sur le lot *No Match*, ce qui peut paraître contre-intuitif ; cela tient à leur nature de métriques de segmentation, qui considèrent symétriquement les segments avec et sans citations. Nous proposons donc de les compléter par une estimation directe du recouvrement en caractères entre citations de référence et détectées, afin de quantifier plus précisément le texte manquant ou en excès, ainsi que par le nombre brut d'opérations de split et de merge.

L'analyse qualitative des sorties générées sur le corpus FRACAS met en évidence plusieurs types de divergences récurrentes entre les citations détectées et les annotations de référence : micro-décalages de frontières, inclusion ou omission de segments narratifs, fragmentation ou fusion de citations discontinues, ainsi que certaines reformulations produites par le modèle génératif. Ces écarts ne correspondent toutefois pas systématiquement à des détections entièrement erronées, plusieurs sorties demeurant partiellement ou majoritairement correctes d'un point de vue discursif. Quelques exemples représentatifs sont présentés en Annexe B.

6 Recommandations pour la communauté

Les enjeux identifiés dépassent la seule détection de citations et concernent plus largement l'évaluation de prédictions structurées fondées sur l'alignement de segments produits par un système avec une référence. Nos résultats montrent qu'un score agrégé unique est insuffisant pour caractériser ces systèmes : un même F1-score peut correspondre à des opérations évaluatives différentes selon le protocole adopté. Nous recommandons donc une évaluation multidimensionnelle distinguant détection, segmentation et fidélité textuelle.

Dimension	Question évaluative	Indicateur	Objectif
Détection	La citation est-elle correctement identifiée comme unité ?	F1@θ	Détection par seuil de recouvrement
Détection	La bonne zone citationnelle est-elle détectée ?	DE-News F1	Tolérance aux décalages de frontières
Détection	Le contenu citationnel est-il globalement récupéré ?	AADS F1	Couverture globale du contenu segmenté
Délimitation	Les limites des citations sont-elles stables ?	Δ(F1@0.3-0.8)	Sensibilité aux décalages de frontières
Délimitation	Quelle est la nature des divergences ?	ZME	Diagnostic des erreurs de segmentation
Délimitation	Les segments se recouvrent-ils globalement ?	Coverage / Purity	Quantification du recouvrement de segments
Contenu	Le texte extrait reste-t-il fidèle au texte initial ?	CER	Reformulation textuelle

TABLE 4 – Indicateurs pour l'évaluation multidimensionnelle des citations.

La Table 4 propose une grille minimale d'indicateurs permettant une comparaison plus interprétable entre systèmes. Un score basé sur les segments (par ex. DE-News) peut fournir une estimation globale de la détection des citations, mais une analyse plus fine nécessite des mesures complémentaires, telles que le F1 au niveau des jetons et les diagnostics d'erreurs d'AADS, éventuellement complétés par Coverage/Purity et, pour les systèmes génératifs, par un indicateur de fidélité textuelle comme le *CER* (taux d'erreur caractère).

Les scores rapportés dépendent fortement du protocole d'évaluation (unité, appariement, critère de recouvrement, agrégation), qui devrait être explicitement documenté. Dans les tâches structurées, il est recommandé de distinguer la détection des segments de l'évaluation des éléments associés (source, cue, relations), évalués conditionnellement aux segments correctement localisés. Dans le cas des modèles génératifs, les segments doivent être réalignés au texte source et les sorties non alignables (hallucinations) ou invalides (json invalides) devraient être rapportées.

7 Conclusion

L'évaluation de la détection de citations constitue un choix méthodologique structurant plutôt qu'un simple détail technique. La décomposition proposée selon les dimensions D0–D3 montre qu'un même indicateur peut correspondre à des opérations évaluatives différentes et que, à prédictions identiques, les scores peuvent varier substantiellement selon la métrique retenue. Les analyses empiriques confirment que ces écarts ne reflètent pas uniquement des différences algorithmiques, mais aussi des choix de protocole d'évaluation

Dans le cas des modèles génératifs, ces enjeux sont amplifiés par les variations de segmentation, les reformulations et les segments non alignables. Il devient dès lors nécessaire de distinguer explicitement localisation, cohérence structurelle et fidélité textuelle, afin de ne pas confondre erreurs de génération et erreurs de détection.

Le cadre proposé invite ainsi à considérer l'évaluation non comme une étape terminale, mais comme un objet méthodologique à part entière dans les tâches de prédiction structurée en NLP.

Remerciements

Ces travaux sont partiellement financés par le projet ANR Pantagruel (ANR-23-IAS1-0001-02) et par France 2030 dans le cadre du projet ArGiMi (n° BPI DOS0238736).

Références

BANERJEE S. & LAVIE A. (2005). METEOR : An automatic metric for MT evaluation with improved correlation with human judgments. In J. GOLDSTEIN, A. LAVIE, C.-Y. LIN & C. VOSS, Éds., *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, p. 65–72, Ann Arbor, Michigan : Association for Computational Linguistics.

BEZERRA Y. F. & WEIGANG L. (2025). Llmquoter : Enhancing rag capabilities through efficient quote extraction from large contexts. *arXiv preprint arXiv :2501.05554*.

BREDIN H. (2017). pyannote.metrics : a toolkit for reproducible evaluation, diagnostic, and error analysis of speaker diarization systems. In *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association*, Stockholm, Sweden.

DURANDARD N., TRAN V. A., MICHEL G. & EPURE E. (2023). Automatic annotation of direct speech in written French narratives. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 7129–7147, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.393.

GALIBERT O., KAHN J. & OPARIN I. (2014). The zonemap metric for page segmentation and area classification in scanned documents. In *2014 IEEE International Conference on Image Processing (ICIP)*, p. 2594–2598 : IEEE.

GU J., JIANG X., SHI Z., TAN H., ZHAI X., XU C., LI W., SHEN Y., MA S., LIU H. *et al.* (2024). A survey on LLM-as-a-judge. *The Innovation*.

JANICKI M., KANNER A. & MÄKELÄ E. (2023). Detection and attribution of quotes in Finnish news media : BERT vs. rule-based approach. In T. ALUMÄE & M. FISHEL, Éds., *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, p. 52–59, Tórshavn, Faroe Islands : University of Tartu Library.

LIN C.-Y. (2004). ROUGE : A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, p. 74–81, Barcelona, Spain : Association for Computational Linguistics.

MAINGUENEAU D. (2012). *Analyser les textes de communication*. Armand Colin.

MICHEL G., EPURE E. V., HENNEQUIN R. & CERISARA C. (2025). Evaluating LLMs for quotation attribution in literary texts : A case study of LLaMa3. In L. CHIRUZZO, A. RITTER & L. WANG, Éds., *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics : Human Language Technologies (Volume 2 : Short Papers)*, p. 742–755, Albuquerque, New Mexico : Association for Computational Linguistics. DOI : 10.18653/v1/2025.naacl-short.62.

NOVÁK M., DOHNALOVÁ B., KONOPIK M., NEDOLUZHKO A., POPEL M., PRAZAK O., SIDO J., STRAKA M., ŽABOKRTSKÝ Z. & ZEMAN D. (2024). Findings of the third shared task on multilingual coreference resolution. In M. OGRODNICZUK, A. NEDOLUZHKO, M. POESIO, S. PRADHAN & V. NG, Éds., *Proceedings of The Seventh Workshop on Computational Models of*

Reference, Anaphora and Coreference, p. 78–96, Miami : Association for Computational Linguistics. DOI : 10.18653/v1/2024.crac-1.8.

PARETI S. (2016). PARC 3.0 : A corpus of attribution relations. In N. CALZOLARI, K. CHOUKRI, T. DECLERCK, S. GOGGI, M. GROBELNIK, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, p. 3914–3920, Portorož, Slovenia : European Language Resources Association (ELRA).

PETERSEN-FREY F. & BIEMANN C. (2024). Dataset of quotation attribution in German news articles. In N. CALZOLARI, M.-Y. KAN, V. HOSTE, A. LENCI, S. SAKTI & N. XUE, Éds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, p. 4412–4422, Torino, Italia : ELRA and ICCL.

RICHARD A., ALONZO CANUL L. C. & PORTET F. (2024). FRACAS : a FRench annotated corpus of attribution relations in newS. In N. CALZOLARI, M.-Y. KAN, V. HOSTE, A. LENCI, S. SAKTI & N. XUE, Éds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, p. 7417–7428, Torino, Italia : ELRA and ICCL.

SAULNIER A. (2026). Annoter et détecter les citations : vers un cadre unifié entre linguistique et humanités computationnelles. In *Proceedings of Humanistica 2026* : Anthology of Computers and the Humanities.

SCHREIEDER T., SCHOPF T. & FÄRBER M. (2025). Attribution, citation, and quotation : A survey of evidence-based text generation with large language models. *arXiv preprint arXiv :2508.15396*.

SOUMAH V.-G., RAO P., EIBL P. & TABOADA M. (2023). Radar de Parité : An NLP system to measure gender representation in French news stories. *Proceedings of the Canadian Conference on Artificial Intelligence*.

VAUCHER T., SPITZ A., CATASTA M. & WEST R. (2021). Quotebank : A corpus of quotations from a decade of news. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, WSDM '21, p. 328–336, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3437963.3441760.

XIAO J., ZHANG B., HE Q., LIANG J., WEI F., CHEN J., LIANG Z., YANG D. & XIAO Y. (2024). Quill : Quotation generation enhancement of large language models. *arXiv preprint arXiv :2411.03675*.

ZHANG T., KISHORE V., WU F., WEINBERGER K. Q. & ARTZI Y. (2019). Bertscore : Evaluating text generation with bert. *arXiv preprint arXiv :1904.09675*.

ZHANG W., GUI L., PROCTER R. & HE Y. (2023). NewsQuote : A dataset built on quote extraction and attribution for expert recommendation in fact-checking. In *Workshop Proceedings of the 17th International AAAI Conference on Web and Social Media (Mediate 2023 : News Media and Computational Journalism Workshop)*. DOI : 10.36190/2023.22.

ZHANG Y. & LIU Y. (2022). DirectQuote : A dataset for direct quotation extraction and attribution in news articles. In N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 6959–6966, Marseille, France : European Language Resources Association.

ZHONG W., NARADOWSKY J., TAKAMURA H., KOBAYASHI I. & MIYAO Y. (2024). Who said what : Formalization and benchmarks for the task of quote attribution. In N. CALZOLARI, M.-Y. KAN, V. HOSTE, A. LENCI, S. SAKTI & N. XUE, Éds., *Proceedings of the 2024 Joint International*

Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), p. 17588–17602, Torino, Italia : ELRA and ICCL.

A Données synthétiques et protocole expérimental

A.1 Construction des lots synthétiques

Les lots synthétiques utilisés dans l'évaluation empirique contiennent chacun cinq citations construites manuellement, afin d'illustrer des configurations contrôlées d'erreurs. Chaque lot repose sur les mêmes phrases de base, pour lesquelles les segments prédits sont modifiés de manière systématique afin de simuler un type d'erreur donné. Figure 1 et tableau 5 montre les examples des lots synthétiques.

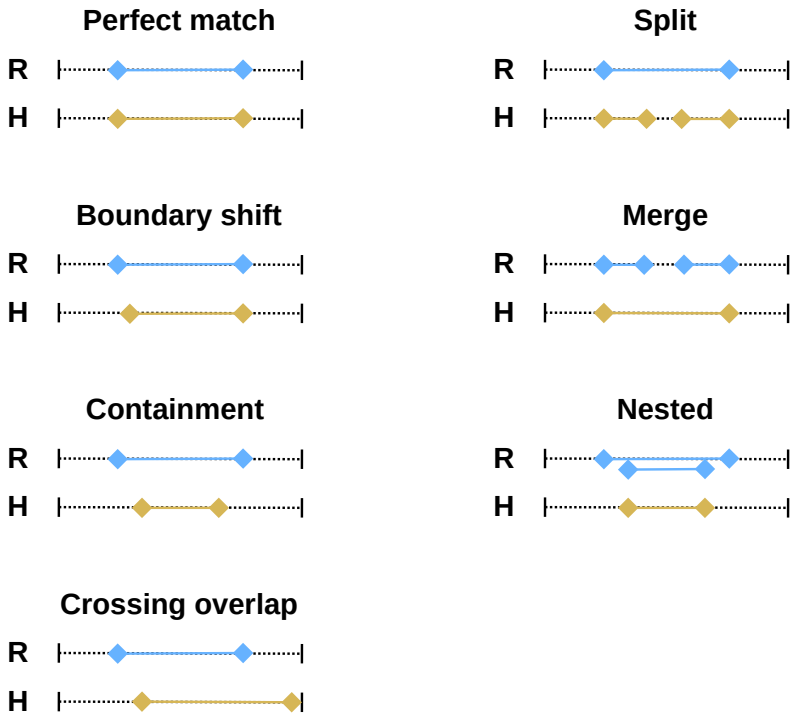

FIGURE 1 – Représentation schématique des lots synthétiques. Les lignes pointillées correspondent au texte (**R** pour la référence et **H** pour l'hypothèse), tandis que les lignes colorées indiquent les segments de citations.

A.2 Détails expérimentaux

Les expériences sur données réelles utilisent le corpus FRACAS (Richard *et al.*, 2024). Les prédictions sont obtenues à l'aide d'un modèle génératif Mixtral 8×7B finetuné pour la détection de citations.

Lot	Citation d'hypothèse
Text Formatting	LA RÉFORME SERA APPLIQUÉE DÈS JANVIER PROCHAIN
Text Modification	la réforme va être appliquée dès janvier prochain
Paraphrase	la réforme sera mise en œuvre à compter de janvier prochain

TABLE 5 – Exemples des lots synthétiques pour le texte et la citation de réference *(en italique)* : « Le ministre affirme que *la reforme sera appliqué dès jenvier prochain* malgré les critiques persistantes. »

Les segments générés sont ensuite réalignés sur le texte source afin de reconstruire leurs offsets. Les métriques structurales (F1@θ, DE-News, AADS, Coverage et Purity) sont calculées sur ces segments alignés.

Les paramètres précis d'entraînement du modèle ne constituent pas l'objet de cette étude, l'objectif étant uniquement d'obtenir des prédictions réalistes permettant de comparer différents protocoles d'évaluation.

A.3 Choix d'implémentation des métriques

Plusieurs détails d'implémentation peuvent influencer les scores obtenus, notamment la stratégie d'appariement entre segments, l'ordre d'application des seuils de recouvrement ou les modalités d'agrégation des scores.

Dans nos expérimentations :

- **AADS F1** est calculé à partir de l'implémentation officielle fournie dans le dépôt GitHub du corpus AADS [3] (Durandard *et al.*, 2023) ;
- **DE-News F1** est calculé à partir de l'implémentation disponible dans le dépôt GitHub associé à Petersen-Frey & Biemann (2024) [4] ;
- **F1@θ** est implémenté avec un appariement 1—1 basé sur le recouvrement maximal (stratégie hongrois [5]). Pour ces métriques, les paires de segments dont le recouvrement est inférieur au seuil θ sont d'abord filtrées, puis l'appariement est effectué sur les correspondances restantes ;
- **Coverage** et **Purity** sont calculés selon la définition standard utilisée dans l'évaluation de la diarisation, fondée sur les proportions de recouvrement entre segments prédits et segments de référence.
- **CER** est calculé à l'aide de l'implémentation standard disponible dans la bibliothèque `jiwer` [6].

3. https://github.com/deezer/aads_french
4. https://github.com/uhh-lt/german-news-quotation-attribution-2024
5. https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.linear_sum_assignment.html
6. https://jitsi.github.io/jiwer/

B Exemples d'erreurs de détection générées par le système génératif

Les exemples suivants illustrent des erreurs produites par le modèle génératif Mistral 8x7B, affiné sur le corpus FRACAS et évalué sur le jeu de test de ce même corpus. Dans les exemples ci-dessous, les citations détectées par le LLM sont soulignées, tandis que les **citations de référence (vérité terrain)** apparaissent en gras.

Micro-décalage de frontières

— A Gaza, le président Yasser Arafat a souligné **qu'il n'avait "aucune objection" à la nomination de Madeleine Albright au poste de Warren Christopher.**

Ambiguïté entre citation et commentaire journalistique

— Aucun chiffre n'a été divulgué, mais la commission note dans un communiqué que, **contrairement à des informations de presse, les positions de BNP Finance sur les deux lignes d'OAT impliquées, la 7,25% 2006 et la 8,50% 2008, "n'étaient pas excessives en terme d'emprise".**

Morceau de citation manquant

— **Mais "même avec ces perspectives de croissance équilibrée, le risque inflationniste ne peut être sous-estimé"**, a-t-elle poursuivi, **dans la mesure où le marché du travail est tendu avec un taux de chômage inférieur à 5,5%.**

Une citation de référence correspond à deux citations détectées

— Dans un communiqué publié en soirée, les intersyndicales du CFF et du Groupe Caisse d'épargne déclarent que **"le dossier du Crédit foncier vient d'être remis à plat". "Dès lors les deux intersyndicales estiment possible de débattre d'une solution durable pour le Crédit foncier, garantissant son intégrité, la poursuite de ses activités, son développement".**

Une citation détectée correspond à deux citations de référence

— Jean Pierre Raffarin, ministre des PME et de l'Artisanat, déclare au Figaro que **"d'autres mouvements de ce type sont envisageables"** et que **"la réussite de la grande distribution à l'international exige des groupes puissants donc des rapprochements de nécessité. Pour l'avenir ce secteur restera de toute façon hautement concurrentiel."**

Hallucination et réécriture

Référence : **"Oui. Ce sera une course très ouverte mais je cours toujours pour gagner. C'est pour ça que je suis là."**

Hypothèse : "Pour gagner ? "Oui. Ce sera une course trèss ouverte mais je cours toujours pour gagner des médailles, c'est ce qui me motive le plus."