Annoter et détecter les citations : vers un cadre unifié entre linguistique et humanités computationnelles

Agnès Saulnier[1]

[1] Institut national de l'audiovisuel, Bry-sur-Marne, France

Abstract

Quotation is a central but theoretically unstable object, situated at the intersection of linguistic, narrative, media, and computational traditions. This plurality is reflected in existing annotated corpora, which rely on heterogeneous scopes and categories, with consequences for automatic detection and evaluation. Based on a review of the literature and an analysis of existing corpora, this article shows why these divergences become problematic in computational contexts. It then proposes a typology of quotation grounded in linguistic descriptions, designed to make explicit the scope choices underlying annotated corpora. The discussion shows how this typology can guide annotation practices, support system comparison, and facilitate extensions to audiovisual data.

Mots-clés: discours rapporté, citation, corpus annotés, corpus audiovisuel, TAL
Keywords: reported speech, quotation, annotated corpora, audiovisual corpora, NLP

1 Introduction et problématique

La citation constitue un objet central des sciences humaines, mais elle ne renvoie pas, en linguistique, à une notion théorique unifiée. Elle est décrite à partir de traditions descriptives distinctes, mobilisant des critères syntaxiques, énonciatifs, narratifs, pragmatiques ou juridiques, dont les périmètres ne se recouvrent que partiellement.

Dans ce contexte, les travaux en traitement automatique du langage (TAL) ont cherché à opérationnaliser la citation à partir de divers corpus annotés, parmi lesquels PARC [12], FRACAS [16], *DE-News* [13] et *QuoteBank* [22] ont joué un rôle structurant pour la détection et l'attribution. Ces ressources reposent toutefois sur des choix de périmètre différents, ce qui devient problématique lorsqu'on cherche à comparer des systèmes ou à interpréter leurs performances.

L'objectif de cet article n'est ni de proposer une définition universelle de la citation, ni d'unifier artificiellement les corpus existants, mais de clarifier les enjeux linguistiques et computationnels liés à son opérationnalisation. Nous proposons une typologie conçue comme un outil de mise en cohérence, visant à expliciter les choix de périmètre et à fournir un cadre commun pour articuler corpus, méthodes de détection et protocoles d'évaluation, y compris en audiovisuel.

2 État de l'art : une notion plurielle et difficile à stabiliser

La notion de citation, ou plus largement de discours rapporté, occupe une place centrale dans plusieurs traditions disciplinaires sans renvoyer à un objet théorique unifié. Selon les cadres mobilisés, elle peut être décrite comme une structure morphosyntaxique, une opération énonciative, un procédé narratif, un acte de langage ou un objet juridique et médiatique. Cette pluralité se reflète dans les périmètres hétérogènes des corpus.

Agnès Saulnier. "Annoter et détecter les citations : vers un cadre unifié entre linguistique et humanités computationnelles." *Actes de la Conférence Humanistica*, éd. par Serena Crespi, Simon Gabay, Martin Grandjean, Ariane Pinche, Marie Puren et Léa Saint-Raymond. Vol. 4. Anthology of Computers et the Humanities. 2026, 37–43. https://doi.org/10.63744/UM96NSIlUNaK.

Les grammaires de référence du français décrivent le discours rapporté à partir de critères morphosyntaxiques et distinguent discours direct, discours indirect et discours indirect libre [9 ; 18]. Le discours direct et le discours indirect sont identifiés par des marqueurs formels explicites (ponctuation, subordination, concordance des temps), tandis que le discours indirect libre repose sur un faisceau d'indices linguistiques et contextuels (temps verbaux, déictiques, point de vue), sans marqueur explicite d'attribution. Dans cette approche, la citation est définie à partir de configurations linguistiques identifiables, sans prise en charge systématique des opérations de reformulation ou de la prise de position du locuteur citant.

La linguistique du discours et de l'énonciation montre que rapporter un discours ne consiste pas à le copier mais à le reformuler, que le locuteur citant prend position vis-à-vis du discours rapporté, et qu'il existe des formes intermédiaires entre citation et interprétation [2 ; 3], fréquentes dans le discours journalistique [19].

En narratologie, le discours rapporté est analysé comme un mode de représentation des paroles et des pensées, centré sur le point de vue et la mise en scène des voix dans le récit [6 ; 7 ; 8]. L'attention porte sur la frontière entre narrateur et personnage, et sur des formes telles que le discours indirect libre ou la pensée représentée.

En pragmatique, cette réflexion est prolongée par la distinction entre *oratio recta*, reprise d'un dire comme tel, et *oratio obliqua*, reprise d'un contenu reformulé [1 ; 15 ; 20]. Cette opposition permet de distinguer la reprise locale d'un dire formulé de la circulation plus abstraite de contenus propositionnels, indépendamment de leur énonciation d'origine.

Enfin, les cadres juridiques définissent la citation comme une reprise partielle d'un extrait préexistant, intégrée dans une œuvre seconde et justifiée par une finalité (critique, scientifique, etc.)[1]. La doctrine du droit d'auteur souligne en outre que cette reprise doit demeurer subordonnée au propos de l'œuvre citante et ne pas se substituer à l'œuvre source [5 ; 23]. Les études médiatiques, de leur côté, s'intéressent à la circulation de fragments discursifs autonomisés (slogans ou petites phrases) dans l'espace public [4 ; 11].

Les approches linguistiques ne proposent pas de typologie directement annotable de la citation, mais décrivent des phénomènes à un niveau conceptuel plus général (opérations de reformulation, continuum des formes, effets de point de vue). Ces descriptions constituent néanmoins un socle théorique indispensable pour élaborer des typologies opératoires adaptées aux corpus et aux méthodes computationnelles.

3 Corpus existants pour l'étude et la détection des citations

Les travaux en TAL ont donné lieu à de nombreux corpus annotés pour l'étude de la citation et de l'attribution, reposant sur des conceptions hétérogènes du discours rapporté.

Certains corpus, comme *QuoteBank* (anglais), se concentrent principalement sur les citations directes et l'identification du locuteur, sans modéliser la diversité des formes de reprise.

À l'inverse, des schémas plus structurels, tels que PARC (anglais), décrivent les relations d'attribution à partir de composantes explicites (Source, Cue, Content) et distinguent citation directe, indirecte et mixte a posteriori. Le corpus FRACAS (français) adopte une approche similaire. Par exemple :

[*Rihanna*]$_{\text{speaker}}$ [*a demandé*]$_{\text{cue}}$ [*de ne pas arrêter la musique*] $_{\text{quote}}$.

La catégorie d'indirect y couvre toutefois des réalisations plus hétérogènes.

PARC annote l'attribution de contenus relevant de la parole, de la pensée ou de l'écrit, là où FRACAS regroupe ces réalisations sous des catégories citationnelles plus larges.

Une ressource comme DE-News (allemand), adopte un schéma plus fin, distinguant types de discours rapporté (dont le discours indirect libre) et médias du dire (parole, pensée, écrit). Cette

1. Code de la propriété intellectuelle, art. L.122-5.

diversité reflète des découpages théoriques distincts, que la typologie proposée vise à rendre comparables.

4 Pourquoi la définition de la citation devient un problème computationnel

La définition de la citation n'est pas seulement un enjeu théorique : elle conditionne le fonctionnement des systèmes de détection automatique. Selon la famille de méthodes mobilisée, la notion de citation est encodée, apprise ou inférée de manière différente, avec des conséquences sur les résultats et leur interprétation.

Les systèmes à règles reposent sur l'identification de marqueurs formels explicites, tels que les guillemets, la ponctuation ou des verbes déclaratifs prototypiques, selon des patrons lexicosyntaxiques définis a priori [14]. Ils détectent efficacement les citations directes et certaines formes d'indirect, mais uniquement dans le périmètre prévu par les règles : ce qui n'a pas été anticipé (formes thématiques, indirect libre, reprises elliptiques) est mécaniquement exclu. Ce type de pipeline reste largement utilisé dans des projets en sciences humaines et sociales, notamment pour l'étude de la représentation des genres dans la presse, comme dans *GenderedNews* [17] ou *Radar de Parité* [21], où l'extraction de citations constitue une étape intermédiaire du traitement.

Les modèles discriminatifs, entraînés sur des corpus annotés, apprennent à reconnaître les classes définies par ces corpus. Leur comportement reflète la typologie locale de l'annotation, qu'il s'agisse de cadres fondés sur l'attribution [12], de corpus à grande échelle comme *QuoteBank* [22], ou de ressources plus fines telles que FRACAS [16]. Plusieurs travaux comparatifs montrent que les différences de performance entre approches à règles et modèles neuronaux tiennent principalement aux choix de définition et d'annotation de la citation, comme le montrent Janicki et al. [10] pour la presse finlandaise.

Les grands modèles de langage (LLMs) introduisent une situation différente. Sans fine-tuning, ils mobilisent une notion implicite de la citation issue de données d'entraînement hétérogènes ; après fine-tuning sur un corpus donné, leur comportement se rapproche de celui des modèles discriminatifs. La formulation du prompt encode alors implicitement une définition de la citation, susceptible d'entrer en tension avec celle du corpus.

Enfin, les scores sont comparables à l'intérieur d'un corpus, mais pas entre ressources fondées sur des définitions différentes ; la typologie proposée vise à rendre ces périmètres explicites.

5 Notre proposition : vers une typologie unifiée et étendue des reprises discursives

L'analyse de l'état de l'art et des corpus montre que la notion de citation pose moins un problème terminologique qu'un problème de périmètre. La typologie proposée vise à expliciter les formes de reprise mobilisées dans les corpus, y compris en audiovisuel. Il s'agit d'un schéma d'opérationnalisation. L'articulation générale de cette typologie est synthétisée dans la Figure 1, qui présente l'organisation des différentes formes de reprises discursives.

Nous utilisons le terme de reprise de discours pour désigner un sous-ensemble du discours rapporté correspondant aux formes donnant accès à un dire ou à une pensée attribuable et annotable. Dans ce cadre, nous excluons les discours narrativisés au sens narratologique [7], qui se bornent à mentionner une activité discursive sans en reprendre le contenu (« il a expliqué sa position »), ainsi que les formes de discours intérieur ou de pensée représentée qui ne permettent pas, dans le contexte local, de reconstruire un dire ou une pensée formulée [6]. Nous excluons également les formes d'hétérogénéité énonciative [2], qui signalent la présence diffuse de discours autres sans attribution ni contenu propositionnel identifiable (« certains diront que… »), ainsi que les évaluations journalistiques du dire qui relèvent d'une interprétation de posture discursive sans reprise effective (« elle dramatise la situation »).

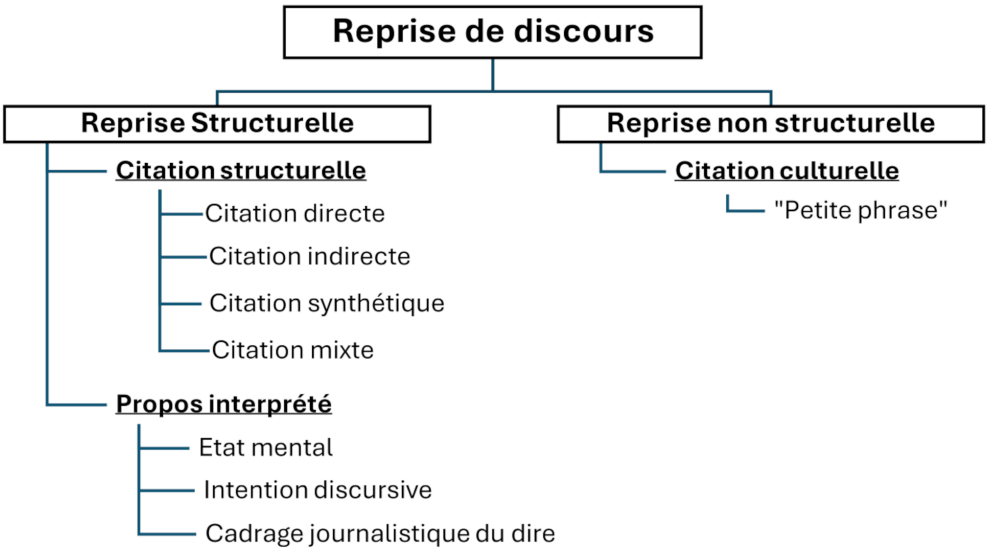

FIGURE 1 – Typologie des formes de reprise de discours.

Notre proposition se concentre sur la reprise d'un dire et repose sur un principe simple : toutes les reprises ne relèvent pas du même régime discursif. Nous distinguons deux domaines complémentaires : les citations structurelles, fondées sur l'attribution explicite d'un dire, et à des fins d'analyse médiatiques, les citations culturelles, fondées sur la reconnaissance intertextuelle de fragments discursifs issus d'un événement discursif identifié, circulant indépendamment de leur attribution locale.

Les citations structurelles peuvent être continues ou discontinues (séparées par des incises ou des commentaires narratifs), et porter sur des propos issus de l'écrit, de l'oral ou de l'audiovisuel, repris dans un texte, une transcription ou un document audiovisuel. Au sein de ce domaine, nous proposons une typologie fondée sur la nature de la reprise du dire : citation directe (reprise verbatim), citation indirecte (reformulation d'un propos formulé), citation synthétique (reprise des idées du propos) et citation mixte (combinaison de reprise littérale et de reformulation).

Cette typologie est pleinement transversale aux supports. En audiovisuel, la citation directe prend deux formes principales : l'insert audio ou vidéo d'un propos antérieur, ou la reproduction verbatim d'un propos par le narrateur. Dans le premier cas, seules les reprises d'extraits déjà rendus publics constituent des reprises de discours, ce qui exclut les prises de parole produites pour l'émission (reportage original, micro-trottoir, direct) ou de leur rediffusion. Dans le second cas, la citation est généralement signalée par des marqueurs oraux (« je cite », « selon ses propres mots »), ou par des indices prosodiques ou multimodaux (variation du registre vocal, gestes iconiques). La délimitation de la citation, notamment de sa fin, constitue un enjeu saillant. En audiovisuel, elle repose sur des indices combinés (retour à la voix initiale, rupture prosodique, changement thématique, montage). À l'écrit, la clôture des citations indirectes dépend souvent du réancrage énonciatif du narrateur plutôt que de marqueurs formels.

Au sein des reprises structurellement marquées, nous distinguons enfin une catégorie de propos interprété. Elle regroupe les cas où le narrateur attribue à une source une position, une intention ou une attitude discursive, sans rapporter explicitement un énoncé formulé, mais en interprétant le dire ou la pensée de la source. Elle recouvre notamment des attributions d'états mentaux (« il pense que… »), des intentions discursives (« elle veut convaincre ») et certaines formes de cadrage journalistique du dire (« il persiste et signe »), lorsqu'elles fonctionnent comme attributions interprétatives, sans ancrage explicite dans un dire rapporté. Cette distinction ne vise pas à améliorer la

détection, mais à isoler des cas interprétatifs difficiles pour l'annotation et l'évaluation.

Les citations culturelles constituent des reprises non structurelles relevant d'un régime de circulation intertextuelle (ex. « petites phrases »), issues d'un dire historiquement situé et reconnues indépendamment de leur attribution locale dans le document.

6 Discussion : portée et usages d'une typologie opérationnelle

La typologie proposée ne vise ni à remplacer les schémas d'annotation existants ni à imposer un standard unique pour la détection de la citation. Son principal apport est de rendre explicites et comparables les choix théoriques structurant les corpus et les tâches computationnelles. En distinguant la reprise d'un dire formulé (citation structurelle) de l'interprétation narrative d'un dire (propos interprété), elle permet de relire des ressources comme PARC, FRACAS ou *DE-News* non comme des ensembles incompatibles, mais comme des découpages différents d'un même continuum du discours rapporté. Un test d'alignement avec ces corpus permettrait ultérieurement d'en examiner la couverture.

Cette explicitation est également déterminante pour l'interprétation des résultats de détection automatique. Les divergences observées entre systèmes à base de règles, modèles discriminatifs et modèles génératifs apparaissent moins comme des limites algorithmiques que comme des effets des définitions implicites de la citation encodées dans les données d'entraînement et d'évaluation. Une typologie explicite permet de distinguer désaccord théorique et difficulté technique, et d'envisager des extensions de corpus cohérentes.

Un troisième enjeu concerne l'extension de la détection des citations à l'audiovisuel. Les pratiques médiatiques contemporaines – inserts de discours, reprises verbatim orales, circulation de « petites phrases » médiatiques ou politiques – ne peuvent être traitées sans une définition explicite des objets annotés. L'introduction de la distinction entre citations structurelles et citations culturelles constitue ici un apport central : elle permet de séparer l'attribution locale d'un dire formulé de la reconnaissance intertextuelle de fragments discursifs circulants issus d'événements médiatiques. La définition proposée de la citation directe audiovisuelle, par insert ou par reprise verbatim signalée, fournit en outre une équivalence fonctionnelle aux guillemets de l'écrit. Cette clarification ouvre la voie à la constitution de nouvelles vérités terrain audiovisuelles et au développement de systèmes de détection adaptés, tout en restant compatibles avec les corpus textuels existants.

7 Conclusion

Cet article soutient que les difficultés de détection et d'évaluation des citations tiennent moins aux méthodes qu'à l'instabilité conceptuelle de l'objet « citation ». Tant que cette notion reste implicite ou variable selon les cadres théoriques, les performances des systèmes demeurent difficiles à interpréter et à comparer.

Pour répondre à cette situation, nous avons proposé une typologie opératoire qui articule plusieurs distinctions complémentaires. Au sein des citations structurelles, la différenciation entre citations directe, indirecte, synthétique et mixte permet de mieux circonscrire le périmètre effectif de la reprise d'un dire. La séparation explicite entre citation et propos interprété clarifie les frontières entre attribution d'un propos formulé et interprétation narrative. L'introduction des citations culturelles permet enfin de rendre compte de la circulation médiatique de fragments discursifs emblématiques.

En intégrant ces distinctions et en prenant explicitement en compte l'audiovisuel, notamment à travers les inserts et les reprises verbatim orales, la typologie proposée fournit un cadre explicite pour comparer les choix d'annotation, guider l'extension des corpus et articuler définition linguistique, détection automatique et évaluation.

Références

[1] AUSTIN, John L. *How to Do Things with Words*. Oxford University Press, 1962. DOI : 10.1093/acprof:oso/9780198245537.001.0001.

[2] AUTHIER-REVUZ, Jacqueline. *Ces mots qui ne vont pas de soi*. Larousse, 1996.

[3] BAKHTINE, Mikhaïl. « Le discours d'autrui ». In : *Marxisme et philosophie du langage*. Trad. par Marina YAGUELLO. Les Éditions de Minuit, 1977, p. 161-172.

[4] BOYER, Henri et GABORIAUX, Chloé. « Splendeurs et misères des petites phrases ». In : *Mots. Les langages du politique 117* (2018), p. 9-17. DOI : 10.4000/mots.23160.

[5] CHAMBAT-HOUILLON, Marie-France et WALL, Anthony. *Droit de citer*. Rosny-sous-Bois : Bréal, 2004.

[6] COHN, Dorrit. *Transparent Minds : Narrative Modes for Presenting Consciousness in Fiction*. Princeton University Press, 1978.

[7] GENETTE, Gérard. *Figures III*. Seuil, 1972.

[8] GENETTE, Gérard. *Palimpsestes*. Seuil, 1982.

[9] GREVISSE, Maurice et GOOSSE, André. *Le Bon Usage*. De Boeck, 2016.

[10] JANICKI, Maciej, KANNER, Antti et MÄKELÄ, Eetu. « Detection and Attribution of Quotes in Finnish News Media : BERT vs. Rule-Based Approach ». In : *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*. Tórshavn, Îles Féroé, 2023. URL : https://aclanthology.org/2023.nodalida-1.6.

[11] KRIEG-PLANQUE, Alice. *La Notion de* formule *en analyse du discours : cadre théorique et méthodologique*. Besançon : Presses universitaires de Franche-Comté, 2009.

[12] PARETI, Silvia. « A Database of Attribution Relations ». In : *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*. Istanbul, Turquie, 2012. URL : https://aclanthology.org/L12-1571.

[13] PETERSEN-FREY, Fynn et BIEMANN, Chris. « Dataset of Quotation Attribution in German News Articles ». In : *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Turin, Italie, 2024. URL : https://aclanthology.org/2024.lrec-main.394.

[14] POULIQUEN, Bruno, STEINBERGER, Ralf et BEST, Clive. « Automatic Detection of Quotations in Multilingual News ». In : *Proceedings of the Internation Conference Recent Advances in Natural Language Processing (RANLP)*, sous la dir. de Galia ANGELOVA, Kalina BONTCHEVA, Ruslan MITKOV, Nicolas NICOLOV et Nikolai NIKOLOV. Borovets, Bulgarie, 2007, p. 487-492.

[15] RECANATI, François. *Oratio Obliqua, Oratio Recta : An Essay on Metarepresentation*. MIT Press, 2000. DOI : 10.7551/mitpress/5163.001.0001.

[16] RICHARD, Ange, ALONZO CANUL, Laura Cristina et PORTET, François. « FRACAS : a FRench Annotated Corpus of Attribution relations in newS ». In : *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Turin, Italie, 2024. URL : https://aclanthology.org/2024.lrec-main.654.

[17] RICHARD, Ange, BASTIN, Gilles et PORTET, François. « GenderedNews : Une approche computationnelle des écarts de représentation des genres dans la presse française ». Rapp. tech. HAL, 2022. URL : https://shs.hal.science/halshs-03604628.

[18] RIEGEL, Martin, PELLAT, Jean-Christophe et RIOUL, René. *Grammaire méthodique du français*. PUF, 1994.

[19] ROSIER, Laurence. *Le Discours rapporté : histoire, théories, pratiques*. Bruxelles / Paris : Duculot, 1999.

[20] SEARLE, John R. *Speech Acts*. Cambridge University Press, 1969. DOI : 10.1017/CBO9781139173438.

[21] SOUMAH, Valentin-Gabriel et al. « Radar de Parité : An NLP System to Measure Gender Representation in French News Stories ». In : *arXiv* (2023). DOI : 10.48550/arXiv.2304.09982.

[22] VAUCHER, Timoté, PRETI, Daniele, TOPÎRCEANU, Camelia-M. et WEST, Robert. « 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 2021)*. 2021. DOI : 10.1145/3437963.3441760.

[23] VIVANT, Michel et BRUGUIÈRE, Jean-Michel. *Droit d'auteur*. 4e édition. Paris : Dalloz, 2019.