Lingue e Linguaggi
Lingue Linguaggi 65 (2024), 385-412
ISSN 2239-0367, e-ISSN 2239-0359
DOI 10.1285/i22390359v65p385
http://siba-ese.unisalento.it, © 2024 Università del Salento

DISCOURS NUMERIQUES ET APPROCHES OUTILLÉES
Quelques réflexions sur les apports des métadonnées

CLAUDIA CAGNINELLI
UNIVERSITÀ DEGLI STUDI DI MILANO

Abstract – This paper examines the role of metadata and annotations in the analysis of digital corpora by linguistic analysis tools. The present study is grounded in a discursive approach to the analysis of linguistic data. This conceptual framework determines both the concept of the corpus adopted for the study and the methodological principles guiding its exploration. The structuring of the corpus through metadata represents a pivotal juncture between the digital constitution of the material observed and the way in which it can be explored by means of corpus analysis tools. The paper initially addresses the methodological transformations, and the epistemological implications produced by digital resources. The evolving conceptualization of empirical data and the nature of the objects of study are also highlighted. A typology of metadata is then proposed, based on two main parameters: the different types of information they represent, on the one hand, and the characteristics of the data, on the other. Specific focus is directed towards the digital discourses of Web 2.0.

Keywords: corpus; digital discourse; corpus analysis tools; corpus structure; annotations.

1. Introduction

Les évolutions technologiques qui accompagnent nos sociétés de même que la recherche scientifique, y compris les études en sciences humaines et sociales, ont transformé les données soumises à l'observation, leurs formats, les objets d'étude ainsi que les ressources et les approches méthodologiques adoptées pour les analyser. En particulier, les corpus en format numérique représentent désormais un outil fondamental et de plus en plus diffus grâce à une accessibilité facilitée aux ressources numériques pour la constitution et l'exploration des corpus (cf. Condamines 2022). Comme le remarquait Mayaffre en 2010, le corpus a gagné un rôle central et incontournable dans les études linguistiques dès le début du nouveau millénaire, au point que « la linguistique *sans* ou *hors corpus* paraît aujourd'hui une spéculation intellectuelle marginale pratiquée seulement par une minorité » (Mayaffre 2010, p. 233).

Si les corpus numériques et le recours à des analyses outillées sont de

plus en plus nécessaires à divers domaines des recherches linguistiques, il faut toutefois bien souligner que la conception, la constitution et les démarches d'analyse du corpus changent selon les cadres théoriques, les objets et les objectifs d'étude ainsi qu'en fonction des méthodes d'exploration adoptées (Condamines 2005 ; Mayaffre 2005 ; Cori *et al.* 2008). Il nous semble ainsi fructueux d'interroger les fondements théorico-méthodologiques qui caractérisent les évolutions de la conception du corpus et de ses principes d'exploration au sein du domaine où se situe la recherche du linguiste, en opérant aussi un retour réflexif à partir d'expériences personnelles d'application.

S'inscrivant dans une approche discursive qui repose sur une « conception *philologique-herméneutique* » du corpus (Rastier 2011, p. 35), notre contribution se propose d'aborder dans une perspective théorico-méthodologique la question des métadonnées dans la structuration des corpus numériques. Nous traiterons notamment leur portée dans la création des conditions nécessaires pour l'appréhension d'observables discursifs dans le cadre d'une analyse outillée. Nos réflexions prennent appui sur la recherche menée pour notre thèse (Cagninelli 2022), ainsi que sur d'autres études que nous avons réalisées sur des corpus numériques relevant de plusieurs types et genres de discours, et les prolongent. Tout d'abord, les transformations méthodologiques provoquées par l'introduction de l'informatique dans la constitution et l'exploration de corpus en linguistique seront mises en relation avec les enjeux épistémologiques qu'elles ont entraînés sur l'appréhension des données linguistiques et des objets d'étude (Section 2). Nous aborderons ensuite le rôle des métadonnées dans la structuration des corpus numériques et nous proposerons une typologie de métadonnées sur la base des informations dont elles rendent compte et de la nature des données qu'elles renseignent (Section 3). Après une bipartition principale entre métadonnées contextuelles et analytiques (3.1), nous approfondirons le cas des métadonnées de décomplexification (3.2), qui s'avèrent nécessaires pour l'étude outillée de textualités numériques du web 2.0 à la lumière de leurs spécificités distinctives. La réflexion théorico-méthodologique et l'opérationnalité de la classification proposée seront enfin illustrées dans la section 4 à travers une exemplification aussi bien de la structuration d'un corpus numérique que de ses possibles applications à des fins d'analyse.

2. Corpus numériques et explorations outillées : transformations méthodologiques et implications épistémologiques

L'informatique a marqué un tournant dans l'évolution du format et de la conception même des corpus, tout comme dans les méthodes et les outils

destinés à leur exploration. De même, elle a eu des répercussions tant sur la nature des données à soumettre à l'observation – envisagées comme les matériaux langagiers sélectionnés et abordés dans une perspective linguistique – que sur celle des objets d'étude d'un point de vue à la fois théorique et méthodologique.

Le passage au format numérique a influencé aussi bien la visualisation des données en corpus que les modes d'accès aux données. La quantité de données à examiner a augmenté notablement et les fonctionnalités de recherche intégrées aux environnements numériques en ont facilité la consultation. Ce sont toutefois les calculs statistiques et les méthodes d'analyse intégrées aux outils d'exploration exploités par la lexicométrie et plus tard par la textométrie et la logométrie qui ont jeté les bases de l'« herméneutique numérique » (Mayaffre 2002). Comme le fait remarquer Mayaffre :

> Par une lecture différente (hypertextuelle plutôt que linéaire, nous l'avons vu, mais aussi paradigmatique plutôt que syntagmatique, quantitative plutôt que qualitative), l'ordinateur voit autre chose pour déranger nos certitudes et élargir l'horizon étroit de nos *modes* (aux deux sens du terme) d'interrogation. (Mayaffre 2002, p. 155)

De surcroît, le format numérique des corpus présente l'avantage de pouvoir enrichir la matérialité linguistique des productions empiriques sélectionnées, en ajoutant toutes sortes d'informations pertinentes pour leur étude sous forme de métadonnées. Les métadonnées sont des informations que l'on peut associer aux matériaux langagiers et leur mise en corpus contribue à ce qu'elles soient appréhendées en tant que données linguistico-discursives à partir du cadre théorico-méthodologique sur lequel repose la conceptualisation et la structuration du corpus.

Prenons l'exemple d'un corpus textuel créé au format XML. Avec ce type de format, il est possible d'insérer les données linguistiques au sein d'une structure arborescente constituée de balises qui permettent d'ajouter des informations à différents niveaux. Le tableau 1 illustre comment ces informations peuvent être ajoutées sous forme de métadonnées en relation avec l'unité textuelle de base envisagée pour la constitution du corpus. La colonne gauche restitue la partie verbo-textuelle d'un discours numérique de nature « technodiscursive » (Paveau 2017) – il s'agit en l'occurrence d'un tweet[1] publié par le compte @EmmanuelMacron le 12 décembre 2020 (cf.

[1] Le tweet représente le format de base des publications via le réseau socio-numérique Twitter, rebaptisé X en juillet 2023. Considérant que les exemples proposés datent de 2020, nous adopterons ici les dénominations répandues dans la littérature académique en vigueur jusqu'au changement de nom de la plateforme. Nous précisons également que plusieurs des

infra Figure 1) –, alors que la colonne droite propose un exemple de sa mise en forme pour l'intégrer dans un corpus au format XML. Dans ce cas, la partie verbo-textuelle du tweet est structurée en relation avec des informations concernant l'identification de l'unité textuelle dans le corpus (balise *texte*), la source énonciative (*locuteur*) ainsi que la date de production (*date*).

Données linguistiques du corpus retenues à partir de la matérialité langagière de la production sélectionnée	**Données enrichies de métadonnées au format XML**
Avons-nous atteint tous nos objectifs #5AnsAprès l'Accord de Paris sur le climat ? Non. En revanche, ne sous-estimons pas le chemin parcouru. Je m'étais engagé à rendre des comptes : c'est ce que je veux faire aujourd'hui.	<texte="1"> <locuteur="EMacron"> <date="12/12/2020">Avons-nous atteint tous nos objectifs #5AnsAprès l'Accord de Paris sur le climat ? Non. En revanche, ne sous-estimons pas le chemin parcouru. Je m'étais engagé à rendre des comptes : c'est ce que je veux faire aujourd'hui.</date> </locuteur> </texte>

Tableau 1
Exemple de mise en forme d'une unité textuelle du corpus au format XML avec l'ajout de métadonnées qui définissent une structuration possible au niveau textuel.

Si la balise *texte* sert ici à reconnaître de manière univoque le tweet en tant qu'unité textuelle spécifique du corpus, les balises *locuteur* et *date* apportent des éléments d'information relevant du contexte de production du discours. Ces métadonnées peuvent ainsi être exploitées pour opérer différents types de regroupements et de comparaisons des données au service de plusieurs perspectives d'observation, à l'instar d'études qui prennent en compte la variation au fil du temps ou qui comparent les discours de locuteurs différents.

Les balises envisagées à titre d'exemple dans le tableau 1 définissent la structure d'une unité du corpus textuel, tout en façonnant en même temps la structure de base du corpus. Par ailleurs, ce dernier pourrait également prévoir des métadonnées supplémentaires aussi bien au niveau macro-textuel (à l'instar de la classification de plusieurs genres textuels – cf. section 3.1) qu'aux niveaux micro- et mésotextuels, comme c'est le cas pour l'ajout d'annotations d'ordre morpho-lexical, sémantique ou syntaxique, qui ne seront toutefois pas traitées dans la présente étude.

fonctionnalités de la plateforme ont été progressivement renommées aussi à partir de juillet 2023.

Comme nous le verrons plus loin (Section 3), il est en effet possible de distinguer plusieurs macro-catégories de métadonnées pour *informer* – au sens de modeler, de donner une forme spécifique – la matérialité langagière des productions socio-discursives. Cependant, le type et la quantité de métadonnées à prévoir dépendent aussi bien de la nature du discours sélectionné que des objectifs de l'étude et des méthodes d'analyse adoptées[2].

La constitution d'un corpus en format numérique et enrichi de métadonnées permet de multiplier non seulement les informations associées aux données, mais aussi les perspectives d'appréhension et les paramètres à prendre compte au cours de l'analyse.

Par ailleurs, les données rassemblées en corpus sont susceptibles de se diversifier, tant du point de vue de la nature de leurs constituants sémiotiques que des textualités dont elles relèvent. En effet, les technologies numériques ont également contribué à la naissance de pratiques communicationnelles nouvelles et en conséquence de données et notamment de formes de textualités tout aussi nouvelles. On peut mentionner à titre d'illustration le cas des « technodiscours » (Paveau 2017), comprenant les productions issues du web 2.0, comme les posts des réseaux socio-numériques, dont la nature est à la fois langagière et technologique, ainsi que le remarque Paveau. Ces textualités créées par le biais de dispositifs connectés présentent des traits distinctifs inédits qui appellent la recherche linguistique à questionner ses cadres épistémologiques et méthodologiques pour les étudier, à l'instar de la proposition avancée par Paveau (2017). Les technodiscours théorisés par Paveau (2017) se caractérisent en effet par un potentiel de *délinéarisation* qui peut ouvrir divers parcours de lecture grâce à la présence de « formes technolangagières », telles que les hashtags ou les liens URL, participant d'ailleurs de l'*augmentation énonciative* de ces productions. L'auteure souligne également leur nature *composite* et *relationnelle*, ainsi que l'*investigabilité* et l'*imprévisibilité* qui est propre à leur matérialité numérique.

Cependant, il est évident que les données qui ne sont pas natives – au sens de Paveau (2017) – d'un environnement numérique peuvent à leur tour donner lieu à des corpus en format numérique. Paveau (2015, en ligne, §13) distingue en effet « trois ordres linguistiques du numérique » en fonction de plusieurs critères tels que « le mode de production de l'écrit, le mode de lecture du texte et son inscription dans l'écosystème du Web ». Aux deux extrêmes de la tripartition proposée par Paveau (2015), les textualités numériques non natives ou « numérisées » s'opposent aux textualités

[2] Il faut également noter que le format du fichier à analyser doit être compatible avec les formats utilisés par le logiciel d'analyse. Cela a donc des répercussions sur la mise en forme des données, sans pour autant affecter nécessairement leur conceptualisation.

numériques natives ou « numériquées » correspondant aux technodiscours du web 2.0. Puisque les textualités numérisées résultent de textes qui sont importés en contexte numérique par des opérations de numérisation, elles ne présentent pas les traits définitoires des technodiscours cités plus haut. En l'occurrence, un processus de médiation est par conséquent nécessaire, comme c'est le cas d'ailleurs pour tout passage du médium d'origine à un autre support, à l'instar de la transcription écrite d'une production orale.

Cela ne signifie toutefois pas que les discours numériques natifs ne soient pas soumis à des opérations d'interprétation lors de la constitution du corpus. Il faut en effet rappeler que le corpus est un objet construit (cf. entre autres, Dalbera 2002 ; Mayaffre 2005 ; Garric, Longhi 2012), qui rassemble des matériaux linguistiques empiriques selon des critères de sélection établis par l'analyste en fonction d'un cadre théorico-méthodologique et d'un objet d'étude déterminés. De fait, tout groupement de textes ne correspond pas nécessairement à un corpus, selon l'acception que ce dernier assume en sciences du langage, comme le montre Duteil-Mougel (2006 ; cf. aussi Rastier 2011). Le corpus dépend des conditions de sa constitution (cf. Rastier 2011 ; Pincemin 1999, 2012). En particulier, il doit respecter un « critère d'interprétabilité » (Pincemin 2012) afin de matérialiser un « terrain » représentatif pour l'objet envisagé (Longhi 2018).

En bref, les transformations engendrées par les technologies, les dispositifs et les outils informatiques ont affecté tant les formes de la matérialité langagière – participant de différents ordres textuels (cf. *supra*) – que les modalités et les moyens de son observation. Ces transformations ont ainsi offert à l'attention du linguiste de nouvelles textualités technodiscursives – façonnées par le numérique et composées du technologique – et de nouveaux objets d'étude, dont l'appréhension est rendue possible par divers outils et méthodes d'exploration informatisée. Plusieurs chercheurs (Mayaffre 2010 ; Rastier 2011 ; Neveu 2016) soulignent les changements épistémologiques profonds entraînés par un « rapport renouvelé à l'empirique » (Rastier 2011, p. 47 ; Neveu 2016, p. 3), rendu également possible par la constitution de corpus numériques[3] et par le développement d'outils d'analyse informatiques. Comme l'affirme Rastier,

> Un noveau rapport à l'empirique change non seulement l'étendue mais la nature des faits et rend nécessaire l'innovation théorique. Il permet de produire de nouveaux faits, qui naissent pour ainsi dire de la rencontre entre de nouveaux modes d'observation et d'explication. (Rastier 2011, p. 47)

[3] Par *corpus numériques*, nous désignons les corpus en format numérique, qu'ils soient constitués ou non de textes nativement numériques.

Mayaffre (2010, p. 240) soutient en outre qu'« [a]vec les grands corpus numériques, c'est à la fois notre rapport à l'empirie du langage qui est modifié et une nouvelle heuristique des phénomènes langagiers ou discursifs qui peut être envisagée ».

Les apports des corpus numériques et des explorations outillées concernent donc plusieurs plans interreliés : (1) la constitution des ressources à observer – ce qui a des répercussions sur la taille et le format des données, le cadrage et les catégorisations théoriques associées ; (2) l'approche des données, transformée par le formatage numérique et l'exploration outillée ; (3) l'émergence des observables ; (4) les apports théoriques de ces nouveaux objets d'observation, méthodes d'observation et observables, et finalement leurs implications épistémologiques pour les études linguistiques. La mise en corpus numérique – c'est-à-dire la structuration en format électronique du matériau linguistique attesté par des réalisations authentiques de l'activité langagière – permet d'obtenir des données linguistiques empiriques analysables en fonction du cadre théorico-méthodologique de l'analyste ainsi que des objectifs de son analyse. Cette mise en corpus constitue par conséquent une étape fondamentale pour les appréhender à partir d'un cadrage et de catégorisations théoriques spécifiques, matérialisées par la structure du corpus.

La structuration du corpus peut en effet avoir des répercussions sur la conceptualisation des observables ainsi que sur leur identification. Sans approfondir la question des observables dans le cadre de cette étude, nous préciserons pourtant l'acception dans laquelle nous utiliserons ce mot qui est susceptible d'emplois variés dans la recherche en sciences du langage, ainsi que le signale Neveu (2016). Suivant les travaux de Neveu (2016) et de Constantin de Chanay et Ferron (2017), nous envisageons l'observable comme ce qui, à partir de la matérialité attestée et observée en corpus, est interprété par l'analyste comme étant pertinent et relevant de l'objet d'étude théorisé. Notre utilisation du terme *observable(s)* vise donc à attirer l'attention sur le lien de pertinence, établi par l'analyse interprétative du linguiste, entre la matérialité attestée en corpus et l'objet d'étude ancré dans un cadre théorique spécifique. Ce lien résulte de la relation entre les dimensions empirique et théorique qui est présupposée par le concept d'*observable(s)* tel que l'envisagent à la fois Neveu (2016), en s'appuyant aussi sur le travail de Culioli (1968), et Constantin de Chanay et Ferron (2017). Ces derniers mettent en outre en évidence la dimension même virtuelle des observables, qui peuvent s'actualiser par l'analyse d'un corpus, lorsqu'ils remarquent l'interrelation étroite des activités théorique et pratique de l'analyse linguistique :

> Parler d'observable implique donc deux choses : non seulement que la linguistique soit considérée comme une science empirique, mais aussi que

> l'observé soit théorisé : qu'il soit défini, qu'il soit extensible à du non encore observé, que l'on puisse établir un lien décisif entre lui et l'objet scientifique que l'on poursuit à travers lui. (Constantin de Chanay, Ferron 2017, p. 8)

L'analyse du corpus donne ainsi accès à la matérialité linguistique à partir d'une perspective déterminée, celle qui est établie par l'analyste en fonction de son objet d'étude. Par l'exploration du corpus, l'analyste met en relation l'empirique avec le théorique, en repérant les traces, les faits et les phénomènes qui caractérisent et déterminent l'objet d'étude. Il s'ensuit que les unités d'observation changent selon l'objet d'étude et que la multiplication des angles d'observation enrichit non seulement l'appréhension des données mais aussi la possibilité d'observables pertinents. La variété de perspectives d'analyse matérialisées par la structure du corpus permet en effet d'avoir accès et de vérifier la théorisation d'objets d'étude plus complexes.

La façon dont la matérialité langagière est structurée en corpus permet donc de l'appréhender différemment de celle d'une observation hors corpus. En combinant plusieurs méthodes d'exploration, on a en outre la possibilité d'étudier les mêmes données à travers des découpages différents et des angles d'approche variés concrétisés par la structure du corpus. Comme nous l'illustrerons dans la section 3, les différents types de métadonnées envisagées peuvent apporter aussi bien des informations factuelles, relatives notamment au contexte des productions ainsi qu'à leur format structurel, que des catégorisations et donc des conceptualisations théorico-méthodologiques des matériaux sélectionnés. Selon les objets et les objectifs de l'étude, les métadonnées d'ordre théorico-méthodologiques donnent corps à plusieurs niveaux d'observation de la textualité – à travers l'articulation du local et du global (cf. Rastier 2001, 2011 ; Mayaffre 2005) – tout comme à différents niveaux possibles d'analyse linguistique et (techno)discursive (cf. section 3.2). L'approche des données devient ainsi multidimensionnelle, plus riche quantitativement et qualitativement, et finalement plus complexe.

3. Structuration des corpus numériques : types et rôles des métadonnées

L'activité de structuration d'un corpus détermine son organisation interne. Pour les corpus en format numérique, cette opération se réalise par l'ajout d'informations et de descripteurs aux données, selon différentes unités d'observation et divers niveaux de découpage et regroupement. Par la mise en corpus, la matérialité langagière sélectionnée est structurée et balisée en fonction des exigences méthodologiques imposées par sa nature, de même que par l'objet d'étude envisagé. Ainsi, si les données du corpus constituent les sources primaires qui font l'objet de l'observation, les métadonnées

structurant le corpus pourraient être considérées comme des sources secondaires qui les enrichissent (cf. section 3.1), voire les complètent dans le cas de textualités technodiscursives et plurisémiotiques (cf. section 3.2).

Par conséquent, les métadonnées informent et orientent l'observation. Elles jouent un rôle clé (1) pour restituer des informations du contexte de production qui sont significatives aux fins de l'étude, (2) pour rendre compte et identifier des éléments constitutifs des textualités[4] sélectionnées, ainsi que (3) pour projeter des catégorisations théoriques sur les données. Le format numérique du corpus et notamment sa structuration par le biais de différents types de métadonnées contribuent à concrétiser une pluralité de formes de contextualisation et d'enrichissement des données. Ces informations associées aux données peuvent être de nature aussi bien factuelle, lorsqu'elles concernent en particulier le contexte de production du discours et son format structurel, que théoriques, lorsqu'elles établissent des catégorisations des données à partir du cadre théorique et de l'approche méthodologique adoptés. La « corporalité » du corpus (Mayaffre 2010) et ses matérialisations au cours de l'analyse s'avèrent donc incontournables pour l'actualisation[5] des observables. Le corpus représente ainsi

> le lieu de réalisation, et le niveau d'accès, à/(de) la prise sociocognitive et langagière du monde par le sujet, dont la matérialité langagière corporéifie la saisie. Le corpus rend compréhensible cette dynamique du niveau le plus local (palier du morphème ou mot) à l'organisation textuelle, jusqu'à l'insertion en discours, selon une pratique sociale et une orientation argumentative donnée. (Longhi, Sarfati 2018, p. 105)

La structuration du corpus contribue de manière significative à sa corporalité, en raison des différentes configurations des données qu'il est susceptible de matérialiser et d'offrir au regard de l'analyste lors de son exploration outillée. Comme nous le montrerons dans la section 3.2, la prise en compte d'éléments intégrés au corpus en tant que métadonnées permet par exemple d'analyser les caractéristiques de la partie verbo-textuelle des technodiscours en fonction de la présence de différents contenus pluritechnosémiotiques qui participent à la création du sens global du message – des images, aux vidéos, aux GIFs ou encore aux mèmes. Différentes similitudes et oppositions peuvent également ressortir des catégorisations théoriques projetées sur les données ainsi que des paramètres et des niveaux d'observation adoptés, qui peuvent se multiplier au cours de l'analyse.

[4] Précisons que notre réflexion porte sur des corpus rassemblant une ou plusieurs formes de textualité.

[5] Kerbrat-Orecchioni (2017, p. 22) rapproche la notion d'observable de celle de « signifiant *actualisé* (car en langue, nous dit Saussure, le signifiant est une "image mentale", donc non observable par définition) ». Dans notre contribution, la notion d'*actualisation* voudrait donc insister sur l'appréhension de l'observable qui, par l'analyse du corpus, acquiert une matérialité.

En ce qui concerne la structure des corpus numériques, il est important de remarquer que « le corpus est avant tout sériel […], et non nécessairement organisé linéairement » (Mayaffre 2010, p. 243). À ce propos, Mayaffre souligne que l'organisation non linéaire du corpus numérique contribue à son pouvoir heuristique. Comme « le numérique permet des lectures hypertextuelles dont la caractéristique est justement de s'affranchir du linéaire »[6] (Mayaffre 2010, p. 244), il rend possible de nouveaux modes d'appréhension des données, favorisant également l'émergence de nouveaux observables.

La structuration des données en corpus contribue ainsi à distinguer et à matérialiser des niveaux et des paramètres d'analyse qui seront essentiels pour l'étude des observables. D'où notre proposition de considérer le corpus comme un « observatoire à observables »[7] (Cagninelli 2023), à savoir un observatoire des actualisations potentielles de l'objet de recherche à partir de données authentiques attestées et contextualisées à plusieurs niveaux. En effet, si le corpus donne accès à un observé empirique potentiellement significatif et pertinent selon le cadre et les objectifs posés, ce sera ensuite l'analyste, par l'exploration du corpus, qui devra (re)construire les observables pertinents, en établissant les faits linguistiques et les phénomènes déterminant leur pertinence.

Ancrant notre recherche dans le cadre de l'analyse du discours française (ADF), une telle conception du corpus s'attache à faire face aux enjeux soulevés par des problématiques discursives. Ces dernières visent à étudier la matérialité linguistique à la lumière des influences exercées par les pratiques sociales et les déterminants extradiscursifs lors de leur production (cf. entres autres Maingueneau 2014). Notre approche présuppose en outre que la constitution du corpus et/ou son exploration adopte un principe d'hétérogénéité (Garric 2012 ; Garric, Longhi 2012) qui donne accès à la variation des pratiques et des usages discursifs. En l'occurrence, la structuration du corpus participe à la matérialisation de niveaux d'hétérogénéité au service d'une analyse combinant à son tour des méthodes diversifiées et complémentaires[8]. Le principe différentiel est d'ailleurs plutôt distinctif de l'exploration outillée de corpus numériques pour en identifier les traits caractéristiques : « À l'instar de toute entreprise de caractérisation, l'exploration de corpus est très souvent **contrastive**, puisqu'elle repose sur

[6] Mayaffre (2010, p. 244) précise toutefois que « la linéarité apparaît irréductible à la textualité et constitue le socle de sa définition ».

[7] Le développement théorique de notre proposition d'envisager le corpus comme un *observatoire à observables*, abordé en relation avec la question des observables, fera l'objet d'une étude ultérieure.

[8] Pour une description plus ample de la conception et constitution d'un corpus hétérogène, exploré par une démarche d'analyse multidimensionnelle combinée, voir par exemple Cagninelli (sous évaluation).

des comparaisons permettant de mettre au jour des différences entre les unités choisies » (Poudat, Landragin 2017, p. 30).

Dans le cadre d'un corpus conçu pour une étude discursive, les métadonnées contribuent finalement à matérialiser des niveaux d'hétérogénéité au service d'un principe différentiel, tout comme à réfléchir sur les conditions théoriques et méthodologiques pouvant favoriser la comparabilité de données hétérogènes.

3.1. Métadonnées contextuelles et analytiques

La nature des informations intégrées aux corpus sous forme de métadonnées nous amène à proposer d'en distinguer deux classes principales ayant des répercussions différentes sur l'appréhension et l'analyse des données linguistiques, mais aussi sur les possibilités d'émergence des observables. Les métadonnées peuvent relever de deux macro-catégories d'informations : des informations empiriques de nature contextuelle aux matériaux linguistiques sélectionnés et des informations analytiques résultant de concepts théoriques ou théorico-méthodologiques.

Les *métadonnées contextuelles* se rattachent à des éléments plus objectifs et empiriques attestés au niveau de l'unité d'observation et de constitution du corpus, à savoir l'unité textuelle dans le cas d'un corpus textuel. À l'intérieur de cette catégorie, nous regroupons deux types d'informations : des paramètres situationnels-factuels qui concernent les conditions de production (par ex. temps, lieu, participants, etc.), ainsi que des paramètres formels-structurels qui déterminent l'unité textuelle dans son ensemble. De fait, les discours produits au sein d'un même contexte peuvent présenter des formats structurels différents, tout en correspondant à une même forme de textualité et à un même genre (techno)discursif, à l'instar d'un tweet-réponse qui ne partage pas les mêmes caractéristiques structurelles d'un retweet. Il s'agit donc d'informations empiriques d'ordre formel-structurel relatives au macro-niveau de l'unité d'observation, qui sont cependant susceptibles de varier en fonction tant de la forme de textualité observée que du genre du discours dont elle relève, à la différence des paramètres situationnels qui ne dépendent que du contexte de production. Si les informations situationnelles et les caractéristiques formelles-structurelles constituent des paramètres objectifs de nature contextuelle-empirique, ces deux types de métadonnées contextuelles peuvent toutefois être exploités différemment dans l'analyse, puisqu'ils peuvent servir de point de départ pour la concrétisation de différents niveaux d'analyse qui ne sont pas forcément superposables (cf. *infra*).

Nous envisageons en revanche comme *analytiques*, les métadonnées qui relèvent de catégorisations et d'annotations effectuées sur la base de notions théoriques d'ordre linguistique, discursif ou technodiscursif. Du point

de vue linguistique, l'une des annotations les plus utilisées est l'étiquetage morpho-syntaxique, qui associe à chaque forme du corpus le lemme et la catégorie grammaticale correspondants. Si ce type d'informations analytiques peut résulter aussi d'une procédure informatisée intégrée au logiciel d'analyse, cela n'empêche pas d'ajouter, manuellement ou automatiquement, d'autres informations lexico-sémantiques (à l'instar de connotations positives / négatives en association à des formes d'intérêt) et syntaxiques (concernant par exemple l'aspect du verbe ou la fonction syntaxique d'un élément dans l'énoncé) sous forme de métadonnées analytiques pour des besoins d'analyse spécifiques. La classification d'un texte comme appartenant à un genre de discours particulier constitue en revanche un exemple de métadonnées analytiques d'ordre discursif. Enfin, dans le cas des textualités numériques natives, des métadonnées analytiques technodiscursives peuvent également être prévues, à l'instar de regroupements des unités textuelles en fonction du rôle primaire ou complémentaire que la partie verbo-textuelle peut assumer dans des technodiscours tels que les tweets (voir la différence entre tweets « simples » et tweets « augmentés » abordée dans la section 4)[9].

Les métadonnées analytiques actualisent un certain cadrage théorique sur les données sous forme de catégorisations théoriques ou d'annotations théorico-méthodologiques pouvant s'appuyer aussi sur d'autres métadonnées. La structuration du corpus par des métadonnées analytiques peut en effet se fonder sur des catégories théoriques préétablies et attestées par la littérature sur le sujet, ou bien sur des catégorisations en phase de théorisation et de vérification. L'intégration de métadonnées analytiques entraîne par conséquent l'activation d'un angle d'appréhension spécifique des données et pose les conditions pour accéder à un certain type de caractérisation et donc d'observables. Dans la mesure où ces métadonnées dépendent des critères de sélection et des notions adoptés par l'analyste, elles prennent appui sur des paramètres plus subjectifs au sens qu'ils sont susceptibles de varier en fonction de l'ancrage théorique. Par les métadonnées analytiques, les données sont ainsi inscrites dans une conception théorique qui concrétise une perspective d'appréhension et d'analyse spécifique.

Néanmoins, comme les métadonnées contextuelles permettent à leur tour de matérialiser des niveaux d'hétérogénéité au service d'une analyse contrastive-différentielle, elles peuvent aussi servir de base à l'élaboration de catégorisations théorico-méthodologiques, déterminant la création de métadonnées analytiques. Il est ainsi possible d'envisager des catégorisations théorico-méthodologiques à partir de métadonnées contextuelles d'ordre *situationnel* pour étudier des variations qui touchent aussi bien le niveau

[9] Pour une illustration plus ample des métadonnées analytiques discursives et technodiscursives ici mentionnées, nous renvoyons à la section 4 et notamment au tableau 2.

intergénérique – à travers des formes de textualités et des genres de discours différents – que le niveau intragénérique – à travers des textes relevant d'une même forme de textualité et d'un même genre de discours. C'est le cas par exemple pour l'étude de la variation diachronique d'un fait linguistique ou bien de la variation selon différentes instances d'énonciation. Dans les deux cas, le moment de production des discours envisagés ou la source énonciative qui en est à l'origine correspondent à des paramètres objectifs[10] pouvant être comparés à travers différentes formes textuelles et différents genres discursifs, bien que leur sélection soit soumise aux choix de l'analyste en fonction de l'objet et des objectifs de la recherche.

En revanche, les métadonnées contextuelles relatives au format *structurel* de la textualité peuvent servir de base pour l'élaboration de métadonnées analytiques visant l'étude de variations qui s'attestent au niveau intragénérique et non nécessairement au niveau intergénérique. Si ces métadonnées d'ordre structurel se rapportent à leur tour à des éléments empiriques propres à la matérialité de la production envisagée dans sa globalité, les informations structurelles ainsi catégorisées sous forme de métadonnées contextuelles peuvent changer en fonction aussi bien de la forme de textualité de la production discursive observée que de ses caractéristiques génériques. À titre d'exemple, on peut mentionner le cas des différents types de tweets existants qui se distinguent par des formats structurels spécifiques, tout en relevant à la fois d'une même forme de textualité et d'un même technogenre de discours. Le format structurel peut ainsi constituer un paramètre factuel servant de base pour établir une catégorisation théorico-méthodologique d'ordre technodiscursif sous forme de métadonnées analytiques. En effet, si le format structurel d'un tweet peut refléter une pratique (techno)discursive spécifique, il se peut qu'il n'existe pas toujours une correspondance biunivoque entre un format structurel – à savoir un type de tweet – et la pratique technodiscursive qu'il réalise (cf. exemple 3 dans la section 4).

3.1.1. Fonctions et applications des métadonnées : vérification d'hypothèses théoriques et caractérisations linguistico-discursives

Les différentes métadonnées peuvent être exploitées pour remplir diverses fonctions au cours de l'analyse du corpus. En premier lieu, l'exploration du corpus permet de vérifier des hypothèses théoriques, en se servant aussi des catégorisations théoriques proposées sous forme de métadonnées. Comme l'affirmait Culioli déjà en 1968 : « l'ordinateur ne peut que vérifier

[10] Il peut arriver qu'il ne soit pas possible de déterminer la datation exacte d'un discours ou d'en identifier la source énonciative de production. Cela n'empêche pas toutefois que ce type d'informations concerne des aspects objectifs.

l'adéquation d'une théorie en vérifiant la pertinence et la consistance d'un jeu de descripteurs, mais ne permettra jamais de faire l'économie du travail théorique » (p. 106).

La saillance de ces catégorisations attestée par l'analyse du corpus n'est pourtant que le point de départ pour caractériser ensuite les unités d'observation ainsi différenciées et comparées. À titre d'exemple, on peut mentionner l'analyse que nous avons réalisée dans un travail précédent (Cagninelli, sous évaluation) sur un corpus d'articles de journal structuré selon une distinction interne entre discours d'information et discours d'opinion, où il a été possible d'identifier des traits linguistiques sur la base desquels ressortent des divergences entre ces sous-parties. Il s'agit en l'occurrence de caractéristiques relatives à la dimension énonciative-pragmatique des productions rassemblées en corpus. Ces résultats liés à la caractérisation pragma-énonciative du corpus seraient d'ailleurs susceptibles de généralisation dans la mesure où ils relèvent de l'influence des spécificités sous-génériques sur ces productions. D'autres textes appartenant à ces mêmes sous-genres de discours pourraient ainsi présenter des caractéristiques similaires au niveau énonciatif-pragmatique. L'une des divergences montrées par cette analyse concerne en effet la présence plus ou moins significative de ressources énonciatives et modales qui participent d'une énonciation « plutôt *subjectivisée* » ou « plutôt *objectivisée* », pour reprendre la distinction proposée par Moirand (2007).

Afin de structurer les textes en corpus en fonction d'une hétérogénéité sous-générique et pragma-énonciative (Cagninelli, sous évaluation), nous nous sommes appuyée sur les catégorisations théoriques avancées par des études précédentes sur le sujet (Adam 2001 ; Moirand 2007 ; Charaudeau 2011), que nous avons prises comme référence. L'analyse effectuée sur le corpus ainsi structuré nous a permis de tester les critères et les hypothèses de classification des textes en fonction de ces deux sous-genres du discours journalistique, en clarifiant les ressources linguistiques qui en sont distinctives. Les résultats confirment et enrichissent les catégorisations théoriques existantes et ils seront utiles lors de comparaisons ultérieures pour tester l'évolution de ces traits distinctifs.

3.1.2. Fonctions et applications des métadonnées : création des conditions d'observation de nouveaux observables

Le corpus enrichi de métadonnées assure différentes perspectives d'observation, contribuant à faire ressortir des éléments distinctifs et des phénomènes qui n'avaient même pas fait l'objet d'hypothèses au stade de sa conceptualisation. L'exploration outillée du corpus exploitant les niveaux d'observation et de comparaison matérialisés par les métadonnées permet en effet d'accéder à de nouveaux observables, qui seraient difficilement

perceptibles autrement (cf. aussi Rastier 2011). Il nous semble important d'insister en l'occurrence sur l'apport heuristique du recours à des outils d'exploration informatisée (Mayaffre 2002, 2005) qui favorisent des approches variées des données (Garric 2012). Ainsi, « l'ordinateur se révèle […] un outil heuristique susceptible de reculer l'horizon de nos investigations » (Mayaffre 2002, p. 159).

En effet, ce sont les diverses formes et modalités d'appréhension et d'observation des données par le biais de l'outil informatique qui transforment le rapport de l'analyste aux données. Précisons-le tout de suite : il ne s'agit nullement de dévaloriser l'approche qualitative des données, bien au contraire. L'analyse qualitative et la démarche interprétative de l'analyste se nourrissent des observations rendues possibles par le recours à l'outil, de même que la réflexion théorique informe le corpus en concrétisant un cadrage spécifique de la matérialité langagière. L'analyste joue donc un rôle essentiel non seulement au cours de l'exploration du corpus, mais aussi lors de sa conception et de sa constitution (cf. aussi Pincemin 2020). Une exploration informatisée qui s'effectue sans sélectionner les outils et les fonctions adéquates aux données et aux objectifs envisagés, de même qu'une exploration qui ne se modèle pas au fur et à mesure que l'étude avance et que de nouveaux questionnements surgissent, peut au contraire se révéler stérile. Les apports de l'exploration outillée se rattachent au rôle de l'analyste tant dans les méthodes adoptées que dans l'interprétation et la mise en relation des résultats obtenus afin d'accéder à des observables pertinents dont l'analyse permet aussi des retombées théoriques.

3.2. Métadonnées de décomplexification

La structuration et l'analyse d'un corpus rassemblant des textualités natives de l'environnement numérique posent des défis supplémentaires, notamment si l'on envisage ces productions comme des « technodiscours » (Paveau 2017), sans pour autant se priver des apports d'une approche outillée. Afin de saisir les technodiscours dans leur complexité, la matérialité langagière dont ils se composent gagne à être appréhendée en relation avec leurs constituants non-langagiers et avec l'environnement natif de production, selon une perspective « écologique » (Paveau 2017).

D'un point de vue méthodologique, il s'avère donc nécessaire de saisir la nature complexe et composite des technodiscours (voir la section 2) à travers un processus de décomplexification initiale qui permet de distinguer les divers constituants dont dépendent au moins en partie leurs propriétés spécifiques, avant de les mettre en relation entre eux. Par conséquent, un corpus constitué de technodiscours requiert une classe supplémentaire de métadonnées pour individualiser ses composantes et les associer à la matérialité linguistique de ces productions : les métadonnées de

décomplexification. À la différence des métadonnées contextuelles qui concernent le niveau macro-textuel de l'unité textuelle retenue comme unité de base du corpus, les métadonnées de décomplexification rendent compte d'éléments constitutifs de cette unité, en décomposant la complexité de la production envisagée dans sa globalité pluritechnosémiotique. La mise au point et l'adoption de métadonnées de décomplexification relève ainsi d'une approche spécifique des textualités complexes de nature pluritechnosémiotique. La dimension linguistique de ces productions reste en l'occurrence au centre de l'attention de l'étude tout en étant observée eu égard aux spécificités technodiscursives distinctives. À titre d'exemple, la dimension relationnelle et l'augmentation énonciative impliquées par les ressources technolangagières (cf. Paveau 2017), telles que les mentions utilisées dans les posts des RSN, constituent l'une des composantes technodiscursives dont la prise en compte permet d'aborder la matérialité langagière et, plus généralement, le sens de ces productions de manière plus complète voire plus complexe (pour une plus ample illustration, voir la section 4).

En restituant les différentes composantes des technodiscours, ces métadonnées contribuent à créer les conditions pour une observation et pour des analyses outillées plus fines et complètes, qui doivent toutefois se combiner avec des démarches contextualisées et avec le retour au technodiscours dans son contexte de production (cf. aussi Longhi 2020, 2021 ; Cagninelli 2022). De fait, l'individualisation des éléments constitutifs des technodiscours intégrés sous forme de métadonnées représente une étape nécessaire pour outiller l'analyse et pour accéder enfin à la complexité de l'observé d'origine. La déstructuration de celui-ci répond ainsi à une nécessité épistémologique et méthodologique qui sert d'appui pour atteindre une approche plus exhaustive.

Dans ce cas aussi, les objectifs de la recherche déterminent les métadonnées de décomplexification à intégrer au corpus. Pour ce faire, une bonne connaissance des matériaux sélectionnés est néanmoins indispensable. Il faut connaître non seulement les spécificités technodiscursives des textualités envisagées, mais aussi leurs propriétés formelles et technologiques, qui sont également liées au format des données retenues. À cet égard, on peut mentionner la distinction posée par Paveau (2013) entre les deux perspectives méthodologiques principalement adoptées pour aborder les discours natifs du web 2.0, qui privilégient respectivement l'« extraction » ou la « contextualisation » du matériau linguistique. Chacune s'appuie, du fait de l'approche adoptée, sur des formats et des traitements informatiques différents des données linguistiques ainsi que d'autres éléments constitutifs des technodiscours pouvant être éventuellement pris en compte. Plus récemment, des études ont par ailleurs proposé d'articuler ces deux approches afin d'aborder ces textualités de manière plus complexe, comme dans le cas

des travaux de Longhi (2020, 2021). L'élaboration de métadonnées de décomplexification et leur application lors de l'exploration du corpus s'inscrivent ainsi dans le sillage de ces études visant une approche outillée plus complexe des discours numériques, où les analyses d'ordre statistique et mesurable sont effectuées et interprétées en relation avec l'observation en contexte du matériau linguistique traité informatiquement (cf. aussi Cagninelli 2022).

À l'instar des métadonnées contextuelles, les métadonnées de décomplexification peuvent également servir de base pour l'élaboration d'annotations et de catégorisations nouvelles des données. En d'autres termes, à partir des métadonnées de décomplexification, l'analyste peut établir des métadonnées analytiques, contribuant à encadrer théoriquement les données et/ou à introduire de nouveaux niveaux d'analyse et de caractérisation. L'ajout de ces annotations peut relever donc d'« une activité interprétative du chercheur. L'annotation s'apparente ainsi à ce que le fait le linguiste tous les jours : analyser, décrire, classer le matériau langagier » (Née, Fleury 2017, p. 85). Il faut toutefois préciser que les annotations auxquelles nous nous référons ici ne correspondent ni aux annotations morphosyntaxiques ou sémantiques (Née, Fleury 2017), ni aux annotations complexes (Poudat, Landragin 2017). Elles matérialisent plutôt des classements d'éléments au niveau des unités textuelles ou des unités d'observation sur la base de l'un ou plusieurs des paramètres ainsi concrétisés. Les métadonnées de décomplexification et les annotations qui s'ensuivent (pour illustration voir *infra* Tableau 2) donnent corps à plusieurs formes de mise en contexte des matériaux verbaux, enrichissant ainsi l'interprétation.

4. Corpus numériques et métadonnées : exemples de structuration et d'application pour l'analyse

Nous proposons ici quelques exemples de métadonnées que l'on pourrait envisager dans le cadre d'un corpus de tweets[11], représentant une forme de textualité que nous avons déjà étudiée[12] (parmi d'autres, voir Cagninelli 2022, 2024). En ce qui concerne les métadonnées contextuelles, on pourrait

[11] Rappelons que, à la suite du changement de nom de la plateforme Twitter en X en juillet 2023, les tweets ont été renommés « posts ». Comme nous l'avons précisé plus haut, nous utiliserons en l'occurrence les dénominations employées à l'époque où remontent les exemples sur lesquels prend appui notre illustration de quelques cas d'analyse.

[12] Comme nous l'avons déjà souligné, l'élaboration des métadonnées analytiques et de décomplexification requiert une bonne connaissance de la textualité – ou des textualités – dont relèvent les données du corpus. D'où la nécessité d'illustrer notre propos à partir de textualités qui ont déjà fait l'objet de nos recherches.

par exemple retenir des paramètres temporels à l'instar de la date de publication. Cela permettrait d'observer la variation diachronique en tenant compte également d'éventuels événements marquant la période sélectionnée. Comme dans tout partitionnement, l'unité sélectionnée pour segmenter le corpus pourra toutefois avoir des répercussions sur le niveau de granularité des résultats (cf. aussi Rastier 2011). Une autre catégorie de métadonnées contextuelles pourrait correspondre au type de dispositif (un ordinateur ou un portable) employé pour la création du message afin d'analyser l'influence du support matériel sur les caractéristiques linguistiques de la production technodiscursive.

Relativement aux métadonnées analytiques, on pourrait en revanche supposer des catégorisations qui relèvent de notions théoriques transversales aux formes de textualités, à l'instar d'un partitionnement du corpus en sous-parties d'ordre générique et sous-générique. En l'occurrence, la création de métadonnées analytiques pourra procéder aussi d'annotations ou de catégorisations élaborées à partir de métadonnées contextuelles (comme dans le cas de la catégorisation des différentes pratiques technodiscursives à partir des types de tweets distingués en fonction de leur format structurel ainsi que de leur dimension pragma-énonciative, cf. *infra* Tableau 2 et exemple 3) et notamment de métadonnées de décomplexification. Celles-ci peuvent rendre compte d'éléments constitutifs des tweets à l'instar de la présence de ressources technolangagières spécifiques telles que les hashtags et les mentions, la présence d'un lien, d'une image ou d'autres contenus médias, ou encore des taux d'engagement relatifs au nombre de réponses, de retweets, de citations et ainsi de suite, pour ne citer qu'eux. Toutes ces informations constitutives pourraient donc devenir de nouveaux paramètres à prendre en considération – seuls ou en combinaison entre eux – pour caractériser la matérialité langagière des technodiscours et examiner les variations qui s'ensuivent.

Ce sera néanmoins la problématique de la recherche qui déterminera le type de métadonnées de décomplexification à envisager ou encore l'ajout d'annotations supplémentaires à partir de celles-ci. L'analyste devra définir et délimiter ces catégorisations qui s'appuient sur des éléments constitutifs attestés. Les métadonnées de décomplexification jouent ainsi un rôle de premier plan dans la déconstruction initiale de la complexité technodiscursive et dans sa reconstruction au fil de l'analyse.

Afin de donner une illustration concrète de la façon dont des métadonnées de décomplexification peuvent servir de base pour l'élaboration de métadonnées analytiques apportant une catégorisation théorique – déjà attestée dans la littérature ou en cours d'élaboration – sur les données empiriques, nous revenons sur la distinction que nous avons proposée entre tweets « simples » et tweets « augmentés » (Cagninelli 2022). Cette distinction vise à rendre compte de deux macro-catégories de structuration

pluritechnosémiotique du tweet en fonction de la présence ou absence d'éléments qui augmentent sa sémiose, à l'instar des liens URL et des différents types de contenus visuels ou audiovisuels pouvant être intégrés dans les tweets. Elle prend appui sur une réflexion théorique concernant le rôle du composant verbo-textuel – c'est-à-dire le message linguistique affiché dans la partie supérieure du tweet – dans la création du sens de l'ensemble du technodiscours.

Notre hypothèse était que les tweets qui ne comportent qu'une partie verbo-textuelle – que nous avons appelés tweets « simples » – pourraient généralement être plus autosuffisants sur le plan sémantique et sémiotique, puisque leur composante langagière n'entre pas en relation avec d'autres éléments sémiotiques pour la création du sens global. Au contraire, la composition plurisémiotique des tweets augmentés pourrait plus fréquemment aboutir à des contenus verbo-textuels sémantiquement incomplets, mais dont le sens est saturé par d'autres contenus sémiotiques. Cette hypothèse de catégorisation théorique a été matérialisée sous forme de métadonnées analytiques – par l'opposition tweets simples / tweets augmentés – à partir de métadonnées de décomplexification rendant compte de la présence de contenus sémiotiques supplémentaires. Les tweets dont les métadonnées de décomplexification attestaient la présence de liens, de contenus iconiques et audiovisuels ont ainsi servi comme point de départ pour la création de la macro-catégorie des tweets « augmentés ». Par ricochet, les métadonnées de décomplexification signalant l'absence de contenus sémiotiques supplémentaires au texte ont permis de classer les tweets « simples ». L'exploration du corpus effectuée sur la base de cette distinction a enfin fait ressortir des divergences entre les deux macro-catégories, permettant de vérifier l'hypothèse en corpus.

Le tableau 2 offre une synthèse des exemples des métadonnées que l'on vient d'illustrer, en les classant en fonction de la typologie avancée. Il montre en outre les croisements entre les métadonnées contextuelles ou de décomplexification, d'une part, et les métadonnées analytiques, de l'autre.

Classes de métadonnées	Types d'information	Valeurs ou catégorisations correspondantes
Métadonnées contextuelles	Informations sur le contexte de production et sur le format structurel du tweet	Date
		Dispositif
		Type de tweet
Métadonnées de décomplexification (constituants du technodiscours)	Constituants pluritechnosémiotiques	Présence/absence de liens
		Présence/absence d'images
		Présence/absence de vidéos
		Présence/absence de GIFs, etc.
	Constituants technolangagiers	Présence/absence de hashtags
		Présence/absence de mentions
	Taux d'engagement	Nombre de retweets
		Nombre de citations
		Nombre de *j'aime*
		Nombre de réponses
Métadonnées analytiques (catégorisations théoriques)	Genre discursif	Technodiscours – tweet
	Structuration pluritechnosémiotique (annotation des métadonnées de décomplexification)	Tweets simples vs tweets augmentés
	Spécificités pragma-énonciatives (annotation des métadonnées contextuelles)	Pratiques technodiscursives initiatives, réactives, enchainântes, citationnelles, etc.

Tableau 2
Exemplification de quelques métadonnées envisageables pour un corpus de tweets selon la typologie proposée.

Afin de mettre en pratique cette typologie, nous montrons des applications possibles à partir d'une sélection d'exemples. Pour chacun, nous rendons compte de plusieurs métadonnées envisageables et classifiables selon les trois classes proposées ci-dessus. Compte tenu de la nature illustrative du propos, les métadonnées ne seront donc pas sélectionnées en vue d'un objectif d'analyse déterminé. Il convient néanmoins de souligner de nouveau que la nature des données à examiner et l'objet d'étude envisagé représentent des éléments fondamentaux pour la constitution, la structuration et l'analyse de tout corpus.

L'exemple 1 illustre un tweet du compte Twitter – renommé X depuis juillet 2023 – d'Emmanuel Macron, dont le nom d'utilisateur correspondant – appelé aussi *pseudonyme* ou *pseudo* – est @EmmanuelMacron, publié le 12 décembre 2020.

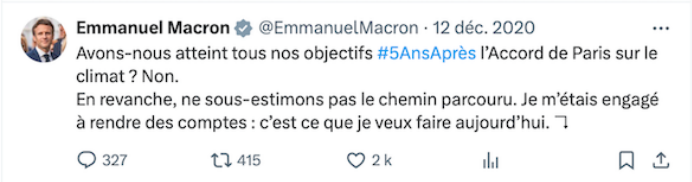

Figure 1
Tweet d'Emmanuel Macron.

Au cas où ils seraient pertinents pour les objectifs de l'étude et pour la mise en place d'analyses contrastives, ces trois types d'informations – le nom du compte, le pseudonyme et la date – peuvent être retenus et catégorisés sous forme de métadonnées contextuelles. En ce qui concerne le format structurel du tweet, il est ensuite possible de classer l'exemple 1 comme un tweet « original ». Sur la base de son format et de sa nature énonciative et pragmatique, ce tweet peut ensuite être catégorisé du point de vue théorico-méthodologique comme une pratique technodiscursive à valeur initiative, dans la mesure où il lance la conversation – à la différence des tweets-réponses et des retweets avec ou sans citation qui procèdent d'un tweet antérieur – sans qu'il soit en même temps prolongé par des tweets consécutifs et enchaînés les uns aux autres, comme c'est le cas pour les threads.

Quant aux constituants des technodiscours qui peuvent donner lieu à des métadonnées de décomplexification, on peut renvoyer, d'une part, à la présence des « technomots » (Paveau 2017) – en l'occurrence, un hashtag (*#5AnsAprès*) – et, de l'autre, aux taux d'engagement du tweet. Les taux d'engagement sont affichés dans la partie inférieure de chaque tweet par le biais d'icônes associées à des valeurs chiffrées. La petite bulle renvoie en effet au nombre de réponses obtenues par le tweet (327), les deux flèches courbes pointant respectivement vers le haut et vers le bas indiquent le nombre de fois que le tweet a été rediffusé sur la plateforme via le retweet (415), alors que l'icône du cœur correspond au nombre de réactions obtenues via le bouton du « j'aime » (2 k). Étant donné que le tweet n'inclut pas d'autres contenus sémiotiques augmentant la partie verbo-textuelle, il est possible de le catégoriser comme un *tweet simple*, selon la distinction mentionnée plus haut, en relation avec les métadonnées analytiques concernant la structuration pluritechnosémiotique.

L'exemple 2 consiste en un tweet du compte du ministère de l'Économie et des Finances – @Economie_Gouv – du 8 décembre 2020. Au-delà des éléments qui peuvent faire l'objet de métadonnées et donc de catégorisations similaires ou superposables au cas précédent (nom du compte, pseudo, date et format du type tweet « original » pour les métadonnées contextuelles ; pratique technodiscursive initiative et structuration du type « tweet simple » pour les métadonnées analytiques ; technomots et taux d'engagement pour les

métadonnées de décomplexification), l'attention se dirige en l'occurrence sur les constituants qui présentent en revanche des différences.

Figure 2
Tweet du ministère de l'Économie et des Finances.

En ce qui concerne les constituants technolangagiers, on remarque la présence de plusieurs hashtags (*#JournéeMondialeDuClimat*, *#climatique*, *#environnement*, *#PlanClimat*) ainsi que d'une mention (*@Gouvernement*). Les hashtags et les mentions constituent deux types de métadonnées de décomplexification distincts, pouvant ou non être regroupés dans une seule classe de métadonnées analytiques en fonction de leurs propriétés technolangagières, selon les exigences de l'étude. Deux autres éléments constitutifs du tweet et ici présents[13] peuvent correspondre à autant de métadonnées de décomplexification, à savoir l'inclusion d'un lien URL et celle d'un contenu iconique. En suivant le tableau précédent, ces deux métadonnées de décomplexification peuvent servir de base pour la création de la catégorie des *tweets augmentés*, relevant de métadonnées analytiques.

Le dernier exemple nous permet enfin d'attirer l'attention sur la catégorisation du type de tweet et de la pratique technodiscursive qu'il réalise, les deux pouvant ne pas correspondre de manière biunivoque.

[13] Il est possible de prévoir ces mêmes métadonnées pour l'exemple 1 aussi. Il est néanmoins évident que dans ce cas les deux catégories signaleraient l'absence de liens et de contenus iconiques.

Figure 3
Tweet-réponse d'Emmanuel Macron.

Du point de vue formel-structurel et donc du type de tweet associé, l'exemple 3 constitue un tweet-réponse, comme l'indique l'expression *En réponse à* affichée juste au-dessous des informations de contextualisation. On remarque toutefois que le pseudonyme auquel la réponse est adressée coïncide avec celui du compte qui est à son origine. Il s'agit ainsi d'un cas particulier qui, tout en correspondant formellement à une réponse, mérite une réflexion lors de sa catégorisation par les métadonnées analytiques. Il est en effet utile de rappeler que cette pratique de répondre à soi-même était utilisée pour développer son propos à travers plusieurs tweets, avant que la plateforme n'introduise une fonctionnalité spécifique : le thread. Comme le remarque Bibié (2019), cet usage détourné de la réponse permettait de contourner la contrainte des 180 caractères par message prévue par la plateforme[14]. Dans la mesure où cette forme d'autoréponse s'apparente pragmatiquement à la pratique technodiscursive incarnée par les threads, il reviendra à l'analyste d'évaluer s'il est plus pertinent pour son étude de classer cette occurrence comme une réponse ou comme un thread.

On remarque ainsi qu'à la fois les métadonnées contextuelles et les métadonnées de décomplexification peuvent servir de base pour l'élaboration de nouvelles catégorisations à annoter sous forme de métadonnées analytiques. En l'occurrence, la métadonnée analytique de nature technodiscursive identifie l'une de deux pratiques technodiscursives (réactive ou enchaînante) dont relève le tweet, en prenant appui, d'une part, sur une métadonnée contextuelle qui catégorise le type structurel du tweet envisagé et, de l'autre, sur l'interprétation de sa nature pragmatique de la part de l'analyste en fonction des exigences et des objectifs de l'étude.

Le tableau 3 met en relation la typologie de métadonnées proposée et l'application qui en a été illustrée à partir des exemples 1-3.

[14] Des changements concernant la limite des caractères par tweet – augmentée à 180 en 2017 – ont été également apportés depuis 2023.

Classes de métadonnées **Type d'information et catégorisations associées**		**Valeurs associées aux catégories de métadonnées à partir des exemples 1-3**		
Métadonnées contextuelles		***Exemple 1***	***Exemple 2***	***Exemple 3***
Contexte de production	Nom du compte	Emmanuel Macron	Ministère de l'Économie et des Finances	Emmanuel Macron
	Pseudo	@EmmanuelMacron	@Economie_Gouv	@EmmanuelMacron
	Date	12/12/2020	08/12/2020	12/12/2020
Format structurel	Type de tweet	Tweet « original »	Tweet « original »	Tweet-réponse
Métadonnées de décomplexification		***Exemple 1***	***Exemple 2***	***Exemple 3***
Constituants pluritechno-sémiotiques	Liens URL	Absents	Lien URL présent	Absents
	Images	Absents	Contenu icono-textuel présent	Contenu icono-textuel présent
Constituants techno-langagiers	Hashtags	#5AnsAprès	#JournéeMondiale DuClimat #climatique #environnement #PlanClimat	Absents
	Mentions	Absents	@Gouvernement	@EmmanuelMacron[15]
Taux d'engagement	Retweets	415	23	50
	J'aime	2k	28	267
	Réponses	327	7	19
Métadonnées Analytiques		***Exemple 1***	***Exemple 2***	***Exemple 3***
Classification générique		Technodiscours – tweet	Technodiscours – tweet	Technodiscours – tweet
Structuration pluritechno-sémiotique	Tweet simple / augmenté	Tweet simple	Tweet augmenté	Tweet augmenté
Spécificités pragma-énonciatives	Pratique techno-discursive	Pratique initiative du type « tweet original »	Pratique initiative du type « tweet original »	Pratique réactive du type « tweet-réponse » Pratique enchaînante assimilable au « thread »

Tableau 3
Tableau résomptif de la catégorisation des métadonnées appliquées aux exemples.

[15] À la différence de la mention présente dans l'exemple 2 qui est insérée en phase d'écriture du message, la mention de l'exemple 3 est générée automatiquement par l'acte de répondre à un tweet antérieur en cliquant le bouton respectif.

Pour résumer, dans le cadre d'un corpus de technodiscours, les différentes classes de métadonnées contribuent à traduire en corpus la complexité des matériaux à analyser et à enrichir leur appréhension. Si les métadonnées décomposent la complexité pour permettre de la saisir, ce sera ensuite à l'analyse et à l'analyste de la reconstituer par la synthèse des résultats obtenus et des interprétations avancées. D'où l'importance de la démarche d'analyse adoptée. Comme le met en évidence Longhi (2020) à partir de la distinction identifiée par Paveau (2013) et citée plus haut (cf. section 3.2), l'analyse des corpus de discours numériques natifs gagne à rechercher un équilibre entre des démarches centrées plutôt sur l'« extraction », qui se focalisent sur la matérialité linguistique, et des démarches plutôt « contextualisantes », qui visent l'appréhension du langagier en relation avec le technologique et l'environnement de production. C'est dans cette direction que vont la notion de « corpus réfléchi » de Longhi (2021) et le modèle théorique du logiciel *Visaneco* (Cagninelli, Taglioli 2022), présenté dans Cagninelli (2022). Une analyse plus fine et plus riche des technodiscours dépend donc aussi bien des conditions d'exploration créées par le corpus et par sa structure que des approches et des méthodes d'analyse adoptées.

5. Conclusion

Considérant que la réflexion théorico-méthodologique préside à la constitution du corpus et qu'elle en est enrichie en retour, la structuration du corpus représente une étape significative de la recherche, d'abord réflexive et ensuite appliquée. Les métadonnées s'avèrent en effet essentielles non seulement pour contextualiser les données, mais surtout pour introduire des niveaux d'analyse qui diversifient les angles et les méthodes pour les appréhender. La phase de structuration du corpus se pose ainsi à l'articulation entre la nature des données, le cadrage théorico-méthodologique de l'analyste et l'objet d'étude.

La classification des métadonnées avancée ici vise à attirer l'attention aussi bien sur la nature attestée ou théorique des paramètres adoptés que sur les spécificités des matériaux observés. S'appuyant sur les métadonnées en tant que paramètres d'analyse, l'exploration outillée du corpus permet d'accéder à de nouvelles configurations des données. Celles-ci offrent la possibilité de varier et de multiplier les perspectives d'approche des données, contribuant à soulever de nouvelles hypothèses. Dans le cadre d'une étude discursive, les métadonnées participent en outre du travail interprétatif mené par l'analyste dans la mesure où elles favorisent l'ancrage matériel des données linguistiques empiriques. L'exploration du corpus devra par ailleurs être également étayée par les compétences et les connaissances encyclopédiques de l'analyste, afin d'articuler les différentes perspectives

d'appréhension des données en corpus en les envisageant selon une approche discursive qui met en relation le linguistique et l'extralinguistique.

La structuration du corpus par le biais des métadonnées favorise finalement une appréhension plus fine et plus riche des matériaux à observer et alimente la réflexion et les questionnements non seulement sur l'objet d'étude mais aussi sur le cadrage théorique et sur la démarche d'analyse. En association avec l'outillage de l'analyse, les métadonnées contribuent à la portée heuristique de l'exploration outillée de corpus numériques.

Bionote : Claudia Cagninelli is post-doctoral researcher at the Department of Languages, Literatures, Cultures and Mediations of the University of Milan La Statale. She holds a PhD in Human Sciences/*Sciences du langage* under the joint supervision of the University of Modena and Reggio Emilia and CY Cergy Paris University. Her research interests lie in the field of digital discourse analysis and institutional discourse analysis, with a particular focus on the semantic, enunciative and argumentative dimensions. Additionally, she is interested in theoretical and methodological issues related to the role of the corpus and computerized tools for discourse analysis.

Adresse de l'auteure : claudia.cagninelli@unimi.it

Références bibliographiques

Adam J.-M. 1997, *Unités rédactionnelles et genres discursifs : cadre général pour une approche de la presse écrite*, in « Pratiques » 94 [1], pp. 3-18.

Bibié L. 2019, *Utilisation de* ok *sur Twitter, entre (re)présentation de soi et stabilisation énonciative*, in « Lexique » 25, pp. 57-75.

Cagninelli C. 2022, *Positionnements discursifs dans le débat public sur la fin de vie : variations génériques entre (inter)subjectivité et interdiscursivité*, Thèse de doctorat, Università di Modena e Reggio Emilia/CY Cergy Paris Université, Modena.

Cagninelli C. 2023, *Corpus et observables discursifs : quelle(s) articulation(s) entre réflexion épistémologique et exigences méthodologiques ? Quelques considérations à partir d'un cas concret d'application*, communication présentée au Colloque jeunes chercheuses et chercheurs de R2DIP 2023, *Observables, observant.es, observés : construire et analyser des objets en analyse du discours*, 18 et 19 décembre, CY Cergy Paris Université.

Cagninelli C. 2024, *Effets rhétoriques de l'augmentation énonciative des discours numériques. Le cas des "retweets avec citation"*, in « Lingue e Linguaggi » 62, pp. 205-226.

Cagninelli C. sous évaluation, *Genres de discours et hétérogénéités en corpus. Une approche méthodologique des variations intra- et intergénériques*.

Cagninelli C. et Taglioli M. 2022, *Visaneco* [Computer software].

Charaudeau P. 2011, *Les médias et l'information. L'impossible transparence du discours*, De Boeck Université, Bruxelles.

Condamines A. 2005, *Linguistique de corpus et terminologie*, in « Langages » 157 [1], pp. 36-47.

Condamines A. 2022, *Présentation*, in « Éla. Études de linguistique appliquée » 208 [4], pp. 391-394.

Constantin de Chanay H. et Ferron S. 2017, *Introduction*, in « Le discours et la langue. Revue de linguistique française et d'analyse du discours » 9 [2], pp. 7-19.

Cori M., David S. et Léon J. 2008, *Présentation : éléments de réflexion sur la place des corpus en linguistique*, in « Langages » 171 [3], pp. 5-11.

Culioli A. 1968, *La formalisation en linguistique*, in « Cahiers pour l'analyse » 9, pp. 106-109.

Dalbera J.-Ph. 2002, *Le corpus entre données, analyse et théorie*, in « Corpus » 1. http://journals.openedition.org/corpus/10

Duteil-Mougel C. 2006, *Groupements de textes et corpus : point de vue linguiste*, in Duteil-Mougel C. et Foulquié B. (éds.), *Actes du Colloque international et école d'été « Corpus en Lettres et Sciences sociales – Des documents numériques à l'interprétation »*, pp. 205-216.

Garric N. et Longhi J. 2012, *L'analyse de corpus face à l'hétérogénéité des données : d'une difficulté méthodologique à une nécessité épistémologique*, in « Langages » 187 [3], pp. 3-11.

Garric N. 2012, *Construire et maîtriser l'hétérogénéité par la variation des données, des corpus et des méthodes*, in « Langages » 187 [3], pp. 73-92.

Kerbrat-Orecchioni C. 2017, *« Observable » et « observer » en sciences du langage*, in « Le discours et la langue. Revue de linguistique française et d'analyse du discours » 9 [2], pp. 21-45.

Longhi J. 2018, *Du discours comme champ au corpus comme terrain : contribution méthodologique à l'analyse sémantique du discours*, L'Harmattan, Paris.

Longhi J. 2020, *Explorer des corpus de tweets : du traitement informatique à l'analyse discursive complexe*, in « Corpus » 20. http://journals.openedition.org/corpus/4567
Longhi J. 2021, *Du corpus réflexif au corpus réfléchi : la plateforme #Idéo2017 pour extraire contextuellement les pratiques citationnelles et analyser la circulation des discours politiques sur Twitter*, in « Le discours et la langue. Revue de linguistique française et d'analyse du discours » 12 [2], pp. 99-113.
Longhi J. et Sarfati G.-E. 2018, *Conception du corpus et méthodologie d'analyse : Pour un renouvellement de l'analyse des discours institutionnels et politiques*, in « Semiotica » 223, pp. 87-110.
Maingueneau D. 2014, *Discours et analyse du discours : une introduction*, A. Colin, Paris.
Mayaffre D. 2002, *L'Herméneutique numérique*, in « L'Astrolabe. Recherche littéraire et Informatique », numéro spécial, pp. 151-161. <hal-00586512>.
Mayaffre D. 2005, *Rôle et place des corpus en linguistique : réflexions introductives*, in « Actes des Journées d'Etude TOUlousaines JETOU 2005 », pp. 5-17
Mayaffre D. 2010, *Corpus et web-corpus. Réflexion sur la corporalité numérique*, in « Cahiers de praxématique » 54-55, pp. 233-248.
Moirand S. 2007, *Les discours de la presse quotidienne : observer, analyser, comprendre*, Presses universitaires de France, Paris.
Née É. et Fleury S. 2017, *Constituer un corpus en trois scénarios*, in Née É. (éd.), *Méthodes et outils informatiques pour l'analyse des discours*, Presses universitaires de Rennes, Rennes, pp. 63-101.
Neveu F. 2016, *Observatoires et observables en linguistique française*, in « Le français moderne – Revue de linguistique française » 84 [1], pp. 1-12
Paveau M.-A. 2013, *Analyse discursive des réseaux sociaux numériques* [Dictionnaire]. *Technologies discursives*. Consulté à l'adresse https://doi.org/10.58079/uowy
Paveau M.-A. 2015, *Ce qui s'écrit dans les univers numériques. Matières technolangagières et formes technodiscursives*, in « Itinéraires » 2014-1. http://journals.openedition.org/itineraires/2313
Paveau M.-A. 2017, *L'analyse du discours numérique : dictionnaire des formes et des pratiques*, Hermann, Paris.
Pincemin B. 1999, *Construire et utiliser un corpus : le point de vue d'une sémantique textuelle interprétative*, in Condamines A., Péry-Woodley M.-P. et Fabre C. (éds), *Atelier Corpus et TAL : pour une réflexion méthodologique* (TALN 99), Cargèse, pp. 26-36.
Pincemin B. 2012, *Hétérogénéité des corpus et textométrie*, in « Langages » 187 [3], pp. 13-26.
Pincemin B. 2020, *La textométrie en question*, in « Le français moderne – Revue de linguistique française » 88 [1], pp. 26-43.
Poudat C. et Landragin F. 2017, *Explorer des données textuelles : méthodes – pratiques – outils*, De Boeck supérieur, Louvain-la-Neuve.
Rastier F. 2001, *Sémiotique et sciences de la culture*, in « Linx » [En ligne] 44. http://journals.openedition.org/linx/1058
Rastier F. 2011, *La mesure et le grain. Sémantique de corpus*, H. Champion, Paris.