Intelligibilité du numérique

N° 7 | 2025
Le 4ème paradigme en question

Construire des connaissances en sciences du langage : données langagières, corpus numériques et instrumenta(lisa)tions

Jérémi SAUVAGE *Professeur des universités*
Département des Sciences du langage
LHUMAIN
Université Paul-Valéry Montpellier

Édition électronique :
URL :
https://intelligibilite-numerique.numerev.com/index.php/articles/revue-7/3210-construire-des-connaissances-en-sciences-du-langage-donnees-langagieres-corpus-numeriques-et-instrumentalisations
Date de publication : 04/05/2025

CertiScience© *Certifié évalué par les pairs*

Cette publication est sous licence **CC BY-NC-ND** (Attribution - No commercial - No derivatives).

Pour **citer cette publication** : SAUVAGE, J. (2025) Construire des connaissances en sciences du langage : données langagières, corpus numériques et instrumenta(lisa)tions. *Intelligibilité du numérique*, (7). https://doi.org/10.34745/

Le présent article interroge le « quatrième paradigme » scientifique fondé sur l’analyse massive de données, en particulier dans le champ des sciences du langage et des humanités numériques. L'auteur questionne l’idée selon laquelle l’augmentation des corpus et l’automatisation des traitements conduiraient mécaniquement à une plus grande objectivité ou à une vérité scientifique renforcée. En s’appuyant sur une réflexion épistémologique inspirée notamment de la pensée complexe de Morin, il montre que toute construction de connaissance reste marquée par la subjectivité humaine, les choix méthodologiques et les cadres théoriques. La donnée langagière n’est jamais le fait brut, mais une représentation transformée, potentiellement biaisée par l’instrumentation, l’étiquetage ou la visualisation. Les traitements automatiques et l’IA peuvent ainsi produire des interprétations erronées ou illusoires. L’article met en garde contre une instrumentalisation de la data à des fins idéologiques ou argumentatives. Il est défendu ici une approche de la pensée complexe, attentive aux limites de la perception et des méthodes. Enfin, une vigilance critique permanente est de mise à propos du statut scientifique des données et sur leur rôle dans la construction des savoirs.

Building knowledge in language sciences: linguistic data, digital corpora and instrumentation

This article examines the “fourth scientific paradigm” based on the massive analysis of data, particularly in the fields of linguistics and digital humanities. The author questions the idea that increasing corpora and automating processing will automatically lead to greater objectivity or strengthened scientific truth. Drawing on an epistemological reflection inspired in particular by Morin’s complex thought, the author demonstrates that all knowledge construction remains marked by human subjectivity, methodological choices, and theoretical frameworks. Linguistic data is never the raw fact, but a transformed representation, potentially biased by instrumentation, labeling, or visualization. Automated processing and AI can thus produce erroneous or illusory interpretations. The article warns against the instrumentalization of data for ideological or argumentative

purposes. It advocates an approach based on complex thought, attentive to the limitations of perception and methods. Finally, constant critical vigilance is required regarding the scientific status of data and its role in the construction of knowledge.

Mots-clés :
Épistémologie, Données de la recherche, Epistémologie de la recherche, Linguistique de corpus, Construction des connaissances, Sciences du langage, Epistemology, Research data, Knowledge construction, Research epistemology, Corpus linguistics, Language sciences

---

1. Introduction

Le développement exponentiel de la numérisation des pratiques sociales aboutit dans le domaine scientifique à ce quatrième paradigme (Hey, Tansley & Tolle, 2009) selon lequel on doit se baser sur l'analyse de données massives pour produire de nouvelles connaissances. Les sciences du langage (Minel, 2017 ; Cornish, 2019 ; Savatovsky, 2019), comme champ disciplinaire correspondant à la section 07 du CNU (« Sciences du langage : linguistique et phonétique générales ») se sont intéressées depuis plusieurs années aux relations entre les êtres humains et les outils / artefacts / objets numériques (Perea, 2022), ce que recouvre en France la dénomination « *humanités numériques* ». Plutôt que de rester au niveau des « *Digital studies* » qui s'intéressent plus aux dimensions technique et technologique du numérique, les « Humanités numériques » s'intéressent en priorité à la place de l'humain et aux comportements humains (Sauvage & Nourrit, 2022) en lien avec le numérique dans différents domaines (éducation, santé, vie quotidienne, etc.). Sont apparues alors des formations universitaires estampillées *humanités numériques*, des revues scientifiques, des intitulés de postes d'enseignants chercheurs, des programmes de recherche. Il semble toutefois sain et pertinent, après quelques années, de se demander si les *humanités numériques* ne sont pas une chimère et ont un avenir, voire s'il ne s'agit pas d'une simple mode ou d'une escroquerie intellectuelle (Mounier, 2018).

Dans cet article, nous proposons de réfléchir aux processus de construction des connaissances, en particulier dans notre champ disciplinaire qui est celui des sciences du langage, en posant comme postulat que construire des connaissances constitue en soi une question (et donc un objet) de recherche. Nous entendons par « construire des connaissances » une (en)quête de vérité que mène le chercheur en adoptant une méthode de pensée et une réflexion à la fois rigoureuse et méthodique.

Nous interrogerons donc dans un premier temps les notions de *corpus*, de *données* et d'*instrumentation*, telles qu'on les utilise en sciences du langage (Azzopardi, 2010). L'augmentation parfois vertigineuse des volumétries de données langagières (grands corpus langagiers) et la création d'outils pour permettre le traitement automatisé ou semi-automatisé de ces données conduit-elle obligatoirement à des résultats toujours plus précis ou bien la rigueur apparente de la démarche ne dissimulerait pas des biais potentiels conduisant le chercheur sur des fausses pistes ? Sur un plan plus général, nous proposons de revenir sur les phantasmagories des chercheurs et de l'existence d'idéologies de tous ordres (scientifiques, économiques, politiques, religieuses, philosophiques…), que celles-ci soient internes ou externes aux disciplines scientifiques. Enfin, nous reviendrons plus spécifiquement sur le statut de la *donnée* (ici la donnée langagière) comme donnée scientifique.

2. Méthodes scientifiques et rigueur de raisonnements

Depuis l'Antiquité, certains humains ont régulièrement cherché à établir une démarche cognitive organisée, se voulant rigoureuse, précise et indépendante de toute influence idéologique. Cette recherche d'une méthode rationnalisée indispensable à une pensée scientifique nous a influencé dans notre démarche intellectuelle et a trouvé une concrétisation dans *La Méthode* de Morin (2008) mais aussi dans sa métathéorie de la pensée complexe (Morin, 1977 ; 1980 ; 2004 ; 2011). Ainsi, dans le troisième volume de *La Méthode*, *La Connaissance de la connaissance*, Morin (1986) défend la nécessité d'une épistémologie de la complexité au regard de l'évolution des connaissances humaines, par exemple en lien avec les apports continus de la Science. Puis le quatrième volume sur *Les Idées* (Morin, 1999) renforce cette conception en insistant notamment (ce qui attire notre attention) sur le langage, la logique et la paradigmologie. Si Descartes (1637) avait insisté sur la nature d'*être pensant* des humains, Morin (1973) s'inscrit dans cette continuité conceptive de l'Homme en expliquant que le langage fait l'homme et non l'inverse.

En Sciences du langage, il est alors intéressant de s'interroger, y compris sur un plan philosophique, sur cette relation complexe entre l'humain et le langage, ce qui fut le cas dans l'Histoire depuis le *Cratyle* de Platon, jusqu'à la philosophie du langage (Austin, 1962 ; Searle, 1972), en passant par Brunot (1901) ou le *Cours de linguistique générale* de Saussure (1916). Ce langage rationalisé correspondant au logos est à la fois un objectif à atteindre et le moyen d'y arriver, ce qui pose immanquablement une limite d'ordre paradoxal, ce qui a pu donner lieu à différentes théories sur le langage au fil du temps (Bronckart, 2019). L'humain reste limité par son essence d'humain et dépendant de sa propre nature humaine pour viser un objectif louable, mais finalement inatteignable : une pensée parfaitement objective, organisé et transmissible par le langage qui nous fait Homme. En d'autres termes, il s'agit là d'une nouvelle quête d'un Graal où le chemin compte plus que le but.

Dans un souci d'objectivation de la pensée, ce qui est couramment appelé « démarche scientifique » avec l'importance de son caractère de réfutabilité (Popper, 1934) regroupe quelques méthodes d'observation d'un objet de recherche, comme l'expérimentation (avec groupe témoin / groupe test), l'enquête (avec entretiens semi-directifs, questionnaires…), la démarche écologique… Souvent, le raisonnement choisi s'inscrit dans une démarche hypothético-déductive, à l'exception de certaines démarches ethnométhodologiques (Garfinkel, 2007). Dès lors, la « coutume » en Sciences humaines et sociales, inspirée par les sciences dites *dures*, est de de s'appuyer sur un corpus de données, ce à quoi nous souscrivons d'un point de vue personnel. Ainsi, la manière de la constitution de ce corpus devient un principe fondamental de la démarche : *qui*, *combien*, *pourquoi*, *comment*… avec l'alternative qui va avec ce raisonnement : pourquoi pas eux, pourquoi pas plus ou moins, pourquoi et comment mais pas autrement… avec comme objectif de protéger le principe de défense d'une thèse scientifique basée sur le débat contradictoire (Popper, 1934 ; Renucci, 2013). C'est pourquoi, en *Sciences du langage*, nous avons l'habitude de nous appuyer sur cette pratique qui consiste à réunir à corpus de données langagières afin de pouvoir analyser telle ou telle pratique langagière, dans le but d'éprouver telle ou telle hypothèse dans une démarche hypothético-déductive. Or, depuis la constitution des premiers grands corpus oraux (comme le corpus ESLO d'Orléans en sociolinguistique dont les données ont été recueillies d'abord entre 1968 et 1974), une autre question a émergé : comment un être humain peut-il traiter un « grand corpus » de données (Abouda & Baude, 2006) ? La réponse est simple : ce n'est pas possible (Habert, Nazarenko et Salem, 1997). On se tourne alors, à juste titre, vers des outils dédiés à cette tâche, qui ont connu un développement continu tout au long de l'évolution de l'informatique. L'idée sous-jacente est que la volumétrie de donnée importante constitue une représentativité plus importante des faits et tend donc vers une objectivation du réel permettant ainsi de toucher du doigt la Vérité. L'analyse de grands corpus langagiers numérisés (Badin *et al*., 2022), leur nature complexe (multimodalité, oral, écrit…) ainsi que l'*open data* (Broudoux & Ghislaine, 2015), aboutissent à une généralisation de nouveaux comportements chez les chercheurs (Squividant & Revelin, 2021) constituant autant d'interrogations en lien avec les méthodes scientifiques usitées (numérisation, *data mining*, étiquetage, transcription…) et la complexité de la réalité observée que personne ne devrait dénaturer en la simplifiant pour la comprendre (Bachimont, Verlaet & Gançarski, 2021). On peut alors légitimement se demander quelles sont les limites et les dévoiements possibles des traitements (semi)automatiques, statistiques, pour lesquels les données sont tellement transformées pour être permettre ces traitements qu'elles ne correspondent plus aux réalités étudiées (Hovy & Prabhumoye, 2021 ; Henderson Ho, Vallebueno & Handan-Nader, 2024) ? Par ailleurs, les techniques de visualisation (Verlaet, 2018) utilisées ou non pour mettre en évidence un aspect plutôt qu'un autre mérite également, selon nous, un temps de réflexion, notamment quant à la possibilité d'un biais sur un plan représentatif (Grandjean, 2022). Redisons-le sans ambages, le corpus de données dans la démarche scientifique est pour nous indispensable, ce qui n'est pas contradictoire avec l'intérêt d'un raisonnement construit à partir d'un essai purement intellectuel sans

s’appuyer sur des données (dans un premier temps). Après tout, les hypothèses heuristiques restent souvent le point de départ pertinent à la construction d’un raisonnement scientifique (Schlanger, 1983 ; Vincent, 2012). Ainsi, la théorie constructiviste du développement de l’intelligence chez l’enfant de Piaget (1923, 1924 ; 1936 ; 1946) trouve son origine dans l’observation d’abord naïve d’un cas quotidien (sa fille), pour déboucher sur la constitution du Centre européen de psychologie génétique de Genève où des milliers d’enfant ont par la suite été observés. Ainsi, pendant un demi-siècle, des milliers d’études expérimentales (avec leurs collectes de données) ont corroboré cette pensée initiale. C’est pourquoi, il nous paraît inconcevable d’être tenté par l’abandon de l’analyse de données à partir d’un corpus constitué aujourd’hui. Cela renvoie, selon nous, à la nécessaire séparation entre le raisonnement scientifique, la collecte et l’analyse de données, dans le sens de Cori & David (2008) à propos de la « linguistique de corpus ».

3. Données langagières et données scientifiques : aux frontières des percepts

D’un point de vue philosophique, nous nous interrogerons sur les frontières possibles des percepts en phénoménologie comme données exploitables pour construire de nouvelles connaissances du monde. Si le chercheur mène des observations, il a besoin de se focaliser sur des indicateurs à étudier, à partir d’un corpus de données qu’il aura collecté ou déterminé puis de déterminer des stratégies d’exploration de ces données, ce qui constituera un nouveau cadre épistémique (Kuhn, 1983 ; Juignet, 2015) en rupture avec une posture simplement empirique. Ainsi, l’observation de l’apparente course du Soleil dans le ciel peut conduire tout un chacun à déduire que l’astre céleste tourne autour de la Terre. Cela pose la question de la perception d’un phénomène observé et du rapport à la Vérité et pose une limite à l’observation empirique. Pouvons-nous poser comme postulat que ce que nous percevons en notre qualité d’être humain correspond à une manifestation exacte et vrai en tout point du réel ? C’est pourquoi les données, langagières ou non, se doivent d’être analysées, sous le prisme de différentes interprétations dans un débat contradictoire. On retrouve ainsi la confrontation et du débat d’idées comme condition nécessaire à l’entreprise d’un début de conclusion.

3.1. Data et vérité

Or, si la donnée est la représentation formalisée d’un fait (Bachimont, 2016), le fait est la seule manifestation du vrai (Nietzsche, 1882), puisque l’esprit humain se trouve dans une sorte d’incapacité à penser le monde de manière objective, c’est-à-dire indépendamment de toute idéologie, indépendamment des influences conscientes ou non-conscientes. Cependant, la donnée n’est pas toujours le fait. En Sciences du langage, il s’agit très souvent d’une représentation du fait vrai, cette représentation ne

pouvant être qu'une altération de ce fait, une illusion (Narcy-Combes, 2010), voire une hallucination pour les Intelligences artificielles. Voilà pourquoi, ce que représente la donnée est aussi important que la donnée elle-même, dans un contexte donné et prédéfini par la tendance scientifique du moment.

Dans la continuité d'une théorie des origines de la connaissance initiés dans l'Antiquité, que veut dire alors aujourd'hui, en Sciences du langage, « construire des connaissances » (Knorr-Cetina, 1999) ? Quelle place accorder aux théories et à la réflexion fondamentale en lien avec ces relations aux corpus de données ? Si on ne peut pas se fier à la simple perception d'un phénomène, pourquoi continuer à recueillir des données et les analyses pour tenter de mieux comprendre des phénomènes ? L'histoire de l'origine (*principes*) des connaissances est en cela éloquente.

Nous pensons que depuis la l'allégorie de la caverne (Platon, *La République*, Livre VII), celui qui est quête de Vérité (le *philosophe* chez Platon, le *chercheur scientifique* aujourd'hui) est celui qui ose se couper de ses chaines, sortir de la caverne, pour affronter la lumière de la vérité et penser par lui-même. La seule raison valable du chercheur, quelle que soit sa discipline académique, de conduire son chemin de chercheur est la recherche de la vérité. Même si celle-ci ne l'arrange pas (lorsque l'hypothèse est non validée). Nous pensons qu'un *chercheur* n'a d'autre objectif que celui de produire des savoirs dans le but de s'approcher de la Connaissance, bien évidemment, indépendamment des idéologies ou des motivations des subventionneurs de ladite recherche.

Et pourtant, nos sens nous trompent souvent sur le plan phénoménologique, à grands coups d'illusions visuelles ou auditives (Hinton, 1967 ; Johnson, 1978), tout comme les IA connaissent aujourd'hui des hallucinations (Dahl, Magesh, Suzgun & Ho, 2024). Ainsi, donner à une IA le pouvoir d'analyser un corpus de données peut conduire à une mésinterprétation, tout comme la démarche empirique de l'Antiquité lorsqu'elle se contente des apparences perçues.

3.2. Pensée statiste vs dynamiste

Aristote a construit sa pensée à partir des philosophes présocratiques comme Parménide, puis Socrate et Platon. Dans sa *Physique*, l'objectif est notamment de déterminer les principes (« arkhè », ἀρχι) fondateurs des connaissances, c'est-à-dire, littéralement, le « commencement ». Parménide, dans la continuité d'une pensée pythagoricienne logique, défend le caractère statique de l'Univers (du Cosmos) : ce qui est, est. Ce qui n'est pas, n'est pas. Même si dans l'absolu cette affirmation fait sens, elle est facilement fragilisée par l'observation et le constat de certains phénomènes (de la « course » du Soleil dans le ciel à la *Physique quantique* aujourd'hui). Or, Platon dans son dialogue intitulée *Parménide*, récuse déjà avant Aristote ce qui est qualifié de « monisme », c'est-à-dire une pensée que l'on pourrait aujourd'hui qualifier de « simpliste ». Qu'en est-il de la Vérité si l'on n'essaie pas de voir au-delà des apparences ? Aristote

enfonce alors le clou en défendant un point de vue beaucoup plus complexe : il existe plusieurs principes (et non un seul) pour expliquer le monde non plus d'un point de vue statique mais d'un point de vue dynamique. Par ailleurs, pour Aristote, tout principe est le résultat d'une interaction, ce qui définit la nature de la dynamique. Si Parménide considère que deux principes sont impossibles (pour lui un seul suffit), Aristote explique bien que quand on voit un cygne noir, la seule conclusion logique possible est que « tous les cygnes ne sont pas blancs ».

Trois principes sont alors nécessaires. 1. Le Principe d'être : ce qui est, est. 2. Le Principe de changement : il existe une interaction entre la chose et la cause. 3. Le principe épistémologique : ce par quoi une chose est connue. La pensée aristotélicienne se caractérise donc par une dynamique générale qui n'est pas sans rappeler les courants épistémologiques du XXème siècle interactionnistes (Habermas, 1987 ; Bronckart, 2017), complexe (Morin, 1977 ; 1980 ; 1986 ; 1999 ; 2004 ; 2011) jusqu'aux théories du chaos et de la complexité (Prigogine & Stengers, 1984 ; Larsen-Freeman, 1997 ; Dewaele, 2001 ; Sauvage, 2015).

3.3. De l'observation biaisée à la causalité de la data

Ainsi, la démarche d'un raisonnement scientifique ayant pour but de faire émerger un nouveau savoir peut déterminer et orienter la nature de ce savoir, jusqu'à en faire un faux-savoir, c'est-à-dire biaiser la nature-même de son essence de savoir. Ce sera le cas d'un raisonnement basé sur une expérience empirique dont les apparences peuvent nous faire croire autre chose que la vérité du phénomène, comme la course apparente du Soleil dans le ciel. En analyse des pratiques linguistiques et langagières, il en va de même pour toute étude s'intéressant aux discours déclaratifs. Dire que l'on dit, dire que l'on pense ne correspond pas toujours (loin de là) à une vérité effective et attestée. Les actes illocutoires et perlocutoires n'ont pas la même signification que l'acte locutoire perçu (Austin, 1962 ; Searle, 1972). Ainsi, « j'ai froid » ne signifie pas forcément que *j'ai froid*, mais peut induire que je souhaite qu'un autre présent ferme la fenêtre.

On retrouve ce type de travaux en sociolinguistique et en analyse du discours lorsqu'on s'intéresse aux décalages possibles entre les usages linguistiques d'un locuteur et les représentations qu'il en a. Ainsi, je peux affirmer avec sincérité dans le cadre d'un entretien semi-directif que je prête une attention particulière dans mes enseignements à respecter la grammaticalité de la langue (l'emploi de la double négation « ne…pas », par exemple) mais lorsque je suis enregistré, je découvre que j'omets quasi-systématiquement « ne » comme la plupart des locuteurs francophones à l'oral.

Un autre aspect « biaisant » est celui où la data elle-même crée le savoir, c'est-à-dire un savoir qui n'existerait plus dans un autre contexte, avec un autre corpus de données ou structuré différemment. Sans toujours aller jusqu'à une démarche frauduleuse, les choix opérés pour visualiser les données peuvent aller jusqu'à servir d'argument dans un raisonnement. C'est par exemple le cas dans le choix d'échelle en abscisse ou en

ordonnées pour accentuer ou au contraire amoindrir une évolution chiffrée (depuis Tufte, 1983 ; Cleveland & Mc Gill, 1984). La donnée n'est plus au service de la quête de vérité mais elle devient un adjuvant argumentatif pour peser dans les résultats et ainsi mettre en évidence ce qu'on souhaite plutôt qu'autre chose. À l'heure des grands corpus et des traitements plus ou moins automatiques de ces *big data*, le recours à l'instrumentation (faire usage d'instruments de mesure par exemple, comme des scripts ou des logiciels) devient une obligation et peut, dans certains cas, devenir un objet d'instrumentalisation de la data pour persuader plutôt que de prouver. En Linguistique, le cas le plus connu de ce type de phénomène est celui de la théorie de Chomsky en grammaire générative, où pour prouver qu'il existe une *Grammaire Universelle* dans toutes les langues naturelles du monde et un dispositif biologique inné permettant aux humains de faire usage d'une langue, il ne se contente que de certains exemples en langue anglaise, exemples qui l'arrangent (Chomsky, 1965 puis Pinker, 1994). Or, il a été rapidement contredit, simplement en prenant d'autres exemples qui, cette fois, n'allait plus dans le sens escompté (Sampson, 1980 ; Harris, 1993 ; Tomasello, 1999 ; 2003 ; Fortis, 2008).

4. Enjeux épistémologiques et pistes de réflexion

Ainsi, le percept rendu (*3e Principe* d'Aristote) peut également contribuer à la constitution du savoir en tant que tel. Le danger serait donc que la data et ses représentations participent à la création des savoirs, ce qui pose une interrogation d'ordre épistémologique fondamentale : est-ce un problème ou non ? Les savoirs (pré)existent-ils en dehors des humains qui les pensent et les font émerger (dans la continuité d'une pensée socratique puis aristotélicienne) ou bien ne trouvent-ils leur réalité que comme résultante d'une réflexion humaine (Descartes, 1637). Les données, du moins en Sciences du langage, sont des constructions humaines, interprétées dans leurs représentations par des humains, ce qui pose la question du statut de ce type de données langagières : ont-elles un statut « scientifique », c'est-à-dire comme résultant d'une démarche scientifique et rigoureuse (mais dans les limites du fonctionnement de l'esprit humain) ?

La tentation d'instrumentaliser l'instrumentation de la constitution d'un corpus et de son traitement devient alors plausible. Lorsqu'on s'intéresse à la *Science poubelle* (ou *Junk Science*, Huber, 1993) ou *pseudoscience* (Rasplus, 2014), c'est-à-dire quand la méthodologie utilisée est biaisée ou les données corrompues plus ou moins volontairement, on y trouve des raisons le plus souvent idéologiques et/ou commerciales (Foucart, Horel & Laurens, 2020).

À l'ère du numérique, un script, un étiquetage de corpus, un choix de visualisation peut abonder dans un sens ou dans un autre pour valider ou non une hypothèse. Intéressons-nous à l'étiquetage. Pour rendre « traitable » un corpus de données de manière informatique, une transformation de ces données est nécessaire pour permettre à un

programme de tourner et en tirer un substrat. Mais bien entendu, la manière de concevoir et d'opérationnaliser cet étiquetage n'est pas neutre quant aux résultats que l'on souhaite obtenir, ce qui pose la question de l'objectivation des données et des résultats dont on sait qu'elles doivent être interprétées (Cori, David & Léon (dir.), 2008 ; Garric & Longhi, 2012). Voilà où toute la subjectivité humaine intervient. De la même manière, en écartant certains sujets en psychologie, en médecine ou pharmacie, on « simplifie » (mais comment faire autrement ?) l'échantillon observé de manière à rendre opérationnalisable le traitement des données et les rendre homogènes donc traitables. Au-delà de ce niveau, il ne semble pas exister de réponses ou de solutions adéquates pour dépasser cette difficulté épistémologique.

Ainsi, dans le cadre d'une étude doctorale que nous avons supervisée, il s'est posé la question toujours délicate pour un linguiste de l'étiquetage du corpus. Il s'agit d'un corpus numérique partagé issu de l'ANR *CoLaJe* (Morgenstern & Parisse, 2012) dans lequel on souhaite analyser l'évolution de l'emploi des catégories grammaticales de la langue. Le projet est de faire analyser par un script l'apparition et de l'évolution des « natures de mots », comme on dit à l'école, dans le développement de la parole d'un enfant francophone. Le problème est : comment étiqueter le corpus pour classer les chaque « mot » sous la bannière d'une dénomination grammaticale puisque les linguistes et autres spécialistes ne sont pas d'accord entre eux. Les catégories grammaticales sont en débat et liées à certaines écoles de grammaire : vaut-il mieux parler de catégories syntaxiques ou de catégories lexicales ? Ce qui peut conduire au fait qu'un verbe puisse avoir une valeur nominale ou adjectivale. On comprend ici que l'offre multiple de la dénomination aura un impact sur l'étiquetage du corpus, étiquetage qui sera nécessaire lié au contexte parfois complexe de sa place dans l'énoncé, tout autant qua la perspective disciplinaire adoptée : syntaxe en linguistique, grammaire du point de vue scolaire ou académique. Puisque choisir c'est renoncer, les choix opérer lors de ce type d'étiquetage en disent long sur l'ancrage théorique qui se situe en arrière-plan de la réflexion.

5. Ouverture et perspectives

Notre principal objectif dans cet article a consisté, finalement, à refuser une simplification extrême de ce qu'est une donnée, ce qu'elle apporte à la réflexion du chercheur en Sciences du langage, en essayant de problématiser ce 4ème paradigme à l'aune d'une perspective complexe. Il nous apparaît indispensable à l'heure de la numérisation des données et de leurs traitements de nous interroger de manière critique sur cette data à la fois indispensable mais également problématique, en lien avec une pensée scientifique rigoureuse et indépendante de toute influence idéologique (dans la mesure du possible). C'est pourquoi cette dimension épistémologique doit, selon nous, attirer toute l'attention d'un chercheur en Sciences humaines et sociales en général, en Sciences du langage en particulier. Nous savons, en Sciences du langage, qu'un discours s'inscrit dans un contexte, que ce qu'on dit signifie parfois beaucoup plus

que ce qu'il paraît et que l'expression de la pensée n'est jamais parfaitement exacte et fidèle à l'intention voulue. Or, dans cette optique de « construire des savoirs » pour atteindre la Connaissance et la vérité de phénomènes réel, il nous faut garder à l'esprit que ce que nous percevons comme phénomènes ou produisons comme discours n'est pas obligatoirement à prendre de manière littérale. L'interprétation des données est en cela une nécessité, dans la mesure où les données ne sont pas forcément vraies en tant que phénomènes réels. Bien entendu, la numérisation, qui est une chance formidable pour le chercheur, ajoute une couche de complexité à un sujet qui n'a cessé d'être questionné depuis l'Antiquité. L'avantage de considérer la *pensée complexe* dans sa dimension épistémique permet de favoriser le débat contradictoire, à la fois dans les choix épistémologiques opérés, mais aussi à propos des corpus de données qui ne sont, même en situation naturelle, que des fabrications humaines.

6. Références bibliographiques

Abouda, L. & Baude, O. (2006). Constituer et exploiter un grand corpus : choix et enjeux théoriques. Le cas des ESLO. Corpus en Lettres et Sciences sociales, *Des documents numériques à l'interprétation*, Albi, France. halshs-01162506

Austin, J.L. (1962). *Quand dire c'est faire*. éd. 2024, Le Seuil.

Azzopardi, S. (2010). Présentation : La linguistique « de » corpus au-delà des champs disciplinaires : questions et enjeux transversaux, *Cahiers de praxématique*, 54-55 | 2010, 11-24.

Bachimont, B., Verlaet, L. & Gançarski, P. (2021). Traces, données et preuves en contexte numérique : quelles acceptions interdisciplinaires ?. *Revue Intelligibilité du Numérique*, 2 |2021. ⟨hal-04821059⟩

Badin, F., Cance, C., Dugua, C., Kanaan-Caillol, L., Minard, A.-L., & Ploog, K. (2022). Les données orales en linguistique. *Bulletin de l'AFAS* [En ligne], 48 | 2022, mis en ligne le 31 décembre 2022, consulté le 09 janvier 2025. DOI : https://doi.org/10.4000/afas.7496

Bronckart, J.-P. (2019). *Les Théories du langage*. Mardaga.

Broudoux E. & Ghislaine C. (2015). *Big Data - Open Dataâ□¯: Quelles valeursâ□¯? Quels enjeuxâ□¯?* Paris, De Boeck Supérieur.

Brunot, F. (1901). *La pensée et la langue*. Gallica BnF.

Bulot, T. (1990). Le générateur de dialogue : la dimension dialogique de la communication homme-machine, Cahiers de Linguistique Sociale, 16, 141-162.

Chanet, C. (1995). Dialogue homme-machine et observables : les actes de langage en question, *Scolia : Sciences Cognitives, Linguistiques et Intelligence Artificielle*, 5, 389-426.

Charaudeau, P. (2010). Dis-moi quel est ton corpus, je te dirai quelle est ta problématique *revue Corpus n°8.* *www.patrick-charaudeau.com/spip.php?page=imprimir_articulo&id_article=103*

Cleveland, W. S., & McGill, R. (1984). *Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. Journal of the American Statistical Association*, 79(387), 531-554.

Corai, M., David, S. & Léon, J. (éds) (2008), *Langages* n° 171 : *Construction des faits en linguistique : la place des corpus*, Paris : Larousse/Armand Colin.
Cori, M. & David, S. (2008) Les corpus fondent-ils une nouvelle linguistique ? *Langages*, 171(3), 111-129.
Cornish, F. (2019). Les données en linguistique : choix, nature et exploitation. Une vue personnelle. *Travaux de linguistique*, n° 79(2), 7-42. https://doi.org/10.3917/tl.079.0007.
Dahl, M. Magesh, V., Suzgun, M. Ho, D. (2024). Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models, *Journal of Legal Analysis*, Volume 16, Issue 1, 2024, 64–93, https://doi.org/10.1093/jla/laae003
Descartes, R. (1637). *Le Discours de la Méthode*. éd. 2024, Flammarion.
Devillers, L. (2013). Les dimensions affectives et sociales dans les interactions humain-robot. *Interfaces numériques*, 2(1). https://doi.org/10.25965/interfaces-numeriques.1760
Dewaele J.-M. (2001). L'apport de la théorie du chaos et de la complexité à la linguistique. *La Chouette*, 32, 76-86.
Dybkaer, L., Bernsen, N.O., Dybkaer, H. (1998). A Methodology for diagnostic evaluation of spoken human-machine dialogue, *International Journal of Human Computer Studies*, vol. 48, 605-625.
Fortis, J.-M. (2008). Le langage est-il un instinct ? Une critique du nativisme linguistique, de Chomsky à Pinker. Texto !, Vol. XIII, n°4, https://shs.hal.science/halshs-01527764/document
Foucart, S. & Horel, S & Laurens, S. (2020). Les gardiens de la raison : Enquête sur la désinformation scientifique, La Découverte.
Garfinkel, H. (2007). *Recherches en ethnométhodologie*, PUF.
Garric, N. & Longhi, J. (2012). L'analyse de corpus face à l'hétérogénéité des données : d'une difficulté méthodologique à une nécessité épistémologique. *Langages,* 187, n°3, 3-11. https://doi.org/10.3917/lang.187.0003.
Grandjean, M. (2022). La visualisation de données, entre usages démonstratifs et heuristiques, in Bohet, B. & Pringuet, V., *Les devenirs numériques des patrimoines*, (pp. 199-217), Editions de la Maison des sciences de l'homme.
Habert, B., Nazarenko, A. & Salem, A. (1997). *Les linguistiques de corpus*, Armand Colin.
Harris, R.A. (1993). *The Linguistics Wars*. Oxford University Press.
Henderson, P., Ho, D., Vallebueno, A. & Handan-Nader, C. (2024). Corpus Enigmas and Contradictory Linguistics: Tensions Between Empirical Semantic Meaning and Judicial Interpretation, *Minnesota Journal of Law, Science and Technology*, 25-2. Available at: https://scholarship.law.umn.edu/mjlst/vol25/iss2/12
Hey T., Tansley S. & Tolle, K. (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Redmond (USA), Microsoft Research.
Hinton J.M. (1967) Visual Experiences, *Mind*, 76, 217-227.
Hovy D, Prabhumoye S. (2021). Five sources of bias in natural language processing. *Lang Linguist Compass*. https://compass.onlinelibrary.wiley.com/doi/full/10.1111/lnc3.12432
Huber, P.W. (1993). Galileo's Revenge: Junk Science in the Courtroom. Reprint.
Johnson, F.H. (1978). The anatomy of hallucinations, Chicago, Nelson-Hall Co.
Juignet, P. (2015). Les paradigmes scientifiques selon Thomas Kuhn. *Philosophie, science et société*. https://philosciences.com/paradigme-scientifique-thomas-kuhn.

Knorr-Cetina, K. (1999). *Epistemic Cultures: How the Sciences Make Knowledge*. London, Harvard University Press.
Landragin, F. (2004). *Dialogue homme-machine multimodal*. Hermès-Lavoisier.
Landragin, F. (2013). *Dialogue Homme-machine. Conception et enjeux*. Lavoisier.
Larsen-Freeman, D. (1997). Chaos/complexity. Science and second language acquisition. In *Applied Linguistics, 18-2*, 141-165.
Minel, J.-M. (2017). La linguistique face à la multiplication des données langagières numériques. Méthodes, risques et enjeux. *7ème Séminaire International de Linguiistique et 3ème Symposium de Linguistique Textuelle*, Université Cruzeiros del Sud, Sao Paulo, Sep 2017, Sao Paulo, Brésil. halshs-01590750
Morgenstern, A. & Parisse, Ch (2012). The Paris Corpus. *Journal of French Language Studies*, 2012, 22 (Special issue 1), 7-12. ⟨10.1017/S095926951100055X⟩. ⟨halshs-01350592⟩
Morin, E. (1973). *Le paradigme perdu*. Le Seuil.
Morin, E. (1977). *La nature de la nature*. Le Seuil.
Morin, E. (1980). *La vie de la vie*. Le Seuil.
Morin, E. (1986). *La connaissance de la connaissance*. Le Seuil.
Morin, E. (1999). *Relier les connaissances*. Le Seuil.
Morin, E. (2004). *Ethique*. Le Seuil.
Morin, E. (2008). *La Méthode*. Le Seuil.
Morin, E. (2011). *L'humanité de l'humanité*. Le Seuil.
Mounier, P. (2018). *Les Humanités numériques*. Editions de la Maison des Sciences de l'homme.
Narcy-Combes, J.-P. (2010). Narcy-Combes. Pratiques, illusion ontologique et recherches en didactique des langues. *Didactique des langues, didactique des sciences*, 17, 33-46.
Nietzsche, F. (1882). *Le gai savoir*, éd. 2007, Garnier Flammarion.
Péréa, F. (2022). *Des objets qui parlent ?* MKF Editions, coll. « Les Essais numériques ».
Perea, F. (2024). Des origines du langage au dialogue naturel avec les machines : un regard anthropologique. *Revue Intelligibilité du numérique*, 6|2024. [En ligne]
Piaget J. 1923. *Le langage et la pensée chez l'enfant*. Delachaux et Niestlé.
Piaget J. 1924. *Le jugement et le raisonnement chez l'enfant*. Delachaux et Niestlé.
Piaget J. 1936. *La naissance de l'intelligence chez l'enfant*, (8e éd. 1977). Delachaux et Niestlé.
Piaget J. 1946. *La formation du symbole chez l'enfant*, (6e éd. 1976). Delachaux et Niestlé.
Pinker, S. (1994). *The Language Instinct: How the Mind Creates Language?* William Morrow and Co.
Platon. *Le Cratyle*, éd. 1998, Flammarion.
Popper, K. (1934). *La logique de la découverte scientifique*. Payot, éd. 2017.
Prigogine I. & Stengers I. 1984. *Order out of chaos : Man's new dialogue with nature*. New York : Bantam.
Rasplus, V. (dir.). (2014). *Sciences et pseudo-sciences : regards des sciences humaines et sociales*, Matériologiques.
Renucci, J.-F. (2013). L'importance du débat contradictoire. *Revue de science criminelle et de droit pénal comparé*, 04, pp.930. ⟨halshs-02245090⟩

Sampson, G. (1980). *Educating Eve: The 'Language Instinct' Debate*. Continuum.
Saussure (de), F. (1916). *Cours de linguistique générale*. Payot.
Sauvage, J. & Nourrit, D. (2022). Humanités numériques et pensée complexe. *LHUMAINE*, 1 | 2022. https://lhumaine.numerev.com/articles/revue-1/2727-humanites-numeriques-et-pensee-complexe
Savatovsky, D. (2019). De quoi les sciences du langage sont-elles le nom ? In Achard-Bayle, G., Krylyschin, M. & Temmar, M. (éds.). *Les Sciences du langage face au défi de la disciplinarisation et de l'interdisciplinarité. Actes du colloque 2019 de l'ASL*, Lambert-Lucas, pp.25-45, 2021, Linguistique et sociolinguistique, 978-2-35935-332-7. ⟨hal-03973668⟩
Schlanger, J. (1983). La question heuristique. *Revue de Métaphysique et de Morale*, 88, n°3, 322-355.
Searle, J.R. (1972). *Les actes de langage. Essai de philosophie du langage*. Hermann.
Squividant, H. & Revelin, F. (2021). S'hybrider sous le 4e paradigme de la "science dirigée par les données massives" : l'ouverture des données favorise-t-elle l'interdisciplinarité et la transdisciplinarité ?, *Communication, technologies et développement* [En ligne], 9 | 2021, mis en ligne le 23 mars 2021, consulté le 08 janvier 2025. URL : http://journals.openedition.org/ctd/3791 ; DOI : https://doi.org/10.4000/ctd.3791
Tomasello, M. (1999). *The Cultural Origins of Human Cognition*, Harvard University Press,
Tomasello, M. (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition*, Harvard University Press.
Tufte, E. R. (1983). *The Visual Display of Quantitative Information. Graphic Press.*
Verlaet, L. (2018). Exploitation et exploration des données : l'intérêt de l'infovisualisation. 10.13140/RG.2.2.23318.40005.
Vincent, C-P. (2012). Heuristique : création, intuition, créativité et stratégies d'innovation, BoD.