Cahiers de praxématique

54-55 | 2010
Corpus, données, modèles

Présentation : La linguistique « de » corpus au-delà des champs disciplinaires : questions et enjeux transversaux

Sophie Azzopardi

Édition électronique
URL : http://journals.openedition.org/praxematique/1108
ISSN : 2111-5044

Éditeur
Presses universitaires de la Méditerranée

Édition imprimée
Date de publication : 1 janvier 2010
Pagination : 11-24
ISBN : 978-2-36781-012-6
ISSN : 0765-4944

Référence électronique
Sophie Azzopardi, « Présentation : La linguistique « de » corpus au-delà des champs disciplinaires : questions et enjeux transversaux », *Cahiers de praxématique* [En ligne], 54-55 | 2010, document 1, mis en ligne le 01 janvier 2013, consulté le 19 avril 2019. URL : http://journals.openedition.org/praxematique/1108

Cahiers de praxématique 54-55, 2010, 11-24

Sophie Azzopardi
Université Paris-Diderot — CLILLAC ARP E.A. 3967

Présentation : La linguistique « de » corpus au-delà des champs disciplinaires : questions et enjeux transversaux

Ce numéro double entend placer les corpus au centre de la réflexion. La problématique n'est pas nouvelle, et cette place centrale est celle que « le » corpus en tant que notion comme « les » corpus en tant qu'objets occupent dans les recherches de plusieurs domaines depuis plusieurs dizaines d'années. En , Damon Maya re met en évidence la prépondérance des corpus en linguistique, à une époque où ils deviennent un enjeu institutionnel et financier :

> Le corpus — la notion et l'objet — risque d'être victime aujourd'hui en France de son succès. Plus une discipline, plus un comité scientifique, plus un chercheur qui n'y fasse référence ; plus un linguiste, surtout, qui ne le manipule, le caresse ou le maltraite.
>
> (Maya re :)

Le corpus est donc un incontournable des analyses linguistiques, à tel point que de nombreuses manifestations et un nombre tout aussi important de publications ont pour objet de réflexion le corpus et les enjeux de son utilisation au sein des sciences du langage. Une revue a émergé de la nécessité de mener une réflexion constante autour de cette notion (*Corpus* et *sq.*). Des recueils (Williams) et des numéros thématiques analysent la place et dressent déjà le bilan de l'utilisation des corpus en linguistique (*Revue française de linguistique appliquée* / , *Langages* /). On est pourtant loin d'avoir fait le tour de la question comme le prouve la sortie très récente du numéro de *Langages* consacré à l'analyse de corpus (*Langages* /). C'est pourquoi cette présentation entend donner un aperçu des problématiques posées par le recours au(x) corpus dans les analyses linguistiques.

Ce premier numéro vise à donner un aperçu des questions et des enjeux qui traversent la linguistique de corpus en dépit des di érentes postures analytiques et des types de corpus constitués et exploités par les chercheurs. C'est pourquoi on se propose ici de définir ce que l'on entend par *linguistique « de » corpus* en l'opposant dans un premier temps à la linguistique « sur » corpus. Ce bref état de la question permettra de mettre en évidence les principales problématiques qui sous-tendent cette approche des études linguistiques et qui sont présentes dans les travaux rassemblés ici : la question de la représentativité, celle du biais inteprétatif du chercheur et enfin celle de la partageabilité et de l'utilisation pluridisciplinaire des corpus.

1. Linguistique « de » corpus et linguistique « sur » corpus

Le déploiement récent de l'intérêt suscité par la notion de corpus et par le recours aux corpus est d'autant plus remarquable qu'il a pris une amplitude considérable en très peu de temps.

L'utilisation de corpus n'est pourtant pas nouvelle. À la suite d'Anne Condamines ( ), on en fait remonter l'origine aux linguistiques du Traitement automatique du langage (TAL). La perspective TAL apparaît dans les années avec pour objectif de modéliser une partie du langage afin de l'automatiser. On a a aire ici à une modélisation du langage à partir du discours et cette modélisation implique le traitement et l'analyse de nombreuses données. Les recherches en TAL supposent donc l'utilisation de corpus de grandes tailles et de formats variés. Cela dit, ce n'est qu'à partir de la fin des années qu'on peut parler d'une généralisation du recours aux corpus à une grande majorité des courants d'analyse linguistique. On peut attribuer ce phénomène à l'essor des technologies qui semblent avoir joué à deux niveaux. D'une part, au niveau de la collecte des données. Les progrès d'Internet ont permis un accès facilité aux données à un grand nombre de personnes. On a en e et assisté ces dix dernières années non seulement à une amélioration considérable de la rapidité des connexions mais aussi au développement de son accessibilité. D'autre part au niveau du traitement et de l'analyse des corpus. Les outils pour traiter les données et les analyser se sont eux aussi rapidement améliorés et la familiarisation du plus grand nombre avec l'outil informatique permet à de nombreux chercheurs de se former à leur utilisation. De ce fait,

la plupart des courants linguistiques qui s'appuyaient sur des énoncés intuitifs ou des exemples construits ont fini par avoir recours à un ensemble de réalisations authentiques de la langue pour appuyer leur démonstration.

Il faut ainsi di érencier deux types d'approches dès lors qu'on s'intéresse à la place du corpus. Celle dont se réclament les linguistiques « de » corpus et celle qui qualifierait la démarche des linguistiques « sur » corpus (Maya re). Les linguistiques « sur » corpus sont celles dont le point de départ est l'intuition d'un fonctionnement systémique de la langue. Les corpus sont alors utilisés pour confirmer au moyen de données authentiques l'analyse menée par le linguiste. Ces approches sont celles que l'on appelle *corpus-based approaches* et procèdent donc depuis la langue envisagée comme système vers les réalisations en discours (approche *top-down*). Les linguistiques « de » corpus, au contraire, revendiquent une approche *bottom-up*, c'est-à-dire une démarche inductive partant du corpus pour aller vers une éventuelle modélisation des régularités émergeant des données authentiques traitées. Il s'agit là de *corpus-driven approaches*, et leur postulat de départ s'inscrit en réaction aux méthodologies intuitives ou introspectives des courants linguistiques traditionnels.

Comme le montre Marie-Paule Jacques ( ), les linguistiques « de » corpus entendent ainsi pallier un certain nombre de di cultés auxquelles sont confrontées les autres types d'approches. D'une part, elles permettent de déceler des phénomènes auxquels l'intuition du chercheur n'aurait pas accès. On pense notamment à des usages spécifiques ou des usages répandus que le linguiste aurait qualifiés de marginaux. D'autre part, elles corrigent les intuitions que le linguiste a du fonctionnement de la langue : tout chercheur, en tant qu'être au monde, a une expérience propre et subjective de la langue qui peut le conduire à appréhender les phénomènes d'une façon particulière. Les courants traditionnels ne seraient donc pas à même de rendre compte des phénomènes de variation et de fréquence des phénomènes et poseraient par conséquent le double problème de la représentativité des régularités dégagées et de la place du chercheur au sein de l'analyse.

Ce double problème reste pourtant au centre des linguistiques « de » corpus. En e et, la démarche *bottom-up*, en partant des données authentiques, entend décrire les usages de la langue à partir de ses réalisations particulières en discours, sans préjuger des résultats. Pour

autant, un corpus est un ensemble construit et repose par conséquent sur une série d'*a priori*. François Rastier ( ) définit le corpus de la façon suivante :

> *Un corpus est un regroupement structuré de textes intégraux, documentés, éventuellement enrichis par des étiquetages, et rassemblés : (i) de manière théorique réflexive en tenant compte des discours et des genres, et (ii) de manière pratique en vue d'une gamme d'applications.*
> Tout corpus suppose en e et une préconception des applications [...].
> (:)

Il semble en e et peu probable de pouvoir analyser l'ensemble des réalisations discursives qui existent et même dans une approche *corpus-driven*, la sélection des données pose le problème de la représentativité du corpus par rapport à l'ensemble des usages d'une langue donnée. Si certains chercheurs prennent le parti de généraliser leurs résultats et acceptent ce problème de représentativité comme étant inhérent à ce type d'approche, d'autres, pour tenter de résoudre cette difficulté décident de poser que les résultats de leurs recherches n'étaient valables que pour le corpus analysé et ne pouvaient en aucun cas être généralisés à l'ensemble de la langue. Cependant, cette approche ne permet pas de rendre compte de la variation ni de la fréquence des phénomènes, pas plus qu'elle ne permet de dégager des régularités afin de modéliser le langage.

Les travaux rassemblés dans ce numéro se situent tous dans la perspective des linguistiques « de » corpus et la question de la représentativité tout comme celle de la place du chercheur au sein de l'analyse donnent encore aujourd'hui matière à réflexion. Le corpus, en tant qu'ensemble de réalisations discursives d'une langue, est un objet en constant mouvement. S'ajoute à cela le fait que les technologies évoluent très rapidement, au même titre que la nature des corpus et les outils pour les appréhender. C'est ce qui nous semble justifier le fait que l'on doive régulièrement faire un point sur ces pratiques, et qu'un nouveau numéro thématique ayant pour objet le(s) corpus ne constitue pas une redite mais plutôt une réflexion nécessaire sur les (nouveaux) enjeux et problèmes posés par le recours au(x) corpus dans l'analyse linguistique.

Les enjeux méthodologiques et épistémologiques abordés dans ce numéro sont de trois ordres. Les articles présentés dans un premier

temps ont pour point central la problématique de la représentativité d'un corpus et abordent les enjeux de la constitution d'un corpus de référence. Une seconde partie s'intéresse à des corpus d'une nature particulière : ceux de la Langue des Signes française. Si les principes de collecte et de transcription des corpus textuels ont fait l'objet de nombreux travaux, la recherche concernant les corpus de L.S.F. n'a émergé que récemment et les deux articles de ce numéro mènent tous deux, bien que de façon distincte, une réflexion sur les aspects problématiques de la transcription de ces corpus non-vocaux. Enfin, on a vu que la place du chercheur au sein de l'analyse constitue encore un aspect problématique de la linguistique « de » corpus. De cette question découle celle de la partageabilité et de l'exploitation pluridisciplinaire d'un corpus donné qui est abordée par les derniers articles.

2. La question centrale de la représentativité

On l'a vu, la question de la représentativité est centrale dès lors que l'on a recours au corpus dans l'analyse linguistique puisqu'elle se pose à deux moments de l'analyse. D'une part, en amont de l'analyse, au niveau de la constitution du corpus et d'autre part, en aval, au niveau des résultats. Tout corpus étant envisagé comme un objet construit par un chercheur empreint de subjectivité, on peut se demander dans quelle mesure il est pertinent de chercher à constituer un corpus qui soit représentatif des usages d'un ensemble de locuteurs donnés. Si l'objectif poursuivi est celui-ci, il s'agira alors de déterminer les paramètres et critères à prendre en compte pour cibler au mieux les données à recueillir. Cependant, cette démarche présuppose une définition intuitive et *a priori* du fonctionnement linguistique de l'ensemble de locuteurs visé, ce qui semble aller à l'encontre de la perspective *bottom-up* prônée dans les approches *corpus-driven*. Du point de vue de l'exploitation des résultats se pose la question de la généralisation des fonctionnements mis en évidence et on peut se demander dans quelle mesure ils sont représentatifs d'une communauté de locuteurs. C'est dans cette perspective que s'inscrivent les trois premiers articles du numéro.

Catharina Pinon aborde les problèmes posés par la constitution d'un corpus en vue d'élaborer une grammaire de l'arabe contemporain. Cette langue a pour particularité d'être une langue pluriglossique.

La question de la représentativité est donc centrale dès lors que la construction du corpus a pour finalité la création d'une grammaire de la langue. On se trouve ici face à un problème récurrent, notamment en linguistique de corpus : celui de la dichotomie entre la langue envisagée comme un ensemble de régularités que l'on fige à un moment donné et les usages en discours qui sont variables et en constante évolution. Dans le cas de l'arabe contemporain, cet écart est particulièrement significatif étant donné que non seulement plusieurs variétés, mais aussi plusieurs états de langue coexistent. L'auteur pose alors le problème inévitable des critères de constitution d'un tel corpus, mais aussi celui des moyens techniques dont elle dispose pour le faire, pour finalement interroger la façon même dont devrait être envisagée la grammaire d'une langue en relation avec ses réalisations discursives.

La question de l'écart entre un système linguistique figé et des usages discursifs variés et en mouvement est également au centre du travail de Stéphanie Lopez. L'auteur mène une analyse comparative de deux corpus de communications radio-téléphoniques entre pilotes et contrôleurs aériens : un corpus de référence, qui constitue une phraséologie normée et préétablie par l'Organisation de l'Aviation Civile Internationale et un corpus d'usages en situation réelle. La question de la représentativité se pose ici à deux niveaux. D'une part, au niveau de chacun des corpus, l'auteur s'interroge sur le choix des critères à retenir en vue de la constitution d'un corpus qui soit représentatif de la langue normée, et d'un corpus qui soit représentatif d'usages réels. D'autre part, la représentativité intervient au moment de la comparaison des deux corpus : dans quelle mesure la norme est-elle représentative des usages et inversement ? Ce type de comparaison permet de mettre en évidence une des limites possibles des linguistiques de corpus : les choix opérés par le chercheur au moment de la construction du corpus peut s'avérer être un facteur de restriction et donner lieu à des résultats qui ne pourront être généralisés à l'ensemble de la langue.

La perspective adoptée par Simon Landron *et al.* est tout à fait di érente. Leur réflexion porte sur les enjeux et problèmes induits par la constitution d'un corpus phonologique du français à des fins didactiques. Il s'agit pour ce groupe de chercheurs de constituer un corpus de réalisations de sons du français d'une part par des locuteurs natifs et d'autres part par des apprenants afin d'identifier les écarts de réalisation constatés. À terme, la mise en évidence de régularités de prononciation par les locuteurs natifs comme par chaque com-

munauté linguistique au niveau des apprenants devrait permettre de modéliser ces écarts. Chaque apprenant pourra ainsi visualiser les écarts de réalisation et pourra entreprendre un travail de correction en s'appuyant sur les réalisations de locuteurs natifs. Le problème des critères de constitution d'un tel corpus se situe donc tant du point quantitatif, à savoir le nombre de locuteurs requis pour prétendre à une certaine représentativité, que du point de vue qualitatif, c'est-à-dire sur les communautés linguistiques identifiées et recensées. Le travail mené par les auteurs étant le fait de plusieurs chercheurs et non d'un seul, la place du chercheur et du but qu'il poursuit dans l'élaboration du corpus est un point sur lequel ils ont été amenés à s'interroger.

On le voit, la représentativité d'un corpus se pose tant au moment de sa constitution qu'au moment de son exploitation du fait de l'inadéquation entre l'objet d'étude (des usages en constante évolution) et le but poursuivi (la modélisation du langage et la mise en évidence de régularités qui conduisent à l'élaboration de principes de fonctionnement figés). Cette question est inhérente à tout type de corpus, quel que soit le but poursuivi par le chercheur. À cette di culté s'ajoute parfois également celle de la collecte et du traitement de données particulières. C'est le cas de la Langue des Signes française dont traitent les deux articles suivants.

3. Le biais interprétatif du chercheur et la conservation de données non scripturales : cas des corpus de L.S.F.

La Langue des Signes française apporte un éclairage particulier sur les di cultés rencontrées par le chercheur lors de la constitution et le traitement d'un corpus du fait de ses spécificités par rapport à la langue parlée. La langue parlée passe par di érents canaux : elle est constituée d'un ensemble d'informations verbales (les énoncés prononcés) et co-verbales (gestes, regard, intonation) qui construisent ensemble le message à communiquer. Selon le type de corpus choisi (écrit, audio, video), le chercheur sélectionne une partie ou l'ensemble de ce message. La L.S.F. de son côté est une langue signée et elle comporte non seulement un message visuo-gestuel codé mais aussi des émissions vocales, des mouvements des lèvres et des gestes qui peuvent faire partie ou bien des informations verbales ou bien des informations co-verbales. Cette spécificité complexifie la tâche du chercheur qui doit souvent avoir recours à un traducteur-interprète s'il n'est pas

lui-même locuteur de L.S.F. La subjectivité inhérente au chercheur dans l'analyse se trouve ici doublée de celle du traducteur-interprète et semble remettre en question la pertinence des résultats issus de l'analyse du corpus.

Raphaël de Courville *et al.* présentent un système permettant de garder une trace des énoncés de L.S.F. autrement qu'en passant par un système d'écriture. La di culté rencontrée par les di érentes tentatives de graphier la L.S.F. est inhérente à la multilinéarité de cette langue (par opposition à la monolinéarité de la langue parlée). Pour tenter de résoudre cette di culté, les auteurs ont envisagé un dispositif d'inscription photocalligraphique permettant de rendre compte de la trace graphique dessinée par le signe gestuel, en partant de l'hypothèse selon laquelle cette trace est porteuse de sens. Ce projet est issu d'une recherche pluridisciplinaire ancrée dans l'approche du design photographique et faisant intervenir la typographie, la photographie et l'ingénierie linguistique. L'exposé de cette méthodologie est l'occasion pour les auteurs non seulement de mettre en évidence les di cultés rencontrées lors de la collecte des données, mais aussi de poser la question de la place du chercheur et de ses *a priori* sur la L.S.F. en amont de cette collecte, lors de la conception des outils.

Le travail de Sophie Dalle-Nazébi *et al.* est lui aussi mené par une équipe pluridisciplinaire : sociologues, experts L.S.F. et informaticiens posent également la question de la conservation des échanges en L.S.F. et des outils envisagés pour constituer des corpus dans cette langue sans passer par une traduction en français qui viendrait fausser les données collectées. Si l'objectif de la constitution de corpus n'est pas ici linguistique mais sociologique, la question de la transcription des données reste la même : comment rendre compte des spécificités spatio-temporelles de la L.S.F. tout en conservant l'authenticité des données ? Le support vidéo semble être le mieux à même de rendre compte des échanges signés, mais l'exploitation de ces données audio-visuelles conduit le chercheur à envisager des pratiques spécifiques d'inscription et d'annotation qu'on ne rencontre que pour ce type de langue signée.

La question de l'accès aux données est par conséquent une question étroitement liée à celle de la place du chercheur dans l'analyse de corpus, quelle que soit leur nature. Le chercheur construit son corpus

dans un but spécifique. On peut donc se demander dans quelle mesure un corpus constitué peut ensuite être utilisé à des fins tout à fait différentes de celles pour lesquelles il avait été prévu. C'est dans cette perspective que l'on a souhaité clore ce numéro par une réflexion sur la partageabilité et l'exploitation pluridisciplinaire d'un corpus.

4. Par-delà les différences : partageabilité et exploitation pluridisciplinaire des corpus

Les contributions réunies dans cette dernière partie abordent des thèmes très diversifiés et présentent l'avantage de mettre en évidence la multitude des disciplines ayant recours à une approche *corpus-driven* dans leurs analyses. Toutes ont en commun le fait de poser la question de la pluridisciplinarité. On a pu voir précédemment que la présence de plusieurs chercheurs pouvait complexifier les critères de choix préalables à la constitution d'un corpus. On se situera ici à l'autre bout du processus pour voir quelles sont les utilisations possibles d'un corpus au-delà de l'objectif de recherche pour lequel il a été initialement conçu. Ce troisième volet ouvre donc la voie à d'autres disciplines et à d'autres questionnements inhérents à la place occupée par le corpus en tant qu'objet construit.

Le premier article de la série est consacré aux relations entre corpus et pathologie du langage. Agnès Witko, linguiste et orthophoniste, montre l'intérêt du recours à une approche *corpus-driven* dans le cadre des études sur les pathologies du langage. La linguistique de corpus y est présentée comme le point de rencontre permettant un travail de recherche interdisciplinaire entre les linguistes et les cliniciens. En e et, le corpus, on l'a vu, est envisagé comme un ensemble de données en mouvement permettant de rendre compte des usages des locuteurs, de leur fréquence et de leur variation. Le recours au corpus est considéré comme le moyen pour le linguiste de découvrir des fonctionnements atypiques, auxquels il n'aurait pas songé de façon intuitive. Or, les pathologies langagières donnent lieu à des réalisations d'un genre tout à fait particulier et l'analyse de corpus permettrait à la fois aux cliniciens et aux linguistes d'avancer chacun dans leur recherche. Les di cultés rencontrées par le chercheur sont du même coup tout aussi spécifiques et l'auteur en aborde certaines, comme par exemple

le cadre législatif qui protège les données des patients et qui de ce fait, constitue un paramètre à prendre en compte lors de la collecte et l'exploitation des données.

Dans un tout autre domaine, Delphine Giuliani et Radia Hannachi s'intéressent à l'une des applications les plus anciennes des linguistiques de corpus : la didactique des langues. L'expérimentation qu'elles ont mise en place a consisté à confronter des étudiants en didactique du F.L.E. à des outils de *data driven learning*, c'est-à-dire l'apprentissage des langues en adoptant une démarche inductive : à partir de l'analyse de données issues d'un sous-corpus médiatique constitué dans le cadre du projet européen *IntUne*, les apprenants d'une langue repèrent certaines régularités grâce au concordancier *Xaira* et comprennent le fonctionnement des structures analysées avant de les mettre eux-même en œuvre. Cette analyse permet de mettre en évidence une fois encore le décalage existant entre la conception normative et figée d'un système linguistique (tel qu'il est présenté par les grammaires) et les réalisations discursives de la langue. On comprend dès lors la pertinence du recours à une approche *bottom-up* dans le cadre de l'apprentissage du F.L.E. et plus généralement dans celui de l'apprentissage des langues.

Le corpus analysé par Aude Wirth est tout à fait singulier et ses visées applicatives le sont tout autant. Il s'agit du *Corpus de documents comptables médiévaux*, un corpus de textes médiévaux qui recensent avec précision les recettes et les dépenses d'une entité administrative. L'originalité de ce corpus réside dans le fait d'une part qu'il n'a jamais été exploité et d'autre part qu'il constitue la source d'une langue spécifique à un genre discursif (les documents comptables) et d'une langue issue d'une époque dont on dispose peu de témoignages (le Moyen Âge). Ce type de corpus est donc tout à fait à même d'intéresser non seulement les linguistes en diachronie mais aussi les historiens. On a donc a aire ici à un corpus qui doit être constitué en vue d'au moins deux types de recherches totalement distinctes et qui doit pouvoir être étendu aux recherches portant sur le droit et l'économie. Il est par conséquent intéressant de voir les critères qui ont été pris en compte préalablement à la constitution de ce corpus à visée pluri- et interdisciplinaire

La mutualisation et la partageabilité des corpus est au centre du travail mené par Lolita Berard et Christophe Benzitoun. Leurs recherches portent sur des corpus oraux. Ce type de corpus fait par-

tie de ceux que l'on constitue avec le plus de di culté. Le temps de la transcription vient s'ajouter à celui de l'enregistrement, et en dépit des outils informatiques dont le chercheur dispose, la constitution de corpus oraux s'inscrit dans la durée. Or, on a pu voir que tout l'intérêt de la linguistique de corpus réside dans l'analyse d'usages discursifs en constante évolution, ce qui peut sembler peu compatible avec une telle inscription dans le temps. Partant de ce constat, les auteurs mettent en évidence la nécessité de mutualiser les corpus oraux du français collectés par des chercheurs issus de di érentes disciplines. Et comme cela a été exposé dans certains des articles de ce numéro, la partageabilité de corpus constitués dans des buts fort di érents les uns des autres n'est pas sans poser de problèmes. Une fois encore, les outils informatiques, s'ils aident à la tâche, ne peuvent tout résoudre : les conventions de transcription, par exemple, doivent être uniformisées, ce qui n'est généralement pas le cas à l'origine. La mutualisation des données est quant à elle, tout comme la constitution d'un corpus, toujours envisagée dans un but bien particulier, et des choix doivent être opérés. C'est alors la question de la place du chercheur et de ses objectifs d'analyse qui se pose à nouveau comme biais à l'exploitation des données.

C'est sur une réflexion sur la place du chercheur et des informations dont il dispose sur le contexte de collecte des données que se termine ce numéro. Alexandra Caria propose une analyse de *data sessions*, réunions au cours desquelles des données audio-visuelles sont soumises à plusieurs reprises à un groupe de chercheurs afin que chacun propose une analyse de ces données. On a vu en e et que l'interprétation des données collectées par le chercheur est empreinte de subjectivité non seulement de celle inhérente au chercheur en tant qu'être humain, mais aussi de celle inhérente à l'objet d'analyse qu'il s'est fixé. Le partage de données permet ainsi de faire émerger di érentes analyses et de diminuer autant que faire se peut ce degré de subjectivité. Comme tout usage en situation réelle, les données analysées ont été recueillies dans un contexte déterminé. L'auteur se propose ici de montrer l'influence de la communication de ces informations contextuelles aux chercheurs préalablement à leur analyse pour déterminer si leur prise en compte permet d'améliorer leur analyse ou si au contraire elle augmente la part de subjectivité de leur interprétation.

Bilan et perspectives

La place du corpus est donc bien centrale dans les travaux rassemblés au sein de ce numéro. Plusieurs questions transversales se sont posées quel que soit le type de corpus et quel que soit le champ disciplinaire ou les objectifs pour lesquels celui-ci a été constitué. Tout d'abord, au niveau de la constitution des corpus, le problème de la représentativité est apparu comme primordial. Ensuite, une fois les corpus constitués, la réflexion doit être menée sur la possibilité de généraliser les résultats obtenus. Enfin, au-delà de ces résultats, la question se pose généralement de la mutualisation ou la partageabilité des corpus, étroitement liée à celle de l'exploitation pluri- et interdisciplinaire d'un corpus avec des objectifs di érents de ceux pour lesquels ils ont été initialement conçus. Une des préoccupations récurrentes des chercheurs en linguistique de corpus est celle de la subjectivité et des intermédiaires entre les données et le chercheur. La place du chercheur, les biais introduits par sa subjectivité comme par les médiations posées entre les données et son regard sont au cœur de chacun des travaux présentés.

Cependant, si ces questions sont transversales, elles ne s'expriment pas de la même façon selon la nature du corpus constitué et exploité. Ce numéro entend mettre en lumière les enjeux et les problématiques qui sous-tendent toute approche *corpus-driven*, toute linguistique « de » corpus, et c'est pourquoi nous avons choisi d'y rassembler des travaux provenant de divers horizons, utilisant des corpus particuliers et interrogeant la possible interdisciplinarité des données ainsi recueillies. Le second numéro quant à lui se propose d'aborder les enjeux et les questions spécifiques à chaque type de corpus. À travers des travaux portant sur des corpus textuels, sur des corpus vidéo et sur des corpus multimodaux, ce sera l'occasion de voir émerger les di érences entre « les » linguistiques de corpus après avoir envisagé tout ce qui les rassemble dans le champ de « la » linguistique de corpus.

Références bibliographiques

C A.,
, « Linguistique de corpus et terminologie », *Langages*
, - .

C M. *et al.* (éd.), / , *Langages* (n°), *Construction des faits en linguistique : la place des corpus*.

G N. & L J., / , *Langages* (n°), *L'analyse de corpus face à l'hétérogénéité des données*.

J M.-P., , « Pourquoi une linguistique de corpus ? », in W G., *La linguistique de corpus*, Rennes, Presses universitaires de Rennes, - .

M D., , « Rôle et place des corpus en linguistique : réflexions introductives », *Texto !* [en ligne], décembre , vol. X, n° . Disponible sur : `www.revue-texto.net/Reperes/Themes/Mayaffre_Corpus.html`. Consultée le octobre .

M S. (dir.) (et *sq*), *Corpus*, Bases, Corpus, Langage — U.M.R. .

R F., , « Enjeux épistémologiques de la linguistique de corpus », in W G., *La linguistique de corpus*, Rennes, Presses universitaires de Rennes, - .

Revue française de linguistique appliquée, / (vol. XII), *Corpus : bilans et perspectives*.

W G., , *La linguistique de corpus*, Rennes, Presses universitaires de Rennes.