SHS Web of Conferences **20**, 01018 (2015)
DOI: 10.1051/shsconf/20152001018

L'hétérogénéité des données provenant du web ; des étapes pour la constitution du corpus complexe

Heterogeneity of datasets from the Web: stages for the constitution of a complex corpus

Camila Pérez Lagos[1,a]

[1]*Sorbonne Nouvelle-Paris 3, EA1484, CIM-ERCOMES*

Résumé. Le corpus issu d'Internet fait émerger de nouvelles problématiques pour les sciences de l'information et de la communication ainsi que pour l'analyse du discours. Au moment de traiter des données multiformes nous risquons de les adapter aux outils déjà existants en contournant les aspects qu'il n'est pas possible de saisir tels que la volatilité des contenus et la multiplicité des signes. Sur une seule page web nous pouvons être confrontés à des photographies, des vidéos, des hyperliens, etc. qui sont constamment actualisés en fonction des contenus. Dans le cadre de cet article nous nous proposons de formuler des réflexions autour de la notion de corpus compris comme une construction de données complexes due à une hétérogénéité de deux types : énonciative et technique. Cet aspect est traité en rapport avec une première analyse de corpus de six sites web de salles de théâtre provenant du Chili, de France et d'Espagne. Une telle démarche nous a permis de dégager les premières conclusions autour des données provenant d'Internet : la diffusion des contenus émanant des sites web et répandus également sur les réseaux sociaux provoque l'amplification du rôle du destinataire, qui devient producteur des contenus ainsi que diffuseur et critique de spectacles de théâtre à l'affiche.

Abstract. The corpus derived from the Internet causes new difficulties for information and communication sciences, as well as for discourse analysis. When analyzing multiform data there is a risk of adapting them to the already existing tools, therefore bypassing aspects that were not possible to take into account as content volatility and sign multiplicity. For example, in a web page we can be confronted by pictures, videos, and hyperlinks that are constantly being actualized according to the content. This paper formulates ideas on the concept of corpus, understood as the construction of complex datadue to two types of heterogeneity: enuciative and technical. Drawing on a preliminary analysis of six websites belonging to theater halls from Chile, France and Spain, the corpus will be fully addressed and discussed. Initial findings regarding data sets from the Internet, reveal how the diffusion of content from websites and its spread to the social networks reates an amplification of the role of receivers who become both content diffusers and theater show critics.

1 Introduction

L'un des principaux inconvénients pour constituer un corpus unifié et représentatif provenant du web concerne l'hétérogénéité inhérente à la toile ainsi que la volatilité et la pérennité des données ; ces caractéristiques s'avèrent cependant essentielles pour l'objet, que sont par exemple l'hypertexte et l'image, « les mutations technologiques des dernières décennies ont profondément modifié les pratiques discursives, ce qui pose un certain nombre de questions inédites. À l'heure où l'image et l'hypertexte se font les corollaires indissociables du texte » (Florea, 2012 : 45 [13]). Cela rend difficile

[a] Auteur de correspondance : perez.lagos.camila@gmail.com

Article available at http://www.shs-conferences.org or http://dx.doi.org/10.1051/shsconf/20152001018

une approche scientifique et nous invite à réfléchir aux méthodes et à la notion même de corpus. D'ailleurs, « un corpus est défini [traditionnellement] comme un ensemble raisonné de textes, structuré par une cohérence interne » (Garric et Longhi, 2012 : 4 [14]) ; cependant d'un point de vue discursif, les données d'ordre textuel (au sens large, la partie écrite du site) se caractérisent déjà par une hétérogénéité multiforme complexe : sémiotique, textuelle, et énonciative (Moirand, 2004 [23]). Cette hétérogénéité liée au texte est comprise comme une propriété du texte et, comme le disent Garric et Longhi (2012 [14]), doit être inévitablement intégrée au dispositif théorique et/ou méthodologique d'analyse.

De plus, face au web, nous risquons d'adapter les données aux outils déjà existants en contournant les aspects qu'il n'est pas possible de saisir. Par exemple la propriété multicanale des sites web, car sur une seule page nous pouvons être confrontés à des photographies, vidéos, entretiens écrits ou non, hyperliens, etc. Comme le déclare Paveau (2015 [25]), « Internet et le Web en particulier, ne constituent pas de simples supports pour une production scripturale qui s'y transporterait, mais bien des environnements qui configurent structurellement les écritures de manière spécifique » (Paveau, 2015 : 2 [25]). Par conséquent, au moment d'analyser des données émanant d'un site Internet il convient d' « étudier les images, les textes, les sons, les parcours de lectures et la relation entre tous les éléments proposés » (Rouquette, 2009 : 10 [26]).

Face à la complexité du corpus il devient indispensable de travailler sur une démarche explicite (Mellet, 2002 [21] et Mayaffre, 2002 [20]) en nous intéressant à tout ce qui est relié à la constitution et à l'analyse du corpus. Ainsi que le déclarent Charaudeau et Maingueneau (2002 [10]) « le mode de constitution du corpus n'est donc pas, en analyse du discours, un simple geste technique répondant aux exigences ordinaires de l'épistémologie des sciences sociales : il est problématique en ce qu'il met en jeu la conception même de discursivité, de sa relation avec les institutions et du rôle de l'analyse du discours » (DAD : 150 [10]).

Cette problématique s'inscrit dans une recherche doctorale plus large concernant les destinataires de sites web de théâtre ; du point de vue de la recherche en communication nous travaillons sur la rubrique de programmation de spectacles de sites web de théâtres provenant de Santiago, Paris et Madrid. Or dans le cadre de cet article, nous allons d'abord formuler des réflexions autour de la notion de corpus comprise comme une construction de données complexes découlant d'une hétérogénéité de deux types : énonciative et technique. Ensuite, nous allons présenter le cheminement suivi pour construire le corpus exploratoire qui nous a permis de dégager les premières conclusions autour de l'étude de données provenant du web. Nous voudrions en bref que notre travail s'avère utile à tous ceux qui travaillent avec des matériaux émanant de la toile, ce qui est sans doute passionnant mais n'est toutefois pas exempt de certaines difficultés inédites au moment de constituer le corpus.

2 Autour de quelques définitions du corpus, vers une procédure explicite

Il nous semble pertinent de faire d'abord le point sur une définition transparente et profondément liée à notre objet d'étude, de ce que nous allons comprendre lorsque nous parlerons de corpus. Dalbera (2002 [12]), dans un article intitulé « Le corpus entre données, analyse et théorie » souligne dans une première partie comment les dictionnaires des sciences du langage définissent ce terme. A savoir, « un corpus est un ensemble d'éléments sur lequel se fonde l'étude d'un phénomène linguistique » (Dalbera, 2002 : 2 [12]). Sur la base de cette définition, parler *d'ensemble d'éléments* est ici très significatif, ce pour deux raisons. D'une part, l'idée d'ensemble renvoie à une composition sérielle qui a évolué de la façon suivante :

> « De la notion de « collection d'objets » réunis parce qu'ils ont en commun, tout du moins superficiellement, une ou plusieurs propriété(s), on passe à un ensemble trié d'objets, c'est-à dire à un ensemble de données filtrées, puis à un ensemble de données construites, c'est-à-dire complété ou remodelé par rapport à l'ensemble précédent d'une telle manière qu'il soit susceptible d'attester des possibilités que l'analyse de l'ensemble précédent a suggérées » (Dalbera, 2002 : 2 [12]).

D'autre part, parler d'éléments au lieu de matériaux linguistiques est également significatif. On sait bien que le corpus peut être construit des textes oraux comme écrits, ce tout particulièrement aujourd'hui, car du fait de l'importance de l'utilisation de ressources provenant du web ces choix se multiplient. Dalbera (2002 [12]) approfondit dans son article l'étude des modes de constitution d'un corpus, plus précisément pour parler de la délimitation et de la clôture. Ces deux aspects sont problématisés au regard de deux exemples, à savoir, les corpus d'échantillons représentatifs et les corpus exhaustifs. Pour ce qui nous concerne, il est adéquat de souligner que l'auteur conclut en formulant l'idée que le corpus est indissociable de la théorie, autrement dit, que plus qu'un ensemble de données, le corpus est une construction des données qui s'avèrent pertinentes pour l'enquête. On comprend donc le corpus comme un *objet heuristique*.

Mayaffre (2002 [20]), dans l'article intitulé « Les corpus réflexifs : entre architextualité et hypertextualité » se propose de donner une définition qui facilite la recherche autour de grands corpus textuels électroniques. Face à cela, le corpus est ainsi compris comme un *objet heuristique* au sein duquel méthode et corpus s'influencent constamment. L'auteur s'écarte de l'idée plus traditionnelle d'un corpus compris comme « un rassemblement inanimé de textes à disséquer sous la lumière crue de projecteurs extérieurs » (Mayaffre, 2002 : 8 [20]). Encore une fois le corpus est compris comme :

> « Une composition relative qui n'a de sens, de valeur et de pertinence qu'au regard des questions qu'on va lui poser, des réponses que l'on cherche, des résultats que l'on va trouver. Ce n'est pas un donné disciplinaire mais un objet heuristique. Le contenu objectif ou matériel d'un corpus textuel n'appartient pas à l'Histoire, à la Linguistique ou à la Philosophie. C'est l'intention du chercheur qui est importante et lui donne son sens » (Mayaffre, 2002 : 3 [20]).

Un aspect important de l'article de Mayaffre (2002 [20]) est le caractère *réflexif* du corpus. Pour lui, un corpus est fait à partir des sous-parties qui renvoient les unes aux autres à travers des liens, tel que fonctionne l'hypertexte, permettant de former un *tout cohérent et auto-suffisant* : « nous entendons par réflexivité du corpus le fait que ses constituants renvoient les uns aux autres pour former un réseau sémantique performant dans un tout cohérent et auto-suffisant » (Mayaffre, 2002 : 5 [20]). C'est surtout le co-texte qui est incorporé au corpus réflexifs.

Dès le début en AD on s'intéresse en plus du co-texte aux *conditions de production* des textes. D'un point de vue historique Guilhaumou (2002 [15]) parcourt l'AD dans le but de mettre en lumière les façons de constituer un corpus. L'auteur traverse la sociolinguistique, la lexicométrie, la linguistique du corpus et l'ethnométhodologie. Nous allons nous arrêter sur un moment en particulier de l'AD qui concerne l'introduction de corpus émanant des discours politiques, puisque cela influence de façon déterminante la conception du corpus et la méthodologie en AD, notamment avec l'intérêt porté aux *conditions de production* :

> « Évoquons rapidement les étapes de l'opération initiale de l'analyse de discours. On puise dans ce que Jean Dubois appelait « l'universel du discours », donc dans la totalité des énoncés d'une époque, d'un locuteur, d'un groupe social. Découpage arbitraire à partir d'intérêts, de thèmes, de jugements de savoir. Dans un second temps, au sein du genre « discours politique » alors promu par les événements de mai 1968, on ne retenait finalement que l'ensemble des phrases contenant, en quelque position syntaxique que ce soit, tel ou tel mot pivot » (Guilhaumou, 2002 : 3 [15]).

En s'intéressant aux mots pivots les analystes du discours se sont proposé de mettre en évidence le sens caché des discours politiques. Pour l'auteur cela pose l'enjeu de constituer un corpus à la fois représentatif et systématique. Dans les années 70 se produit un changement de terrain d'inspiration marxiste qui au niveau de la constitution d'un corpus « rendait obligatoire l'étude des conditions de production du discours, en liaison avec l'histoire des formations sociales » (Guilhaumou, 2002 : 5 [15]). Les notions de *formation discursive* et d'*interdiscours* de Michel Pêcheux (1975) deviennent centrales. Dans les années 80 l'AD est comprise comme une « discipline interprétative à part entière ». Le tournant interprétatif devient une notion fondamentale, comprise comme l'acte consistant à interpréter le discours au centre d'un acte de communication.

Du même point de vue discursif, Moirand (2004 [23]) dans un article publié par la revue *Tranel*, intitulé « L'impossible clôture des corpus médiatiques », traite ainsi le problème de la constitution du corpus à partir d'un exemple construit autour d'un « moment discursif ». Pour elle, le corpus est

toujours réalisé sur un travail de contextualisation à partir des mêmes textes, de sorte que la clôture devient problématique. Selon Moirand (2004 [23]), on ne peut pas « fermer le corpus a priori […], avant de procéder à l'analyse, mais [on peut] constituer au contraire un corpus en boule de neige au fur et à mesure de l'avancée de l'analyse » (Moirand : 2004 : 74 [23]).

Venons-en maintenant à deux notions fondamentales : *sous corpus* et *corpus de référence*. Les premiers d'ordre textuel et peuvent être construits par exemple à partir de la nomination et des discours rapportés. Une telle démarche permettra ensuite de les comparer à d'autres moments discursifs. Pour sa part, le corpus de référence « renvoie aux extérieurs du discours […]. Cela nous renvoie à l'histoire, aux savoirs et à la perception que l'on a des choses du monde » (Moirand, 2004 : 90). Autrement dit c'est tout ce qui « intervient dans l'interprétation sémantique des données » (Moirand, 2004 : 90 [23]). En ce sens, la notion de *corpus de référence* est beaucoup plus ample que celle du *corpus réflexif* de Mayaffre (2002 [20]).

Sur la même ligne, Charaudeau (2009 [9]), dans un article publié par la revue *Corpus* intitulé « Dis-moi quel est ton corpus, je te dirai quelle est ta problématique » fait remarquer que la définition qu'on donne au corpus va dépendre du positionnement théorique et de l'objectif d'analyse que l'on vise : c'est-à-dire, de la problématique. Il divise en trois les types de problématiques, à savoir, *cognitive, communicationnelle* et *représentationnelle*. Dans le premier cas, le corpus peut être constitué aléatoirement, dans le deuxième la construction du corpus est faite en fonction d'un type de situation discursive, dans le dernier cas, il s'agit d'un ensemble de textes ou de signes à valeur emblématique. Pour Charaudeau (2009 [9]) le corpus est donc une construction, « résultant de divers types de regroupements : corpus selon le paratexte (de mots, d'énoncés, de modes d'énonciation), corpus selon l'interdiscours (savoirs de connaissance, savoirs de croyance), corpus selon la situation (locuteurs, finalité et dispositif). Ces regroupements se font en fonction de la problématique d'analyse et de la mise en contraste choisies » (Charaudeau, 2009 : 48 [9]).

Cependant, le problème apparaît avec la question de la clôture et de la représentativité. Pour Moirand (2004 [23]) tel que pour Charaudeau (2009 [9]) « le corpus n'est jamais définitivement fermé » (Charaudeau, 2009 : 56 [9]). Dans le cadre de l'AD, on a besoin à travers une démarche interprétative de se confronter à d'autres corpus, que :

> « consiste à mettre en relation les résultats d'une analyse descriptive avec ceux d'autres analyses : ceux d'autres corpus connexes (confrontation des articles de différents journaux pour en interpréter les ressemblances et différences) ; ceux de corpus de textes d'un même domaine mais de situations différentes (confrontation des écrits journalistiques de différentes époques) ; ceux, enfin, des analyses proposées par d'autres disciplines sur le même domaine discursif (philosophie, histoire, sociologie, psychologie sociale), sur, par exemple, le domaine politique » (Charaudeau, 2009 : 56 [9]).

L'auteur propose donc de travailler sur le corpus « escargot », pour Moirand « boule de neige » qui va se nourrir à « partir d'un premier corpus noyau déterminé selon des paramètres de temps, d'espace, de genres, de dispositifs, de locuteurs, de thèmes, etc., et ce en fonction des objectifs d'analyse que l'on se propose ; puis [d']étendre progressivement ce corpus en le confrontant à d'autres, autant que de besoin, en fonction des questions qui surgissent au fur et à mesure des analyses » (Charaudeau, 2009 : 57 [9]). Lorsque les questions seront plus ou moins révélées nous allons pouvoir repérer des échantillons compris comme « un ensemble de fragments de texte qui peut être considéré comme représentatif au regard des catégories qui serviront à l'analyser de façon qualitative : la parole des acteurs, les caractéristiques du dispositif, le traitement de la thématique » (Charaudeau, 2009 : 63 [9]). En définitive, travailler sur la base d'une procédure explicite est un aspect indispensable pour la recherche scientifique, plus encore s'il s'agit de corpus échantillons complexes.

3 Des données hétérogènes, un corpus complexe

Dans le cadre de cette journée organisée par le laboratoire ICAR dont l'intitulé est « corpus complexes et enjeux méthodologiques : de la collecte de données à leur analyse », il semble adéquat de se demander : en quoi notre corpus s'avère-t-il complexe ? En bref la réponse est qu'un corpus provenant

de la rubrique de programmation de sites web de théâtre est complexe du fait de l'hétérogénéité énonciative et technique des données, ainsi que l'évolution de ces dernières.

L'hétérogénéité est comprise comme constitutive de matériaux langagiers, « un discours n'est jamais homogène : il mêle divers types de séquences textuelles, fait varier la modalisation, les registres de langue, les genres de discours, etc. » (DAD, 2002 : 292 [10]). Sur ce point, Authier-Revuz (1984 [1]), dans l'article intitulé « Hétérogénéité(s) énonciative(s) » signale que « la complexité énonciative […] rend compte de formes linguistiques, discursives ou textuelles altérant l'image d'un message monodique » (Authier-Revuz, 1984 : 98 [1]). Il y a donc d'une part un type d'hétérogénéité qui est constitutif du discours, c'est-à-dire, « le discours […] se constitue à travers un débat avec l'altérité, indépendamment de toute trace visible de citation, allusion, etc. » (DAD, 2002 : 293 [10]). Et d'une autre part un type de hétérogénéité qui est montré compris comme les « formes linguistiques représentant des modes divers de négociation du sujet parlant avec l'hétérogénéité constitutive de son discours » (Authier-Revuz, 1984 : 99 [1]), comme par exemple, l'usage de l'autonymie, des guillemets et des italiques. Pour le propos de cet article nous allons surtout travailler sur les derniers.

D'ailleurs, comme nous l'avons vu dans la partie précédente, nous sommes d'accord sur le fait que dans le cadre de l'AD il faut tenir compte des conditions de production des discours, cela renvoie à un autre aspect de l'hétérogénéité. Cislaru et Sitri (2012 [11]) dans « De l'émergence à l'impact social des discours : Hétérogénéités d'un corpus », signalent que « c'est l'hétérogénéité du corpus qui semble en fin de compte inéluctable si l'on admet que l'opération de contextualisation et la prise en compte des déterminations interdiscursives sont constitutives du caractère interprétatif de la discipline » (Cislaru et Sitri, 2012 : 70 [11]). En ce sens l'hétérogénéité contribue à l'analyse.

Jusque là, l'hétérogénéité n'est effectivement pas une nouveauté pour ceux qui étudient le discours. Cependant, c'est l'aspect de l'hétérogénéité technique qui demeure le moins étudié, ce qui ajoute d'autres difficultés dès lors qu'il s'agit de traiter les données. Moirand (2004 [23]), par exemple, parle d'une *hétérogénéité multiforme* qui est propre au corpus :

> « Les corpus ainsi constitués se caractérisent par une hétérogénéité multiforme: sémiotique (dans la composition des émissions ou dans l'aire de la page), textuelle (présence de genres différents et de modes discursifs différenciés, tels le conseil, la description, l'explication, le récit) et Énonciative (textes apparemment monologiques ou exhibant au contraire leur dialogicité à travers des dires rapportés, empruntés ou imaginés et produits par différents acteurs ou différentes communautés langagières impliqués dans l'événement) » (Moirand, 2004 : 73 [23]).

Vis-à-vis du corpus, nous pouvons ajouter que l'hétérogénéité proprement textuelle est corollaire d'une hétérogénéité des formes sémiotiques ainsi que de la technique que chaque site offre aux destinataires. Un exemple est l'étude de Maingueneau (2013 [19]), dans le 4ème chapitre du *Manuel d'analyse du web en Sciences Humaines et Sociales*, où est traitée la question des genres de discours face à internet. Il comprend le genre comme « une activité communicationnelle socialement identifiable, saisie dans sa globalité » (Maingueneau, 2013 : 75 [19]). Pour lui le web ne change pas seulement le support car la *scénographie verbale* est en rapport à la *scénographie numérique*. Un texte dans le web « devient à la fois une image sur un écran, un support d'opérations (par exemple si l'on peut cliquer sur tel ou tel mot ou groupe de mots), un module dans l'architecture d'un site. Autant d'éléments qui interagissent fortement avec la scénographie proprement verbale » (Maingueneau, 2013 : 80 [19]). La toile rassemble d'une part, des ressources proprement verbales, et de l'autre, des ressources multimodales (image fixe, mouvement, son) ainsi que les opérations hypertextuelles qui impliquent une technique. Pour sa part, Barats (2013 [3]), dans l'introduction au *Manuel d'analyse du web* avertit que l'hétérogénéité devient aujourd'hui presque une condition intrinsèque quand on veut travailler sur des données provenant du web, ce qui dévoile « la nécessité de construire des méthodes adaptées aux objets et phénomènes étudiés et aux particularités du web » (Barats, 2013 : 6 [3]).

Pour sa part, Garric et Longhi (2012 [14]), dans l'introduction du numéro de la revue *Langages* intitulé « L'analyse de corpus face à l'hétérogénéité des données : d'une difficulté méthodologique à une nécessité épistémologique » travaillent sur la base de données émanant de différents matériaux sémiotiques et proposent de définir le corpus comme une « construction, potentiellement hétérogène et donc évolutive » (Garric et Longhi, 2012 : 6 [14]). L'évolution nous semble un aspect original et

pertinent au moment de traiter des donnés qui proviennent de la toile. En ce sens, apparaissent de nouveaux problèmes telles que la volatilité et la pérennité des données. Barats, Leblanc et Fiala (2013 [4]) se demandent si « la volatilité, la complexité, la disparité, l'immédiateté, rendent difficile une approche scientifique intégrée ? L'architecture multidimensionnelle et multimodale du web, sa croissance exponentielle, ses innovations permanentes et difficilement prédictibles, invitent à réfléchir à l'adaptation de méthodes stabilisées aux objets instables du web » (Barats, Leblanc et Fiala, 2013 : 100 [4]). Notre cas fait parfaitement écho à cette description, il s'agit de données qui sont constamment actualisées en termes de contenus et de formats, de sites web qui offrent un portrait public de chaque institution et gardent dans la plupart des cas un registre de l'histoire du théâtre, mais ce sont aussi des sites web à caractère pratique (Rouquette, 2009 [26]) puisqu'on peut y acheter des places pour un spectacle en particulier. Sur ce plan, l'importance de ce type de sites réside dans les informations sur l'affiche par le biais de descriptions des œuvres, de liens vers la compagnie ou les comédiens, de photographies du spectacle, de vidéos du spectacle, d'articles de presse, de liens vers les réseaux sociaux numériques entre autres. Pour se faire une idée, il suffit de consulter par exemple le site web officiel du *théâtre de la ville* : theatredelaville-paris.com.

4 La démarche explicite pour la construction d'un corpus exploratoire

Avant tout il faudrait mentionner que cette tâche s'inscrit dans le cadre de notre thèse doctorale concernant le destinataire de sites web de théâtre dans laquelle ces résultats sont croisés avec d'autres données provenant d'une part, des entretiens menés auprès de personnes chargées de sites web et d'autre part, des politiques culturelles des théâtres de trois pays. Cette étude comparative est faite en collaboration avec l'université Complutense de Madrid et la Sorbonne Nouvelle-Paris 3 et financée par CONICYT-Chile. En termes simples, nous travaillons, d'un point de vue communicatif, sur la problématique de la diffusion de spectacles de théâtre face à la diminution de la fréquentation des salles.

Comme nous l'avons vu dans les parties précédentes, la collecte, l'enregistrement et le traitement de données hétérogènes émanant du web peut être une démarche complexe. C'est pourquoi nous nous sommes inspirés de la notion de Moirand (1992 [22]) de *corpus exploratoire*, dont l'analyse permet d'établir des catégories ajustées aux données, ainsi que d'en mettre en place de nouvelles censées être traitées dans un corpus plus large. Nous allons donc présenter explicitement la procédure que nous avons suivie pour la constitution du corpus exploratoire.

Tout d'abord, concernant l'impératif d'avoir une vision générale des théâtres, nous avons constitué une liste exhaustive des salles se trouvant à Santiago, Paris et Madrid (tous les pays ne disposent pas de registres des théâtres). Nous avons donc fait appel aux institutions culturelles de chaque pays, comme les ministères de la culture et les associations culturelles, ainsi qu'aux programmes culturels émanant de diverses sources (sur papier et en ligne). Ainsi, nous avons demandé conseil aux comédiens qui travaillent sur place, ce qui nous a permis d'enrichir la liste. De plus, nous avons utilisé un logiciel qui permet d'identifier les liens entrants et sortants d'un noyau des sites initiaux : Navicrawler nous a servi à augmenter légèrement le catalogue avec quelques théâtres moins reconnus au niveau institutionnel. D'après les premiers résultats, nous avons pu constater l'existence de liens surtout vers les réseaux sociaux et d'autres plateformes comme Youtube et Dailymotion. Cela nous permet de souligner un premier diagnostic, le fait que toutes ces autres ressources convergent vers le site et engendrent d'autres types d'échanges que nous allons approfondir par la suite. Enfin, nous avons pu constituer un catalogue exhaustif que sert de panorama des théâtres dans ces ville : 39 à Santiago, 46 à Madrid et 114 à Paris. Puis, dans le cadre de la construction et analyse du corpus exploratoire nous avons travaillé sur deux sites web par pays, les sites officiels de Matucana 100 et d'Estación Mapocho pour le Chili ; le Théâtre de la ville et le Théâtre de la Bastille pour la France ; et le Centro Dramático Nacional et Teatro Español pour l'Espagne, ce qui a représenté un total de 44 spectacles ainsi répartis : 8 à Santiago, 17 à Paris, et 19 à Madrid.

Afin de stocker les données et de faciliter l'accès à ces dernières nous avons utilisé ScrapBook, une application gratuite qui nous a permis d'enregistrer les sites et de procéder à quelques opérations

textuelles comme la notation et la recherche de mots. Dans notre cadre, cet outil ne semble que partiellement pertinent puisqu'il ne permet pas d'enregistrer automatiquement les actualisations faites sur le site. De même, cette application ne permet pas de conserver une copie des vidéos, photos ni d'aucun des liens que peut offrir une page dans la plupart des cas. Les formes les plus classiques pour l'enregistrement des pages web s'avèrent ainsi utiles, comme la sauvegarde du site web, néanmoins, dans certains cas les sites fonctionnent avec *flash,* ce qui ne permet pas d'enregistrer avec certitude les vidéos par exemple ni l'environnement où se trouve disposée la page. Aujourd'hui, face au manque d'un outil adéquat -que puisse nous permettre à la fois d'enregistrer texte, image, vidéo et liens, ainsi que noter le corpus-, nous travaillons en simultané sur les enregistrements listés auparavant.

Enfin, à partir de la lecture textuelle et numérique en profondeur du corpus exploratoire nous avons fait émerger des catégories d'analyse discursive, essentiellement les énonciateurs-destinataires (Kerbrat-Orecchioni, 2009 [16] ; Charaudeau, 1995 [8] ; Maingueneau, 2007 [17]), les désignations (Mortureux, 1993 [24] ; Beacco et Moirand, 1995 [5]) et les discours rapportés (Authier-Revuz, 1992 [2] ; Maingueneau 2011 [18]). Encore une fois, ces catégories essentielles pour l'AD sont dans la toile imbriquées à des formes techniques (Rouquette, 2009 [26] ; Barats, 2013 [3]), de sorte que certaines sont revisitées et d'autres définies différemment. Nous allons à présent exposer quelques résultats.

5 Les premiers résultats, l'ensemble discursif et numérique

En ce qui concerne les ressources du web et la façon d'inscrire le destinataire, d'un point de vue général, on retrouve sur la page de chaque spectacle une multiplicité de signes qui fonctionnent ensembles, les vidéos et les photographies étant les plus saillantes. En termes de contenus, la plupart des vidéos et des photos montrent des extraits soit du spectacle soit des répétitions. Le plan est fait du point de vue du spectateur assis au théâtre. Or il est possible d'interpréter cela comme une façon de faire participer à l'avance les destinataires du spectacle. Nous avons ainsi observé les types d'interactions que le web offre aux internautes. Comme l'explique Cardon (2010 [6]) il y a au moins deux points de vue sur ces supports : ceux qui partagent le premier point de vue considèrent le web comme un média traditionnel, c'est-à-dire qu'ils « pensent que sa démocratisation reflète une simple extension de son audience ; ils souhaitent amener de nouveaux spectateurs devant des pages bien éditées et encadrées de bandeaux publicitaires » (Cardon, 2010 : 49 [6]). Les autres jugent que le web s'apparente à une « révolution abolissant la frontière entre lecteurs et rédacteurs et imaginent un espace numérique dans lequel chacun serait amené à devenir tour à tour un média d'information et un producteur de connaissances » (Cardon, 2010 : 50 [6]). Ces deux points de vue sont, à notre avis, les extrêmes d'un continuum complexe de rapports de participation et d'interactions entre internautes. Dans notre cas, sur les six sites étudiés concernant la rubrique de programmation, un seul présente les commentaires sur la même page, autrement dit, la voix des destinataires n'est généralement pas mise en avant dans cette rubrique. Au contraire, la participation fonctionne sur les réseaux sociaux numériques, car tous les théâtres inclus dans cette étude exploratoire ont des profils officiels sur les réseaux sociaux qui invitent les destinataires à venir se joindre à eux, en visant à toucher le plus grand nombre d'usagers. À notre avis même si l'on s'intéresse à un site web en particulier, il est fondamental d'inclure dans l'analyse ces plateformes qui sont notamment ancrées dans le site. Autrement dit, nous ne pouvons pas nous arrêter seulement sur la page web sans prendre en compte les diverses autres interfaces, comme le déclare d'ailleurs Cardon (2013 [7]) :

> « Les nouvelles interfaces du Web proposent des formats de « publication » beaucoup plus variés que l'écriture hypertextuelle d'un texte ou d'un post sur un site ou un blog : phrases de statut, tweets, conversations autour de photos ou de vidéos, simples marques d'appréciation, de recommandation et de signalement exprimées à travers les boutons « I like », « +1 » et les outils de partage des liens. Les nouveaux publics d'Internet disposent ainsi de formats d'énonciation brefs, immédiats, simples qui rapprochent considérablement l'écriture en ligne des formes oralisées de la conversation ordinaire. Beaucoup moins exigeantes et imposantes, ces formes peu coûteuses d'appréciation ne requièrent plus les compétences scripturaires, cognitives et culturelles de l'écriture distanciée qui conféraient un caractère oligarchique à l'espace public traditionnel » (Cardon, 2013 : 179-180 [7]).

En ce sens on peut signaler qu'à travers ces outils les usagers ont pris l'habitude de commenter les spectacles, ils deviennent à la fois producteurs et destinataires de l'information, ce qui permet même d'établir des dialogues entre les spectateurs. Par exemple, sur l'évènement Facebook d'un spectacle donné les usagers posent des questions concernant la pièce (son prix, sa durée, etc.) mais ils échangent aussi leurs avis sur le spectacle, sur la trajectoire de la compagnie ou sur le style de mise en scène. De même, à travers le partage d'un lien sur le spectacle les usagers en modifient le contenu. À travers des modalisateurs ils invitent à voir le spectacle, expriment leur envie d'y assister ou laissent une trace de leur propre expérience du spectacle. De plus, les usagers sont producteurs de contenus ainsi que de critiques. Sur ce dernier aspect, le cas de Twitter est le plus parlant, puisqu'on observe la recommandation de la part des usagers d'utiliser des Hashtag, soit avec le nom du spectacle, soit avec celui de la compagnie ou de la salle. On observe ainsi le rassemblement de conseils de spectateurs sur Twitter avec un hashtag qui varie en fonction du pays, notamment #tuiteatreros #teatro #théâtre #théâtrices. Par le biais de technomots et technosignes définis par Paveau (2015 [25]) comme des « éléments (mot, segment, phrase) cliquables qui mèneraient à d'autres documents en ligne par le biais technodiscursif » (Paveau, 2015 : 6 [25]), il est possible d'approfondir le terrain de lecture des usagers, par exemple en allant vers le site web de la compagnie, vers des entretiens sur les journaux ou des vidéos faites par le même théâtre et mises en ligne sur des plateformes comme Youtube et Dailymotion. Ce type d'usage invite ainsi à revisiter la notion d'énonciateur et de destinataire, en particulier avec l'utilisation d'un Pseudo Twitter qui attire le destinataire vers les comédiens, le metteur en scène, le compte officiel de la compagnie, etc. Tout cela accroît le nombre de destinataires d'un Tweet à travers la visualisation qu'ont les suiveurs de chacun des @. En ce sens, l'audience devient très large.

Intéressons-nous maintenant au point de vue économique c'est-à-dire, posons-nous la question suivante : « le site cherche-t-il et, si oui, a-t-il une stratégie pour capter et faire revenir les internautes ? Y a-t-il une programmation de fidélisation des visites ? … » (Rouquette, 2009 : 40 [26]). Cela leur permet de viser l'usager afin de lui envoyer par la suite des informations plus personnalisées : par exemple s'il s'agit d'un théâtre multidisciplinaire, ils arrivent à distinguer les usagers par discipline en fonction des achats. De même, tous les sites signalent les dates, les horaires, le lieu, les tarifs des spectacles et cela se reproduit à travers les réseaux sociaux numériques, parfois sous la forme de concours pour des places gratuites, ou de promotions ou réductions destinées aux premiers acheteurs. Les mécanismes pour la fidélisation dans les six cas étudiés concernent les abonnements aux newsletters ainsi qu'à la salle, le suivi de toutes les plateformes mentionnées auparavant, de même tous les sites permettent l'achat en ligne de tickets pour les spectacles. À travers l'achat qui dans la plupart des cas se fait sur le même site, les responsables obtiennent un registre général concernant l'identité des acheteurs, leur âge, leur quartier, leur historique d'achat.

Enfin, dans les trois pays les destinataires sont inscrits d'un point de vue plus explicite à travers deux désignations les plus récurrentes : spectateurs et public. D'ailleurs, tout au long des descriptions des spectacles, on observe l'utilisation du pronom « Nous » de type inclusif, que fonctionne pour rapprocher l'énonciateur du destinataire ; le premier se donne comme un énonciateur spectateur de l'œuvre tandis que le second va vraisemblablement être à son tour un spectateur de l'œuvre. Ce phénomène est accentué par l'utilisation de certains types de verbes ayant un effet sur l'objet, dans ce cas les destinataires et spectateurs potentiels. Cela peut être observé dans des expressions comme « ils nous montrent », « il nous accueille », « il nous offre ». Autrement dit, l'énonciateur émet des hypothèses sur les sentiments que va ressentir le destinataire une fois qu'il sera devenu, tout comme lui, spectateur, c'est comme si, par le biais de ce « Nous » et de certains verbes, l'énonciateur faisait ressentir des choses aux destinataires qui n'ont pas encore vu l'œuvre. Enfin, aux descriptions s'intègre la parole d'autrui à travers le discours rapporté direct avec mise entre guillemets, ce qui, dans la plupart des cas, vise à citer la voix du dramaturge ainsi que de la presse. Ce dernier type de parole citée, celle de la presse, apparaît dans le cas de spectacles émanant de la même troupe du théâtre. Cela fonctionne comme une parole d'objectivisation de la description.

6 Conclusions

Nous avons pu observer qu'il existe un type de discours qui décrit les spectacles de théâtre en utilisant, pour faire venir les spectateurs, des stratégies de type énonciative et technique lesquelles sont toujours présentes et imbriquées. Cela génère des inconvénients, certains plus classiques et d'autres inédits pour l'AD dès lors qu'il s'agit de construire un corpus de données analysables.

D'après une première approche au corpus exploratoire on conclut qu'une démarche explicite pour la construction du corpus est devenue indispensable en raison de la complexité des données d'ordre énonciatif et technique. Sur une seule page nous sommes face à différents matériaux sémiotiques qui grâce à la technique de la toile nous amènent vers d'autres lignes de lecture. Par exemple, dans la description du spectacle, nous observons qu'il y a des citations concernant des articles de journaux qui parlent du spectacle en plus du lien vers l'article. Il s'agit donc d'une hétérogénéité montrée (du point de vue technique) qui ne reste pas seulement dans le cadre de l'énoncé cité mais qui permet au lecteur d'accéder au texte entier de la citation. Un autre exemple est celui de Twitter : le compte officiel de chaque théâtre, à travers l'arrobas ou le hashtag, peut renvoyer l'usager vers la compagnie, les comédiens ou le metteur en scène. A l'inverse, chacun de ces comptes peut générer des contenus autour d'un spectacle, qui à la fois, peuvent être repris par le compte officiel du théâtre. De ce fait, les destinataires augmentent en devenant ainsi producteurs de contenus.

En ce cadre, la normalisation de ces données, qui permettra ensuite de les traiter, est un des inconvénients les plus évidents car la profondeur des lectures devient très difficile à transcrire ; non seulement par leur nombre très vaste de liens mais encore par leur volatilité. D'ailleurs, comme les sites web de théâtre actualisent leurs contenus constamment, la volatilité est donc l'une des caractéristiques propres à la toile. Les sites compris dans cette étude vont évoluer, il y aura de nouvelles techniques qui vont adapter les contenus et des pages qui seront supprimées. Cela oblige à revisiter la notion de corpus pour le comprendre « comme un état de données provisoire, évolutif, prélevé au sein d'une archive vivante, gardant une marge d'incertitude non négligeable » (Barats, Leblanc et Fiala, 2013 : 105 [4]).

7 Bibliographie

1. Authier-Revuz, J. Hétérogénéité(s) énonciative(s). *Langages*, ***19*(73),** 98–111. (1984)
2. Authier-Revuz, J. « Repères dans le champ du discours rapporté (I) », *L'information grammaticale*, **55,** 38‑42. (1992)
3. Barats, C. *Manuel d'analyse du web en Sciences Humaines et Sociales*. Armand Colin. (2013)
4. Barats, Ch., Leblanc, J.-M. et Fiala, P. « Approches texicométriques du web : corpus et outils ». Dans Barats, Ch. *Manuel d'analyse du web*. Armand Colin, Paris. 100-124. (2013)
5. Beacco, J.-C., et Moirand, S. Autour des discours de transmission des connaissances. *Langages*, ***29*(117),** 32–53. (1995)
6. Cardon, D. *La démocratie Internet*. Paris: Seuil. (2010).
7. Cardon, D. Du lien au like sur Internet. *Communications*, ***93*(2),** 173–186. (2013)
8. Charaudeau, P. Une analyse sémiolinguistique du discours. *Langages*, ***29*(117),** 96–111. (1995)
9. Charaudeau, P. Dis-moi quel est ton corpus, je te dirai quelle est ta problématique. *Corpus*, **(8),** 37–66. (2009)
10. Charaudeau, P., Maingueneau, D. *Dictionnaire d'analyse du discours [DAD]*. Paris : Seuil. (2002)
11. Cislaru, G., et Sitri, F. De l'émergence à l'impact social des discours : hétérogénéités d'un corpus. *Langages*, ***187*(3),** 59. (2012)
12. Dalbera, J.-P. Le corpus entre données, analyse et théorie. *Corpus* 1 [En ligne], mis en ligne le15 décembre 2003, consulté le 01 juin 2015. URL : http://corpus.revues.org/10 (2002)
13. Florea, M.-L. Faire une thèse d'analyse du discours « troisième génération ». *Langage et société*, ***140*(2),** 41–56. (2012)
14. Garric, N., et Longhi, J. L'analyse de corpus face à l'hétérogénéité des données : d'une difficulté

méthodologique à une nécessité épistémologique. *Langages*, **187(3),** consulté le 01 juin 2015. URL : www.cairn.info/revue-langages-2012-3-page-3.htm. (2012)
15. Guilhaumou, J. Le corpus en analyse de discours : perspective historique. *Corpus 1* [En ligne], mis en ligne le15 décembre 2003, consulté le 01 juin 2015. URL : http://corpus.revues.org/8. (2002)
16. Kerbrat-Orecchioni, C. *L'énonciation de la subjectivité dans le langage*. Paris: Armand Colin. (2009)
17. Maingueneau, D. *Analyser les Textes de Communication*. Armand Colin. (2007)
18. Maingueneau, D. *L'énonciation en linguistique française: No30 2ème édition*. Hachette Éducation. (2011)
19. Maingueneau, D. « Genres de discours et web : existe-t-il des genres web ? ». Dans Barats, Ch. *Manuel d'analyse du web*. Armand Colin, Paris. 74-97. (2013)
20. Mayaffre, D. Les corpus réflexifs : entre architextualité et hypertextualité. *Corpus* 1[En ligne], mis en ligne le15 décembre 2003, consulté le 01 juin 2015. URL : http://corpus.revues.org/11. (2002).
21. Mellet, S. Corpus et recherches linguistiques. *Corpus 1* [En ligne], mis en ligne le 15 décembre 2003, consulté le 01 juin 2015. URL : http://corpus.revues.org/7. (2013)
22. Moirand, S. Des choix méthodologiques pour une linguistique de discours comparative. *Langages*, ***26*(105),** 28–41. (1992)
23. Moirand, S. L'impossible clôture des corpus médiatiques : La mise au jour des observables entre catégorisation et contextualisation. *Tranel*, **(40),** 71–92. (2004)
24. Mortureux, M.-F. Paradigmes désignationnels. *Semen **8** [En ligne], mis en ligne le 06 juillet 2007, consulté le 01 juillet 2015. URL : http://semen.revues.org/4132.* (1993)
25. Paveau, M.-A. Ce qui s'écrit dans les univers numériques. *Itinéraires 2014-1* [En ligne] mise en ligne le 12 janvier 2015, consulté le 01 juillet 2015. URL : http://itineraires.revues.org/2313. (2015)
26. Rouquette, C. *L'analyse des sites internet. Une radiographie du cyberspace.* Bruxelles: De Boeck Université. (2009)