Saisir la parole du citoyen / usager / apprenant en interaction sur les réseaux

Thierry Chanier

Thierry Chanier (2017). Saisir la parole du citoyen / usager / apprenant en interaction sur les réseaux. In Ciara R. Wigham & Gudrun Ledegen (dir.) Corpus de communication médiée par les réseaux : construction, structuration, analyse. L'Harmattan. pp 211-222

SAISIR LA PAROLE DU CITOYEN / USAGER / APPRENANT EN INTERACTION SUR LES RÉSEAUX

Thierry CHANIER, LRL, Université Clermont Auvergne – France

1 ENTRE HISTOIRE DES ÉTUDES SUR LA CMR ET CELLE RÉCENTE DES COMMUNAUTÉS DE CORPUS CMR

Cet ouvrage présente une sélection d'écrits construits à partir de présentations données lors des journées d'études d'octobre 2015 à Rennes sur ce qui commence à être dénommé la "communication médiée par les réseaux" (CMR). Le terme français vient caractériser plus justement et précisément le terme anglais traditionnel « *Computer-Mediated Communication* » (CMC), anciennement traduit par « communication médiée ou médiatisée par ordinateur » (CMO).

1.1 UN BREF APERÇU DES ANCIENNES ÉTUDES SUR LA CMR

En effet, si depuis les années 80 (Harvey, 1995), l'ordinateur a été l'unique moyen donné à des personnes non spécialisées de communiquer à distance sur des réseaux, d'abord uniquement par l'intermédiaire d'un clavier, puis avec l'apparition d'Internet également avec des microphones, il est en passe d'être supplanté par d'autres artefacts, mobiles cette fois, tels le téléphone ou la tablette. Cette mobilité nous est offerte par le réseau téléphonique, libéré de son support filaire, tout comme le protocole Wifi pour les réseaux informatiques. Bien sûr, ces évolutions technologiques, emportées par la révolution de l'Internet et des réseaux, conditionnent fortement la façon dont chacun de nous peut rentrer en interaction avec d'autres humains. Elles nous offrent de nouvelles façons de

communiquer que les chercheurs essayent de saisir pour ensuite les décrire et les interpréter.

La description et l'interprétation de tels phénomènes de communication a déjà une histoire. Ces chercheurs, qu'ils soient sociologues, linguistes (appliqués ou non), informaticiens se sont d'abord intéressés à des communautés de tailles d'échelles diverses allant de celle de quartiers ou régions d'Amérique du Nord avant l'Internet (Harvey, ibid.), à celle des groupes d'apprentissage (Lamy, Goodfellow, 1998). Cependant la langue, les interactions mises en scène dans ces environnements technologiques différents n'ont pas été « saisis », au sens où les données, en partie langagières, sur lesquelles se sont appuyés les chercheurs ne sont pas disponibles, et/ou ont disparu à jamais si tant est qu'elles aient fait l'objet d'une collecte systématique. Même dans des études, comme celle de Cherny sur des clavardages dans des mondes textuels de type MUD (*Multi-User Dimension*) (1999) où apparaissent de façon plus systématique les matériaux linguistiques, les mots des usagers sont devenus insaisissables, fixés sur du papier au milieu du discours du savant. C'est une situation un peu particulière que celle de ce domaine d'études qui cherche pourtant à s'inscrire dans une démarche scientifique. Pour l'imager, essayons une comparaison avec des paléoanthropologues d'aujourd'hui qui n'auraient accès sur les études antérieures qu'aux écrits de leurs collègues, sans les artefacts, ni autres restes des terrains étudiés, sans les traces des couches des terrains fouillés (cf. les fonctionnalités et interface de nos systèmes de communication).

1.2 ET CELLE RÉCENTE SUR LES CORPUS CMR

Pourtant, depuis plusieurs dizaines d'années, un pan entier de la linguistique a décidé d'étudier la langue (les langues) à partir de ses usages attestés et non plus seulement des introspections des linguistes, ramassées dans de simples exempliers. Ils ont quitté l'ère de l'*exemplum* pour saisir celui du *datum* (Laks, 2010), ensemble de données qui, si elles sont collectées, organisées, diffusables suivant certains critères spécifiques (Chanier & Ciekanski, 2010), peut recevoir le nom de

« corpus ». Il aura fallu attendre une vingtaine d'années depuis la naissance de l'Internet (ou plutôt de la Toile, *World Wide Web*, en 1993), alors que cette technologie a changé le quotidien des terriens, pour que la linguistique de corpus s'intéresse de façon systématique à la « langue de l'Internet ».

C'est ainsi que des réseaux nationaux de chercheurs tel *Empirikom* en Allemagne, se sont fixés pour objectif de compléter les corpus de référence de leur langue telle qu'écrite au 19ème et 20ème siècle par celle, encore plus conséquente par la taille, des productions provenant de l'Internet. Si l'attention s'est traditionnellement portée d'abord sur les écrits des sites de l'Internet (sorte de prolongement des écrits de presse ou de livres recueillis dans le passé), la communication écrite en ligne y a rapidement trouvé sa place sous la dénomination de CMC. Ainsi les actes du premier évènement scientifique s'inscrivant dans la lignée de celui de Rennes (CMC'13, 2013) a donné la parole aux collègues impliqués dans la constitution de corpus de référence de l'allemand, du néerlandais, du flamand, qui ont jugé indispensable de saisir les premiers échantillons de communications tels celles provenant de courriers électroniques, forums de discussion, clavardage.

La stimulation pour ce faire est en effet importante, car c'est sans doute la première fois dans l'histoire de l'humanité que nous avons l'occasion de saisir les traces à grande échelle d'interactions entre des humains « ordinaires » (pour reprendre le qualificatif utilisé par les linguistes créateurs du corpus de correspondances de soldats de la première guerre mondiale (Corpus 14, 2014), exemple trop rare de telles communications antérieures à l'ère des réseaux). Nos sociétés ont ainsi construit des discours sur la langue, à partir d'études faites sur des recueils de paroles (écrites ou orales) d'écrivains, journalistes, personnages politiques, ou généraux. Ces discours académiques ont trop rarement été contrastés avec celles d'études sporadiques et parcellaires sur celles du reste de l'humanité, usagers, citoyens, apprenants. Nous avons tellement l'habitude de considérer cette situation comme normale qu'il peut être intéressant de tenter une comparaison avec celui du déséquilibre rencontré en histoire. Dans cette discipline, ce sont les

historiens appartenant à des sociétés qui ont possédé tout à la fois l'écrit et les médias pour le convoyer qui nous expliquent ce que devaient penser, ressentir les individus des sociétés actuelles ou passées, sociétés basées sur la seule oralité et/ou n'ayant pu trouver le moyen d'inscrire leur langue dans des supports matériels durables.

Saisir de façon systématique et organisée les données de la communication sur les réseaux constitue un enjeu majeur qu'a voulu relever, en particulier la communauté de chercheurs rassemblés lors des rencontres de 2013 (CMC'13, ibid.) et 2014 (CMC'14, 2014). Ils ont créé une communauté de recherche CMC-corpora (2016) qui a pour but la création, la diffusion de corpus de communication médiée par les réseaux dans des langues variées à des fins de descriptions et analyses, ainsi que la diffusion des résultats des recherches au travers d'évènements scientifiques tels les deux précités, celui présent de Rennes (CMC'15, 2015) ou encore celui de Slovénie (CMC'16).

2 DEUX ÉCHELLES D'ESPACE DE COMMUNICATION ET DE COMMUNAUTÉS : DE L'APPRENANT À L'USAGER

Les journées de 2013 (CMC'13, ibid.) virent se rencontrer des chercheurs issus de domaines différents, la linguistique de corpus appliquée aux CMR, bien sûr, le traitement automatique du langage (TAL) et celui de l'apprentissage des langues, déjà mentionné précédemment à propos des années 80-90. Dans ce dernier milieu, des progrès avaient été accomplis indépendamment afin de constituer des corpus d'interaction en ligne en situation d'apprentissage. Ceux-ci incluent non seulement les interactions, mais leur contexte de façon détaillée, à savoir les scénarios pédagogiques, la description des environnements de communication et les protocoles de recherche. Il s'agit des corpus d'apprentissage, autrement dénommés *Learning & Teaching Corpora* (LETEC) (Chanier &

Wigham, 2016). Ceux-ci sont librement accessibles (Mulce, 2012). Outre l'apport d'une méthodologie de recherche nouvelle au sein de l'apprentissage des langues, ils viennent compléter et enrichir les problématiques de la linguistique de corpus CMR. Les communautés impliquées étant de taille plus réduite (échelle du groupe d'apprentissage comprenant apprenants et enseignants, souvent répartis en mini-groupes), une méthodologie appropriée a permis de renseigner les profils des participants afin de pouvoir suivre leurs interactions propres, de travailler, par exemple, au niveau des richesses lexicales. Les dialogues sont structurés, autorisant des recherches précises sur les échanges. Les interactions des individus se déroulent au fil des semaines dans des environnements différents mêlant synchrone et asynchrone. Enfin, ces échanges mêlent non seulement des interactions verbales de nature textuelle et orale (reprenant et prolongeant ainsi les traditions en linguistique de corpus oraux), mais également non-verbales (actions dans des mondes synthétiques avec avatars, dans des traitements de textes collaboratifs, etc.), inscrivant ainsi les interactions dans un cadre multimodal.

Le présent ouvrage illustre bien la présence de la problématique apprentissage des langues grâce aux contributions de Mayne et Petersen (cet ouvrage), même si leurs études ne sont pas basées sur des corpus structurés et diffusés. Les corpus de CMR constitués par les autres auteurs de cet ouvrage sont de nature uniquement textuelle, tout comme l'ensemble des interactions discutés dans ce volume. Toutefois, la multimodalité est un sujet bien présent dans les différentes conférences organisées par la communauté CMC-corpora (ibid.). Même si l'écrit constitue ici la base de toutes les communications, l'oralité et le non-verbal sont bien pris en compte dans les modèles d'interactions auxquels renvoient nombre d'auteurs ici, comme nous le verrons un peu plus loin.

3 LA DYNAMIQUE CONSTITUTION DE CORPUS ET ANALYSES

3.1 CONSIDÉRONS D'ABORD LES OPPORTUNITÉS DE NOUVELLES ANALYSES

L'opportunité de pouvoir saisir, comme nous l'avons mentionné précédemment, les productions en situation d'échanges de monsieur (ou madame) tout le monde sur l'Internet ouvre bien entendu de nouvelles perspectives et motivations de recherche.

Viennent à l'esprit, tout d'abord, les enjeux patrimoniaux. L'empan temporel entre l'apparition d'une nouvelle situation de communication sur les réseaux et sa disparition peut-être très court, moins de 15ans. Ainsi en va-t-il du corpus *88milSMS* décrit par Ghliss et André (ce volume), tout comme des deux autres corpus de textos / SMS accessibles sur le site CoMeRe (2014 ; 2016). Les messages collectés sur le territoire métropolitain ou à la Réunion, à l'occasion de communication survenues entre les années 2008 et 2011 ont déjà un intérêt patrimonial. En effet, la très grande majorité des scripteurs, ont utilisé des téléphones disposant d'une technologie aujourd'hui dépassée avec des claviers peu ergonomique et de fonctionnalités limitées (tout au mieux celle de l'algorithme T9 de complétion au niveau du mot seul). Aujourd'hui la majorité (voire très grande majorité, déjà, dans certains pays) des téléphones sont des *smartphones* disposant d'une interface de frappe entièrement différente, de fonctionnalités provenant du TAL, permettant de rappeler des occurrences du vocabulaire de l'usager, offrant des suggestions à partir de calcul de bigrammes, sans oublier l'affichage de fil de discussions bien organisées. La nouvelle nature des interactions, des structures langagières (lexique, syntaxe, sans oublier les émoticônes et émojis) va donc apparaitre dans les nouveaux corpus. Ce qui ne veut pas dire, pour autant, que les échanges saisis dans les corpus de textos précédemment cités soient tous conformes aux clichés de style télégraphique, si souvent cités dans les médias.

L'insistance d'une partie notable des scripteurs à composer, malgré les difficultés techniques, des messages bien formés a d'ailleurs peu fait l'objet d'études. Pourtant les données saisies sont bien là, disponibles et les perspectives d'autant plus intéressantes du fait de la structuration de ces corpus, des échanges, plus intéressantes encore lorsque le profil des participants, de leurs usages du téléphone y sont décrits : voir en particulier la version (Panckhurst et al., 2016) du corpus présenté par Ghliss et André (ce volume).

La richesse des saisies, tout autant que la façon de les organiser, de les structurer en corpus, nous éloigne des façons de considérer les corpus comme de simples "sac de mots" comme le rappelle Longhi (ce volume) et mettent à la portée des chercheurs des analyses de type sémiotiques, discursives, comme les discours d'escorte et d'accompagnement dans les tweets (Simon et al. , ce volume), ce qui élargit brusquement la façon traditionnelle de considérer les citations hors des problématiques CMR. Les motivations des groupes ou individus commencent à pouvoir être étudiées dans les réseaux sociaux (cf. Blanchard, ce volume) à partir d'approches combinant aussi bien des modélisations de type SNA (*Social Network Analysis*) où la participation de chaque personne, sous-groupe, clusters est calculée à un niveau global, que des études de contenus des messages.

De même l'étude des controverses entre internautes, voire groupes d'internautes est au centre de la thématique de plusieurs chapitres avec un cadre original de travail où vont pouvoir être contrastés modèles et analyses dans des communications à visée d'élaboration de savoirs dans Wikipédia (Poudat et al., ce volume ; Ho Dac & Laippala, ce volume) ou dans les fils de tweets (Jackiewicz, ce volume). Sans avoir, loin s'en faut, épuisé toutes les thématiques émergentes de recherche (au côté de celles traditionnelles sur les aspects lexicaux, syntaxiques, sémantiques assistés par une première série d'étiquetages morpho-syntaxiques du contenu comme expliqué par Ho Dac & Laippala (ibid.)), notons celles des affects, émotions matérialisés par les connotations positives et négatives,

particulièrement importantes dans les CMR à courts messages du type tweets (Fišer, Erzavec & Ljubešic, ce volume).

Dans ce dernier chapitre, il est intéressant de constater que les auteurs travaillant sur une langue, le slovène, qui recouvre une communauté plus réduite par la taille de locuteurs, ont su rapidement saisir un corpus de CMR écrites variés (tweets, forums, commentaires, blogues), et tout aussi rapidement y adapter des techniques et ressources TAL traditionnelles. L'explosion de l'espace d'expression sur les réseaux pour les locuteurs de cette langue a sans doute constitué un enjeu de premier plan pour les chercheurs du pays. Il a été couplé avec les efforts européens destinés à soutenir le développement de grandes infrastructures de données au bénéfice des sciences humaines, en l'occurrence celles concernant les corpus de langue, telles les structures CLARIN, et leur équivalent français ORTOLANG (2016).

3.2 PUIS LES RELATIONS AVEC LA SAISIE DES DONNÉES

A la lecture des articles précités, le lecteur s'apercevra rapidement que les thématiques de recherche évoquées ici sont en devenir. Beaucoup d'auteurs expliquent en effet comment ils ont saisi les données, constituer les corpus, et les diffusent en libre accès sur la Toile, toutes étapes indispensables, pensent-ils, avant les analyses.

Ce positionnement atteste d'une véritable mue du milieu de la recherche dans ces domaines des sciences du langage. Ces auteurs, en effet, délaissent les anciennes habitudes qui consistaient à collecter de façon artisanale des données, à les organiser de façon idiosyncratique, à commencer sans attendre quelques analyses partielles pour les publier au plus vite. Aucun relecteur ne pouvait alors voir les données sur lesquelles les analyses avaient été élaborées, en apprécier la portée. Les autres chercheurs ne pouvaient accéder aux données de l'auteur pour les mêler aux siennes afin de se livrer à des analyses plus étendues.

Les auteurs rassemblés dans ce volume travaillent souvent de façon collaborative avec des collègues de différentes disciplines, dont l'informatique, appliquent des méthodes de constitution et diffusion des corpus qui ont tendance à devenir des standards, après l'œuvre pionnière des acteurs de la linguistique de corpus oraux. Ils contribuent à développer les standards et à les diffuser au travers des réseaux nationaux tel CORLI (2016) en France, *Empirikom* en Allemagne, voire de réseaux européens. Cette façon de travailler est une réponse au défi de la saisie des données de CMR, un préalable pour éviter les analyses qui relèvent de l'anecdote.

4 LE DÉFI DE LA SAISIE DES DONNÉES DE CMR

4.1 VOUS AVEZ DIT « MÉGADONNÉES » ?

La quantité de données disponibles dans les systèmes CMR de type réseaux sociaux en accès libre, tels que les discussions Wikipédia ou les tweets, pourrait laisser penser que le chercheur se doit d'adopter une approche de type Mégadonnées (*Big Data*) pour les saisir. Cela supposerait de collecter les données, puis appliquer de nouveaux types d'algorithmes de fouilles sans organisation préalable. Boyd et Crawford (2011) ont fort bien souligné le danger d'une telle approche, en insistant sur les défis posés à l'approche Mégadonnées de traitement de l'information.

4.2 ÉLIMINER LE BRUIT, (RE)CONSTUIRE LES ÉCHANGES

Plusieurs auteurs dans ce volume illustrent bien les étapes de saisie qu'il leur a fallu mettre au point en éliminant tout d'abord le bruit, à savoir, par exemple, dans les tweets (Jackiewicz, ce volume), ceux fabriqués automatiquement par les robots, ceux quasiment vides de mots, ou dans Wikipédia (Poudat et al., ce volume), les nombreuses répétitions rencontrées dans les pages gigantesques d'archives des discussions passées.

Passé ce stade, l'objet de recherche doit être construit en sélectionnant les parties en rapport, en remontant les liens hypermédia dans cette toile. Alors commence à apparaitre les prémisses de ce qui va devenir un corpus. Longhi (ce volume) décrit comment reconstruire le fil des interactions de tweets autour d'un sujet politique : sélectionner les comptes d'auteur (twittos) ayant contribué de façon significative, trouver tous les fils de discussion, éliminer progressivement ce qui n'est pas significatif. De même, dans Wikipédia, une fois les thèmes de controverses choisis, il faut sélectionner les pages de discussions en rapport direct, tout comme celles de niveaux méta portant sur les régulations de conflits internes à Wikipédia, sélectionner les contributeurs principaux, leurs pages, les articles de Wikipédia reliés et leurs versions, etc.

4.3 STRUCTURER À L'AIDE DE MODÈLES APPROPRIÉS

Lors de cette (re)construction, le chercheur s'efforce de conserver, voire remodeler, la structure des données, au contraire de l'approche basique en Mégadonnées où tout est aplati. Comment prétendre, en effet, effectuer des analyses sur les interactions CMR, sans distinguer, les fils de discussions, les messages, leurs auteurs, les dates d'émission, les allocutaires ou les messages, voire commentaires de réponse ? Un effort particulier a été fait pour 3 des corpus présentés dans ce volume (*Wikiconflits*, *Polititweets*, *88MilSMS*) afin de les structurer de façon homogène, en utilisant le modèle développé dans le projet CoMeRe (Chanier et al., 2014), modèle qui s'applique aussi bien aux courriels, forums de discussions, textos, tweets, mondes synthétiques, environnements audio-graphiques et à leurs combinaisons. Ce modèle multimodal de description de CMR, qui affiche au même niveau action textuelle, action orale ou non verbale, est le produit d'un travail européen (Beißwenger et. al, 2015) dont l'objectif est d'étendre le standard TEI (*Text Encoding Initiative*), largement utilisé en linguistique de corpus, à la CMR.

4.4 LES PARADOXES DE LA SAISIE DE LA COMMUNICATION DE MONSIEUR OU MADAME TOUT LE MONDE

Avant de diffuser les corpus, à travers les grandes infrastructures qui en garantissent un accès libre non labile, il manque une étape que nous rappellent Ghliss et André (ce volume), celui de l'anonymisation. Celle-ci s'inscrit dans un cadre plus large, celui de l'éthique et des droits qu'il faut avoir étudié avant de commencer les saisies de données. Le traitement correspondant ne doit pas effacer l'information, mais la transformer en en préservant la sémantique, tout en évitant qu'on puisse identifier un participant particulier ou, à tout le moins, éviter ce qui pourrait constituer des nuisances à sa vie personnelle. Les chercheurs renseignent cette modification d'information du corpus initial au sein de l'entête du corpus de façon à la rendre lisible et compréhensible pour les analyses ultérieures. Ce chapitre témoigne de la préoccupation actuelle des chercheurs en CMR qui s'inscrit en prolongement de méthodologies traditionnelles en linguistique appliquée et de celle provenant des corpus d'apprentissage LETEC (Reffay, 2013).

Ghliss et André abordent une autre question qui mériterait de plus amples discussions, à savoir celui de la non saisie d'une partie notable des textos lorsqu'ils revêtent des caractères offensants, injurieux (discussions qui ne peuvent avoir lieu qu'en supposant au préalable que ces messages aient été anonymisés afin que l'on ne puisse pas identifier les personnes concernées). Imaginons des chercheurs de la prochaine génération technologique qui essayeraient de se faire une idée de la façon dont nous communiquions sur les réseaux au début du 21ème siècle à partir de nos corpus qui omettraient une partie, malheureusement, significative des données. Pour prendre un exemple plus récent, peut-on imaginer d'opérer une étude des échanges qui ont eu lieu lors de l'élection étasunienne qui a porté Trump au pouvoir en constituant des corpus à partir des seuls médias traditionnels, alors que l'on sait que la majorité des

citoyens du pays ne les ont pas regardés, ni écoutés, mais se sont au contraire tournés vers les réseaux sociaux, réseaux dans lesquels on peut supposer qu'une fraction notable des communications reprenaient des termes et thématiques dans l'esprit de celles rejetées ici ?

Enfin, la question de l'éthique soulève des problèmes au tout début de la saisie des données. Les textos du corpus *88MilSMS*, tout comme les deux autres figurant dans CoMeRe (2014, 2016) ont suivi le même protocole de collecte des données, en passant par des opérateurs téléphoniques. Les procédures de dons par les usagers de leurs productions et de collecte via ses opérateurs ne permettent pas de disposer des réponses au message de l'émetteur ou, lorsque celui-ci est par hasard présent, de le mettre en rapport avec celui de l'émetteur. Par ailleurs, les dates d'émission des messages ont été substituées par celle du dépôt chez l'opérateur. Autrement dit, il est impossible d'effectuer une recherche sur les interactions entre correspondants de textos. Ce protocole était sans doute inévitable à l'époque et n'empêche pas d'effectuer sur les données structurées nombre de recherches intéressantes. Mais il est temps d'imaginer des protocoles différents si nous désirons, par exemple, collecter des échanges entre usagers de *WhatsApp*. Le caractère personnel et privé est quasiment un axiome de départ, aussi faudra-t-il peut-être imaginer de nouveaux types de rapports, à l'échelle de la Toile, entres les internautes donateurs et les chercheurs, protocoles dont les modalités devront être soumises à discussion entre le milieu clos de la recherche et celui du citoyen internaute.

Ainsi ce volume rassemble un ensemble de contributions qui illustrent précisément les perspectives actuelles de recherche en CMR sur des données, d'un type absolument nouveau, déjà recueillies et diffusées. Il pointe vers les défis à relever dans un avenir proche si l'on veut être capable de mieux saisir les données des citoyens, usagers, ou apprenants, émetteurs des communications sur les réseaux.