Wigham, C.R. & Ledegen, G. (2017). Introduction. In, Wigham, C.R. & Ledegen, G. (eds) *Corpus de communication médiée par les réseaux : construction, structuration, analyse*. Paris : L'Harmattan.

Les journées internationales de recherche (JIR) sur les Médias sociaux et les corpus de communication médiée par les réseaux (CMR[1]) se sont tenues à Rennes les 23-24 octobre 2015. Portant sur la communication et les interactions découlant de réseaux tels que Internet ou les télécommunications, communications de types mono- ou multimodales, synchrones ou asynchrones, elles étaient ouvertes aux études sur tous les genres de communications sur les réseaux : la communication dans les clavardages, les forums de discussion, les blogues, les courriers électroniques, les messages SMS et WhatsApp, des environnements multimodaux et/ou 3D, les textos, et les interactions ou discours produits sur les réseaux sociaux numériques, tels que Facebook, Twitter, Linkedin, wikis (type Wikpédia), etc.

L'objectif de ces Journées Internationales de Recherche était de réunir des chercheurs qui ont collecté des données sur ces interactions et communications et qui souhaitent les organiser et les partager en vue de développer des recherches communautaires. Les JIR ont ainsi abordé les questions relatives au processus de construction de tels corpus, leurs annotations et analyses, les questions d'éthique et de droits en vue de la publication de tels corpus sous forme de données libres (OpenData).

Les articles réunis dans cet ouvrage s'organisent en 3 volets. Dans un premier volet, cinq premiers articles présentent la méthodologie et le développement de corpus, ainsi que l'annotation et l'analyse : les deux premiers articles présentent ainsi la constitution du corpus et son encodage en TEI, et ouvrent sur des perspectives de recherche (le corpus *Wikiconflits*, par Céline POUDAT et ses collègues, et le corpus *Polititweets*, par Julien LONGHI). Le troisième article présente la méthodologie de constitution d'un corpus de tweets pour analyser le discours d'escorte (Justine SIMON et ses collègues) et présentent une première classification des discours d'escorte, prenant appui sur une approche quantitative et qualitative interdisciplinaire. Ensuite, Yosra GHLISS et Frédéric ANDRÉ présentent le travail de mise à disposition du corpus 88MILSMS à travers l'anonymisation et l'élimination de discours répressibles (incitation à la haine, racisme). Enfin, le projet CLAPOTY (Contacts de Langues : Analyses Plurifactorielles assistées par Ordinateur et conséquences Typologiques) est présenté par Pascal VAILLANT dans sa conception, sa mise aux normes TEI et son analyse.

Les deux articles suivants portent sur une analyse complète, allant du développement du corpus à l'analyse effective et ses résultats : dans un premier temps, Lydia-Mai HO DAC et Veronika LAIPPALA présentent leur corpus de discussions Wikipédia, à travers la constitution du corpus et une analyse contrastive entre ce corpus et des forums de santé. Ensuite, le *Janes v0.3* corpus est présenté par Darja FIŠER et ses collègues. La création du corpus, son annotation et une première présentation quantitative sont complétées par une analyse automatisée des niveaux linguistiques et techniques des écrits dans le sous-corpus de Tweets.

Le dernier volet réunit les quatre derniers articles portent sur l'analyse de micro-domaines : Agata JACKIEWICZ présente son corpus de tweets *Filiation* où elle décortique le discours polémique, à un niveau notionnel et linguistique. Jette Milberg PETERSEN analyse finement la présentation des répertoires linguistiques dans un Réseau Social d'Entreprise.

[1] Selon leurs objets de recherche, certains contributeurs adoptent le terme CMR (communication médiée par les réseaux), tandis que d'autres continuent à employer CMC provenant des recherches anglo-saxonnes (Computer-Mediated Communication ; communication médiée par ordinateur, CMO).

Jean-François BLANCHARD étudie la présence du breton sur Internet. Enfin, Liz MAYNE présente l'analyse des pronoms d'adressage *tu-vous* proposé sur le forum WordReference, une communauté en ligne qui réunit des natifs et des apprenants.

Enfin, une postface de la main de Thierry CHANIER clôt l'ouvrage en éclairant la place de ce colloque au sein des études sur la CMC.

La contribution de Céline POUDAT, Natalia GRABAR, Camille PALOQUE-BERGES, Thierry CHANIER et Kun JIN présente les choix de structuration ainsi que des procédures d'annotation appliquées lors de la constitution d'un corpus de pages de discussions éditoriales associées aux articles de Wikipédia, *Wikiconflits,* et plus spécifiquement sur des pages ou la présence de conflits entre contributeurs ont été constatés. Le chapitre expose les étapes de constitution de corpus et notamment décrit le choix de pages à inclure dans le corpus, l'extrait des clusters de page et la structuration des fils et des messages qui prend en considération leur aspect hybride. Les auteurs proposent ensuite une discussion autour de comment le corpus a été encodé en TEI-CMC avant de présenter les perspectives de recherche dans lesquelles ils vont mobiliser le corpus structuré afin de contribuer au développement des études linguistiques sur Wikipédia en France.

La contribution de Julien LONGHI porte sur la méthodologie de constitution de corpus et décrit les enjeux philologiques, herméneutiques ainsi que institutionnels lors de la constitution du corpus de tweets politiques *Polititweets* et sa structuration en TEI. Elle met en avant les synergies entre un projet de recherche interdisciplinaire au sein d'une université et une infrastructure de recherche nationale. A travers des exemples, l'auteur montre comment la méthodologie a dû tenir compte des éléments spatio-temporels, contextuels, technologiques ainsi qu'interactionnel dans les messages produits pour offrir à l'analyste de discours un terrain d'analyse du lexique politique par la suite. L'article propose une discussion des aspects juridiques liés au recueil des tweets ainsi que les enjeux de la mise à disposition du corpus avec une licence Creative Commons du type CC-BY pour mettre le corpus dans le domaine public.

Le chapitre de Justine SIMON, Bénédicte TOULLEC et leurs collègues proposent également un retour réflexif sur la méthodologie de constitution d'un corpus de tweets et soulignent l'importance d'une approche mixte composé d'un traitement quantitatif pour classer les données à une échelle macro et une phase qualitative, plus micro, qui dans le cadre de leur étude avait pour objectif d'extraire des corpus pertinents pour l'analyse de discours d'escorte. Les auteurs soulignent l'apport d'une équipe interdisciplinaire vis-à-vis les choix méthodologiques auquel leur projet a été confronté. Après avoir résumé chaque phase de la constitution du corpus, les spécialistes en analyse du discours et analyse des médias proposent une première classification des discours d'escorte autour de trois classes pas exclusives : d'opinion, d'éditorialisation et de communication. Ce qui ressort de l'analyse du discours d'escorte sur Twitter est comment les théories du discours ne sont pas outillées pour appréhender le discours médié.

Ensuite, Yosra GHLISS & Frédéric ANDRÉ présentent les enjeux éthiques et juridiques de la mise à disposition du corpus 88MILSMS : afin de permettre la publication des travaux qui prennent appui sur le corpus en question, mais aussi la mise en ligne du corpus lui-même, sa diffusion dans la communauté scientifique afin qu'il puisse être exploitable par différents chercheurs, l'étape de l'anonymisation, par le biais de remplacements des données personnelles par des balises est indispensable. Par ailleurs, ce corpus juridiquement « sensible » contient des discours qui incitent à la haine ou relèvent du racisme, lesquels constituent une autre nature de données à éliminer avant publication du corpus ou d'extraits.

Enfin, Pascal VAILLANT présente le projet CLAPOTY (Contacts de Langues : Analyses Plurifactorielles assistées par Ordinateur et conséquences Typologiques), un corpus de langue orale dans diverses situations de contact de langues. Du recueil à l'intégration dans le format TEI en passant par l'analyse et la mise à disposition de la communauté, le programme de recherche examine sous plusieurs angles les contacts de langues, une base riche pour un travail de comparaison avec les corpus de CMR.

Darja FIŠER, Tomaž ERJAVEC et Nikola LJUBESIC présentent le Janes v0.3 corpus, le premier corpus pour la langue slovène avec un contenu généré par l'utilisateur ; il réunit plus de 135 millions de mots et presque 5 millions de textes. Sont ici présentés la diversité de textes réunis (tweets, des discussions de forums, des commentaires sur les informations et des blogs), l'annotation linguistique appliquée sur le corpus, une première analyse quantitative, puis une présentation plus détaillée du sous-corpus de Tweets, dont par exemple, l'analyse des sentiments (positif, négatif, neutre). Une contribution originale consiste en l'attribution automatique d'un niveau (1 = standard à 3 = non standard) à chaque texte, à un niveau linguistique (prenant en compte le choix lexical, l'orthographe, la morphologie et l'ordre des mots) et à un niveau technique (l'utilisation des majuscules, de la ponctuation, des espaces …) ; cette analyse préfigure le développement d'outils pour l'analyse du slovène non standard.

Lydia-Mai HO DAC et Veronika LAIPPALA étudient toutes les discussions associées à un article de Wikipédia : elles présentent un premier état des lieux de la ressource en proposant une analyse déductive, orientée par des caractéristiques linguistiques généralement associées aux CMR, puis une analyse inductive pour analyser les caractéristiques lexicales et syntaxiques des discussions Wikipédia. Ces discussions constituant des textes libres, relativement bien renseignés, présentant un niveau d'écriture correct, forment un terrain d'expérimentation indéniable pour caractériser un certain type de communication médiée par les réseaux, ici en l'occurrence, la discussion Wikipédia. L'étude contrastive opposant les discussions aux articles Wikipédia, et à des forums de santé, a mis à jour que les discussions Wikipédia présentent un niveau d'écriture et de langage, ainsi que des traces de l'expression de l'opinion du scripteur, la description des changements effectués dans les articles et enfin un lexique propre au processus de rédaction des articles et de la politique éditoriale de Wikipédia.

Twitter est également le terrain d'observation pour Agata JACKIEWICZ qui propose aux lecteurs une analyse des interactions autour du sujet polémique de la filiation à partir du corpus de tweets Filiation et selon deux axes d'exploration. Le premier est fondé sur les notions qui renvoient aux aspects caractéristiques des controverses et permet l'observation de l'évolution du processus polémique selon différents aspects ciblés. Le deuxième est guidé par les procédés linguistiques propres aux échanges polémiques, par exemple la négation polémique et l'utilisation des connecteurs argumentatifs, et permet une lecture fine des interactions. Le prolongement fort prometteur de ce travail consistera à outiller cette méthode, sous forme d'une plateforme d'annotation et d'analyse, pour pouvoir articuler les deux modèles.

Analysant dans le détail les aspects multilingues dans un Réseau Social d'Entreprise, et plus particulièrement les manières d'indiquer un répertoire linguistique dans le profil SolidarNet, présentées en typologie, Jette Milberg PETERSEN en vient à proposer des modifications et améliorations fines : elle préconise ainsi de concevoir un profil RSE qui reflète tous les répertoires linguistiques et les niveaux de langue de ses membres ; en effet, la distinction 'langue maternelle/langues parlées' ne pouvant suffire et prêtant à de multiples confusion ; par ailleurs, afin que les membres deviennent davantage actifs, et afin de contrer l'insécurité linguistique ambiante, surtout dans le monde de l'écrit, elle propose d'autoriser une pratique langagière fondée sur le concept d'intercompréhension.

Jean-François BLANCHARD étudie la présence du breton sur internet (Wikipédia, réseaux sociaux, microblogging, blogs, médias en ligne), renaissance apparente qui contraste avec la diminution inexorable de l'effectif des locuteurs. Revitalisation de la langue bretonne dans de nouveaux espaces d'interlocution, ou bien nouvelles pratiques de médiactivisme et de militantisme tentant de surseoir à la disparition annoncée d'une langue minoritaire ? Pour Wikipédia comme pour le groupe Facebook en breton, les acteurs sont militants, représentants associatifs ou nouveaux locuteurs, au sein de ces médias d'information qui ne constituent donc pas un véritable espace d'interlocution en breton.

Liz MAYNE analyse la variable pragmatique du pronom d'adressage à la 2° personne, tu, le pronom informel versus vous, le pronom formel, sur le WordReference forum. Cette communauté en ligne en accès libre réunit des locuteurs natifs et des apprenants lors de discussions sur l'utilisation des langues. Les analogies avec d'autres langues ou d'autres situations sociolinguistiques donnent un éclairage interculturel riche sur cet apprentissage qui reste difficile pour les apprenants L2, et enrichissent utilement les échanges au-delà d'une situation de classe entre apprenants.

Les différents contributeurs apportent ainsi leur expertise sur la CMC, en particulier la collecte des données et les questions d'éthique inhérentes, leur annotation et analyse, et, *in fine*, leur mise en partage pour la communauté.

La tenue du colloque autant que la publication ont bénéficié du soutien du laboratoire de recherche PREFICS (Plurilinguismes, représentations, expressions francophones, information, communication, sociolinguistique) - EA 4246 de l'Université de Rennes 2, ainsi que du Laboratoire de Recherche sur le Langage - EA 999 de l'Université Clermont Auvergne.