Résumés

Le sens des données

Statut du corpus et herméneutique à l'aune des humanités numériques

ANDREACOLA Florence, Université Grenoble-Alpes - IUT 1,

florence.andreacola@univ-grenoble-alpes.fr

Les données d'internet : Le sens d'un corpus d'images numériques par la prise en compte de son contexte informatique de production

La construction d'un corpus est très fréquemment marquée par le contexte disciplinaire dans lequel la recherche s'inscrit. Le constat que les postures théoriques et méthodologiques, par exemple, en ethnologie, en histoire de l'art ou en sémiologie dirigent la construction d'un corpus nous interroge sur la façon dont, à l'inverse, le corpus spécifique des données numériques peut agir sur les approches théoriques et méthodologiques disciplinaires qui les exploitent. L'objet de cette communication est de présenter les façons dont on peut, dans une approche interdisciplinaire qui associe informatique et sciences de l'information et de la communication, donner du sens à un corpus de données numériques exploitables dans l'analyse d'usages sociaux d'Internet. Cette approche s'appuie sur une définition singulière des corpus de données numériques qui prend en compte les conditions nécessaires au façonnage des écrits informatiques, soit le contexte informatique de leur production : les protocoles d'échanges d'Internet dans sa dimension décentralisée face aux formats clos d'échanges sur des plateformes centralisées (Google ou Facebook). La spécificité de cette approche est qu'elle s'appuie aussi sur la mise en relation de conditions de façonnage des écrits d'Internet avec les dimensions visibles, pour l'usager, de ses propres formes écritures informatiques. Cette communication sera articulée en trois points. Après une introduction sur un état des lieux de définitions des corpus de données informatiques, nous présenterons notre définition d'un corpus de données numériques qui se compose, selon nous, d'écrits *contextuels* et *éditorialisés*. Ensuite, nous présenterons les enjeux épistémologiques et éthiques face auxquels le chercheur est confronté lorsqu'il travaille avec des données, en particulier des images, issues d'un modèle décentralisé et/ou centralisé du web. Enfin, nous présenterons quelques résultats sur les usages sociaux d'Internet issus d'un corpus qui se fonde sur des données numériques et qui a mobilisé plusieurs outils de récolte différents.

***Mots-clés** : données numériques, corpus, traces informatiques, interdisciplinarité, participation informatique*

BERNAT Justyna, Université Paris-Sorbonne,

justynabernat.pro@gmail.com

Accéder aux savoirs cachés du corpus

Aujourd'hui, être chercheur en SHS, et plus précisément en linguistique, demande malgré tout des connaissances mathématiques et informatiques. L'automatisation de certains processus, loin d'être contraignante, apporte une nouvelle dimension au travail des chercheurs en sciences du langage. Les projets interdisciplinaires, comme le CASK (*Computer-aided Acquisition of Semantic Knowledge)*, qui était à l'origine de la conception du logiciel SEMANA, permettent à la fois d'analyser des corpus beaucoup plus importants en taille et de découvrir des savoirs invisibles dans le travail "à la main". SEMANA est un outil interactif d'exploration automatique des données. Il permet non seulement de construire et d'annoter le corpus mais aussi accompagne le chercheur dans la construction de sa méthodologie. La collaboration machine-humain permet de combiner les capacités uniques de l'une et de l'autre afin de faire évoluer la méthode d'analyse jusqu'au point où elle devient satisfaisante. Le logiciel emploie des méthodes d'extraction des données (*data mining*) et d'analyse des données symboliques KDD (*Knowledge Discovery in Databases*), déjà répandues en sciences de l'information. Le chercheur a ainsi accès non seulement aux informations relatives au contenu linguistique du corpus mais aussi aux méta-données relatives au système d'annotation et à la méthodologie. La modélisation de ces données révèle des nouvelles relations entre les catégories d'annotation découvrant ainsi une sémantique parallèle au contenu linguistique. Ce type de procédés s'avère très utile lorsqu'il s'agit de croiser et comparer un grand nombre de variables, notamment dans le domaine de la sociolinguistique où les facteurs de natures différentes s'influencent mutuellement. Le cas d'étude sur la politesse verbale (étude comparée de l'emploi des termes d'adresse) montre comment SEMANA permet d'extraire les savoirs cachés dans les méta-données du corpus spécialement annoté et préparé et comment ces savoirs impactent l'analyse linguistique.

***Mots-clés :** data mining, sociolinguistique, politesse verbale, linguistique du corpus, métalangage grammatical*

BIGOT Jean-Edouard, MABI Clément, Université de Technologie de Compiègne,

jean.edouard.bigot@gmail.com ; clement.mabi@gmail.com

“L’équipement” des humanités numériques : analyse des outils, réflexivité des usages et construction du sens

Notre communication vise, dans une optique définitionnelle, à tenter de préciser le statut sémiotique des dispositifs numériques qui équipent aujourd’hui le travail des chercheurs dans le domaine émergeant des humanités numériques, et d’évaluer les effets de leurs médiations sur les pratiques savantes en SHS. Ces outils, de plus en plus mobilisés pour accompagner la constitution et l’étude de vastes corpus de données (p. ex. les pratiques de la cartographie des controverses développée à Sciences Po et de la Digital Methods Initiative de l’Université d’Amsterdam), importent une rationalité particulière qui structure les modalités de construction du sens dans un contexte donné. Pour le dire autrement, chaque étape du processus de traitement des données (de la collecte à la visualisation) introduit des médiations techniques et sémiotiques dans le rapport à l’observable dont il convient d’analyser les enjeux. En nous inscrivant dans le cadre d’une techno-sémiotique des médias informatisés, nous pensons qu’il convient d’appréhender les outils de visualisation et de traitement statistique de données (numériques ou numérisées) en tant qu’ils participent à l’élaboration d’une posture phénoménologique singulière qui conditionne les modalités interprétatives des phénomènes dont ils permettent l’observation et l’analyse. Dans le contexte de la recherche en SHS, on partira de l’hypothèse selon laquelle ces dispositifs médiatiques s’inscrivent dans la longue lignée des techniques de l’intellect (Goody, 1979) qui transforment les rapports empiriques entre un chercheur et son objet d’étude et par là même dessinent de nouveaux cadres épistémologiques, circonscrivent l’espace dans lequel une certaine pratique scientifique est possible. Si sur le plan cognitif ces dispositifs déterminent un regard « situé » sur le réel, sur le plan symbolique ils appellent également des représentations de la science : ils réactivent notamment le mythe d’une scientificité appareillée, une forme d’« objectivité mécanique » (Daston & Galison, 2012). Pour explorer cette hypothèse nous nous appuierons sur l’analyse sémiotique du logiciel de visualisation de graphes *Gephi*. Il s’agira de montrer que cet outil, aujourd’hui largement utilisé pour représenter des réseaux sur le web (cartographie numérique), véhicule un rapport singulier au social qui oriente les analyses qui le mobilisent pour comprendre les interactions sociales en ligne.

Mots-clés : *méthodes numériques, outils de visualisation, techno-sémiotique, techniques de l’intellect, médiations*

CHOLET Céline, Université Bordeaux-Montaigne,

celine1903a@yahoo.fr

Explorer l’image par les données numériques. Enjeux sémiotiques : faire voir pour faire sens

Pour reprendre les propos de Rastier (*La Mesure et le Grain*, 2011), la numérisation des textes permet un retour réflexif sur leur élaboration et leurs parcours d’interprétation. Pour l’auteur, elle pose la question de l’émergence de nouvelles formes d’élaboration de connaissances. Dans le cadre de cette présentation, nous souhaitons envisager la question du corpus à partir de deux corpus : le premier, notre corpus de référence, est basé sur les images de découverte d’espèces botaniques publiées dans les revues scientifiques du Muséum national d’histoire naturelle de Paris depuis 1804 ; le second, l’un de nos corpus de travail, est basé sur la sélection de trois images représentatives des trois types d’image en usage dans notre premier corpus (dessin, photographie et herbier). Ce second corpus, ainsi circonscrit, a fait l’objet d’une recherche exploratoire basée sur un projet d’eye-tracking. L’objectif est d’étudier la corrélation entre nos deux corpus, et leurs problématiques, à partir de deux visées : (i) pratique et méthodologique : il s’agit d’aborder notamment la question de l’élaboration d’hypothèses et de traitement des données. Ce travail nous conduira à interroger le rapport entre corpus qualitatif et quantitatif, le rapport syntagme et paradigme ; (ii) herméneutique : il s’agit de questionner la caractérisation et l’évolution de certains concepts comme ceux de lecteur, de performance ou encore d’interprétation.

Mots-clés : *Analyse de l’image, corpus visuel, eye-tracking, sémiotique structurale, sciences naturelles*

COMPAGNO Dario, Université Sorbonne Nouvelle,

dario.compagno@gmail.com

Analyse sémiotique de corpus et analyse sémiotique de textes : pouvons-nous utiliser les mêmes instruments pour deux modèles hétérogènes ?

Celui de texte est sans doute un des concepts les plus importants issus de la réflexion sémiotique des années 1960 et 1970. Il a permis d'identifier un premier niveau de pertinence optimal pour étudier le sens et a été le lieu de confrontation capable de montrer les points de force et les faiblesses des différentes approches disciplinaires (Barthes, Greimas, Eco). Au-delà de leurs différences, pour tous les pères fondateurs de la sémiotique le modèle de texte a eu la fonction d'organiser de manière homogène des données hétérogènes. La priorité accordée au modèle plutôt qu'aux phénomènes « nus » a cependant risqué de négliger les spécificités de ces derniers. Plus récemment, plusieurs chercheurs ont mis en question la validité universelle du modèle textuel pour analyser et interpréter tout type de phénomène signifiant. Les données signifiantes peuvent être organisées à plusieurs niveaux de pertinence (Fontanille) et leur constitution en objets d'analyse peut dépasser le format textuel. François Rastier, spécialement, met en avant les différences entre textes et corpus : un corpus n'a pas toutes les caractéristiques d'un texte, et vice versa. Surtout, les instruments nécessaires pour analyser des corpus ne sont pas les mêmes que pour l'analyse textuelle : le niveau d'intégration du corpus rend observables et pertinents d'autres traits (« nouveaux observables ») qu'on ne pourrait pas isoler avec le modèle textuel. Pouvons-nous dire, par exemple, qu'un corpus montre une narrativité telle qu'on le dirait d'un texte ? La logique de constitution et d'analyse d'un corpus est « planaire », « réversible », on peut collecter et trier les données de façon multiple. Quels rapports entre la structure communicative d'un corpus et celle des textes qui le composent ? Si tout texte a un lecteur modèle, un sens prévu, le sens des textes d'un corpus est censé être plus indépendant de l'intention du chercheur qui les a collectés. Les corpus ont un précis rapport à l'empirie, à leur « plan de référence » : ils doivent être représentatifs pour pouvoir permettre de justifier des généralisations (au contraire l'analyse d'un texte a une valeur même si on ne peut pas généraliser les résultats à un plus ample discours). Texte et corpus sont deux modèles bien différents, qui demandent d'être construits et analysés de manière différente. Nous avons travaillé à la constitution de grands corpus de données textuelles et multimodales et à leur analyse. L'approche sémiotique est censée pouvoir traiter ces données signifiantes. Cependant, nous avons eu besoin d'acquérir d'autres compétences pour pouvoir efficacement aborder un format de données de facto très éloigné de celui de texte. Premièrement, il n'est pas possible d'analyser un corpus sans utiliser des instruments statistiques. Toute analyse de corpus qui n'utilise pas de statistiques risque de n'être qu'une analyse textuelle déguisée. Si on traite des petits corpus, nous pouvons réaliser des séries d'analyses textuelles et, juste à la fin, combiner les résultats dans un système. Mais dès qu'on travaille avec des bases de plusieurs millions de textes et des corpus de plusieurs centaines de milliers d'éléments, on s'aperçoit tout de suite qu'une segmentation en textes individuels n'est pas réalisable ni souhaitable. Nous proposons de développer une analyse sémiotique de corpus fondée sur des procédés d'opérationnalisation (Moretti), distincte de l'analyse textuelle bien qu'issue des résultats consolidés de cette dernière. Le premier défi d'une analyse sémiotique de corpus est qu'un corpus n'est pas interprétable. Des milliers de textes n'ont tout simplement pas un sens tel qu'on pourrait le trouver dans chacun d'entre eux. L'analyse de corpus nécessite alors d'abord d'une textualisation : pour pouvoir interpréter, on doit produire du texte là où il n'y en avait pas. La textualisation est une forme de production sémiotique, comme telle caractérisée par des règles de transformation : il faut savoir choisir quels procédés statistiques et quels algorithmes peuvent produire des diagrammes signifiants à partir des documents du corpus, sans détruire leur sens au passage. L'identification de ces règles de transformation constitue la compétence la plus importante pour une méthodologie d'analyse sémiotique de corpus. L'aspect méthodologique de nos travaux vise justement à identifier des transformations expressives capables de garder leur rapport originaire avec des contenus. Nous présenterons les résultats de certains de nos travaux sur des corpus issus de grandes bases de données textuelles et multimodales (ANR OT-Media, ANR Info-RSN, Inter-MSH TEE2014). L'objectif sera de discuter la validité de certaines transformations algorithmiques et l'interprétation des diagrammes produits par ces transformations. Nous mettrons en avant la valeur des concepts sémiotiques utiles pour l'interprétation de corpus de données textuelles et multimodales, et la nécessité d'intégrer des techniques statistiques dans la « boîte à outils » du sémioticien.

Mots-clés : *sémiotique, corpus, statistiques, opérationnalisation, textualisation*

DE ANGELIS Rossana (Université de Paris), **MOUTAT Audrey** (Université de Limoges),

rossana.deangelis@gmail.com ; audrey.moutat@yahoo.fr

La linguistique à l'épreuve du numérique : vers des nouveaux types de sémioses

Les premières techniques d'exploitation des corpus ont été développées pendant les années 1970 au sein des recherches sur l'analyse lexicale. Au début des années 1990, les textes deviennent ses objets spécifiques. À cette époque Sinclair (1991, p. xvii) analyse des corpus constitués de textes (*extended texts*) écrits en langue anglaise, afin d'identifier les modèles lexico-grammaticaux sous-jacents. Toutefois, comme le fait remarquer Meyer (2012, p. 24), la technologie a changé le rapport aux textes. La possibilité de transformer un écrit dans un ensemble de données permet une nouvelle modalité d'exploitation de celui-ci. Le numérique propose une nouvelle vision du texte et de la textualité. « Un texte anti-naturel donc, dématérialisé – virtuel pourrait-on dire commodément –, dont les contours physiques tels que perçus depuis des siècles sont abolis, et la structure et le contenu – entendons, pour faire simple : la textualité – reconsidérés » (Mayaffre 2007a, p. 17). L'impact des technologies numériques sur les méthodes d'exploitation des corpus provoque une remise en question du concept de *texte.* Celle-ci se présente, par exemple, sous la forme du « *dépassement/complément de la linéarité* » (Mayaffre 2007a, p. 17), ce qui comporte par conséquent aussi une redéfinition du concept de *textualité*. Au vu de ces « nouveaux observables », on pourrait s'interroger sur les différents types de sémioses engagées et sur la manière dont le processus interprétatif (Rastier, 1987) est mené par l'analyste en sachant que ce processus est d'abord instauré par un traitement automatique des données. Dans la mesure où les pratiques (Fontanille, 2008) d'exploitation ne sont plus les mêmes que celles relevant de la linguistique de corpus à ses débuts et que l'exploitation de données n'engage plus les mêmes objets au départ (textes rassemblés dans le corpus) et à l'arrivée (documents alors produits), comment pourrait-on prendre en compte ces transformations dans un processus interprétatif qui s'avère si différent ? Comment pourrait-on donc intégrer cette nouvelle pratique d'exploitation dans la construction du parcours interprétatif de l'analyste ? Comment le traitement automatique des données articule-t-il de nouveaux types de sémioses ?

Mots-clés : *sémiose, interprétation, texte, texture, document*

ELIE-DESCHAMPS Juliette, Université de Limoges,

juliette.elie-deschamps@unilim.fr

Un exemple de base de données en psycholinguistique : CHILDES

Le domaine de la psycholinguistique peut être défini comme « l'étude des processus psychologiques par lesquels un sujet humain acquiert et met en œuvre le système d'une langue naturelle » (Caron, 1983). Dans le cadre des recherches menées dans ce domaine, les chercheurs se basent principalement sur deux types d'approches méthodologiques. Ils peuvent, tout d'abord, utiliser des situations expérimentales afin de tester diverses compétences linguistiques des enfants, ce qui permet d'aboutir à des corpus semi contrôlés. Ils peuvent également collecter les données à partir de situations écologiques, et donc travailler à partir de corpus issus d'un recueil de langage spontané. Nous proposons de nous intéresser plus particulièrement au deuxième type de corpus à partir d'un exemple de base de données, nommée CHILDES. Le Child Language Data Exchange System fut créé en 1984 par Brian MacWhinney et Catherine Snow dans le but de pouvoir échanger des corpus de langage d'enfants au sein d'une communauté scientifique. MacWhinney et Snow ont développé un système de codage (CHAT) pour standardiser les transcriptions et un certain nombre d'outils et de programmes (logiciel CLAN) afin de faciliter les analyses de ces données. Cette base de données, accessible par internet, est constituée de quatre grands types de corpus : 1) corpus d'enfants monolingues en suivi longitudinal, 2) corpus d'enfants bilingues, 3) corpus de narration et 4) corpus clinique issus de données recueillies auprès d'enfants avec un trouble du langage. Pas moins de 39 langues sont représentées dans cette base de données (anglais, français, danois, arabe, mandarin, turc, hongrois, farsi, basque, etc. pour ne citer que quelques exemples). Nous vous présenterons les grands principes d'un tel système à partir d'un exemple de recherches françaises l'utilisant, avant d'en dégager ses avantages.

Mots-clés **:** *base de données partagée, corpus écologiques, psycholinguistique*

EMERIT-BIBIÉ Laëtitia, Université Bordeaux Montaigne,

emerit.laetitia@gmail.com

La notion de lieu de corpus : enjeux théoriques et propositions méthodologiques pour la prise en compte des données nativement numériques

Dans cette communication nous tenterons de dépasser la question du corpus afin de l'adapter aux particularités des terrains numériques. Constituer un corpus demande de travailler à partir de données stabilisées, c'est-à-dire pour les données numériques d'extractions ou, au mieux, de captures d'écran. Dans ce cas le terrain premier, c'est-à-dire l'environnement numérique où se situe la recherche, n'est plus accessible. Renoncer à l'instabilité des environnements numériques c'est perdre la nature technolangagière (Paveau 2012) des données pour n'en conserver qu'une image. C'est également renoncer à l'interactivité que permettent ces objets de recherche en les isolant des locuteurs-scripteurs-utilisateurs et de leur écosystème numérique d'apparition. Le lieu de corpus représente une alternative au figement total des données numériques. Il s'agit d'un espace numérique délimité mais dont les données possèdent trois caractéristiques incompatibles avec la notion de corpus : l'instabilité, la mixité et l'incomplétude. Pour définir et illustrer la notion de lieu de corpus nous nous appuierons sur la description du compte Facebook « Ma Thèse Sdl ». Ce lieu de corpus a été créé dans l'objectif d'observer le discours produit sur le réseau social numérique Facebook. La notion de lieu de corpus n'exclut pas celle de corpus, celui-ci devenant une potentialité qui lui est subordonnée. À partir d'un lieu de corpus il sera possible de créer plusieurs corpus différents constitués de figements focalisés sur une partie des données accessibles. Nous proposerons de représenter cette hiérarchisation au travers d'une modélisation arborescente. La notion de lieu de Corpus permet de rendre compte des aspects évolutifs, polysémiotiques et collaboratifs inhérents aux environnements numériques. Elle remet en question les rôles des locuteurs-scripteurs-utilisateurs, du chercheur−utilisateur et de l'environnement numérique. Elle soulève également des interrogations concernant sa représentation et l'existence, en raison de la personnalisation grandissante des environnements numériques, d'un corpus idionumérique qui reste hors de portée du chercheur.

Mots-clés : *analyse du discours numérique, corpus plurisémiotiques, lieu de corpus, réseaux sociaux numériques, représentation arborescente.*

FEWOU NGOULOURE Jean-Pierre, Université de Toulouse Jean Jaurès, **NGAMCHERA YANGOUO** Anne Aimée, Université de Yaoundé,

jpngouloure55@yahoo.fr

Corpus numérique ou le sens en sursis

De nombreux travaux ont déjà permis de montrer la grande complexité du corpus numérique à la fois en termes de collectes des données (manque d'unité thématique : Labbe, Marcoccia, 2005), d'énonciations composites (association des formes linguistiques et d'artefacts technologiques ; hyperliens, hashtag : Paveau 2013 ; ou notion de multifocalisation chez Herring, 1999), de transgressions de tout genre : néographies telles qu'étudiées par Anis (1998), voire de contraintes de l'univers informatique (capitalisme linguistique qui a déjà permis à Google d'engranger des milliards en investissant uniquement sur des mots dont la plupart sont d'ailleurs générés à la place des utilisateurs : Kaplan 2014). Sur ce dernier point, il suffit parfois d'appuyer simplement sur une touche de clavier d'ordinateur pour voir défiler une série de mots qu'on peut ensuite sélectionner à souhait, sans avoir à fournir le moindre effort. Tous ces travaux montrent à l'évidence combien à ce jour, il semble encore difficile de parler de représentativité, encore moins d'exhaustivité pour définir le corpus numérique, que l'on se situe d'un point de vue qualitatif ou quantitatif, tout semblant montrer *a postériori* son caractère davantage lacunaire et sa dimension éminemment symbolique. L'objet de notre réflexion est de poser la problématique du corpus numérique du point de vue de la signification et du sens, sous fond d'une sémiotique des pratiques (Fontanille, 2008) et dans le prolongement de ce que Pignier appelle « sociale expérience », qu'elle considère à juste titre comme une « scène de vie culturelle qui engage chacun de nous […] en tant qu'homocommunicans ». (Pignier, 2009 : 101). Dès lors si le corpus numérique pose comme nous le démontrerons la question de la textualité, de la discursivité, de la généricité et de la rhétoricité, notamment à travers des formes parfois ambiguës et non linguistiques comme les scripts et les émoticônes, comment envisager dans ces conditions des parcours interprétatifs qui répondent mieux aux critères de pertinence, de cohérence et d'objectivité ? Il s'agit finalement de montrer que la question du sens, lorsqu'il s'agit d'appréhender le corpus numérique, nous rapproche parfois de l'insignifiance, pas au travers d'élucubrations philosophiques, mais dans sa conception greimassienne, qui correspond en fait à une sorte d'imperfection et d'impensé de la signification. En somme, parce que le corpus numérique se veut particulièrement en situation, toujours dans d'une logique de co-production et de co-construction, devons-nous sans doute limiter parfois nos gestes interprétatifs à de simples hypothèses, loin des parcours achevés pour rester dans le sillage de la tension (tendre vers) et de la surface du sens. C'est le but majeur de cette réflexion, bien nourrie de nombreux retours d'expériences sur nos travaux antérieurs et en cours.

Mots-clés : *corpus numérique, sens, imperfection, usages, sociale expérience*

HALTÉ Pierre, Université François Rabelais de Tours,

pierre.halte@orange.fr

Pour l'intégration des émoticônes et des interjections acronymiques aux outils automatiques de traitement de corpus de textes numériques « pluricodes » : enjeux sémiotiques et pragmatiques

L'objectif de cette communication est d'abord d'interroger les effets de la plurisémioticité des textes numériques (en l'occurrence, les interactions entre du texte et des pictogrammes) sur les lignes de partage entre la sémantique et la pragmatique. Une approche (et une comparaison) sémiotique de l'interjection et de l'émoticône permet à notre avis de faire passer cette ligne non plus entre ce qui est codé linguistiquement et ce qui ne l'est pas, mais plutôt entre ce qui relève d'un sens indexical, énonciatif et ce qui relève d'un sens symbolique, référentiel, ou encore propositionnel. Partant de ce constat, qui se révèle particulièrement dans le cadre des technodiscours, nous montrerons ensuite quelles problématiques soulèvent ces marques en termes d'énonciation et de construction d'un locuteur, confondu avec le sujet parlant, et quels intérêts ont ces considérations pour les recherches concernant l'investigation de la subjectivité des locuteurs sur de vastes corpus numériques. Actuellement, les logiciels de traitement automatique des corpus sont malheureusement, le plus souvent, faits pour traiter des signes linguistiques en ne considérant que leur sens référentiel. Enfin, nous proposerons, sous la forme d'un « cahier des charges », quelques réflexions autour de l'intégration de ces marques modales à des outils de traitement automatique des corpus (TXM, ou encore TextObserver, par exemple). Nous réfléchirons, au-delà des considérations techniques et des limitations catégorielles qu'imposent ces logiciels, à un modèle permettant de rendre compte des aspects énonciatifs du sens des émoticônes et les interjections, automatiquement. Nous proposerons aussi des moyens d'interpréter de façon pertinente des résultats statistiques, en prenant appui sur une étude statistique des émoticônes au sein de trois séquences didactiques réalisées via t'chat.

Mots-clés : *interjections, émoticônes, textométrie/TAL, t'chat, plurisémioticité*

KRAZEM Mustapha, Université de Bourgogne,

mustapha.krazem@wanadoo.fr

Grands corpus ou quand le nombre de mots nous détourne du traitement linguistique des données

Même si je discuterai davantage de la seconde interrogation contenue dans l'axe 3 point 4, il me paraît incontournable dans l'introduction de prendre mes distances sur le contexte socio-scientifique du corpus tel qu'il est depuis quelques années en linguistique. Il n'y a pas si longtemps, on mettait en valeur son cadre théorique en commençant une contribution. Aujourd'hui on avance son nombre de mots, son logiciel d'annotation. Fréquemment, on entend les contributeurs s'excuser du faible nombre d'occurrences trouvées, puis promettre, tout en excuses, d'affiner les conclusions quand ils auront davantage de données. Nos illustres prédécesseurs n'avaient pas besoin d'un tel attirail pour faire de la linguistique, y compris d'ailleurs Blanche-Benveniste, dont l'ouvrage théorique majeur de 1975 ne comprend que des exemples forgés ! Cela justifie que l'on se pose la question suivante : *qu'est-ce que le corpus au sens socio-scientifique du terme a permis de découvrir que le linguiste seul n'aurait pu trouver* ?
Or, à ma connaissance, personne n'a jamais entrepris d'y répondre. Libéré de ce que je tiens pour une idéologie hégémonique du corpus, Je m'intéresserai donc surtout à la deuxième partie de l'axe 3 et 4 à propos des études grammaticales (je resterai dans ce cadre-là). Je m'emploierai à montrer d'une part que les corpus, s'ils sont utiles, ne sont pas nécessaires pour étudier une langue vivante, d'autre part que la stricte dimension quantitative des données est peu pertinente, voire nuisible, pour la description qualitative des faits de langue. Voici les principaux points que j'aimerais développer rapidement (pas forcément dans l'ordre ci-dessous) en m'appuyant sur des exemples extraits de corpus personnels et de travaux d'autres linguistes.
- Les corpus ne décrivent pas la totalité du système linguistique. On ne peut prouver que la totalité des données correspond à la totalité des potentialités du système linguistique. C'est le contraire qui est le plus simple à argumenter.

- Un même corpus (ou type de corpus) peut conduire à des descriptions théoriques très différentes (voir Berrendonner pour un modèle binaire de la « phrase » contre un modèle ternaire chez Blanche-Benveniste). Le corpus ne conduit donc vers aucune « évidence » théorique que les données auraient permis de faire émerger.
- Une approche quantitative du langage devrait tenir compte de ce que le système est capable quantitativement de produire (voir Gross 1975) en le comparant avec ce qui a effectivement été produit en français. Ce faisant, on s'aperçoit que la représentativité en valeur absolue du quantitatif est bien plus faible que celle des sondages d'opinion.
- On peut produire des données importantes par construction systématique ciblée sur un fait précis (voir Gross 1975, aujourd'hui oublié). Un grand corpus peut donc être constitué de données construites.
- Les corpus ne prennent en compte que les données produites sans considérer les données perçues. Toutes les données ne se valent donc pas du point de vue de la compréhension globale du système.

J'insisterai particulièrement, dans ma communication, sur le poids important des genres de discours, poids qui est largement sous-estimé alors que de nombreux faits de langue y sont sensibles. Dans ce cas, c'est moins le nombre d'occurrences qui importe que le nombre de genres (et la description des propriétés de ceux-ci) d'où elles proviennent. J'exemplifierai ce point à partir de faits de langue précis. Mon propos ne signifie pas que les corpus (lesquels ont le droit d'être artisanaux sans apparat numérique) sont inutiles, mais peut-être conviendrait-il qu'ils ne soient plus le point de passage obligé pour être crédibles, surtout si le nombre des données doit fonctionner comme une certification institutionnelle.

Mots-clés *: grammaire, statistiques, carcan scientifique, agrammaticalité, genres de discours*

LARCHÉ, Mélanie, Université de Pau et des Pays de l'Adour, Ethnopôle InOc Aquitaine,

ms.larche@gmail.com

Décrire sans figer : une ontologie pour l'inventaire du PCI en France

Suite à la ratification de la Convention pour la sauvegarde du patrimoine culturel immatériel de l'Unesco par la France en 2006, l'inventaire français du PCI débute dès 2008 sous la responsabilité du département du pilotage de recherche et de la politique scientifique (DPRPS) du ministère de la culture et de la communication (MCC). Dans un premier temps, en a découlé un corpus de fiches d'inventaire consultables au format pdf depuis le site internet du ministère ; format qui s'avère donner un caractère légal à ce nouvel inventaire, mais semble peu pratique quant à la valorisation et à l'accessibilité de celui-ci auprès du public. En effet, l'inventaire du PCI constitue un nouvel axe de recherche qui autorise une (re)découverte et une (re)lecture des pratiques posées, non pas en termes de formes fixes à conserver, mais comme des formes dynamiques dessinant un patrimoine vivant. Sa mise en œuvre conduit notamment à constater une imbrication de pratiques, de patrimoines, qui font anthropologiquement sens dès lors qu'on les aborde de façon systémique. De plus, le besoin intuitif d'un recours à d'autres outils se faisait sentir : géolocalisation, mise en réseau de fiches, sérendipité, etc. Autant d'outils que le développement des nouvelles technologies ont peu à peu démocratisé. Le recours aux technologies numériques permet par ailleurs une double fonction : celle, intrinsèque, de valorisation des pratiques, mais aussi une fonction de mise en réseau des données aboutissant à une « valorisation augmentée ». Ainsi, en 2014 l'InOc Aquitaine a proposé au DPRPS de poursuivre l'expérimentation en créant *PciLab*, un outil se basant sur la technologie du web sémantique qui permette de croiser les pratiques de l'inventaire du PCI en France et les protocoles de médiations par recherche sémantique afin d'accéder aux données de l'inventaire. Ce travail a été guidé par deux axes, dont la définition d'une base de connaissances structurées avec, notamment la création d'une ontologie. Cette dernière crée un nouveau langage tant syntaxique que sémantique (Bachimont, 2007) reposant sur un système conceptuel issu, pour *PciLab*, de Wikidata. L'organisation du corpus en notices interrogeables a obligé à réfléchir à la classification spatio-temporelle (Bowker, 1999) des pratiques de l'inventaire, tout en rendant compte du caractère vivant de celles-ci.

Mots-clés *: patrimoine culturel immatériel, web sémantique, ontologie, concepts, wikipédia*

PAVEAU Marie-Anne, Université de Paris 13 Sorbonne Paris Cité,

ma.paveau@orange.fr

Subjectivité des corpus numériques natifs : du microscope à la caméra

Les discours numériques natifs sont marqués par un trait structurel : leur relationalité. Tous les énoncés produits en ligne sont en effet liés *hic et nunc* i.e. au moment de leur écrilecture, les liens concernant les appareils, les énonciateurs et les énoncés. On montrera que cette relationalité, descriptible en termes de contextualisation technorelationnelle (Paveau 2015, 2016), de personnalisation (Merzeau 2009) et d'affiliation diffuse (Zappavigna 2012), implique la subjectivité des corpus construits à partir de données discursives numériques natives : ils sont idionumériques (Émerit 2016). Les corpus numériques natifs n'ont pas d'existence objective ni stabilité, comme les corpus prénumériques de discours médiatiques, politiques ou littéraires par exemple, ce qui exige des analystes du discours de repenser leur constitution et leur traitement.

PRUNET Anne, Université de Caen-Normandie, Université de Paris III,

anne.prunet@unicaen.fr

Quel usage du corpus en français sur objectif interuniversitaire ?

Notre proposition s'inscrit dans l'axe 2 et plus particulièrement dans la problématique des critères de représentativité, d'exhaustivité et d'homogénéité d'un corpus. Notre terrain de recherche se situe dans le domaine de la didactique des langues, en français sur objectif universitaire. Les recherches dans ce domaine ont pointé un certain nombre de faits de langue à enseigner, qui se démarquent des curricula de l'enseignement d'un FLE « tout public » définissant une progression qui puisse répondre à ce double objectif : celui d'une formation universitaire, en langue étrangère. Nous présenterons dans cet article les directions que nous avons retenues : l'approche énonciative et par genre de discours, les phénomènes de co-illocutions, les notions de dialogisme discursif, d'intertextualité, et d'écriture expansive s'appliquant spécifiquement au champ universitaire. Afin d'évaluer si ces faits de langues sont effectivement saillants dans les productions universitaires, nous avons constitué un corpus contrastif de productions d'étudiants français et étrangers. Nous exposerons les modalités de constitution du corpus : ce qui a procédé à notre choix pour la formulation de sujet de production écrite, le profil des étudiants (langue première, champ disciplinaire, et nombre d'années d'études, nombre de productions), la justification du corpus de référence – les productions des étudiants français – représentatif de la réalité du public étudiant. Nous préciserons ensuite le mode d'exploration des données : spécificités lexicales, grammaticales, comparaison des segments répétés. Nous proposerons enfin quelques pistes d'interprétation et d'exploitation (limite d'une approche « traditionnelle » de la grammaire en FLE, exploitation didactique permettant à l'enseignant de privilégier certains faits de langues, possibilité d'utilisation pour une réécriture des productions des étudiants), et nous nous attarderons sur les limites d'une telle étude, dont l'écueil résiderait en une acceptation exclusive et univoque, au nom d'une scientificité prêtée aux travaux sur corpus, alors qu'elle serait en définitive plutôt à considérer comme source d'information préalable à un travail de conception didactique qui ne dispense pas d'un cadre théorique permettant la construction d'observables.

***Mots-clés** : français sur objectif universitaire, interlangue, littéracie, construction, genre*

VALETTE Mathieu, Institut National des Langues et Civilisations Orientales,

mvalette@inalco.fr

Quelle sémantique pour quels corpus ?

Les humanités numériques (HN) constituent une mutation importante des SHS. La banalisation du support numérique et les grands chantiers de dématérialisation des textes anciens offrent de nouvelles opportunités en termes d'accès et d'analyses des données. L'utilisation d'outils informatiques permet la création et l'observation de nouveaux objets sémiotiques. S'inscrivant dans une problématique de *dématérialisation des documents*, les premiers pas des HN, relevaient d'ambitions à la fois patrimoniales, éditoriales et documentaires. Beaucoup d'initiatives consistaient en collectes, numérisations et collations de documents. Les documents étaient ainsi interrogeables au moyen d'outils d'analyse des données textuels souvent rudimentaires (concordanciers, par exemple). Les projets ont ensuite porté sur la normalisation des bases textuelles avec l'établissement de formats d'échange (XML) et de normes d'encodage (TEI). La normalisation a facilité les travaux d'annotation et d'étiquetage et a permis de complexifier certains outils d'interrogation. Si des techniques existantes commencent à intégrer le terrain (fouille de textes, textométrie), la linguistique offre principalement aux HN un ensemble de propositions philologiques. Elles participent notamment à l'établissement des textes existants dématérialisés. Cette linguistique étant *ressourciste* (*i.e.* pourvoyeuse de corpus normalisés, de lexiques et d'annotations), elle apparaît davantage comme une linguistique de *production* et non d'*analyse* des données. La théorisation y est secondaire. Or, l'adaptation et la création de méthodologies d'analyse constituent un enjeu pour les HN. À la puissance des algorithmes statistiques qui s'imposent aujourd'hui dans l'analyse des données textuelles (ADT) et dans le traitement automatique des langues (TAL), il est nécessaire d'associer un appareil méthodologique linguistique pour l'analyse sémantique des textes. Jadis *science-pilote* des SHS, la linguistique pourrait prétendre aujourd'hui au statut de *science-pivot* des HN en étant notamment prescriptrice de méthodologies. La linguistique bénéficie en effet d'une expérience du texte, tant d'un point de vue théorique (linguistique textuelle, analyse du discours, sémantique interprétative, philologie) que pratique, avec la linguistique de corpus qui l'autorise à prendre position face aux enjeux théoriques et méthodologiques naissants, et de ne pas laisser à d'autres disciplines le soin de décrire, seules, ces nouveaux objets sémiotiques. Notre proposition ici sera de faire le point sur la contribution possible d'une sémantique de corpus (Rastier, 2011, Valette & Eensoo, 2015) aux HN, en nous appuyant sur différents exemples.

Mots-clés : *sémantique de corpus, analyse statistique des données textuelles, text mining, TAL*